Journal of Educational Psychology
2015, Vol. 107, No. 4, 1051–1074

© 2015 American Psychological Association
0022-0663/15/$12.00 http://dx.doi.org/10.1037/edu0000036

How Readability and Topic Incidence Relate to Performance on Mathematics Story Problems in Computer-Based Curricula

Candace Walkington
Southern Methodist University

Virginia Clinton
University of North Dakota

Steven N. Ritter
Carnegie Learning Inc., Pittsburgh, Pennsylvania

Mitchell J. Nathan
University of Wisconsin, Madison

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Solving mathematics story problems requires text comprehension skills. However, previous studies have found few connections between traditional measures of text readability and performance on story problems. We hypothesized that recently developed measures of readability and topic incidence measured by text-mining tools may illuminate associations between text difficulty and problem-solving measures. We used data from 3,216 middle and high school students from 10 schools using the Cognitive Tutor Algebra program; these schools were geographically, socioeconomically, racially, and ethnically diverse. We found that several indicators of the readability and topic of story problems were associated with students' tendency to give correct answers and request hints in Cognitive Tutor. We further examined the individual skill of writing an algebraic expression from a story scenario, and examined students at the lowest performing schools in the sample only, and found additional associations for these subsets. Key readability and topic categories that were related to problem-solving measures included word difficulty, text length, pronoun use, sentence similarity, and topic familiarity. These findings are discussed in the context of models of mathematics story problem solving and previous research on text comprehension.

Keywords: readability, algebra, data-mining, intelligent tutoring system, topic interest

Mathematics story problems, or word problems, are prevalent in mathematics curricula and assessments from kindergarten to undergraduate courses, and there is little evidence that this trend is changing (Jonassen, 2003). They are valuable in psychological research on complex reasoning because they involve both language processing and mathematical reasoning demands. Story problems also represent a primary mechanism through which school-based math is connected to actions and events in everyday and professional life. International comparisons, specifically the Programme of International Student Assessment (PISA; Kelly et al., 2013), accentuate that using mathematics to model contextualized situations that occur in everyday life, society, and in the workplace is a critical skill for economic attainment. However, students in the United States consistently score below international averages on the mathematics PISA.

To date, our understanding of the fine grained influences of *readability* (such as the use of action verbs or the semantic overlap between sentences) on mathematics story problem solving have proven to be elusive, as the structure and variability of natural language is vastly more complex than the grammars for mathematical domains like arithmetic and algebra. *Topic* (such as whether the problem is about banking or traveling) is another aspect of story problems that impacts student engagement and performance, presumably because it is through specific topics that problem solvers' interests are triggered and relevant knowledge is instated. Here again, the sheer scale of topics touched by mathematical story problems can be a formidable barrier to relating language to mathematical reasoning. As Weaver and Kintsch (1988) note, "a knowledge base for college algebra problems would have to comprise a significant segment of human world knowledge. No one today knows how to build such a knowledge base" (p. 6).

This article was published Online First April 27, 2015.

Candace Walkington, Department of Teaching and Learning, Southern Methodist University; Virginia Clinton, Department of Psychology, University of North Dakota; Steven N. Ritter, Carnegie Learning Inc., Pittsburgh, Pennsylvania; Mitchell J. Nathan, Department of Educational Psychology, University of Wisconsin, Madison.

This study was conducted in collaboration with Carnegie Learning, and we would like to particularly thank Steve Fancsali for his contributions to this research. This study was conducted using the online Datashop Repository, available at http://pslcdatashop.org (Koedinger et al., 2010), and we would like to acknowledge the DataShop team. This study made use of the free, online Coh-Metrix tool (McNamara, Louwerse, Cai, & Graesser, 2013), available at http://cohmetrix.com/. We would also like to thank Elizabeth Howell and Alyssa Holland for their assistance with data entry. A portion of the data in this article was previously presented at the 35th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education.

Correspondence concerning this article should be addressed to Candace Walkington, Department of Teaching and Learning, Southern Methodist University, 3101 University Boulevard, Ste. 345, Dallas, TX 75025. E-mail: cwalkington@smu.edu

Recent advances that have combined progress in quantifying the complexities of language structure and topics with the rapid growth of computing power now enable researchers to perform analyses on large corpora of texts that were previously intractable. This has led to the development of powerful, but user friendly text analysis systems, which support hypothesis- driven investigations of the relationship between language and mathematical performance. In this article we use two text analysis systems, CohMetrix (Graesser, Dowell, & Moldovan, 2011; Graesser, McNamara, Louwerse, & Cai, 2004) and Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007) to explore how language within mathematics story problem texts relates to student performance. Here we define story problems as problems involving people, places, and concrete objects from the world—such as problems about the speed of different vehicles, the area of sections of land, or the accumulation of money over time (consistent with Mayer's, 1981 definition of a story problem as having a "story line"). We investigate the language of mathematics word problems in a widely used curriculum using measures of text readability and topic incidence. These measures offer guidance for the development of testable hypotheses regarding the cognitive processes involved as students learn from a math curriculum.

We look at traditional measures of readability like number of words and sentences, as well as measures of text coherence coming from recent research in linguistics and psycho-linguistics. Also critical to students' use of everyday knowledge and the activation of interest is the topic of the story text—whether it references people, places, and things that are familiar to them. A theoretical framework describing the key cognitive and motivational factors associated with story problem solving, as well as a focused review of the prior literature on readability and topic measures, launches our investigation.

Theoretical Framework

Cognitive Factors

The strong relationship between reading comprehension and mathematics problem solving is well known (Hecht, Torgesen, Wagner, & Rashotte, 2001; Lerkkanen, Rasku-Puttonen, Aunola, & Nurmi, 2005). Solving mathematics story problems can be a challenging endeavor because it involves navigating several different types of information. First, the *surface model* of the text is the reader's representation of the text's literal wording (Kintsch & Van Dijk, 1978). Nathan, Kintsch, and Young (1992) proposed a model of mathematics story problem solving where learners coordinate three additional levels of representation: (a) the *textbase* or the propositional information given in the problem text represented as a network of relations; (b) the *situation model* or a mental representations of the relationships, actions, and events in the problem that connects to the reader's prior knowledge to "fill in the gaps left by a sparse story" (Nathan et al., 1992, p. 333); and (c) the *problem model* of formal mathematical operands, numbers, and variables. It is through the coordination of levels that a learner moves from a surface model to the sense-making that mediates the formation of a meaningful answer.

Cognitive theory provides guidance in understanding how learners can be supported in navigating these levels. Such navigation is demanding and can be constrained by cognitive capacity. Cognitive load theory was developed as a framework for understanding and predicting the processing demands that students face during curricular tasks by combining information about the task demands with limitations and strengths inherent in the cognitive architecture (Van Merrienboer & Sweller, 2005). Cognitive load theory differentiates between *extraneous* cognitive load that stems from activities not related to schema acquisition, *intrinsic* cognitive load which relates to the inherent difficulty from interactivity of knowledge elements, and *germane* cognitive load which is the effort students expend to acquire the desired schema (Sweller, van Merrienboer, & Paas, 1998). Reductions in extraneous cognitive load should enhance learning by freeing up cognitive resources, if the schemas are sufficiently challenging. Thus, English language in mathematics story problems that is difficult to read may introduce extraneous cognitive load, which monopolizes working memory resources that could otherwise be devoted to mathematical schema acquisition. This may be especially true when the difficulty with language is unrelated to the mathematical relations. For example, Walkington (2010) relates how seeing the unfamiliar word "greenhouse" disrupted the problem solving of urban youth in an intelligent tutoring system. Language that is clear, consistent, and easy to understand may reduce extraneous cognitive load.

Human capacity for information processing in-the-moment (i.e., working memory) is limited while capacity for storing knowledge schemas in long-term memory is virtually unlimited. Working memory capacity is an important factor in determining performance and achievement, serving as a significantly stronger predictor of academic attainment than standard measures of ability (Alloway & Alloway, 2010). Long-term memory schemas allow multiple elements of information to be categorized as part of a single, higher-level knowledge structure. High-level elements require less working memory for processing than the sum of their constituent low-level elements, thus when used in practice they can reduce the burden on working memory (Kalyuga, Ayres, Chandler, & Sweller, 2003). Story problems that allow learners to draw upon relevant prior knowledge may free up cognitive processing capabilities, allowing for more resources to be devoted to the germane load of learning new concepts.

Goldstone and Son (2005) describe such connections to prior knowledge as providing "grounding" for abstract ideas. One example would be an algebra problem that activates prior knowledge schemas related to accumulating something at a constant rate of change in everyday life (such as distance traveled when driving a car at a constant speed). Integrating prior knowledge of everyday activities with formal knowledge of algebraic structures has a variety of benefits (Koedinger, Alibali, & Nathan, 2008). Grounded representations are more easily accessed in long-term memory and are less prone to errors given the redundant semantic elaborations in long-term memory that can support and verify inferences.

The learners' level of prior knowledge is an important determinant of how various design elements (like the readability or topic of problems) will impact their cognitive processing (Kalyuga et al., 2003). Indeed, Mayer's (2001) *individual differences principle* describes how design elements intended to reduce cognitive load are more important for low knowledge learners because high knowledge learners are better able to use prior knowledge to compensate for less support in the environment. Mayer and Moreno (2003) describe situations where learners experience

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

"cognitive overload" when "processing demands evoked by the learning task may exceed the processing capacity of the cognitive system" (p. 45). Indeed, younger students with mathematics difficulties have a more limited processing speed (Bull & Johnston, 1997). Thus, considerations related to readability and topic may be most important for weaker students and for the solving of more difficult problems, where cognitive overload may be a concern.

Motivational Factors

Story problems and their references to people, events, and actions in the world may also elicit students' interest, enhancing motivation (Walkington, 2013). Hidi and Renninger (2006) define interest as the state of engaging and predisposition to reengage with particular events, objects, or ideas over time. Two types of interest have been identified. *Situational interest* is a spontaneous and transitory reaction to particular features of a learning environment, such as personal relevance, salience, novelty, surprise, or imagery. *Individual interest* refers to enduring, self-initiated predispositions toward engaging with particular objects or ideas. Notably, situational interest appears to be more malleable than individual interest and can transform into individual interest through repeated and prolonged engagement.

A number of text characteristics have been found to be associated with the activation of situational interest, including coherence, completeness, informational complexity, concreteness, ease of comprehension, imageability, suspense, importance and relevance of information, and identification with characters in the text (Schraw, Flowerday, & Lehman, 2001; Schraw & Lehman, 2001). Situational interest can promote persistence and focused attention (Ainley, Hidi, & Berndorff, 2002; Ainley, Hillman, & Hidi, 2002). Features that trigger situational interest, like evocative story titles (Ainley et al., 2002), colorful graphics, or distinctive fonts, are referred to as "catch" interventions, in that their purpose is to temporarily elicit students' interest (Durik & Harackiewicz, 2007). Another method of increasing situational interest is second-person pronouns (e.g., you, your) that place the student as an actor in the text and cue a relaxed, conversational style of language (Mayer, 2009), which can lead to deeper engagement (Mayer, Fennell, Farmer, & Campbell, 2004).

Once triggered, situational interest needs to be maintained over time. Interventions that are designed to "hold" situational interest often are designed to reveal the value of the content to students' lives or goals, or to empower students (Hulleman et al., 2010; Hulleman & Harackiewicz, 2009; Mitchell, 1993). For example, Mitchell (1993) proposed that activities involving group work, computers, and puzzles function as "catch" mechanisms in mathematics, while meaningfulness and involvement "hold" interest. The activation and maintenance of interest has been associated with performance and learning gains (Boscolo & Mason, 2003; Harackiewicz, Durik, Barron, Linnenbrink-Garcia, & Tauer, 2008; Schiefele, 1991). Situational interest may be helpful with story problems because higher levels of interest are positively associated with making connections across different sentences (Clinton & van den Broek, 2012), which strengthens the textbase (Kintsch, 1998).

Eliciting situational interest may be particularly important to two situations. First, when students are confronting an especially challenging task, the focused attention and increased engagement facilitated by situational interest may be especially key in allowing students to persist. Second, activation of interest may be especially important for students with low academic achievement in mathematics classes and low expectations of success in mathematics classes (see Hulleman et al., 2010; Hulleman & Harackiewicz, 2009), as these students are most in need of the attentional and persistence-related resources the activation of interest provides.

Literature Review

Text Readability

Readability measures in nonmathematical texts. Research on text comprehension outside of mathematics suggests that text characteristics influence readability at the surface (wording and syntax), textbase (explicit ideas), and situation model (meaningful representation) levels (Graesser & McNamara, 2011). At the surface level, word difficulty is important to readability. Word polysemy is the number of meanings a word has (e.g., bank can mean a financial institution or the side of a river), and using words with more meanings is negatively associated with comprehensibility (Hagoort, Hald, Bastiaansen, & Petersson, 2004). Word age of acquisition is how early in life one typically learns a word, with words learned later in life being more difficult (Gilhooly & Logie, 1980; Zevin & Seidenberg, 2002). Word concreteness is the level at which one can interact with the concept represented by a word through the senses. A word like *ball* is high in concreteness and a word like *truth* is low in concreteness (Fliessbach, Weis, Klaver, Elger, & Weber, 2006). Concrete words are easier to understand because they are easier to imagine (Paivio, 1991; West & Holcomb, 2000).

At the textbase level, readability is improved with greater ease in connecting different words and ideas from the surface model—by using referents or similar sentences. Referents, such as pronouns, facilitate connections as the learner refers the pronoun with its antecedent, improving reading comprehension (Gernsbacher, 1989; White, 2012). Similarity between sentences increases cohesion (McNamara, Louwerse, McCarthy, & Graesser, 2010) because it is easier to connect ideas in sentences with similar words, meanings, and syntactic structures.

The situation model consists of characters, objects, space, and goals (Zwaan, 1999). Readability measures relevant for the situation model involve the consistency and clarity in which time, space, and cause-and-effect relationships are presented (Graesser & McNamara, 2011). Connectives increase text cohesion because they guide the reader to connect ideas or signal that there will be a discontinuity from previous ideas (Louwerse, 2001), and guide the construction of the situation model (Graesser, McNamara, & Kulikowich, 2011).

Readability measures in mathematical texts. Mathematics story problems came to the attention of many stakeholders in education following the 1983 National Assessment of Educational Progress, which showed that U.S. students had difficulty solving nonroutine problems (Carpenter, Matthews, Lindquist, & Silver, 1984). Around the same period, research on the *Cognitively Guided Instruction* program revealed that slight variations in problem wording result in children using distinct strategies, with the degree to which the story described *action* being an important factor (Carpenter, Fennema, Franke, Levi, & Empson, 1999; Carpenter & Moser, 1984). Concurrent research on elementary school

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

students solving story problems called attention to the issues that young children have with text comprehension, showing that students' mistakes often represent correct answers to misinterpreted stories (Cummins, Kintsch, Reusser, & Weimer, 1988).

However, researchers have found minimal evidence linking traditional measures of readability with problem-solving performance in mathematics (Wiest, 2003). For example, the number of words in a sentence did not affect mathematics problem-solving accuracy nor did the familiarity of the vocabulary words as measured by their frequency in a corpus of texts (Paul, Nibbelink, & Hoover, 1986). Similarly, a meta-analysis of math problems indicated little connection between word and sentence length and accuracy (Hembree, 1992). Traditional measures of readability may provide coarse estimates of a text's difficulty, but they do not capture the *cohesion* of text—the "linguistic glue" that joins the events and ideas, helping readers to understand connections and relationships (McNamara, Graesser, McCarthy, & Cai, 2014).

Recent research on large-scale mathematics assessments has begun to highlight the importance of fine-grained readability measures. Using a large bank of mathematics standardized test items from Grades 4, 7, and 10, Shaftel, Belton-Kocher, Glasnapp, and Poggio (2006) found that use of mathematics vocabulary, polysemous words, and comparative words was associated with greater problem difficulty across all grades. Using widely available data from large-scale assessments (like NAEP and TIMSS) was not the approach used for the research here, however, because readability and topic measures are likely to be strongly associated with student-level motivational factors like situational interest. In the context of a standardized compulsory assessment, the student must complete all problems in a set amount of time, whether or not they are interested in them. This is contrasted with an online curriculum where a student can ask for hints or enter in an incorrect answer to get feedback or a different problem. Rich, in-the-moment traces of student cognition and learning may be best captured by interactions with a mathematics curriculum. We next discuss results relating to readability and topic in this context.

Language comprehension issues are an important determinant of mathematics problem solving for secondary and even postsecondary students (Hall, Kibler, Wenger, & Truxaw, 1989; Koedinger & Nathan, 2004; Walkington, Sherman, & Petrosino, 2012). In a data-mining study of a small number of algebra students working through an intelligent tutoring system, Doddannara, Gowda, Baker, Gowda, and De Carvalho (2011) found that extraneous problem text in algebra story problems, as well as references to concrete people, places, or things, were associated with less concentration and more confusion. However, in a similar study, Baker et al. (2009) found that extraneous text, which often was intended to provide more real world context to increase interest, was associated with fewer unproductive "gaming the system" behaviors. Prior research with students at a low achieving school found that many algebra story problems contained ambiguous wording or confusing vocabulary, and that interpretation issues occurred regularly (Walkington et al., 2012). This underscores the importance of readability and topic measures for students who are struggling with mathematics.

The role of topic for mathematics story problems. The topic of mathematics story problems—the specific kinds of objects and events they reference from the world—may also be related to the formation of a situation model. Algebra students working through online curricula are more likely to accurately solve a story problem with a topic selected to be relevant to their interests (e.g., a "personalized" scenario on playing video games) than a matched story problem that is not selected to be relevant (e.g., a story about harvesting wheat from a field; Walkington, 2013). This is because students may more easily construct a situation model representing the actions and relationships in a story problem if it is about a familiar topic. Such relevant contexts may also elicit students' interest (Hidi & Renninger, 2006), which can facilitate engagement, focus of attention, and use of learning strategies. Inserting familiar referents into a story problem, like the name of a best friend or a favorite food, can improve performance (Anand & Ross, 1987; Cordova & Lepper, 1996; Davis-Dorsey et al., 1991). Using personalized contexts in mathematics curricula has been shown to improve long-term learning (Walkington, 2013). Personalizing mathematics story problems to be relevant to the interests of each individual student may not be possible for a teacher with a heterogeneous classroom of students (Hidi, 1995), or in a curriculum intended to be used with large groups of students. Therefore, it would be useful to identify broad problem topics that support problem solving across many students.

Algebra Story Problem Solving

Algebra is an especially important domain to examine readability and topic measures because Algebra I is considered a gatekeeper course. Indeed, one study found that students who fail Algebra 1 are four times more likely to drop out of high school than those who pass (Orihuela, 2006), and another suggested that students who complete Algebra II are more than twice as likely to graduate from college as students who do not (National Mathematics Advisory Panel, 2008). Many studies have revealed the difficulties students have adopting algebraic symbol systems (Filloy & Rojano, 1989; Herscovics & Linchevski, 1994; Stacey & MacGregor, 1999). One concept that is especially challenging is using symbolic expressions to model real world situations (Bardini, Pierce, & Stacey, 2004; Koedinger & McLaughlin, 2010; Swafford & Langrall, 2000). Students often view algebraic equations as strings of operations rather than statements about equality and have difficulty operating on variables that must be conceptualized as both fixed unknown quantities (e.g., $x + 3 = 7$) and quantities that can vary (e.g., $y = x + 3$; Filloy & Rojano, 1989; Humberstone & Reeve, 2008; Stacey & MacGregor, 1999). When writing expressions, students struggle to navigate the structure and "grammar" of expressions, to symbolically represent their verbal or implicit understandings in expressions, and to understand the utility of using a symbolic expression to represent a relationship (Bardini et al., 2004; Heffernan & Koedinger, 1997; Koedinger & McLaughlin, 2010; Nathan et al., 1992; Swafford & Langrall, 2000). Thus, algebra story problems represent an important class of problems in which issues of constructing situation models and coordinating situational understanding with the problem model are paramount. The task of writing a symbolic expression for a story scenario might be an especially ripe area for investigation, and we will examine this area in the present study.

Research Questions and Hypotheses

Here we examine the texts of a set of algebra story problems solved by students during instruction in schools that are geograph-

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

ically, ethnically, racially, and socioeconomically diverse. Students used an *intelligent tutoring system* known as Cognitive Tutor Algebra (CTA; Ritter, Anderson, Koedinger, & Corbett, 2007). The Cognitive Tutor series is currently used in 3,500 high schools by 650,000 students, and is designed based on 30 years of research on cognitive models of problem solving. CTA tracks students' interactions in detailed logs, and provides as-needed hints and feedback. We address four research questions; the first three involve all students in our sample, while the final research question involves only students from particular schools. Our first two questions are:

1. How is the ***readability*** of algebra story problems associated with problem-solving measures in CTA?

2. How is the ***topic*** of algebra story problems associated with problem-solving measures in CTA?

When solving story problems in CTA, the tutoring program asks students to generate a general algebraic expression and to solve for various specific x and y values. The action of writing an expression from a story—the process of *symbolization*—is a difficult skill that involves the explicit coordination of the situation and problem models, and working directly with the story context. Understanding symbolization motivates our third research question.

3. How do associations for readability and topic vary when focusing on ***symbolization***: the act of translating text into algebraic language?

Finally, the important societal role of algebra as a gatekeeper to advanced studies and college access leads us to consider the special case of students from the lowest performing schools when framing our fourth research question.

4. How do associations for readability and topic vary when focusing on students from the ***lowest performing schools***?

Hypotheses

Based on our theoretical framework, we pose four broad hypotheses about the readability and topic measures that may promote or inhibit performance in CTA. The first three hypotheses discuss three different aspects of the readability and topic that may be important when considering research questions 1 and 2, while the fourth hypothesis discusses differential effects that are relevant for research questions 3 and 4.

Hypothesis 1: Reading the surface model. When reading a story problem, surface features of the text may reduce or increase extraneous cognitive load. These include traditional readability measures like number of words and sentences and the number of words per sentence, as well as measures of word difficulty (like word concreteness and polysemy). We hypothesize that a variety of surface features will reliably predict problem-solving success with these specific predictions: problem solving will be facilitated by shorter and simpler text structure, higher concreteness of words, earlier age of word acquisition; and problem solving will be impaired by the presence of polysemous words.

Hypothesis 2: Forming the textbase. There are also features that may facilitate the construction of the textbase—the length of the text may be important here as well, along with measures of the overlap between sentences and the consistency of information. We hypothesize that problem-solving success will increase when story problem texts are short and contain sentences with overlapping information and consistent and predictable structures, as well as pronouns that are connected to their referents.

Hypothesis 3: Formulating a situation model. When forming a situation model, the student needs to understand, qualitatively, the actions and relationships in the problem text, even if they are not explicitly stated (Nathan et al., 1992). We hypothesize that situation model construction is facilitated when stories are connected to topics the student is interested in and is familiar with, or when students are placed in the story through the use of second person pronouns. Stories with higher incidence of causal verbs, intentional actions, connectives, and active voice, may also promote situation model construction by providing descriptions of clear action.

Hypothesis 4: Differential effects. Students who are already proficient with particular algebra skills have high level schemas in their long-term memory that can be applied to a wide variety of problems; features relating to readability may be less paramount in this situation (Mayer, 2001). Low-achieving students might struggle most with the initial decoding of the problem, given that reading skills track closely with math skills. Thus, our first prediction is that readability measures relating to the surface model and textbase will more strongly predict successful problem solving for students from schools with lower achievement levels. Further, as students write algebraic expressions from stories, they must work directly with the situation model and coordinate it with a problem model, and this is considered a particularly difficult skill in school algebra. Thus, our second prediction is that for the skill of algebraic expression-writing, situation model measures (which include both readability and topic measures) will more strongly predict successful problem solving.

Method

Participants

Data were collected from students across nine high schools and one middle school that use CTA in Algebra I classes. An initial list of 18 schools was selected from CTA database such that the schools had diverse geographic locations and large sample sizes for the 2010 school year. Then this list was narrowed to 10 schools, which were selected to have diverse demographics (see Table 1). These schools contained $N = 3216$ students with active CTA accounts. Carnegie Learning (the company that produces CTA) recommends that students spend 2 class days per week working on the software, with the other 3 days being more typical classroom instruction. The 10 schools were in 10 different states. Three schools had 0%–33% of students eligible for free/reduced price lunch, five had 33%–66% eligible, and two had 66%–100% eligible. Five schools had student populations that were predom-

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 1
Demographic Characteristics of 10 Schools Included in the Study

ID	Math prof %	Math state prof %	Caucasian (%)	African American (%)	Hispanic (%)	F/R lunch (%)	Region	Setting	School type	Reading prof %	Reading state prof %	Number of students included	% English language learners
1	88	70	72	7	15	21	South	Suburb	Middle	98	89	280	7.6
2	81	47	90	4	2	4	South	Suburb	High	96	67	143	1.3
3	95	84	84	10	3	6	South	Urban-suburb	High	94	83	213	4
4	55	46	99	1	1	41	South	Suburb-rural	High	85	66	366	0.4
5	27	Not Avail	20	4	72	77	West	Urban	High	74	84	478	27.5
6	68	59	9	2	88	41	South	Urban	High	79	53	740	19
7	2	31	1	99	1	82	Midwest	Urban	High	21	53	260	0
8	76	84	36	60	2	48	South	Urban	High	90	91	281	0.3
9	39	46	97	1	0	47	South	Rural	High	35	48	182	0.04
10	68	79	38	51	11	62	South	Rural	High	70	83	273	15.9

Note. The "Math prof %" gives the percentage of students from the school who were proficient on the state mathematics exam from 2010, 2011, or 2012 as available from GreatSchools.Org (Algebra I proficiency is given if reported specifically), and the "Reading prof %" column gives the percentage of students proficient in reading or English Language Arts (Grade 9 when available). The subsequent column shows average level of math/reading proficiency for the state. The other columns give the demographics of the entire student body, including the percentage receiving free/reduced (F/R) price lunch.

inantly Caucasian, three were predominantly African American, and two were predominantly Hispanic. Schools also varied in state standardized mathematics assessment scores—three had under 30% of students proficient, four had between 50% and 80% proficient, and three had 80% or more proficient. Based on demographic information, 7% of our sample were English Language Learners—this is similar to the U.S. average. For analyses of the low performing schools (Research Question 4 only), we used students in the three schools with the lowest math proficiency—Schools 5, 7, and 9 (highlighted in Table 1); these schools also had reading proficiency scores below their state's average. Conducting analyses for one of these schools alone resulted in low sample sizes, so the three schools were kept together. CTA does not collect student-level demographic characteristics (like gender or language status); however, two of the three low performing schools (5 and 7) were high-poverty urban with significant minority populations. School 5 had the highest proportion of English language learners (27.5%) in the sample.

Study Environment

Data were collected from the first eight units in CTA that used story problems with linear functions (see Table 2). CTA is adaptive to student needs, so not all students received all problems. More students completed the problems in earlier units (e.g., linear patterns) than in later units (e.g., systems of linear equations), because not all students made consistent progress through the tutor. Story problems with data from fewer than 20 students were omitted—these were largely in the final included unit. In addition, these omitted problems were some problems that were given to very few students, because they were intended only for specific types of remediation. Story problems that had the same cover story, but slightly different numbers were typical and represented different versions of the same problem. Data were collapsed using weighted averages of measures (corrects, incorrects, hints) so all versions of the same problem were included as a single point. After demonstration story problems and story problems with data from fewer than 20 students were omitted, 151 unique story problems remained. On average, each problem was solved by 742 students (SD = 495).

Problem-Solving Measures

The CTA log files from students in the 10 schools were uploaded into Datashop (Koedinger et al., 2010; https://pslcdatashop.web.cmu.edu/), an online repository of student interaction data. In Datashop, data are stored in a consistent format across a number of different technologies, and Datashop provides analysis and modeling tools that operate on such data. The "Performance Profiler" tool compiles summary information for each problem, including how many students solved the problem. The Profiler supplies the percentage of students who on their first attempt to complete a step of the problem, got the step correct, incorrect, or requested a hint.

Each story problem in CTA involves completing multiple steps. For example, the problem in Figure 1 requires 14 steps (to complete cells in two columns and seven rows). The story problems required students to verbally describe the independent and dependent quantities, write a symbolic expression stating the linear relationship(s) in the story, and fill out a table of numerical x and y values. In some units, students were also required to construct graphs. Students attempt to complete each cell and receive immediate feedback on correctness, as well as a diagnosis of their errors. Students may also request hints at each step of the problem, which become progressively specific, eventually "bottoming out" to a hint that tells the student the answer. Because students may attempt a step multiple times and ask for multiple hints on a step, all students will eventually complete each step in the problem. CTA uses mastery learning to control student pacing and problem selection, with mastery determined by Bayesian knowledge tracing (Corbett & Anderson, 1995). In essence, the mastery learning approach means that the relationship between mathematical complexity of problems and student mathematical knowledge is intended to remain constant throughout the tutor, with both increasing at the same rate over time, regardless of what unit the student is in. This differs from an adaptive test where student mathematical knowledge is assumed to remain constant.

For the story problems included, students got the step correct on their first attempt 79.9% of the time (SD = 7.6%), asked for a hint 3.3% of the time (SD = 2.6%), and gave an incorrect answer 16.6% of the time (SD = 5.7%). Because students can request hints on their first attempt, percent correct and percent incorrect

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 2
Random Effects Entered into Regression Models

CTA unit	Sections within unit	Avg. student per prob	# of Prob	Numbers used (# of problems in unit)	Numbers used 2 - from KC models (# of problems in unit)
Linear patterns	1: Finding linear patterns with positive rates of change	914	10	Whole (7)	Large simple (14)
	2: Finding linear patterns with positive and negative rates of change			Large whole (6)	Small simple (7)
		1,167	11	Extra large whole (8)	
Linear models and first quadrant graphs	1: Graphing with positive rates of change	1,384	6	Whole (7)	Large simple (5)
	2: Graphing with positive rates of change and negative starting points			Extra large whole (5)	Simple (2)
	3: Graphing with negative rates of change and positive starting points	1,209	3		Small simple (5)
		1,642	3		
Linear models and independent variables	1: Finding independent variables with positive rates of change	598	8	Whole (9)	Difficult (2)
	2: Finding independent variables with starting points			Extra large whole (3)	Difficult small (12)
	3: Finding independent variables with negative rates of change and starting points	1,322	8	Fraction (3)	Small (12)
		752	12	Decimal (13)	Not specified in KC model (2)
Linear models and ratios	1: Modeling linear functions with ratios	786	11	Fraction (11)	Not specified in KC model (11)
Linear models and four quadrant graphs	1: Graphing with positive integer rates of change			Whole (3)	Difficult (1)
	2: Graphing with positive fractional rates of change	821	6	Extra large whole (8)	Difficult small (1)
	3: Graphing with negative rates of change			Fraction (3)	Large (2)
		870	6	Decimal (3)	Simple (3) Small (4)
		1,272	5		Not specified in KC model (6)
Linear models and slope-intercept graphs	1: Graphing Given an integer slope and Y-intercept			Whole (5)	Difficult (12)
	2: Graphing given a fractional slope and Y-intercept	568	4	Extra large whole (1) Fraction (5)	Simple (4)
		281	12	Decimal (5)	
Linear models and the distributive property	1: Modeling with integer rates of change	677	5	Whole (4)	Difficult (2)
	2: Modeling with fractional rates of change	616	5	Large whole (1)	Simple (5)
	3: Modeling using the distributive property over division	499	6	Extra large whole (4)	Not specified in KC model (16)
	4: Modeling more complex equations	347	7	Fraction (5) Decimal (9)	
Systems of linear equations	1: Solving linear systems involving integers	246	16	Whole (1) Large whole (2)	Not specified in KC model (23)
	2: Solving linear systems involving decimals	321	7	Extra large whole (13) Fraction (2) Decimal (5)	

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

are not redundant, and separate predictions can be made for each. The correlation between % correct and % incorrect was −0.89, the correlation between % hint and % incorrect was 0.49, and the correlation between % correct and % hint was −0.69. The median amount of time spent on each problem was 314 s (approximately 5 min).

Coh-Metrix Analysis of Text Readability

Coh-Metrix is a software tool that provides numerous, varied, and precise measures of text readability (McNamara, Louwerse, Cai, & Graesser, 2013). The development of Coh-Metrix was grounded in theories of text comprehension regarding cohesion and coherence (Graesser, Singer, & Trabasso, 1994; Graesser et al., 2004). Coh-Metrix provides measures related to the surface code, textbase, and situation model (McNamara et al., 2014). Measures related to the surface code assess the difficulty of the words and syntax. Measures related to the textbase assess the ease of connecting different ideas in the text to each other. Measures related to the situation model assess the consistency of various dimensions of the mental representation of the text such as causation, time, and space (McNamara, Graesser, McCarthy, & Cai, 2014). The text of the introduction of each story problem was entered into the Coh-Metrix 3.0 software. The introduction describes the main text of the story problems where the mathematical relationships are developed (see Figure 1, top left). The remainder of the text instructs students to write an expression, and poses

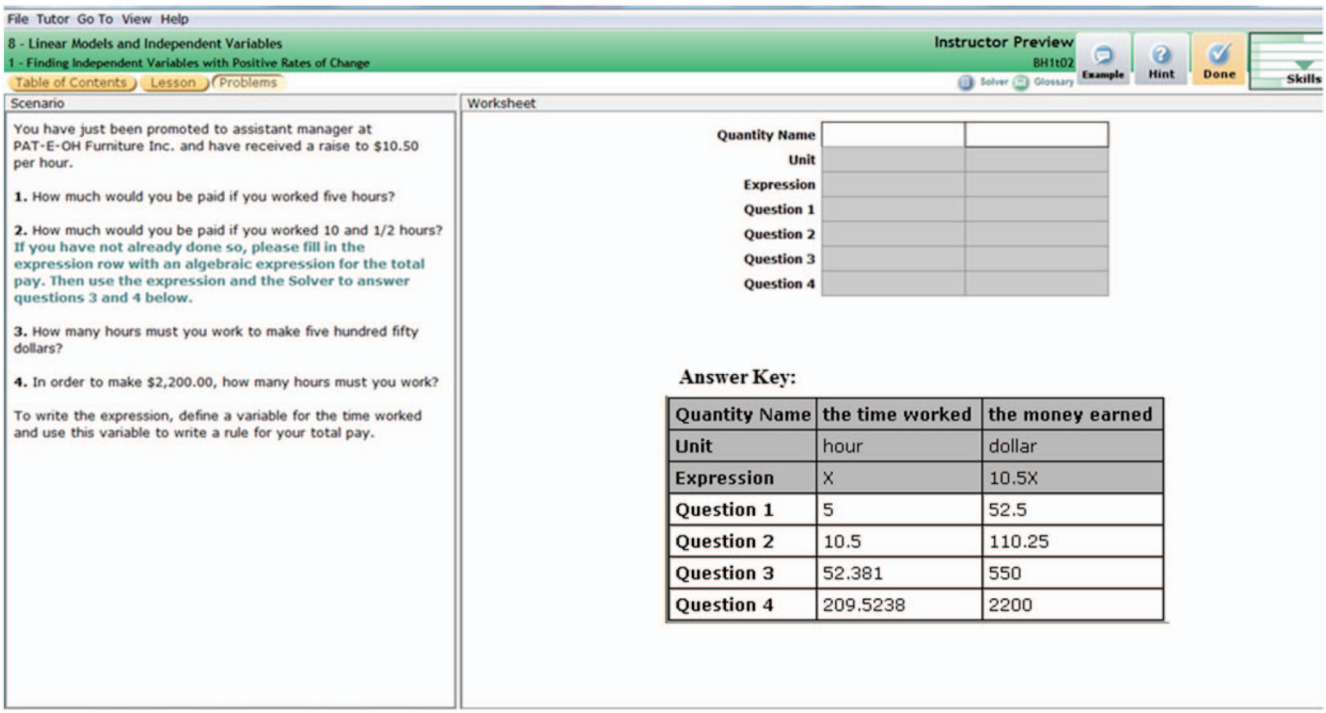

Figure 1. Screenshot of story problem in CTA. The "Answer Key" table has been superimposed over the screenshot to show correct answers to each step. In addition to answers shown, the tutor will accept equivalent expressions, numeric values or linguistic phrases. See the online article for the color version of this figure.

questions giving specific x and y values. Problems were cleaned such that all periods not denoting the end of a sentence (e.g., decimals, abbreviations) were removed.

Because Coh-Metrix provides so many measures of readability, we first examined which measures had significant correlations with problem-solving measures (% correct, % incorrect, % hint). The readability measures that had significant correlations only were then tested for significance in regression models (described later). For some Coh-Metrix analyses, we omitted the 31 story problems in our sample that had an introduction that was only one sentence long, leaving us with a final sample size of 120 story problems for these analyses. This is because some interesting and important measures of readability presuppose multiple sentences (e.g., measuring the degree of semantic overlap between sentences). Appendix A.1 contains tables with all significant correlations between Coh-Metrix measures and corrects, incorrects, and hints.

LIWC Analyses of Topic Incidence

To identify topics of problems, we used LIWC (Pennebaker et al., 2007), a dictionary-based computerized text analysis program that counts words in over 70 categories. Categories are grouped into three superordinate categories: grammatical (e.g., verbs, pronouns, articles), psychological (e.g., affect, cognitive, perceptual), or personal concerns (e.g., family, work, death). Words in each category were compiled from English dictionaries, Roget's Thesaurus, and rating scales used in psychological research (e.g., PANAS by Watson, Clark, & Tellegen, 1988), as well as through brainstorming sessions by small groups of judges. Word lists were then reviewed by judges who determined through majority whether a word should be included.

We used LIWC to identify the topics of the story problems by determining whether each story problem contained words that related to the different topic categories. Specifically, we used LIWC to identify story problems about the following topics: social processes (family, friends, people), affective processes (positive emotions and negative emotions), biological processes (body, health, ingestion), cognitive processes (insight, causation, discrepancy, tentativeness, certainty, inhibition, inclusive/exclusiveness), perceptual processes (see, hear, feel), relativity processes (motion, space, time) and personal concerns (work, achievement, leisure, home, and money). This is a novel, yet appropriate use of LIWC given previous work regarding word use and life domain topics (e.g., Robinson, Navea, & Ickes, 2013; Tov, Ng, Lin, & Qiu, in press).

To more accurately identify the topics, we first needed to remove certain words from the LIWC dictionaries. These words were typically polysemous with meanings intended to be relevant to mathematical content, not the life domain topic. We removed foot and feet from biological processes, value and values from affective processes, names of shapes (square, triangle, circle, and rectangle) from perceptual processes, and words beginning with "numb" from biological processes and affective processes because of confounding issues with "number." Then, the percentages of words in each of the life domain topics were calculated. If LIWC identified a nonzero percentage of words in a particular topic, we

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

dummy coded a story problem as pertaining to that particular topic. This coding was especially important given the general brevity of the story problems in contrast to texts typically analyzed by LIWC (Chung & Pennebaker, 2012). In addition, when the LIWC categories are used in this binary manner, their reliability is much higher (Pennebaker et al., 2007). We first looked for significant correlations between LIWC measures and problem-solving measures, and then included only measures with significant correlations as candidates for the regression models. Appendix A.2 contains tables with all significant correlations between LIWC measures and problem-solving measures.

Development and Implementation of Regression Models

The Coh-Metrix and LIWC measures that had significant correlations to one or more of the problem-solving measures were entered into their respective mixed-effects linear regression models. The models were fit using the *lmer* command (Bates, Maechler, Bolker, & Walker, 2014) in the R software package. This function is useful for analyzing CTA datasets because it has the flexibility to handle mixed effects data that is partially crossed, partially nested, and unbalanced (see Bates, 2010). The dependent measures in the models were average percent correct, average percent incorrect, and average percent hints on each problem. The sample size was not the $N = 3{,}216$ students in the study; it was the $N = 151$ problems (analysis of all problems) or $N = 120$ problems (multisentence only analysis). For Research Questions 1–3, Coh-Metrix and LIWC predictors were fit in separate models, as the Coh-Metrix analysis was done only with multisentence problems, and the LIWC analysis was done with all problems. However, analyses with the Coh-Metrix measures in the multisentence dataset were also done with the inclusion of significant LIWC predictors, and analyses with the LIWC measures in the dataset that included all problems were also done with the inclusion of significant Coh-Metrix predictors. Although results of these joint models are not included here for brevity, generally the same pattern of results held when the models included both types of measures simultaneously.

Each problem's problem-solving measures were averaged for all students who solved that problem and across all problem parts (except in the analysis of expression-writing, which involved only a single problem part) before being entered into the models and thus each problem was a single data point. Although this is a conservative approach to the analysis, we believe that the generalizability and significance of our results ultimately should be limited most by the number of story problems we had at our disposal to test; thus using the number of story problems as a reduced sample size seems appropriate.

Control variables that were entered as candidates into regression models included aspects of the story's mathematical characteristics, including the structure of the linear function (e.g., positive slope/no intercept, negative slope/negative intercept) and the type of numbers used (e.g., fractions, large whole numbers) modeled as fixed effects, and the unit and section the problem came from, modeled as random intercept terms. Random intercepts are assumed to be independent and identically distributed with mean 0 and variance τ^2. QQ-plots for random effects were examined; the first and last units in CTA sometimes showed departures from the reference line on the QQ-plots because they were substantially more difficult than the other units (in Unit 1, students are learning to use the software, and in Unit 12, they are solving complex problems about systems of equations). To examine the degree to which these two more difficult units were driving effects for readability/topic, we conducted all analyses without problems from these units included, and the pattern of results did not change. We examined VIFs for quantitative predictors in the final models to check for multicollinearity—all VIFs were below 2.0. We also examined relevant residual plots and found no violations of the independence assumption. For the homoscedasticity assumption, four of the final models had fan-shaped residual plots when residuals were plotted against fitted values. In each case, a log transformation of the dependent variable fixed the issue. We report untransformed models here for ease of interpretability, as we found that in the transformed models the results were unchanged. Finally, we examined the predictors for linearity with the outcome measures; although relationships were generally weak, we did not see any compelling evidence that a fit other than a linear model would be appropriate, except in the case of one predictor—number of sentences in the story problem.

The number of sentences in a story problem ranged from one to nine ($Mean = 2.76$, $SD = 1.51$), and there was not much differentiation in performance among higher values of number of sentences. Thus, we created a variable that collapsed between levels: one sentence, two sentences, three sentences, and four or more sentences. We also controlled for two additional aspects that captured unique ways in which story contexts interact with linear functions. First, linear functions with negative intercept terms are especially challenging because the "start" value in the story is negative. Often these problems had a special sentence clarifying that the intercept was negative (e.g., instructions clarifying that distance below the ground was considered negative). Similarly, sometimes the calculated answers were not realistic in the context of the story (e.g., 3.4 people), and in these cases there would be a sentence in the text reminding students to answer with respect to the "algebraic model" (rather than their own assumptions of feasibility—e.g., in one such problem, according to the table, an oceanographer is lowered below the bottom of a sea shelf). This was added as a control. These fixed effects were only kept in the model when they were significant.

We also attempted to add in additional effects related to which "knowledge components" (KCs) the story problem corresponded to in CTA's cognitive model. The only KCs that were meaningfully associated with problem-solving measures related to the numbers (e.g., difficult, small) used in the problem. This is not surprising, given that unit and section information gives most of the information on the KCs a problem covers. Similarly, the fixed effect for the structure of the linear function was never significant in a model, likely because this information is well captured by unit and section; this predictor is not discussed further. Both fixed effects for the type of numbers used in the story problem were entered into the models (see Table 2).

Other predictors included the measures from LIWC and Coh-Metrix that had a significant correlation with problem-solving measures. Predictors were tested for inclusion in the models using the *anova()* command in R on nested models, using the full maximum likelihood estimator (FEML), as recommended by the *lmer()* documentation for model selection procedures (Bates,

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

2010). This is a likelihood ratio test that uses a chi-square distribution to test for significant reductions in model deviance. Once model selection was complete, the reduced maximum likelihood estimator (REML) was used for the reporting of the final model. The REML estimator has the advantage of reducing error in the estimate of the variance component (see Bates, 2010).

First, random effects were added to the models, and all random effects that passed the likelihood ratio test for significance were retained.[1] Next, fixed effects were added in one by one. These included predictors that had significant correlations with the problem-solving measure being modeled as well as additional control variables (e.g., Numbers). The *anova()* command was used on each predictor to test whether it reduced the model deviance significantly. When a predictor was significant, it was added to the model, and all remaining predictors were tested again to determine if their inclusion was now warranted in the new model. This process continued until none of the remaining predictors significantly improved model fit. Interactions between predictors were not considered due to the limited sample size.

Ninety-five percent confidence intervals were computed for the regression coefficients. It was also of interest to quantify the proportion of variance explained by the readability or topic measures, in order to estimate the size of the effect. Xu's (2003) metric Ω^2, which gives the percentage of reduction in residual variance between a null model and a full model, was used. Here, the null model was a model with the random effects and any significant control variables. The full model had these variables and the readability and topic[2] predictors. It is important to note that we can only find significant effects depending on the degree to which our 151 story problems actually vary on readability and topic categories. If there were fewer than 10 problems that fell into a category, that category was eliminated. In Appendix B, we provide summary data for how many of our story problems fell into different categories.

Results

We organize our results according to our four research questions, and return to each of our hypotheses in the Discussion section.

RQ1: How is Readability of Story Problems Associated With Problem-Solving Measures?

For the story problems with multiple sentences[3] ($N = 120$), correlations were calculated between Coh-Metrix indices and percent correct, incorrect, and hint (Appendix A.1). A total of six, 12, and 16 predictors had significant correlations to corrects, incorrects, and hints, respectively. Regression results (see Table 3) use the reference category of a two-sentence problem. There are significant differences between three and four or more sentences not shown in the table for correct ($B = -4.20$, $p = .0019$), incorrect ($B = 2.43$, $p = .0190$), and hints ($B = 1.60$, $p = .0003$). Moving from a three sentence problem to a four or more sentence problem seems to be a critical transition—it is associated with a reduction in correct responses by an estimated 4.2% (95% CI [1.59, 6.80]), an increase in incorrects by an estimated 2.43% (95% CI [0.41, 4.46]), and an increase in hints by an estimated 1.60% (95% CI [0.74, 2.45]). Also, there are significantly more hint requests for a four or more sentence problem, compared with a two-sentence problem ($p < .001$). When interpreting regression coefficients, recall that each model contains a variety of covariates, and that the coefficients are partial coefficients.

Third-person singular pronouns (e.g., he, she) are associated with significantly more correct answers and significantly fewer hints and incorrects. This predictor is the number of third-person singular pronouns that occur for every 1,000 words; for most problems it varied from 0 to 100. Comparing a problem with no third-person singular pronouns to a problem that has 10% of its words as third-person singular pronouns, correct answers are higher by an estimated 4.0% (95% CI [1.00, 6.90], incorrect answers are lower by an estimated 2.5% (95% CI [0.10, 4.84]), and hint requests are lower by 1.6% (95% CI [0.63, 2.58]). An example of one such story problem with a pronoun incidence of 107 per 1,000 words is: "A training sumo wrestler Tu Fatmo weighs 470 pounds. He is 80 pounds below his ideal fighting weight. He can safely gain four and one half pounds per week." However, an important caveat is that our data are limited with respect to the pronouns the problems contained—there were not a lot of problems with first-person pronouns (see Appendix B). Thus, we can only conclude that third-person singular pronouns are associated with higher accuracy and lower hint seeking than the comparison group, which contained problems with no pronouns, third-person plural pronouns, second-person pronouns, and (a few) first-person pronouns.

Further, a higher standard deviation of the amount of semantic overlap between adjacent sentences is associated with significantly fewer correct answers. In other words, if some adjacent sentences are very similar to each other and contain similar words and types of words, while other adjacent sentences are very different from each other, accuracy tends to be lower. The standard deviation of the semantic overlap between adjacent sentences in the data set varied from 0 standard deviations[4] to approximately 0.3 standard deviations. Interpreting the coefficient of −15.25 (95% CI [−2.42, −28.09]), a story where the amount of overlap between sentences varied greatly ($SD = 0.3$) was compared with one with little variation in overlap ($SD \approx 0$), the model estimates that correct answers would be 4.6% lower for the problem with greater variation in overlap. However, the confidence interval suggests the size of the effect is difficult to predict with precision.

[1] We also fit models where all control variables (structure, numbers, unit, section) were modeled as random effects and were retained in the model regardless of their significance level (see Barr, Levy, Scheepers, & Tily, 2013). Results were similar, so here we report only models where effects resulted in significant reductions in deviance.

[2] For LIWC analysis, number of sentences was considered a control variable and was included in the null model.

[3] Conducting this analyses for all problems (to include single-sentence problems) had similar results for measures that did not presuppose multiple sentences, but we were not able to look at any of the measures that presuppose multiple sentences.

[4] This measure was exactly 0 for all two-sentence stories in the data set, because there could only be one overlap between sentences, and therefore no potential variance. However, as we had controlled for number of sentences, this was not driving the effect.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 3
Regression Results for Using Coh-Metrix Readability Measures to Predict Problem-Solving Performance Measures (Corrects, Incorrects, Hints) for Story Problems With Multiple Sentences Only (N = 120 Problems)

	% Correct	% Incorrect	% Hint
Random components			
Unit (variance)	18.84	10.05	4.30
Residuals (variance)	26.73	16.98	2.87
Fixed effects			
(Intercept)	80.43 (2.06)***	17.20 (1.56)***	2.99 (0.86)**
Algebraic model language (control variable)			−1.67 (0.83)*
Negative intercept term (control variable)		3.18 (1.37)*	
Numbers—Whole (control variable)	(ref.)	(ref.)	(ref.)
Numbers—Large whole (control variable)	−4.74 (2.32)*	3.35 (1.83)	0.24 (0.76)
Numbers—Extra large whole (control variable)	0.37 (1.41)	−0.69 (1.13)	−0.94 (0.46)*
Numbers—Fraction (control variable)	−0.27 (1.73)	−1.08 (1.37)	0.91 (0.57)
Numbers—Decimal (control variable)	3.73 (1.49)*	−2.80 (1.19)*	−0.75 (0.49)
Two sentences	(ref.)	(ref.)	(ref.)
Three sentences	1.94 (1.41)	−1.28 (0.97)	0.44 (0.40)
Four or more sentences	−2.25 (1.74)	1.16 (1.13)	2.0 (0.48)***
Third-person singular pronoun incidence	0.040 (0.015)**	−0.025 (0.012)*	−0.016 (0.005)**
Standard deviation of semantic overlap of adjacent sentences	−15.25 (6.48)*		
Overall reduction in residual variance due to readability predictors (Ω^2)	21.10%	9.18%	20.18%

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

The stories that have high variation in the amount of overlap tended to have a lone sentence that was conceptually dissimilar from all other sentences. An example of a story problem that scores highly on this indicator is: "Ms. Williamson woke up one morning to find her basement flooded with water. She called two different plumbers to get their rates. The first plumber charges $75 just to walk in the door plus $25 an hour. The second plumber charges a flat $40 an hour." The first sentence is disconnected from its adjacent sentence—the first sentence has the proper noun of the character's name, incorporates the action of waking up and the time of day of morning, and gives the basement as a location. All of this information is dropped in the next sentence. However, the second and third sentence and third and fourth sentence have more overlap in their ideas, and the final two sentences in particular are quite similar. Thus, the variance in the amount of overlap is high ($SD = 0.375$). This can be contrasted with an example of a story problem that has a low score on this indicator ($SD = 0.025$): "You have just become CEO (chief executive officer) of a company that is heavily in debt. The company's balance sheet currently shows a balance of −$525,000. The company is paying the debt off at the rate of $12,500 per month." Here the sentences contain more consistent information about business, finance, and debt. In summary, problems with three or fewer sentences, third-person singular pronouns, and consistent sentence overlap are associated with higher performance levels.

RQ2: How is Topic of Story Problems Associated With Problem-Solving Measures?

For all problems ($N = 151$), we calculated the correlation between LIWC measures of topic and corrects, incorrects, and hints (see correlations in Appendix A.2; see number of problems in each category in Appendix B). A total of four, three, and eight predictors had significant correlations to corrects, incorrects, and hints, respectively. Regression results are shown in Table 4, and examples of problems in the different LIWC topic categories can be found in Appendix C. Contexts involving work words are associated with significantly fewer correct answers and stories with motion words are marginally associated with fewer correct answers.[5] Mayer's (1981) analysis of types of algebra problems commonly found in textbooks suggests that these are common topics in curricula. Having at least one word relating to work is associated with 2.01% (95% CI [0.05, 3.98]) fewer correct answers.

A story containing any words involving social processes is associated with significantly fewer incorrect answers; having at least one of these words is associated with a predicted 1.79% (95% CI [0.11, 3.47]) reduction in incorrects. Social words include references to family, friends, and humans, as well as socializing, having parties, making calls, sending messages, and so forth. However, the list of words in this category also includes pronouns—this may be problematic, given that the Coh-Metrix analyses showed that third-person singular pronouns have a significant negative association with incorrects. We thus ran the models again with a control for third-person singular pronouns. Third-person singular pronouns were not significantly associated with incorrect answers ($p = .203$).

Health-related words and inhibition words are both associated with hint requests, and the presence of any such words was associated with an increase in hints by a predicted 1.66% (95% CI [0.65, 2.67]) and 1.01% (95% CI [0.30%, 1.73%]), respectively. Stories involving health words were about diseases and medication, while inhibition words were used in stories about saving or losing money or getting a discount and stories involving safety. These are dry contexts that may be disconnected from adolescent experiences. In summary, work and motion topics are associated

[5] Model selection warranted including a predictor for motion words, but the p value in the final model was $p = .086$.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 4
Regression Results for Using Linguistic Inquiry and Word Count (LIWC) Topic Measures to Predict Problem-Solving Performance Measures (Corrects, Incorrects, Hints) for All Story Problems (N = 151 Problems)

	% Correct	% Incorrect	% Hint
Random components			
Unit (variance)	22.86	6.28	4.28
Residual (variance)	27.68	17.71	2.40
Fixed effects			
(Intercept)	82.15 (2.70)***	13.96 (2.10)***	2.65 (0.84)**
Numbers—Whole (control variable)	(ref.)	(ref.)	(ref.)
Numbers—Large whole (control variable)	−5.93 (2.51)*	4.99 (1.97)*	0.82 (0.63)
Numbers—Extra large whole (control variable)	−0.77 (1.62)	0.75 (1.28)	−0.36 (0.38)
Numbers—Fraction (control variable)	−2.60 (1.78)	0.09 (1.37)	1.08 (0.48)*
Numbers—Decimal (control variable)	1.16 (1.73)	−1.06 (1.37)	−0.40 (0.40)
Numbers2—None (control variable)	(ref.)	(ref.)	
Numbers2—Simple (control variable)	−0.66 (2.09)	1.68 (1.88)	
Numbers2—Small (control variable)	5.44 (2.36)*	1.25 (2.10)	
Numbers2—Large (control variable)	6.28 (4.37)	1.86 (3.50)	
Numbers2— Simple small (control variable)	−8.34 (3.64)*	8.92 (2.75)*	
Numbers2—Difficult small (control variable)	1.52 (2.71)	3.84 (2.10)	
Numbers2—Simple large (control variable)	−4.96 (3.55)	4.85 (2.60)	
Algebraic model language (control variable)			−1.72 (0.76)*
Negative intercept term (control variable)		2.86 (1.37)*	
One sentence	(ref.)		(ref.)
Two sentences	2.99 (1.45)*		−0.41 (0.41)
Three sentences	3.86 (1.53)*		−0.55 (0.44)
Four or more sentences	−1.09 (1.74)		0.97 (0.50)
Social		−1.79 (0.85)*	
Work	−2.01 (0.99)*		
Motion	−1.77 (1.02)		
Health			1.66 (0.51)**
Inhibition			1.01 (0.36)**
Overall reduction in residual variance due to topic predictors (Ω^2)	4.18%	3.93%	11.83%

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

with lower accuracy, whereas social contexts were associated with higher accuracy. Health and financial contexts were associated with increased hint seeking.

RQ3: How do Associations Vary When Focusing on Writing Symbolic Equations From a Story?

We next examined the relationship between readability and topic measures and problem-solving measures for the single step of writing the algebraic expression only. In order to do these analyses, we had to export a step-level dataset from Datashop (rather than use the Performance Profiler) where one row of the data set was one student solving one step of one problem. We focus on the Coh-Metrix results for this analysis; the results for LIWC were similar to the analysis of the full data set, so we just cite them briefly. For the Coh-Metrix analyses, we use the 120 problems with multiple sentences, and for the LIWC analysis we use all 151 problems.

When writing the algebraic expression, students entered the correct answer on their first attempt only 42.5% (*SD* = 18.6%) of the time. This illustrates that throughout CTA, expression-writing is one of the most difficult KCs. They requested a hint 7.3% (*SD* = 6.7%) of the time and entered an incorrect answer 50.2% (*SD* = 21.9%) of the time. Regression results are in Table 5.

For the expression-writing step, number of sentences was only related to hints, with significantly fewer hints requests for three sentence problems compared with four sentences or more sentences (contrast not shown in table; 5.06% fewer hint requests, 95% CI [2.2, 7.9]). Using words that have many different meanings (i.e., polysemous words) was associated with both significantly fewer correct answers and significantly more incorrect answers. Consider the problem in our dataset: "An open pit copper mine is 1,550 feet deep and the company estimates that it is getting deeper at the rate of seven feet per month. Assume the number of feet below the surface is a negative number." Here the average number of meanings of each content word is 6.375—"mine" can mean something that belongs to you, an explosive, or something in a mountain; "pit" can mean the center of a plum, or a large hole; "feet" can mean a body part or a unit of measurement, and so forth. By comparison, in the problem "On Tuesday morning at 7 a.m. the residents of Bar Harbor Maine awoke to six inches of snow on the ground. The snow fell at the average rate of one half inch per hour during the storm;" the average number of meanings of each content word is 2.342—each word is relatively specific and well defined. Average word polysemy, or the number of meanings on average each content word in the story had, ranged from two to six. The *B* values shown in Table 5 suggest that *each* additional meaning the average content word has is associated with correct answers decreasing by an estimated 4.60% (95% CI [0.7, 8.5]) and incorrect answers increasing by 4.64% (95% CI [1.1, 8.2]). In the extreme case in the contrasting examples we supplied where there

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 5
Regression Results for Using Coh-Metrix Measures to Predict Algebraic Expression-Writing Performance (Corrects, Incorrects, Hints) for Multisentence Story Problems (N = 120 Problems)

	% Correct	% Incorrect	% Hint
Random components			
Unit (variance)	0.71	0.32	
Section (variance)			0.074
Residual (variance)	3.49	2.92	0.325
Fixed effects			
(Intercept)	82.8 (9.6)***	23.48 (8.0)**	5.97 (1.11)***
Two sentences			(ref.)
Three sentences			−1.97 (1.54)
Four or more sentences			2.13 (1.84)
Word polysemy	−4.60 (1.98)*	4.64 (1.78)*	
Gerund density	−0.261 (.083)**		
Standard deviation of words per sentence[1]	−1.01 (.48)*		
Standard deviation of semantic overlap of adjacent sentences[1]			17.08 (7.17)*
Standard deviation of content word overlap of adjacent sentences[1]	−54.6 (18.9)**		
Overall reduction in residual variance due to predictors (Ω^2)	19.84%	4.21%	4.34%

[1] For two sentence problems, these measures were again always 0 because there was no standard deviation. However, the hint model controlled for number of sentences, and forcing the predictor for number of sentences into the model for correct answers (even though it was not significant) did not change the results.
* $p < .05$. ** $p < .01$. *** $p < .001$.

is a difference of 4 points in polysemy, an associated 18% difference in accuracy would be predicted. However, as can be seen from the confidence interval, the size of this effect is quite noisy and difficult to predict with precision.

A gerund is a verb ending in *–ing* that functions as a noun. An example of a problem with gerunds is "We are *going* on a trip on the Pennsylvania Turnpike *traveling* east. We get on the Turnpike at the Monroeville entrance which is at milepost number 56 *meaning* that it is located 56 miles from the western (Ohio) end of the Turnpike. We drive at 55 mph *starting* at milepost number 56" (gerund density = 74.1). More gerunds in a story were associated with significantly fewer correct answers. The incidence score of gerunds (out of 1,000 words) ranged from 0 to 100. The *B* coefficient suggests that every 10 points the incidence score increases is associated with correct answers decreasing by 2.61% (95% CI [0.96, 4.25]). Taking the extreme example we gave, moving from a story with no gerunds to a story with a gerund score of 70 would be associated with a predicted 18.2% decrease in correct answers. Again, the confidence interval for the size of this effect is quite large.

We found several sentence measures associated with accuracy on expression writing. Having stories where there is a high standard deviation in the number of words in each sentence is associated with significantly fewer correct answers. In other words, stories with some sentences that are short and other sentences that are long are associated with poorer performance, with each standard deviation of difference associated with a reduction in correct answers by an estimated 1.01% (95% CI [0.07, 2.0]). Higher standard deviations of semantic overlap are associated with significantly more hints, with a one standard deviation difference associated with an increase in hints by an estimated 17.08% (95% CI [2.9, 31.3]). However, this measure only actually ranged from 0 to 0.4 standard deviations in our data set and the size of this effect is quite noisy and difficult to predict with precision. This finding suggests that stories are associated with more hints if they contain some sentences that are semantically similar, and others that are very dissimilar. We also found that higher standard deviations of content word overlap between sentences is associated with fewer correct answers. This measure varies slightly from the semantic overlap measure, in that it considers only exact word matches and controls for the length of the sentences. A one standard deviation difference was associated with decreasing correct answers by an estimated 54.6% (95% CI [17.2, 92.0]). Standard deviation of content word overlap ranged from 0 to 0.4 standard deviations, and the confidence interval is again quite large and noisy.

LIWC analyses showed that inhibition words (i.e., financial contexts) were associated with more incorrect answers ($B = 9.99$, $SE = 3.90$, $p = .0117$), more hints ($B = 6.22$, $SE = 1.28$, $p < .001$), and fewer correct answers ($B = -13.83$, $SE = 4.77$, $p = .005$). Motion words were associated with more hints ($B = 4.18$, $SE = 1.05$, $p < .001$). In summary, for expression-writing problem parts, polysemous words, gerunds, dissimilar and inconsistent sentences, and financial and motion contexts were associated with lower accuracy and/or more hint seeking.

RQ4: How do Associations Vary When Focusing on the Lowest Performing Schools?

Our final analysis involves looking at the relationship between readability and topic measures and performance for students at the three lowest performing schools in our sample. Here the sample size is smaller, because fewer students at these schools had attempted some of the problems—in all, we had 129 problems remaining in our sample. Because of this reduced sample size, for our Coh-Metrix analyses we decided to keep all the problems together, rather than analyze only multisentence problems. Although this would not allow us to detect findings related to relationships between sentences, it would give us the most power and generality when detecting other readability relationships. This

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

also allowed us to analyze LIWC and Coh-Metrix data in the same model. Performance was slightly lower for students at these schools—students entered an incorrect answer on their first attempt 21.2% of the time ($SD = 6.8\%$), requested hints 3.4% of the time ($SD = 2.3\%$), and entered correct answers 75.4% of the time ($SD = 8.5\%$); $N = 866$ students from these schools were included.

Although we had expected readability and topic measures to be more important for students at low performing schools, overall we found far fewer measures with significant correlations that we could test for inclusion in our models. It may be that at these schools, student-level variation is so high, that detecting effects for characteristics of problems, especially nonmathematical characteristics, is particularly difficult. As shown in Table 6, regressions results show that third-person singular pronouns are associated with fewer incorrect answers, more correct answers, and fewer hints. Stories with more words are associated with more hints, with each additional word in the story increasing the predicted probability of seeking a hint by 0.0207% (95% CI [0.008, 0.033]). Further, the level of concreteness of the words in the story has a negative association with hints, with more concrete words associated with fewer hints. Word concreteness ranged from 300 to 500, and a 100-point increase in concreteness was associated with a predicted 0.62% drop in hint requests (95% CI [0.15, 1.10]). An example of a problem with high concreteness is: "A huge mirror for a telescope is being moved by a truck with 13 axles and 50 tires from Erie Pennsylvania to Raleigh North Carolina. The truck averages 15 mph and has already traveled 60 miles" (word concreteness = 514). An example of a story problem with low concreteness is: "A company has total assets of $575,000. It estimates that these assets are increasing at the rate of $6,500 per week" (word concreteness = 329).

For LIWC predictors, we found that tentative words (e.g., if, assume, approximate) are associated with a predicted 2.12% increase in incorrect answers (95% CI [0.141, 4.10]) and a predicted 2.34% decrease correct answers (95% CI [0.064, 4.62]). Story problems with tentative words would often verbally ask students to make some sort of mathematical assumption, such as assuming there are 365 days in a year, that a distance is negative, or that a rate of change would continue to be constant. These instructions seemed especially challenging for students in these schools. Also, although not shown in Table 6, we found that when predicting incorrects, models that included the presence of social words as a predictor were roughly equivalent to models that included third person singular pronouns as a predictor in terms of model fit parameters. However, when both predictors were in the model together, neither was quite significant. Thus, third-person singular pronouns and social words seem to be measuring similar aspects of the story problem. In summary, for students in low-performing schools, stories with third-person singular pronouns were associated with higher accuracy, while having fewer words overall, and concrete words were associated with fewer hints. Tentative words that asked students to make assumptions were associated with lower accuracy.

Replication Study

To check whether the results generalized to another curriculum, we sought out data from *MATHia* (Carnegie Learning, 2012), an intelligent tutoring system that also implements mastery learning approaches for middle school mathematics. We used data from three units in Course 3 (Grade 8): Linear Models and Slope-Intercept Graphs, Linear Models in General Form, and Linear Models and Multiple Representations. These units cover functions of the form $y = mx + b$ and $ax + by = c$, and they were selected because they were the only *MATHia* units on linear functions that Carnegie Learning had available data for.

The *MATHia* problems were structurally similar to CTA problems—students wrote symbolic equations and solved these equations for particular x and y values after being given a problem introduction that described a linear relationship. However, the characteristics of the text of these problems were often different. These story problems were written more recently, and *MATHia* makes explicit attempts to personalize its instruction to students' out-of-school interests. The stories often had more informal and even humorous language, updated pop culture references, and fewer complex "realistic" applications of algebra. The *MATHia* dataset came from seven middle schools and two combined middle/high schools in six states. Like our original dataset, schools varied in their achievement levels, racial/ethnic makeup, and students eligible for free/reduced lunch. Once problems with fewer than 20 students were omitted, $N = 60$ problems remained, and

Table 6
Regression Results for Using Coh-Metrix Measures to Predict Performance Measures (Corrects, Incorrects, Hints) for Students in Low-Performing Schools (N = 129 Problems)

	% Correct	% Incorrect	% Hint
Random components			
Unit (variance)	27.37	12.96	3.24
Residual (variance)	35.60	27.32	1.30
Fixed effects			
(Intercept)	76.63 (2.01)***	19.95 (1.44)***	5.37 (1.23)***
Negative intercept term (control variable)	−5.27 (1.82)*	5.02 (1.59)**	
Number of words			.0207 (.0062)**
Third-person singular pronoun incidence (WRDPRP3s)	.0411 (.0163)*	−.035 (.0142)*	−.0100 (.0032)**
Word concreteness			−.0062 (.0024)*
Tentative	−2.34 (1.15)*	2.12 (1.00)*	
Overall reduction in residual variance due to predictors (Ω^2)	7.77%	7.48%	16.58%

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

each problem was solved by 103 students on average (SD = 43 students).

Although this is a small sample size for a replication, it is ideal to have problems relating to only one mathematical topic, as this means less variance in their difficulty. We again looked for significant correlations; however, the criteria for a significant correlation for a sample size of 60 problems (r = .255) was too large to be feasible for most readability and topic measures. Because we were just looking into the promise of previously found results, rather than generating new results and hypotheses, we kept the magnitude of the correlation needed for significance the same as it had been in the full model of 151 problems—we considered correlations above r = .16 to be significant. Given the small sample size, we did not attempt to fit regression models and only examined significant correlations to identify promising associations.

The results from the replication of the Coh-Metrix analyses are shown in Table 7. One of the most interesting results is that it appeared in this new data set that longer story texts were associated with *higher* performance. However, this seemed to be due to a skewed distribution of story text length among the 60 problems—one problem had one sentence, 42 problems had two or three sentences, 12 problems had four sentences, and five problems had five to six sentences. Given the relatively small sample size of problems with four sentences, and the scarcity of problems at the tails (with one sentence or more than four sentences), this was not an ideal set for replication. In addition, recall that our original analysis revealed a curvilinear trend (we found the transition from a three to four or more sentence problem is most critical for performance) that might not be well captured by a simple linear correlation in a small dataset. In the dataset, four-sentence problems did have the highest rate of hints, but overall trends were inconsistent.

The number of sentences is a surface-level characteristic of the text with implications for cognitive load that should matter even in the context of a large-scale standardized assessment; thus, this measure may be one of the few that would actually function similarly on a standardized test versus an online curriculum. Consequently, we investigated this finding using N = 190 released eighth grade problems from the National Assessment of Educational Progress.[6] We found that our original finding held, with more sentences and words being associated with fewer correct answers (r = −0.228 and r = −0.318, respectively). Also similar to our prior findings, we found that the move from three sentences to four or more sentences is associated with a large drop in accuracy (from 55% to 41%).

Coh-Metrix results relating to standard deviation in sentence overlap, third person singular pronouns, and concrete words all replicated in the *MATHia* data set. However, results relating to polysemous words and gerunds did not, and actually went in the opposite direction as the original data set.[7] While the original result for gerunds may have been spurious, we note that another large study of math problems found that polysemous words are associated with lower performance (Shaftel et al., 2006), so the original result may have been valid despite the lack of replication. In particular, *MATHia*'s attempts to use more familiar, interesting, and relatable contexts may have made these measures less important. In another study, we found that Flesh-Kincaid readability measures had a smaller impact on student performance when the context of the story problem is selected to be relevant to students out-of-school interests (Walkington & Sherman, 2012).

The results from the replication of the LIWC analysis are shown in Table 8. Although there was not a large enough variety of problems to investigate the replication for social words, we found a similar effect for problems with home words—references to home life are associated with fewer hints. Similarly, work topics were associated with more hint seeking. Inhibition and tentative words had been associated with more hint seeking and lower accuracy in the original dataset, respectively; here results are mixed. Although these topics are associated with more hints, they are also associated with more corrects and/or fewer incorrects. Finally, results for motion words are not replicated. We can conclude that our main finding that familiar contexts (with social or home references) improve problem-solving measures, while less familiar business or work contexts are associated with greater difficulty as evidenced by more hint seeking, generally holds up in another data set.

Discussion

We used a database of problems in Cognitive Tutor Algebra and associated measures of problem readability and topic incidence with student problem-solving measures. As the database contained 151 problems and the analysis was correlational (rather than an intervention), there are limitations to this method. In addition, although we offered replications and only tested predictors with significant correlations to outcome measures rather than all possible predictors (methods similar to those used in other LIWC/Coh-Metrix studies), Type I error is an issue in these analyses. We sought to carefully balance Type I and Type II error; Type II error is particularly important to minimize in our context given our purpose of generating hypotheses about potentially important readability measures.

Another limitation of this study is that we only examined algebra story problems involving linear functions. It is unclear whether our findings would generalize to story problems of other types. Analyzing story problems that all cover a similar concept has advantages—we were able to specify at a fine-grained level different aspects of the mathematical structure. It is important to do readability and topic studies on both broad sets of story problems covering many concepts, and on narrow sets of story problems covering particular concepts. Given that our bank of problems was limited, we did not have complete coverage on all readability/topic measures, and were not able to examine interactions. However, we identified a number of relationships that give directions for future study. We now frame our discussion according to our hypotheses, highlighting the relationships that were in accordance with hypotheses and that replicated in multiple datasets as being the most powerful.

[6] We used only eighth grade problems that were in "real world" contexts. However, results did not change regardless of the subset of problems used.

[7] In the *MATHia* dataset, the variance of the gerund and polysemy measures were slightly lower than in the Cognitive Tutor dataset.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 7
Description of Replication of Coh-Metrix Findings in MATHia Dataset of N = 60 Problems

Relationship tested	Coh-Metrix finding in original data set(s)	Replicate?	Coh-Metrix finding in new data set
Text length vs. performance	**Longer texts** are associated with lower accuracy and more hints.	Partial	The trends in this data set were actually the opposite, with correlations suggesting longer texts were better. However, dataset was small, with few observations at tails (1 sentence and 5+ sentences), especially given that this is likely a curvilinear trend. As a result, we used NAEP problems instead and replicated original result.
Deviation of sentence similarity vs. performance	Having some sentences that are **highly dissimilar** and others that are **highly similar** is associated with lower accuracy more hints.	Yes	A higher standard deviation of the latent semantic overlap between sentences was associated with more hints (r = .283). There were 31 problems with three or more sentences.
Third person singular pronouns vs. performance	**Third person singular pronouns** are associated with greater accuracy fewer hints.	Yes	Of the 60 problems, 31 contained third person singular pronouns, and this measure was associated with fewer hints being sought (r = −0.300)
Concrete words vs. performance	**Concrete words** are associated with fewer hints.	Yes	More concrete words were associated with less hint seeking (r = −0.299).
Polysemous words vs. performance	**Polysemous words** are associated with lower accuracy.	No	Polysemous words were actually associated with fewer hints being sought (r = −0.193).
Gerunds vs. performance	**Gerunds** are associated with lower accuracy.	No	Gerunds were in 35 problems and were actually associated with more correct answers (r = .206) and fewer hints (r = −0.243).

Hypothesis 1: Reading the Surface Code

We found that ***word difficulty*** was associated with problem-solving measures, with word polysemy associated with decreased accuracy and word concreteness associated with less hint seeking. Polysemous words may be troublesome given that many mathematics terms have different meanings outside of mathematics (Lager, 2006). Although the polysemy finding did not replicate in the MATHia dataset, we still highlight this finding as promising because it is in line with the findings of a recent study of student performance on large-scale standardized mathematics assessments (Shaftel et al., 2006). There was also evidence that using more concrete words reduced hint-seeking behaviors, which would suggest they were easier for students, and contradicts Doddannara et al.'s (2011) smaller study in CTA. These findings correspond to well-known effects in the text comprehension literature (Nagy & Townsend, 2012) that concrete words can be easier to comprehend because they are easier to imagine (Fliessbach et al., 2006).

We also found that the ***length of the text*** in algebra story problems—the number of sentences and number of words—seemed to have an important relationship to performance, with longer texts typically associated with more hint seeking and lower accuracy. This corresponds to more general research on the difficulty of reading passages, and longer texts may impact both the surface model and the textbase. In traditional readability analyses, the number of words in a sentence is positively associated with difficulty (Deane, Sheehan, Sabatini, Futagi, & Kostin, 2006); more words indicates more ideas in a sentence the reader must track (Daneman & Carpenter, 1980). However, here there was a curvilinear relationship in which an increase in the number of sentences appeared to have little relationship to performance until the number of sentences was four or greater, which inhibited performance. Three sentences may provide enough context to ground the story problem, but not so much text as to overwhelm the student.

Hypothesis 2: Forming a Propositional Textbase

Pronouns are referents that show connections among different parts of the textbase; here, ***third-person singular pronouns*** were associated with higher accuracy and less hint seeking. The use of pronouns likely facilitated connections between sentences as the students determined the referent of pronoun (Graesser et al., 2011; Graesser & McNamara, 2011; White, 2012). Third-person singular pronouns are helpful to comprehension as they are relatively simple to resolve assuming there is only one possible antecedent that the third-person singular pronoun could be referring to (Gernsbacher, 1989). Third-person singular pronouns may be indicative that there is a single character in a story problem, which may ease comprehension as students need only track the actions of one person or entity (Gernsbacher, Robertson, Palladino, & Werner, 2004).

We also found that ***measures of the variability in sentence similarity*** have an important association with accuracy and hint seeking for multisentence problems. Having a problem in which some sentences are very similar to each other and one or two sentences are very different—in terms of length, semantics, and words used—is associated with lower accuracy and more hint seeking. This finding is likely because these differences make forging connections among all of the sentences in a story problem difficult (Graesser & McNamara, 2011) and may inhibit the construction of the textbase and the situation model (McNamara et al., 2010).

Hypothesis 3: Forming a Situation Model

We found that ***familiarity of the problem topic*** to adolescents is associated with performance. Contexts involving work and finance are associated with lower accuracy, while contexts involving home life/socializing are associated with higher accuracy. Health care and finance contexts are associated with more hint seeking. Al-

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 8
Description of Replication of Linguistic Inquiry and Word Count (LIWC) Findings in MATHia Dataset of N = 60 Problems

Relationship tested	LIWC finding in original data set(s)	Replicate?	LIWC finding in new data set
Social words vs. performance	**Social** words associated with higher accuracy.	Yes	The correlation of **social** words with correct answers was $r = .24$; however, 52 of the 60 problems contained social words.
Work words vs. performance	**Work** words associated with lower accuracy.	Yes	Work words had a positive correlation with seeking hints ($r = .177$), and appeared in 40 of the 60 problems.
Inhibition words vs. performance	**Inhibition** words associated with lower accuracy more hints.	Partial	**Inhibition** words were positively associated with seeking hints ($r = .191$), but were negatively associated with incorrect answers ($r = -0.187$). They were in 10 problems.
Motion words vs. performance	**Motion** words were associated with lower accuracy and more hints.	No	**Motion** words were in 28 problems, but did not have significant associations with performance.
Tentative words vs. performance	**Tentative** words were associated with lower accuracy.	Partial	**Tentative** words were in 15 problems. They were associated with more hints ($r = .214$), but they were also associated with more corrects ($r = .186$) and fewer incorrects ($r = -0.223$)
Health words vs. performance	**Health** words are associated with more hints.	Not enough data	Only one problem contained a **health** word.

though the idea that "relevant" contexts support learning has a long history in mathematics education (see Walkington et al., 2012, for a review), little research has examined what specific contexts might be more or less relevant to large groups of adolescents. This finding offers support for research on personalized learning, which has found improved performance when matching problem topics to student interests (Walkington, 2013). We did not find evidence that use of the active voice, connectives, intentional actions, or causal verbs facilitated performance—these factors may be less important in mathematical contexts where the situation model is coordinated with quantitative information, than they have been in reading research.

Hypothesis 4: Differential Effects

When we limited our analyses to students at schools with the lowest achievement, or to the problem parts that involved algebraic expression-writing, we did find some differences in the readability measures that were important. In accordance with our hypothesis, number of words and word concreteness—measures relating to the surface model and textbase—were only important for lower achieving students, in that they were associated with hint seeking. These surface code measures may be especially important as students who are weaker readers are initially evaluating a problem text and determining if they need the support of a hint. LIWC measures relating to the topic of the problem (and perhaps to the formation of the situation model) were less important for these students. For students struggling with mathematics, immediate reductions in cognitive load through simpler, easier to parse problem texts may be most critical. Contrary to our expectations, situation model measures (including topic) were not particularly important when students wrote algebraic expressions—instead we found that word polysemy became important, which is a surface level characteristic of the text. When students have a weaker math background or are solving a difficult problem part (like expression-writing), the mathematical demands of the situation might cause cognitive overload. Although we would expect more relevant, accessible contexts to provide access for these weaker students, these supports might not have been enough to move students over the first hurdle of formulating a surface model and textbase.

Related Research Studies

Although this is the only major analysis of specific readability and topic measures in the context of a mathematics curricula that we are aware of, other research is examining these issues in the context of standardized testing. In ongoing analyses of the math story problems from the fourth and eighth grade released NAEP and TIMSS tests, we have replicated many of our readability results from Coh-Metrix. In particular, we have found again that word difficulty, length of the text, and pronouns all are important factors that influence student performance (see Walkington, Clinton, Shivraj, & Yovanoff, 2015). We are also examining interactions between student and problem characteristics and the impact of readability measures on the NAEP and TIMSS. Prior work suggests that gender (Boaler, 1994; Cordova & Lepper, 1996), socioeconomic status (Ladsen-Billings, 1995; Cooper & Harries, 2005), attitudes toward and achievement in mathematics (Walkington, Petrosino, & Sherman, 2013; Walkington, Cooper, & Howell, 2013), and English language proficiency (Khisty, 1995) may all impact how students respond to story problems. In addition, the difficulty of the problem and the particular skill it assesses may moderate effects (Walkington et al., 2013). Finally, the impact of readability may differ based on the students' familiarity with the problem's topic (Walkington & Sherman, 2012).

The LIWC findings regarding topic do not seem to replicate in a standardized testing context—likely because, as we described earlier when reviewing the literature, issues of interest activation are less critical. However, the result that familiar contexts or topics are associated with higher performance on algebra word problems, has been recently replicated in the context of online curricula (Walkington, 2013). This study did not look at readability measures or particular problem topics, though, and only did broad comparisons based on whether the problem topic had been specifically selected to be relevant to students' interests.

Implications

There are many extratextual factors that undoubtedly influence performance when solving mathematics story problems, including the student's mathematical background and prior knowledge, the mathematical characteristics of the problem, the characteristics of

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

the school and teacher, and the type of instruction received. Being able to detect significant, measureable effects for readability and topic incidence measures given the myriad of other potentially important factors that influence achievement is important. The value of this kind of exploratory study is to develop testable hypotheses about cognitive processes involved in mathematical reasoning. A key future direction will be to conduct intervention studies where readability and topic measures are modified for groups of students within curricula, and the impact on performance, interest, and learning outcomes is examined. These studies will move the findings away from being strictly correlational and provide stronger evidence for the effect of readability and topic factors on performance. We close by describing some implications.

First, this research makes strides to expand current work on computerized text analysis tools to understand and improve performance. Text analysis tools have been under used in mathematics education, and have great utility for understanding how mathematical language is used in large data sets without manual coding. Mathematics story problems represent specific literary and pedagogical genres (Gerofsky, 2009) whose language follows important norms and regularities. With this research we follow Nathan et al.'s (1992) work in forging stronger connections between the fields of text comprehension and mathematics education. We found that readability and topic measures have relevance to understanding students' mathematical performance, and point to the need for more research that bridges these two fields.

Second, these analyses offer teachers and curriculum designers some initial ideas for how to write story problems that their students find accessible and understandable. It is important that students leave their math classes being able to handle lengthy, semantically complex, high-vocabulary story problems that describe topics that they do not necessarily find interesting. However, when introducing an important, new mathematical idea to a student, keeping the readability measures at a manageable level and using accessible and engaging topics might be particularly valuable. This way, students can move immediately to grappling with and gaining an understanding of the mathematics itself, rather than struggling with the verbal language. Engaging topics that reflect familiar social and home-based situations appear to particularly help students to be autonomous (seek out fewer hints). Early successes using situational interest as an appealing entry point may have long-term advantages for achievement (see Walkington, 2013).

In the language of our theoretical framework, problems with high levels of readability and relevant topics have the potential to reduce cognitive load and elicit situational interest. These stories may allow students to more easily construct a situation model of the actions and relationships, which can in turn support and enhance their problem model of the mathematical processes. Perhaps, as students gain expertise with the mathematical concepts, additional layers of difficulty in the form of complex readability distractors and less familiar topics can be added on to story problems. Students can gain expertise in formulating a situation model from a very complex text, perhaps relying more and more on the problem model to support situation model construction. Similarly, as expertise is gained, the layers of language features that support access can also be completely stripped away, leaving only mathematical notation and abstraction.

We envision curricular sequences for learning math concepts where students begin with simple, verbal problems on concrete and familiar topics, and then transition to both more complex and semantically difficult stories as well as completely abstract symbolic formats. This type of progressive item design could play a key role in the development of learning trajectories for students' mathematics development (Confrey & Maloney, 2010; Simon, 1995). Such a sequence could be a powerful mechanism to ensure that all students, regardless of their background characteristics and prior knowledge, gain the critical, initial access to the mathematical ideas that will provide a foundation for future learning.

References

Ainley, M., Hidi, S., & Berndorff, D. (2002). Interest, learning, and the psychological processes that mediate their relationship. *Journal of Educational Psychology, 94,* 545–561. http://dx.doi.org/10.1037/0022-0663.94.3.545

Ainley, M., Hillman, K., & Hidi, S. (2002). Gender and interest processes in response to literary texts: Situational and individual interest. *Learning and Instruction, 12,* 411–428. http://dx.doi.org/10.1016/S0959-4752(01)00008-1

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology, 106,* 20–29. http://dx.doi.org/10.1016/j.jecp.2009.11.003

Anand, P., & Ross, S. (1987). Using computer-assisted instruction to personalize arithmetic materials for elementary school children. *Journal of Educational Psychology, 79,* 72–78. http://dx.doi.org/10.1037/0022-0663.79.1.72

Baker, R. S., de Carvalho, A. M. J. A., Raspat, J., Aleven, V., Corbett, A. T., & Koedinger, K. R.(2009, July). Educational software features that encourage and discourage "gaming the system." *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 475–482.

Bardini, C., Pierce, R., & Stacey, K. (2004). Teaching linear functions in context with graphics calculators: Students' responses and the impact of the approach on their use of algebraic symbols. *International Journal of Science and Mathematics Education, 2,* 353–376. http://dx.doi.org/10.1007/s10763-004-8075-3

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278. http://dx.doi.org/10.1016/j.jml.2012.11.001

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R*. Retrieved from http://lme4.r-forge.r-project.org/book

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7*. Retrieved from http://CRAN.R-project.org/package=lme4

Boaler, J. (1994). When do girls prefer football to fashion? An analysis of female underachievement in relation to 'realistic' mathematics contexts. *British Educational Research Journal, 20,* 551–564. http://dx.doi.org/10.1080/0141192940200504

Boscolo, P., & Mason, L. (2003). Topic knowledge, text coherence, and interest: How they interact in learning from instructional texts. *Journal of Experimental Education, 71,* 126–148. http://dx.doi.org/10.1080/00220970309602060

Bull, R., & Johnston, R. S. (1997). Children's arithmetical difficulties: Contributions from processing speed, item identification, and short-term memory. *Journal of Experimental Child Psychology, 65,* 1–24. http://dx.doi.org/10.1006/jecp.1996.2358

Carnegie Learning. (2012). *Carnegie learning math series: Carnegie learning MATHia software*. Retrieved from http://mathseries.carnegielearning.com/product-info/software

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Carpenter, T., Fennema, E., Franke, M., Levi, L., & Empson, S. (1999). *Children's mathematics: Cognitively guided instruction*. Portsmouth, NH: Heinemann.

Carpenter, T., Matthews, W., Lindquist, M., & Silver, E. (1984). Achievement in mathematics: Results from the national assessment. *The Elementary School Journal, 84,* 484–495. http://dx.doi.org/10.1086/461379

Carpenter, T., & Moser, J. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal for Research in Mathematics Education, 15,* 179–202. http://dx.doi.org/10.2307/748348

Chung, C. K., & Pennebaker, J. W. (2012). Linguistic Inquiry and Word Count (LIWC): Pronounced "Luke,". . . and other useful facts. In P. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 206–229). Hershey, PA: IGI Global.

Clinton, V., & van den Broek, P. (2012). Interest, inferences, and learning from texts. *Learning and Individual Differences, 22,* 650–663. http://dx.doi.org/10.1016/j.lindif.2012.07.004

Confrey, J., & Maloney, A. (2010, June). The construction, refinement, and early validation of the equipartitioning learning trajectory. In K. Gomez, L. Lyons, & J. Radinsky (Eds.) *Learning in the disciplines: Proceedings of the 9th International Conference of the Learning Sciences* (Vol. 1, pp. 968–975). Chicago, IL: International Society of the Learning Sciences.

Cooper, B., & Harries, T. (2005). Making sense of realistic word problems: Portraying working class 'failure' on a division with remainder problem. *International Journal of Research & Method in Education, 28,* 147–169. http://dx.doi.org/10.1080/01406720500256228

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4,* 253–278. http://dx.doi.org/10.1007/BF01099821

Cordova, D., & Lepper, M. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology, 88,* 715–730. http://dx.doi.org/10.1037/0022-0663.88.4.715

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20,* 405–438. http://dx.doi.org/10.1016/0010-0285(88)90011-4

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19,* 450–466. http://dx.doi.org/10.1016/S0022-5371(80)90312-6

Davis-Dorsey, J., Ross, S., & Morrison, G. (1991). The role of rewording and context personalization in the solving of mathematical word problems. *Journal of Educational Psychology, 83,* 61–68. http://dx.doi.org/10.1037/0022-0663.83.1.61

Deane, P., Sheehan, K. M., Sabatini, J., Futagi, Y., & Kostin, I. (2006). Differences in text structure and its implications for assessment of struggling readers. *Scientific Studies of Reading, 10,* 257–275. http://dx.doi.org/10.1207/s1532799xssr1003_4

Doddannara, L. S., Gowda, S. M., Baker, R. S., Gowda, S. M., & De Carvalho, A. M. (2011). Exploring the relationships between design, students' affective states, and disengaged behaviors within an ITS. *Proceedings of the 16th International Conference on Artificial Intelligence and Education*, 31–40.

Durik, A., & Harackiewicz, J. (2007). Different strokes for different folks: How individual interest moderates effects of situational factors on task interest. *Journal of Educational Psychology, 99,* 597–610. http://dx.doi.org/10.1037/0022-0663.99.3.597

Filloy, E., & Rojano, T. (1989). Solving equations: The transition from arithmetic to algebra. *For the Learning of Mathematics, 9,* 19–25.

Fliessbach, K., Weis, S., Klaver, P., Elger, C. E., & Weber, B. (2006). The effect of word concreteness on recognition memory. *NeuroImage, 32,* 1413–1421. http://dx.doi.org/10.1016/j.neuroimage.2006.06.007

Gernsbacher, M. A. (1989). Mechanisms that improve referential access. *Cognition, 32,* 99–156. http://dx.doi.org/10.1016/0010-0277(89)90001-2

Gernsbacher, M. A., Robertson, R. R., Palladino, P., & Werner, N. K. (2004). Managing mental representations during narrative comprehension. *Discourse Processes, 37,* 145–164. http://dx.doi.org/10.1207/s15326950dp3702_4

Gerofsky, S. (2009). Genre, simulacra, impossible exchange, and the real: How postmodern theory problematises word problems. In B. Greer, L. Verschaffel, W. Van Dooren, & S. Mukhopadhyay (Eds.), *Word and worlds: Modelling verbal descriptions of situations* (pp. 21–38). Rotterdam, the Netherlands: Sense Publishers.

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods and Instrumentation, 12,* 395–427. http://dx.doi.org/10.3758/BF03201693

Goldstone, R., & Son, J. (2005). The transfer of scientific principles using concrete and idealized simulations. *Journal of the Learning Sciences, 14,* 69–110. http://dx.doi.org/10.1207/s15327809jls1401_4

Graesser, A. C., Dowell, N., & Moldovan, C. (2011). A computer's understanding of literature. *Scientific Study of Literature, 1,* 24–33. http://dx.doi.org/10.1075/ssol.1.1.03gra

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

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

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,* 193–202.

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

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304,* 438–441. http://dx.doi.org/10.1126/science.1095455

Hall, R., Kibler, D., Wenger, E., & Truxaw, C. (1989). Exploring the episodic structure of algebra story problem solving. *Cognition and Instruction, 6,* 223–283. http://dx.doi.org/10.1207/s1532690xci0603_2

Harackiewicz, J., Durik, A., Barron, K., Linnenbrink-Garcia, E., & Tauer, J. (2008). The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest, and performance. *Journal of Educational Psychology, 100,* 105–122. http://dx.doi.org/10.1037/0022-0663.100.1.105

Hecht, S. A., Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2001). The relations between phonological processing abilities and emerging individual differences in mathematical computation skills: A longitudinal study from second to fifth grades. *Journal of Experimental Child Psychology, 79,* 192–227. http://dx.doi.org/10.1006/jecp.2000.2586

Heffernan, N. T., & Koedinger, K. R. (1997). The composition effect in symbolizing: The role of symbol production vs. text comprehension. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the nineteenth annual meeting of the Cognitive Science Society* (pp. 307–312). Mahwah, NJ: Erlbaum, Inc.

Hembree, R. (1992). Experiments and relational studies in problem solving: A meta-analysis. *Journal for Research in Mathematics Education, 23,* 242–273. http://dx.doi.org/10.2307/749120

Herscovics, N., & Linchevski, L. (1994). A cognitive gap between arithmetic and algebra. *Educational Studies in Mathematics, 27,* 59–78. http://dx.doi.org/10.1007/BF01284528

Hidi, S. (1995). A reexamination of the role of attention in learning from text. *Educational Psychology Review, 7,* 323–350. http://dx.doi.org/10.1007/BF02212306

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Hidi, S., & Renninger, K. (2006). The four-phase model of interest development. *Educational Psychologist, 41,* 111–127. http://dx.doi.org/10.1207/s15326985ep4102_4

Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackiewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology, 102,* 880–895. http://dx.doi.org/10.1037/a0019506

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science, 326,* 1410–1412. http://dx.doi.org/10.1126/science.1177067

Humberstone, J., & Reeve, R. (2008). Profiles of algebraic competence. *Learning and Instruction, 18,* 354–367.

Jonassen, D. (2003). Designing research-based instruction for story problems. *Educational Psychology Review, 15,* 267–296. http://dx.doi.org/10.1023/A:1024648217919

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38,* 23–31. http://dx.doi.org/10.1207/S15326985EP3801_4

Kelly, D., Nord, C. W., Jenkins, F., Chan, J. Y., & Kastberg, D. (2013). *Performance of US 15-Year-Old Students in Mathematics, Science, and Reading Literacy in an International Context. First Look at PISA 2012. NCES 2014–024*. Washington, DC: National Center for Education Statistics.

Khisty, L. L. (1995). Making inequality: Issues in language and meanings in mathematics teaching with Hispanic students. In W. G. Secada, E. Fennema, & L. B. Adajian (Eds.), *New directions for equity in mathematics instruction* (pp. 279–297). Cambridge, UK: Cambridge University Press.

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

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85,* 363–394. http://dx.doi.org/10.1037/0033-295X.85.5.363

Koedinger, K. R., Alibali, M. W., & Nathan, M. J. (2008). Trade-offs between grounded and abstract representations: Evidence from algebra problem solving. *Cognitive Science, 32,* 366–397. http://dx.doi.org/10.1080/03640210701863933

Koedinger, K. R., Baker, R. S. J. D., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. D. Baker (Eds.), *Handbook of educational data mining* (pp. 43–56). Boca Raton, FL: CRC Press. http://dx.doi.org/10.1201/b10274-6

Koedinger, K., & McLaughlin, E. (2010). Seeing language learning inside the math: Cognitive analysis yields transfer. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 471–476.) Austin, TX: Cognitive Science.

Koedinger, K., & Nathan, M. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *Journal of the Learning Sciences, 13,* 129–164. http://dx.doi.org/10.1207/s15327809jls1302_1

Ladsen-Billings. (1995). Making mathematics meaningful in multicultural contexts. In W. Secada (Ed.), *For equity in mathematics education* (pp. 126–145). New York, NY: Cambridge University Press.

Lager, C. A. (2006). Types of mathematics-language reading interactions that unnecessarily hinder algebra learning and assessment. *Reading Psychology, 27,* 165–204. http://dx.doi.org/10.1080/02702710600642475

Lerkkanen, M. K., Rasku-Puttonen, H., Aunola, K., & Nurmi, J. E. (2005). Mathematical performance predicts progress in reading comprehension among 7-year olds. *European Journal of Psychology of Education, 20,* 121–137. http://dx.doi.org/10.1007/BF03173503

Louwerse, M. (2001). An analytic and cognitive parametrization of coherence relations. *Cognitive Linguistics, 12,* 291–316.

Mayer, R. E. (1981). Frequency norms and structural analysis of algebra story problems into families, categories, and templates. *Instructional Science, 10,* 135–175. http://dx.doi.org/10.1007/BF00132515

Mayer, R. (2001). *Multimedia learning*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139164603

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511811678

Mayer, R. E., Fennell, S., Farmer, L., & Campbell, J. (2004). A personalization effect in multimedia learning: Students learn better when words are in conversational style rather than formal style. *Journal of Educational Psychology, 96,* 389–395. http://dx.doi.org/10.1037/0022-0663.96.2.389

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38,* 43–52. http://dx.doi.org/10.1207/S15326985EP3801_6

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511894664

McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (2013). *Coh-Metrix version 3.0*. Retrieved from http://cohmetrix.com

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47,* 292–330. http://dx.doi.org/10.1080/01638530902959943

Mitchell, M. (1993). Situational interest: Its multifaceted structure in the secondary school mathematics classroom. *Journal of Educational Psychology, 85,* 424–436. http://dx.doi.org/10.1037/0022-0663.85.3.424

Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly, 47,* 91–108. http://dx.doi.org/10.1002/RRQ.011

Nathan, M., Kintsch, W., & Young, E. (1992). A theory of algebra-word-problem comprehension and its implications for the design of learning environments. *Cognition and Instruction, 9,* 329–389. http://dx.doi.org/10.1207/s1532690xci0904_2

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U. S. Department of Education.

Orihuela, Y. (2006). *Algebra I and other predictors of high school dropout*. Retrieved from http://digitalcommons.fiu.edu/dissertations/AAI3249717/

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology, 45,* 255.

Paul, D. J., Nibbelink, W. H., & Hoover, H. D. (1986). The effects of adjusting readability on the difficulty of mathematics story problems. *Journal for Research in Mathematics Education, 17,* 163–171. http://dx.doi.org/10.2307/749299

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. Austin, TX: LIWC. Net.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review, 14,* 249–255. http://dx.doi.org/10.3758/BF03194060

Robinson, R. L., Navea, R., & Ickes, W. (2013). Predicting final course performance from students' written self-introductions: A LIWC analysis. *Journal of Language and Social Psychology, 32,* 469–479. http://dx.doi.org/10.1177/0261927X13476869

Schiefele, U. (1991). Interest, learning, and motivation. *Educational Psychologist, 26,* 299–323. http://dx.doi.org/10.1080/00461520.1991.9653136

Schraw, G., Flowerday, T., & Lehman, S. (2001). Increasing situational interest in the classroom. *Educational Psychology Review, 13,* 211–224. http://dx.doi.org/10.1023/A:1016619705184

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Schraw, G., & Lehman, S. (2001). Situational interest: A review of the literature and directions for future research. *Educational Psychology Review, 13,* 23–52. http://dx.doi.org/10.1023/A:1009004801455

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11,* 105–126. http://dx.doi.org/10.1207/s15326977ea1102_2

Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education, 26,* 114–145. http://dx.doi.org/10.2307/749205

Stacey, K., & MacGregor, M. (1999). Learning the algebraic method of solving problems. *The Journal of Mathematical Behavior, 18,* 149–167. http://dx.doi.org/10.1016/S0732-3123(99)00026-7

Swafford, J., & Langrall, C. (2000). Grade 6 students' pre-instructional use of equations to describe and represent problem situations. *Journal for Research in Mathematics Education, 31,* 89–112. http://dx.doi.org/10.2307/749821

Sweller, J., van Merrienboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10,* 251–296. http://dx.doi.org/10.1023/A:1022193728205

Tov, W., Ng, K. L., Lin, H., & Qiu, L. (in-press). Detecting well-being via computerized content analysis of brief diary entries. *Psychological Assessment*.

van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17,* 147–177. http://dx.doi.org/10.1007/s10648-005-3951-0

Walkington, C. (2010). "Playing the game" of story problems: Situated cognition in algebra problem solving (Doctoral dissertation). University of Texas, Austin, TX.

Walkington, C. (2013). Using learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology, 105,* 932–945. http://dx.doi.org/10.1037/a0031882

Walkington, C., Clinton, V., & Howell, E. (2013). The associations between readability measures and problem solving in algebra. In M. Martinez & A. Castro Superfine (Eds.), *Proceedings of the 35th annual meeting of the North American chapter of the international group for the psychology of mathematics education* (pp. 86–89). Chicago, IL: University of Illinois at Chicago.

Walkington, C., Clinton, V., Shivraj, P., & Yovanoff, P. (2015). Association between readability and topic of mathematics word problems and performance on large-scale assessments. In M. Martinez & A. Castro Superfine (Eds.), *Proceedings of the 35th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 86–89). Chicago, IL: University of Illinois at Chicago.

Walkington, C., Cooper, J., & Howell, E. (2013). The effects of visual representations and interest-based personalization on solving percent problems. In M. Martinez & A. Castro Superfine (Eds.), *Proceedings of the 35th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 533–536). Chicago, IL: University of Illinois at Chicago.

Walkington, C., Petrosino, A., & Sherman, M. (2013). Supporting algebraic reasoning through personalized story scenarios: How situational understanding mediates performance and strategies. *Mathematical Thinking and Learning, 15,* 89–120. http://dx.doi.org/10.1080/10986065.2013.770717

Walkington, C., & Sherman, M. (2012). Using adaptive learning technologies to personalize instruction: The impact of interest-based scenarios on performance in algebra. In J. van Aalst, K. Thompson, M. Jacobson, & P. Reimann (Eds.), *Proceedings of the 10th International Conference of the Learning Sciences* (Vol. 1, pp. 80–87). International Society of the Learning Sciences (ISLS): Sydney, NSW, Australia.

Walkington, C., Sherman, M., & Petrosino, A. (2012). "Playing the game" of story problems: Coordinating situation-based reasoning with algebraic representation. *The Journal of Mathematical Behavior, 31,* 174–195. http://dx.doi.org/10.1016/j.jmathb.2011.12.009

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063–1070. http://dx.doi.org/10.1037/0022-3514.54.6.1063

Weaver, C. A., & Kintsch, W. (1988). *The conceptual structure of word algebra problems*. Boulder, CO: University of Colorado, Institute of Cognitive Science Tech. Rep. No. Series.

West, W. C., & Holcomb, P. J. (2000). Imaginal, semantic, and surface-level processing of concrete and abstract words: An electrophysiological investigation. *Journal of Cognitive Neuroscience, 12,* 1024–1037. http://dx.doi.org/10.1162/08989290051137558

White, S. (2012). Mining the text: 34 text features that can ease or obstruct text comprehension and use. *Literacy Research and Instruction, 51,* 143–164. http://dx.doi.org/10.1080/19388071.2011.553023

Wiest, L. (2003). Comprehension of mathematical text. *Philosophy of Mathematics Education Journal, 17,* 458.

Xu, R. (2003). Measuring explained variation in linear mixed effects models. *Statistics in Medicine, 22,* 3527–3541. http://dx.doi.org/10.1002/sim.1572

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language, 47,* 1–29. http://dx.doi.org/10.1006/jmla.2001.2834

Zwaan, R. A. (1999). Embodied cognition, perceptual symbols, and situation models. *Discourse Processes, 28,* 81–88. http://dx.doi.org/10.1080/01638539909545070

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

(*Appendices follow*)

Appendix A1

Pearson Correlation Between Coh-Metrix Readability Measures and Measures of Problem Solving Performance, for Story Problems With Multiple Sentences Only ($N = 120$ Problems)

	Correct	Incorrect	Hint
Number of sentences (DESSC)	−.180*		.432**
Number of words (DESWC)	−.234*		.465**
Second language readability score (RDL2)	−.193*	.183*	
Word information			
Third-person singular pronouns (WRDPRP3s)	.234**	−.205*	−.181*
Word concreteness (WRDCNCc)	.191*		−.225*
Word imageability (WRDIMGc)			−.247**
Word meaningfulness (WRDMEAc)			−.180*
Connective words			
Incidence of causal connectives (e.g., because, so, therefore) (CNCCaus)	.184*	−.190*	−.190*
Incidence of adversative/contrastive connectives (e.g., but, although, however) (CNCADC)	−.186*		.227*
Situation model support			
Incidence of intentional actions, events, and particles (SMINTEp)		−.200*	
Word diversity and similarity			
Type-token ratio for content words (LDTTRc)	.211*		−.303**
Type-token ratio all words (LDTTRa)	.250**		−.397**
Similarity of words (i.e., minimal edit distance score) (SYNMEDwrd)	.233*	−.210*	−.199*
Phrases			
Incidence of preposition phrases (DRPP)		.189*	
Incidence of gerunds (DRGERUND)	−.199*		
Overlap between sentences			
Standard deviation of content word overlap of adjacent sentences (CRFCW01d)			.263**
Standard deviation of content word overlap of all sentences (CRFCW01ad)			.242**
Standard deviation of overlap of adjacent sentences (LSASS1d)	−.185*		.328**
Standard deviation of overlap of all sentences (LSASSpd)			.301**
Average givenness of each sentence, compared to other sentences (LSAGN)			.210*

Note. In the above table, "content words" include nouns, verbs, adjectives, and adverbs in a story that carries nonlinguistic meaning—these are distinguished from function words (e.g., articles, prepositions, conjunctions) that express grammatical relationships.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Appendix A2

Pearson Correlation Between LIWC Topic Measures and Measures of Problem Solving Performance for All Story Problems ($N = 151$ Problems)

	Correct	Incorrect	Hint
Social processes (e.g., mate, talk, child)		−.241**	
Humans (e.g., adult, baby, boy)			
Insight (e.g., think, know, consider)			.214**
Tentative (e.g., maybe, suppose, assume, guess)	−.216**	.186*	.206*
Inhibition (e.g., safe, save, stop, contain)			.208*
Inclusive (e.g., and, with, include)		−.179*	
Exclusive (e.g., but, without, exclude)			.181*
Motion (e.g., arrive, car, go)	−.199*		.249**
Time (e.g., minute, second, day)	−.184*		.164*
Work (e.g., job, firm, company)	−.177*		.178*
Health (e.g., clinic, flu, pill)			.212**

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

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix B

Descriptive Statistics of Problem Readability and Topic

	# of problems nonzero in this category	Mean	*SD*	Min–Max
Readability category (Coh-Metrix)				
Number of sentences	151	2.76	1.52	1–9
Number of words	151	42.47	24.91	11–155
Words per sentence, mean	151	15.86	4.53	6.7–32
Word length, number of syllables, mean	151	1.46	0.47	1.1–2
Word length, number of syllables, standard deviation	151	0.77	0.20	0.29–1.5
Word length, number of letters, mean	151	4.44	0.43	3.4–5.7
Word length, number of letters, standard deviation	151	2.35	0.46	1.4–3.8
Incidence of all connective words, per 1,000 words	119	53.82	40.48	0–185.2
Causal connectives per 1,000 words	85	24.66	27.70	0–106.4
Logic connectives per 1,000 words	85	25.57	28.82	0–121.2
Temporal Connectives per 1,000 words	52	10.40	19.41	0–133.3
Adversative/contrastive connectives per 1,000 words	25	5.03	13.14	0–90.9
Nouns per 1,000 words	151	299.51	64.81	111.1–500
Verbs per 1,000 words	150	123.39	42.44	0–255.8
Adjectives per 1,000 words	128	61.46	45.96	0–218.8
Adverbs per 1,000 words	96	26.76	27.42	0–120
Personal pronouns per 1,000 words	103	42.48	42.43	0–228.6
First-person singular pronouns per 1,000 words	4	1.07	9.51	0–100
First person plural pronouns per 1,000 words	6	2.81	14.53	0–117.6
Second person pronouns per 1,000 words	21	7.12	20.66	0–125
Third person singular pronouns per 1,000 words	39	16.99	32.96	0–125
Third person plural pronouns per 1,000 words	27	7.61	23.44	0–228.6
Incidence of noun phrases	151	402.82	67.42	214.3–550
Incidence of preposition phrases	150	132.64	53.87	0–277.8
Incidence of adverbial phrases	70	15.21	20.4	0–81.6
Incidence of verb phrases	151	198.64	72.58	55.6–413.8
Incidence of gerunds	77	19.09	26.14	0–187.5
Concreteness of content words	151	417.22	42.64	305.4–514.1
Imagability of content words	151	443.25	38.91	327.6–533.4
Meaningfulness of content words	151	440.99	30.98	323.5–515.5
Word polysemy of content words	151	4.16	1.07	2.1–9.1
Incidence of causal verbs	127	40.76	25.96	0–120
Incidence of causal verbs and causal particles	130	47.60	29.96	0–142.9
Incidence of intentional actions, events, and particles	96	24.60	25.40	0–100
Type–token ratio for content words	151	0.87	0.10	0.55–1
Type–token ratio all words	151	0.80	0.12	0.48–1
Coh-Metrix L2 readability	151	16.69	11.62	−10.7–52.8
Topic category (LIWC)				
Contains social process words (0/1)	111	0.73	0.44	
Contains family words (0/1)	13	0.09	0.28	
Contains humans words (0/1)	30	0.20	0.40	
Contains positive emotion words (0/1)	54	0.36	0.48	
Contains negative emotion words (0/1)	15	0.10	0.30	
Contains cognitive mechanism words (0/1)	139	0.92	0.27	
Contains insight words (0/1)	51	0.34	0.47	
Contains causation words (0/1)	63	0.42	0.50	
Contains discrepancy words (0/1)	43	0.29	0.45	
Contains tentative words (0/1)	58	0.38	0.49	
Contains certainty words (0/1)	61	0.40	0.49	
Contains inhibition words (0/1)	28	0.19	0.39	
Contains inclusive words (0/1)	99	0.66	0.48	
Contains exclusive words (0/1)	36	0.24	0.43	
Contains perceptual process words (0/1)	30	0.20	0.40	
Contains feeling words (0/1)	15	0.10	0.30	
Contains body words (0/1)	21	0.14	0.35	

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix B (*continued*)

	# of problems nonzero in this category	Mean	*SD*	Min–Max
Contains ingestion words (0/1)	18	0.12	0.33	
Contains motion words (0/1)	74	0.49	0.50	
Contains spatial words (0/1)	135	0.89	0.31	
Contains time words (0/1)	128	0.85	0.36	
Contains work words (0/1)	89	0.59	0.49	
Contains achievement words (0/1)	59	0.39	0.49	
Contains leisure words (0/1)	64	0.42	0.50	
Contains home words (0/1)	25	0.16	0.37	
Contains money words (0/1)	75	0.50	0.50	
Contains health words (0/1)	11	0.07	0.26	
Contains biological process words (0/1)	32	0.21	0.41	
Contains affect words (0/1)	61	0.40	0.49	
The following readability categories (Coh-Metrix) presupposed multiple sentences, thus the total possible sample size for them is 120				
Similarity of words (i.e., minimal edit distance score; SYNMEDwrd)	120	0.85	0.08	0.52–1
Anaphor/pronoun overlap between sentences (CRFANP1)	58	0.34	0.41	0–1
Standard deviation of content word overlap of adjacent sentences (CRFCW01d)	66	0.07	0.10	0–0.39
Standard deviation of content word overlap of all sentences (CRFCW01ad)	76	0.08	0.08	0–0.32
Standard deviation of overlap of adjacent sentences (LSASS1d)	76	0.09	0.10	0–0.41
Standard deviation of overlap of all sentences (LSASSpd)	76	0.10	0.10	0–0.34
Average givenness of each sentence, compared to other sentences (LSAGN)	120	0.19	0.07	0.004–0.374
Words per sentence, standard deviation (DESSLd)	116	4.99	3.70	0–25.5

Note. There were 31 one-sentence problems, 44 two-sentence problems, 39 three-sentence problems, 20 four-sentence problems, 7 five-sentence problems, 6 six-sentence problems, 3 seven-sentence problems, and 1 nine-sentence problem.

Appendix C

Examples of Story Problems From the Data Set That Incorporated Topics From LIWC That Were Significant in the Regression Models

Topic	Example problem
Work	You have just been promoted to assistant manager at PAT-E-OH Furniture Inc. and have received a raise to $10.50 per hr.
Motion	A machine called the Crawler which moves space shuttles travels at the rate of 29 feet per second. The Crawler is currently 100 feet from the hanger moving toward the launching pad.
Social	A bride is making nameplates to put on the tables at her reception. She can make them at the rate of 25 per hr. She works for 2 hrs and quits for the night realizing that she cannot complete this many nameplates herself. The next day she calls her mother and they both work together. Her mother can make 35 nameplates per hr.
Health	According to the American Heart Association approximately 145,000 women die every year from smoking-related diseases. In fact lung cancer has become the leading cause of cancer death among women.
Inhibition	During the school year teachers save money for use during the summer when they're not being paid. This year due to some unexpected expenses one teacher was able to save only $879. He figures he will need $23 a day for personal spending money.

Received January 20, 2014
Revision received February 13, 2015
Accepted February 14, 2015 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.