View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by UND Scholarly Commons (University of North Dakota)

University of North Dakota
UND Scholarly Commons

Educational Foundations and Research Faculty Publications

Department of Educational Foundations and Research

4-2018

How Readability Factors Are Differentially Associated With Performance for Students of Different Backgrounds When Solving Mathematics Word Problems

Candace Walkington

Virginia Clinton
University of North Dakota, virginia.clinton@und.edu

Pooja Shivraj

Follow this and additional works at: https://commons.und.edu/efr-fac

 Part of the Science and Mathematics Education Commons

Recommended Citation

Walkington, Candace; Clinton, Virginia; and Shivraj, Pooja, "How Readability Factors Are Differentially Associated With Performance for Students of Different Backgrounds When Solving Mathematics Word Problems" (2018). *Educational Foundations and Research Faculty Publications*. 5.
https://commons.und.edu/efr-fac/5

This Article is brought to you for free and open access by the Department of Educational Foundations and Research at UND Scholarly Commons. It has been accepted for inclusion in Educational Foundations and Research Faculty Publications by an authorized administrator of UND Scholarly Commons. For more information, please contact zeineb.yousif@library.und.edu.

How Readability Factors Are Differentially Associated with Performance for Students of Different Backgrounds When Solving Mathematics Word Problems

Candace Walkington

Department of Teaching and Learning

Southern Methodist University

cwalkington@smu.edu

Virginia Clinton

Department of Education, Health, and Behavior Studies

University of North Dakota

virginia.clinton@und.edu

Pooja Shivraj

Coppell Independent School District

Corresponding Author: Candace Walkington, 512-417-9975, cwalkington@smu.edu

Please cite as the following:
Walkington, C., Clinton, V., & Shivraj, P. (2018). How readability factors are differentially associated with performance for students of different backgrounds when solving math word problems. *American Educational Research Journal, 55*(2), 362-414. doi: 10.3102/0002831217737028

Abstract

The link between reading and mathematics achievement is well known, and an important question is whether readability factors in mathematics problems are differentially impacting student groups. Using 20 years of data from the National Assessment of Educational Progress and the Trends in International Mathematics and Science Study, we examine how readability factors – such as length, word difficulty, and pronouns – interact with student background characteristics – such as race/ethnicity, mathematics achievement, and socioeconomic status. Textual features that make problems more difficult to process appear to differentially negatively impact struggling students, while features that make language easier to process appear to differentially positively impact struggling students. It is critical that readability along various dimensions be considered when designing instruction and assessment.

Keywords: achievement gap; language comprehension/development; NAEP; mathematics education

How Readability Factors Are Differentially Associated with Performance for Students of Different Backgrounds When Solving Mathematics Word Problems

The strong relationship between reading and mathematics achievement is well known (e.g., Crawford, Tindal, & Stieber, 2001; Hecht, Torgesen, Wagner, & Rashotte, 2001; Jiban & Deno, 2007; Lerkkanen, Rasku-Puttonen, Aunola, & Nurmi, 2005). For example, analyses of student performance on the Programme for International Student Assessment (PISA; Kelly, Nord, Jenkins, Chan, & Kastberg, 2013), known to have a particularly high reading demand in its mathematics items, shows a correlation of 0.95 between PISA mathematics country mean scores and PISA reading country mean scores (Wu, 2010). This correlation was higher than the country-by-country correlation between PISA mathematics scores and mathematics scores on the international assessment the Trends in International Mathematics and Science Study (TIMSS). This suggests that measures of reading achievement may actually predict mathematics achievement well, and in fact be *better* predictors than some complementary measures of mathematics achievement.

One reason why mathematics achievement may be so closely linked to reading is that many mathematics problems involve considerable reading demands. Mathematical information is often presented in verbal (rather than symbolic) formats, with significant unraveling and decoding of the English language needed to extract relevant relations. Indeed, mathematics word problems (i.e., problems where a mathematical scenario is posed using language rather than or in combination with symbols) have long been considered notoriously difficult (Cummins, Kintsch, Reusser, & Weimer, 1988) and U.S. mathematics teachers cite word problems as a major weakness of students (Loveless, Fennel, Williams, Ball, & Banfield, 2008). A subset of word problems are story problems, which are situated in "real world" contexts that reference concrete people, places, and objects. International comparisons suggest U.S. students struggle with mathematics word problems (OECD, 2010).

Large-scale mathematics assessments in the United States, such as the National Assessment of Educational Progress (NAEP), have also revealed enduring achievement gaps

between different student groups. Relatively small achievement gaps between males and females, favoring males, persist in some grade levels of mathematics (Robinson & Lubienski, 2011). Large, persistent, and widening achievement gaps between students of low- and high-socioeconomic status (SES) are particularly troubling (Reardon, 2011), as are continued gaps between Caucasian and Hispanic or African-American students (NCES, 2013; Provasnik et al., 2012). Achievement gaps also exist with respect to English Language Learners (ELLs; Fry, 2007), a fast-growing segment of the U.S. population. Finally, research has placed increasing emphasis on the importance of students' attitudes towards mathematics – particularly their interest in learning mathematics – and its positive association with achievement (Kim, Jiang, & Song, 2015).

These gaps in mathematics achievement may in part be explained by differences in reading and language background between student groups. Indeed, recent work highlights that students who have weak language skills have difficulty understanding the text in word problems (Vilenius-Tuohimaa, Aunola, & Nurmi, 2008), and that accommodations that reduce the reading demands of mathematics problems can result in higher performance for struggling students (e.g., Helwig, Rozek-Tedesco, Tindal, Health, & Almond, 1999). When considering that all students, including those from different demographic backgrounds, should understand mathematics and be assessed on their progress, it is critical to investigate how readability characteristics of mathematics problems may be differentially associated with performance.

In the present study, we use almost 20 years of mathematics achievement data from the NAEP and TIMSS to examine how the reading level of mathematics word problems is differentially associated with performance for students from different demographic backgrounds. Examining approximately 1000 problems solved by three-quarters of a million 4th and 8th grade U.S. students, we look at the interaction of word problem readability, focusing on several key text-based indicators identified in prior research, and student background characteristics, focusing on characteristics where achievement differences are well-established. Pilot work we review suggests that readability measures do matter for students' performance on mathematics

word problems. Thus, investigating whether these readability factors are differentially impacting students from different backgrounds is important for understanding and acting upon persistent achievement differences between groups.

Literature Review

Theoretical Framework

Theoretical work on cognitive models for solving word problems has explicated why readability may be important. Early research revealed that slight variations in mathematics problem wording result in children using different strategies (Carpenter, Fennema, Franke, Levi, & Empson, 1999; Carpenter & Moser, 1984). Kintsch and Greeno (1985) developed a model of story problem solving where students first translate from a problem statement to a *propositional textbase*, which is a conceptual representation of the relationships in the text. Students then form a *problem model* or *situation model* that infers the information needed to solve the problem based on knowledge of the domain. Later research (Hegarty, Mayer, & Monk, 1995) recognized that unsuccessful problem solvers use *direct translation strategies*, operating on numbers and keywords from the text and bypassing intermediate formation of a model of the situation. In contrast, successful problem solvers use problem model strategies where they form a mental representation of the situation and use this model to plan and assess their strategies.

Following this work, Nathan, Kintsch, and Young (1992) proposed a model of story problem solving where students coordinate three levels of representation: (1) the textbase or the propositional information given, (2) the situation model or mental representations of the relationships, actions, and events, and (3) the problem model of formal mathematical operands, numbers, and variables. The situation and problem models are thought to be mutually supportive, with students iteratively moving between the two representations. Thus, if students are able to extract a meaningful situation model from the problem's text, this situation model can support and improve their formal mathematical computations. However, forming this situation model is heavily dependent on comprehension of the text itself, which is impacted by readability.

Cognitive load theory also gives an important and related way of understanding the

demands of solving mathematics word problems. Cognitive load refers to the amount of mental effort expended in a learning or assessment task as working memory is utilized (Sweller, van Merrienboer, & Paas, 1998). While *intrinsic cognitive load* is the inherent difficulty level associated with understanding and processing particular mathematical concepts in a word problem, *extraneous cognitive load* describes the way in which cognitive load is further impacted by the manner in which the concepts are presented. Problem texts that are difficult to read may increase extraneous cognitive load as students struggle to decode the written language in order to form a situation model. For example, Walkington, Clinton, Ritter, and Nathan (2015) describe a story problem where a character wakes up to find their basement flooded, and ends up comparing the rate of two plumbers. The contextual information at the beginning of the problem may have contributed to extraneous cognitive load, as such factors may not be directly related to learners' processing and retrieval of relevant schemas. Thus, extraneous cognitive load from readability factors that are unrelated to retrieval of schemas may monopolize working memory, making these schemas more difficult to access. However, it is important to note that readability characteristics are not always extraneous – it may be unavoidable to add reading demands when explaining a complex mathematical situation, and confronting such problems may intrinsically involve reading skills.

Relationship between Readability Measures and Performance on Large-Scale Assessments

Several notable studies on readability and student achievement have been conducted using large-scale assessments. Using mathematics state standardized test items from Grades 4, 7, and 10, Shaftel, Belton-Kocher, Glasnapp, and Poggio (2006) found that use of vocabulary specific to mathematics, complex verbs, polysemous words (words with multiple meanings), pronouns, prepositions, and comparative words were associated with greater problem difficulty across all students in particular grades. The reading demands of TIMSS mathematics items have also been examined using 175 4th grade items, drawn from a year when the TIMSS was administered to the same students who received a reading assessment (Mullis, Martin, & Foy, 2013). The reading demand of mathematics items was measured by the number of words, use of

technical vocabulary, amount of symbolic language, and the complexity of visual displays. The number of words was the most substantial contributor to reading demand based on discriminant function analysis. In addition to the number of words indicating the amount of information to process, the number of words may also indicate syntactic complexity (e.g., greater average number of clauses per sentence; Arnold, Losongco, Waswow, & Ginstrome, 2000). Overall, students with better reading proficiency performed better on mathematics items across all levels of reading demand than students with lower reading proficiency. However, the magnitude of the difference in performance on mathematics items by students of varying reading proficiency differed as a function of reading demand. For proficient readers, accuracy was consistent across levels of item reading demand. For poor readers, accuracy was higher for items with low reading demand.

It is important to consider that reading difficulty and mathematical difficulty are often inexorably linked – problems with more complex mathematics may in turn require more complex language to describe the problem situation (see Walkington et al., 2015). This is not always the case – for example, there could be two nearly identical versions of a story problem about growth over time, one that models the growth using an exponential equation (e.g., increases by 50% every year) and one using linear growth (e.g., increases by 50 every year). Here, the reading demands would be similar, but the mathematical difficulty would be substantially different. Conversely, one could hold the mathematical difficulty of a word problem constant, but create new versions that reduce readability demands by, for example, using familiar or concrete words (Abedi & Lord, 2001). This latter case may be particularly useful for understanding readability factors.

Abedi, Lord, and Hoffstetter (2001) researched 8th grade students who received versions of the 1996 NAEP test booklets that had either 29 standard mathematics items or matched items with their linguistic demand reduced. They found that students performed better on the linguistically modified items, and that item length seemed to be an especially important factor in linguistic demand. In a similar study, Abedi and Lord (2001) used 20 items from the 1992 NAEP

and modified them to be less linguistically complex. In both studies, they manipulated familiarity/frequency of non-mathematics vocabulary, active/passive voice, conditional and relative clauses, question phrases, and made abstract presentations more concrete. The items were administered to 1174 8th graders, and they found improved performance, with performance gains differentially impacting particular groups of students (described later). We next turn to a review of the literature on how the association of readability and performance may differ based on student characteristics, which is our primary focus.

Relationship between Readability Measures and Student Background Characteristics

There are reasons to expect that associations between readability and performance would vary by student characteristics, such as gender, cultural and linguistic background, SES, mathematics achievement, and mathematics attitudes. Prior research has mainly examined reading skill and ELL status as factors that moderate the relationship between readability and performance. We review existing evidence for each of these factors.

Gender

Numerous studies have indicated that girls typically have stronger reading skills than boys; this finding is noted across grade levels and in most countries, including the U.S. (Chatterji, 2006; Department for Education and Skills, 2007; Halpern, Benbow, Geary, Gur, Hyde, & Gernsbacher, 2007; Logan & Johnston, 2010; Mullis, Martin, Foy, & Drucker, 2012; Twist, Schagen & Hodgson, 2007). Text characteristics that ease readability may be more beneficial for boys than girls. For example, because girls are generally more fluent readers (Barth, Tolar, Fletcher, & Francis, 2014), girls may have less difficulty with longer texts.

There are also gender differences in how language is perceived that may interact with reading word problems. Corpus analyses have indicated that females use more pronouns, with the exception of second-person pronouns, than males (Newman, Groom, Handelman, & Pennebaker, 2008). Girls may also be more likely to connect different ideas in a text together than boys (Clinton et al., 2014); therefore, girls may receive less benefit from pronouns in story problems than would boys, as pronouns can serve as linking mechanisms between sentences or

clauses. In addition, females may be more inclined to consider a word "concrete" and less inclined to consider a word "abstract" than males (Bauer & Altarriba, 2008). This may influence how different genders process words of varying levels of concreteness in word problems.

Linguistic, Cultural, and Socioeconomic Background

It is well known that language and culture influence the interpretation of language (August & Shanahan, 2008; Delpit & Dowdy, 2002). If a student's linguistic background differs from the standardized American English in which assessment problems are written, the language in a story problem may be particularly challenging. ELLs are more likely to struggle with reading compared to native speakers of English (Lesaux & Kieffer, 2010), so ELLs may have more difficulty with the language in a story problem (Fuentes, 1998). ELLs may perform less well on mathematics achievement tests because of problems with reading the problems despite understanding the mathematics (Abedi, Hofstetter, Baker, & Lord, 2001; Abedi, Leon, & Mirocha, 2001). Lengthy problems may also be especially troublesome for ELLs because of the increased effort involved in reading something in a second language (Bernhardt & Kamil, 1995). In addition, it is easier to learn concrete words than abstract words in a second language (De Groot & Keijzer, 2000). Given this, it is not surprising that ELLs more quickly comprehend concrete words than abstract words (Jin, 1990). Previous findings indicate that ELLs benefit more than native speakers of English from problems reworded to increase readability (Abedi & Lord, 2001).

ELL students may also struggle due to difficulties adopting the "mathematics register" – the unique meanings and structures of everyday language, such as using "left" to mean either a direction or what is remaining (Khisty, 1995). In addition, students may think of mathematical terms, such as "face" or "product" as typical language and apply non-mathematical meanings. Prepositions may be particularly challenging—for example, one finds the area "of" a triangle, rather than "inside" a triangle (Pimm, 1987). Moreover, because mathematics register involves words with mathematical purposes (Pimm, 1994), it is particularly complex because language must be integrated with the symbols and notations of mathematics (Yore, Pimm, & Tuan, 2007).

The notion of the mathematics register is related to the idea of *disciplinary literacy* – an understanding of the technical vocabulary, language structures, and discursive elements that are specific to content areas like mathematics, and the associated specialized reading routines (Shanahan & Shanahan, 2008). Disciplinary literacy skills are less generalizable than basic reading skills, can be difficult to learn, are rarely explicitly taught, and are prominent in the Common Core State Standards (CCSS, 2010; Zygouris-Coe, 2012). Obtaining disciplinary literary is difficult for native speakers, and may present a particular challenge for ELLs.

Martiniello (2008) analyzed six mathematics word problems that showed Differential Item Functioning (DIF) favoring non-English Language Learners (non-ELLs) over ELLs on a 4th grade standardized test. She found that stories that included complex, multiple clauses, as well as long noun phrases, led to comprehension difficulties for ELLs. In addition, unfamiliar vocabulary words, especially words that English-speaking students might learn at home (e.g., chores) or words relating to mainstream American culture (e.g., spelling bee) were difficult for ELLs, as were polysemous words. In a similar study that used a large bank of mathematics standardized test problems from Grades 4-8, Wolf and Leon (2009) found significant associations between DIF and number of words, amount of and reliance on verbal language in the problem (versus, for example, visuals, and graphs), and use of academic vocabulary. Abedi, Lord, Hofstetter, and Baker (2000) found that modifying mathematics items to be less linguistically complex may reduce the achievement gap between ELLs and non-ELLs.

Moving beyond strictly language issues, cultural differences in how mathematics is utilized in school versus community settings may impact problem solving (Saxe, 1988; Taylor, 2005). Boundaries between cultural knowledge and mathematical domain knowledge can obscure students' mathematical understanding (Nasir, Hand, & Taylor, 2008). Gerofsky (2009) describes how word problems are a literary and pedagogical genre; the language they utilize and the knowledge they draw upon are inherently ambiguous, localized, and conditional, which has implications for students from different cultural backgrounds. Nasir et al. (2008) describe how when basketball players solved mathematics problems in a basketball context, they had greater

success and were able to invent powerful mathematical strategies. However, when the same problem was in an abstract mathematical context, they often misapplied formal algorithms. These authors make the case that school mathematics is a cultural activity that is structured to privilege certain communities and students over others. Mathematics word problems may thus privilege the experiences of the idealized Caucasian middle class (Ladsen-Billings, 1997) or incorporate a Eurocentric perspective (Tate, 1994).

At the intersection of language and culture, some research suggests students who speak African American Vernacular English in the home may struggle with reading standard English if they have a lack of familiarity with standard English (Charity, Scarborough, & Griffin, 2004). This is similar to previously discussed issues with other populations who do not speak standard English in the home (e.g., ELLs). Students who are African American and use African American Vernacular English may read story problems differently than students whose linguistic background is a more standardized form of English. For example, African American Vernacular English is more vivid and imageable than standard American English dialects (Ball, 1996). Students who have more experience with such concrete language could benefit when vivid and imageable terms are used in story problems, and struggle when language is abstract and decontextualized.

Language, race, and culture are also inexorably tied to SES. SES is known to be an important predictor of academic achievement, including reading performance (Perry & McConney, 2010; Sirin, 2005). Students from low-SES backgrounds do less well on reading assessments than students from middle- or high-SES backgrounds (Noble, Farah, & McCandliss, 2006). Given these issues, it is likely that students from low-SES backgrounds would benefit more from more readable problems. Research has shown that low-SES students benefited more from modifications to 8th grade mathematics problems to make them easier to read than other students (Abedi & Lord, 2001). Differences in responses to realistic constraints of story problems have also been found between working class versus more privileged students (Cooper & Harries, 2005). Frankenstein (2009) discusses how word problems contain “hidden messages”

(p. 111) about the sometimes taken-for-granted norms of society. For instance, a word problem about adding up prices at the grocery store contains the implicit message that it is normal to have money to pay for food, even though many children suffer from hunger.

Ladsen-Billings (1995) describes differences in the way suburban and inner city students responded to the story problem, "It costs $1.50 to travel each way on the city bus. A transit system fast pass costs $65 a month. Which is the more economical way to get to work, the daily fare, or the fast pass?" (p. 131). Suburban students assumed that a person would commute to work 5 days a week, and concluded that the daily fare would be more economical. However inner city students opted for the transit pass, posing questions like "How many jobs does this person have?" "Do they have part-time jobs or full-time jobs?" Urban students also recognized that if the transit pass was purchased, family members could use it on evenings and weekends to go to stores, church, etc. Other studies have found that many "unrealistic" or "incorrect" responses students give to story problems represent unanticipated but valid interpretations of the story context based on their everyday cultural knowledge and diverse sense-making activities (Inoue, 2005; Kazemi, 2002).

Mathematics Achievement and Mathematics Attitudes

The association between readability measures and performance may also differ for students with different mathematics achievement levels. Mayer's (2001) *individual differences principle* states that design effects, like making a text more readable, tend to be stronger for low-knowledge learners because high-knowledge learners are better able to use prior knowledge to compensate for less support. Accordingly, research has shown that students in lower-level mathematics courses benefit more from mathematics problems designed to be easier to read than students in higher-level mathematics courses (Abedi & Lord, 2001). In addition, students struggling with algebra tend to benefit more from an intervention that personalized story problem texts to topics they found interesting (Walkington, 2013). Finally, mathematics and reading achievement are very highly correlated (Wu, 2010), so students with higher mathematics skills are likely to have stronger reading skills as well, and struggle less with difficult-to-read

problems.

Students' level of interest in learning mathematics may also moderate the effect of readability on performance. Interest is positively associated with making connections both within a text and between a text and a student's background knowledge (Clinton & van den Broek, 2012), and thus may facilitate situation model construction. Durik and Harackiewicz (2007) found that an intervention designed to trigger interest by adding decorations was most effective for learners with low interest in mathematics, but hampered learners with high interest. Conversely, they found that an intervention that informed students of the value of the content was beneficial for high mathematics interest students, and detrimental for low interest students. In another study, Walkington, Cooper, and Howell (2013) found that adding relevant contexts and illustrations to mathematics story problems was most effective for students who had mixed attitudes towards mathematics – students who were not particularly positive or negative about learning math. Although there is not much research directly relating to readability, it is reasonable to assume that the effect of readability factors may differ based on students' interest in mathematics.

Pilot Studies

There are potentially limitless ways that the readability of a mathematics story problem could be quantified. Coh-Metrix (McNamara, Louwerse, Cai, & Graesser, 2013), a widely used computer-based readability tool, calculates 108 different measures of readability, and there are other such tools available (e.g., the Linguistic Inquiry and Word Count (LIWC) software; Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007). Historically, the use of quantitative measures of readability has been controversial (e.g., Bailin & Grafstein, 2001; Bertram & Newman, 1981; Kirkwood & Wolfe, 1980). Some criticisms are specific to older measures of readability, such as the concern that traditional measures oversimplify word difficulty without considering semantics (Kirkwood & Wolfe, 1980) or the complexity of the words (Bertram & Newman, 1981). These concerns are addressed by contemporary computerized tools, which provide more detailed measures (e.g., Coh-Metrix), and thus are considered preferable to

traditional formulas (Benjamin, 2012). However, other concerns about quantitative readability measures include that they are limited in that they do not take into account the interaction between the reader and text (Schulz, 1981), and they may remove the deep meaning and beauty from language (McNamara, Graesser, McCarthy, & Cai, 2014).

With these concerns in mind, there remain many advantages to quantitative measures of readability: they are objective, relatively quick to obtain, and provide assistance in modifying texts (Zamanian & Heydari, 2012). However, in order to conduct a study of how readability influences mathematics problem solving without significant inflation of Type 1 error, the list of potential readability factors to be tested must be narrowed down to a small set. We present two pilot studies that describe how we narrowed down readability factors to consider.

Readability Factors in Algebra Curricula

In prior work (Walkington et al., 2015) we examined 151 story problems from a widely used Algebra I curriculum, *Cognitive Tutor Algebra*, as well as a set of 60 algebra story problems from a middle school tutoring software *MATHia*. We entered the text of the story problems into two text analysis programs – Coh-Metrix and LIWC. We found that controlling for the mathematical difficulty of the problem, several measures of readability were significantly associated with performance. The difficulty of the words in the problem, measured by word concreteness and word polysemy was important, as was the length of the story text, measured by number of sentences and number of words. In addition, the presence of third-person singular pronouns was positively associated with performance, as was problem topics that were concrete and relatable (e.g., home, socializing) rather than topics related to work, finance, health, or business. However, these data were analyzed at the problem level, using a summary measure of performance for each problem. No information about student demographic characteristics was present. This study also examined only story problems (i.e., mathematics problems in real world contexts) and only problems that involved the mathematical content area of linear functions.

Readability Factors in NAEP/TIMSS

We also examined performance on the 4th and 8th grade released items on the NAEP and

TIMSS (Walkington, Clinton, Shivraj, & Yovanoff, 2015). These datasets simply gave an overall accuracy statistic for each problem averaged for every student who took the test. Thus, these datasets also did not give any information about how student background characteristics might impact problem solving, but did give enough information to correlate readability characteristics of the different problem texts to performance levels on those problems. For all available released problems (757 NAEP problems and 445 TIMSS problems), we calculated all measures of readability from Coh-Metrix and LIWC. We correlated these measures of readability with overall accuracy measures for each problem to narrow down a list of readability measures that were significantly correlated to performance. We then entered this narrowed list of readability measures as predictors into mixed-effects regression models predicting performance, including other problem characteristics (e.g., grade level, answer format, difficulty) as covariates. Results showed that on both the NAEP and TIMSS, a story problem that had more sentences or used second-person pronouns was significantly associated with lower performance levels. On the NAEP, as the average age of acquisition of content words in the story increased, performance significantly decreased. Also on the NAEP, as density of pronouns increased, performance significantly increased as well. Table 1 shows examples of NAEP problems that vary on problem length and use of pronouns.

Taken together, this study and the previous curricula study allowed us to narrow down a set of readability predictors to examine in the study we report here. Specifically, these studies suggest that ***word difficulty*** – expressed by measures such as concreteness, age of acquisition, and polysemy – is a critical factor. In addition, ***pronouns*** are an important readability factor; however, some evidence emerged that second-person pronouns seem to act in a contradictory manner to other pronouns. The ***length*** of the word problem – as measured by sentences or word count – is also an important readability measure. Finally, the study of curricula suggested that the problem's topic mattered; however, this study included *only* problems in real world contexts. We did not see such effects for problem topic in the initial NAEP/TIMSS study, perhaps because the nature of standardized tests causes interest-eliciting characteristics to be less critical. So, in the

present study, we simply examine whether or not the problem was in a ***real world context***, rather than any specific problem topic predictors. Prior research suggests that placing problems in real world contexts can improve accuracy and change strategy use (e.g., Koedinger & Nathan, 2004; Walkington, Sherman, & Petrosino, 2012), and these implications may differ for different student groups. Our final list included five factors: word concreteness, pronoun density, presence of second-person pronouns, word count, and presence of a "real world" context. Although there are a host of other readability measures that research has shown to be generally predictive of text comprehension, these measures were chosen because there was evidence that they were important when solving 4th and 8th grade mathematics word problems specifically.

Research Questions

Although previous studies have established the importance of our five key factors, the contribution of the study reported here is to look at how the associations between these readability factors and performance vary based on student background characteristics. We examine this interaction in a very large bank of story problems collected over decades of testing with nearly a million students. The specific background characteristics we examine include gender, race, SES, mathematics attitudes, language or birth status, and mathematics achievement, all of which have become important to national conversations surrounding equity, access, and achievement gaps. We compare and contrast results from two important and widely used test datasets – the NAEP and the TIMSS – across 4th and 8th grade. Our research questions are as follows:

1) How are key readability factors differentially associated with performance for students of different demographic characteristics (i.e., race, language, SES, gender, mathematics attitudes and achievement) when solving mathematics word problems on the NAEP and TIMSS?
2) How do results change when removing controls for mathematics achievement, which may under-represent the achievement gaps that students from different groups experience?

Method

Overview of Sample

Analyses drew upon student-level datasets from two sources – performance of 4th and 8th grade students on mathematics questions on the NAEP and on the TIMSS. A student-level dataset is a dataset that shows whether each student performed correctly on each problem they received, rather than giving summary measures for each student or problem. While the student-level TIMSS datasets are publicly available, the student-level NAEP datasets are restricted use and were obtained through a data licensing agreement. Along with mathematics test items, the NAEP and TIMSS both include background questionnaires for students and teachers that assess contextual factors that impact student learning. Neither the NAEP nor the TIMSS come with direct consequences for students who take them, as they are not mandated. For this reason, readability may function differently on these assessments than on compulsory assessments, as students may, for example, be more likely to give up if a problem is difficult to read.

The TIMSS, developed by the International Association for the Evaluation of Educational Achievement, provides written test data on the mathematics and science achievement of students in the U.S. and over 60 other countries. It has been administered every 4 years since 1995, and we include mathematics data from 1995, 1999, 2003, and 2011 (2007 released items were not available). The TIMSS aims to capture the breadth and richness of the mathematics taught in participating countries, and uncover improvement or decline in student performance over time. Students receive a sample of the total set of available mathematics items (see Mullis et al., 2009). The selection procedures for schools/classrooms for TIMSS are given in Jones and Foy (2012). Between 10,000 and 20,000 U.S. students typically take the TIMSS in each administration. The TIMSS data we used were limited to U.S. students taking English mathematics items.

The NAEP, administered by the National Center for Education Statistics, provides written test data on the achievement of U.S. students in a variety of subjects, including mathematics. NAEP results are intended to serve as a common and continuous measure of performance for states over time. Long-term assessments began in 1986 and generally take place

every 2 years – here we analyze data from 1996, 2003, 2005, 2007, 2009, 2011, and 2013, as these are the years that released items and student-level data were available. Items given to an individual student are again drawn from a larger sample. The selection procedures for schools/students for NAEP are given in NCES (2009). Between 300,000 and 400,000 U.S. students typically take the NAEP in each administration. Because we were examining student-level data and the NAEP dataset was so large, we took a random 10% subset of the data from each of the included years. This allowed for it to be computationally feasible to fit complex, mixed-effects models with precision. Supplementary analyses that selected multiple or different random subsets, including larger random subsets analyzed with less precise model-fitting techniques, yielded similar results to those presented here. A small percentage of students (7.6%) received some kind of accommodation while taking NAEP. Analyses were run with and without these students and results were similar, so they were left in the sample. Analyses were also run with and without the subset of students who received reading accommodations (4.3%) specifically– results were again similar either way, so students were left in the dataset. For the TIMSS, students with disabilities or students who had received instruction in the language of the test for less than one year were not tested, and thus were not included in the sample – this was typically around 4-7% of all students initially considered for TIMSS testing.

Selection of Problems and Measures of Readability

Our analyses only contain 4th and 8th grade mathematics problems from the aforementioned years that are released (i.e., made publicly available on the TIMSS website, nces.ed.gov/TIMSS/educators.asp, or the NAEP website, nces.ed.gov/NationsReportCard/nqt). In addition, analyses include only items that were multiple choice or short constructed response, and only items that had at least one sentence worth of words. As we used an automated text analysis software, we did not consider the visual representations or the symbols/equations – these were deleted when the problem was entered into the software. We also did not consider any text that occurred in the visuals or answer choices. Coh-Metrix is designed to be used with continuous text and the guidelines recommend that words in visuals be removed (McNamara et

al., 2014). As answer choices usually did not involve full sentences, they were also not considered. In supplementary analyses, we coded the visual along a variety of dimensions (number, type and location of visuals, and whether visual contained words), as well as presence of equations, but generally found these fine-grained predictors to be non-significant in exploratory analyses. We defined "one sentence worth of words" as a problem that, with visuals and symbols removed, was not a fragment and contained a subject and a predicate conveying a question, statement, or command. Also note that our narrowed list of factors contained only word-level readability variables (rather than cohesion measures that examine discourse level constructs), which we believe is appropriate given the relatively short length of most mathematics word problems. Only around 9% of TIMSS items and 6% of NAEP items did not have at least one sentence worth of words.

The final set of problem texts included 428 TIMSS problems and 565 NAEP problems. Although we could have further decreased the problem bank by only considering problems from recent years on the NAEP and TIMSS, we considered this number of problems an appropriate but minimal sample size for full coverage of the readability measures, and thus did not want to decrease it further. Problems were entered into Coh-metrix and LIWC. Coh-Metrix is a computerized text-mining tool that provides a broad set of fine–grained readability measures (McNamara et al., 2013). Coh-Metrix provides 108 different indicators of text readability organized into categories that relate to surface features of the text, such as word concreteness, features of the textbase, such as pronoun density, and deeper features of the text, such as propositional structure or cohesiveness. The LIWC software (Pennebaker et al., 2007) is a dictionary-based computerized text-analysis program that counts words in more than 70 categories, such as social process words (e.g., words relating to family or friends) and cognitive process words (e.g., words describing causation or certainty). LIWC's output consists of the percentage of words in a text from each dictionary.

As mentioned previously, for the analysis reported here we narrowed the Coh-Metrix and LIWC measures under examination to four – word count, pronoun density, word concreteness,

and presence of second-person pronouns. Word count (Coh-Metrix) is a numerical count of how many words the text contains, while pronoun density (LIWC) is the percentage of the total number of words in the problem that are pronoun words (including them/they/it). Word concreteness (Coh-Metrix) is a measure of the level at which one can interact with the concept represented by a word through the senses; for example, the word "ball" would have high concreteness, and the word "truth" would have low concreteness. Concreteness is calculated through ratings compiled in the Medical Research Council Psycholinguistic Database (Coltheart, 1981). Concreteness values for each word are discrete and in the range of 100 to 700 with higher values indicating more concreteness. Coh-Metrix provides a measure of the average concreteness compiled across all content words[1] in the text. All three of these measures were normalized, although we provide summary measures on their original scales in Table 2. Finally, presence of at least one second-person pronoun (i.e., the word "you") was measured through a simple 0/1 indicator, since there were relatively few problems containing these pronouns (around 5% of problems).

In addition to the LIWC and Coh-Metrix categories, each problem was also coded for whether it contained a "real world" context (i.e., any reference to using mathematics in the world to understand objects or events), rather than using mathematics for abstract or academic purposes. For example, a problem asking for the area of a 3-by-4 unit rectangle would not be a real world problem, but a problem asking for the area of a 3-foot-by-4-foot "poster" would. Fifty-three problems were double coded by two coders for whether there was a real-world context and whether the problem contained at least one sentence worth of words. Raters agreed 92.5% of the time about real-world contexts, and 100% for whether the problem contained one sentence of words. Summary statistics for these 5 measures are in Table 2.

Student Variables used in Analyses

When choosing student demographic variables to include in analyses, a primary concern

[1] "Content words" include nouns, verbs, adjectives, and adverbs which have linguistic meaning. These are distinguished from function words (e.g., articles, prepositions, conjunctions) that express grammatical relationships.

was including measures that were available across all years that each test had been administered. If a measure (like racial category) had not been collected in one year, this would result in a large subset of our released problems being omitted from the analyses. This limited the demographic categories we could consider to those shown in Tables 3-4; these tables also give summary measures of how the student population in our final sample varied on each demographic factor, as well as how much data were missing. We discuss each demographic variable in turn.

Gender. Gender information (male/female) was collected all years for both tests. There was very little (less than 0.1%) missing data.

Race. Racial category was collected each year of NAEP, but only 2 of the 4 years were included for TIMSS. For this reason, our analysis of the TIMSS data does not include the Race variable. Racial categories were Caucasian, Black, Hispanic, Asian/Pacific Islander, and Other (American Indian/Alaska Native, two or more races) on the NAEP. For all analyses, categories were collapsed to Caucasian, African-American, Hispanic, and Other, due to small sample sizes. Race was missing for a small percent (1.7%) of students on NAEP (Table 3).

Language. On the NAEP, a variable was available each year for "How often do people in your home talk to each other in a language other than English?" and it was rated on a 4-point scale from "Never" (1) to "All or Most of the Time" (4). This variable was collapsed into a dichotomous predictor of whether they indicated only English (1) or any amount of non-English (2-4). This value was missing for a small percentage (2.6%) of participants (see Table 3). We used this variable instead of the ELL status variable because it was better matched to the "Birth Status" variable available on the TIMSS. Note that ELL status has a specific meaning relating to a student currently having low English proficiency and being served in an assistance program.

Birth Status. On the TIMSS, there was no similar variable available for language. However, a variable was available for whether the student was born in the U.S. or outside of the U.S. This value was missing only for a small percentage (2.0%) of participants (see Table 3).

Socioeconomic Status. To capture SES, the only measures that were relatively consistent across years on NAEP and TIMSS were number of books and a presence of a computer in the home.

We called this indicator "Luxury Items" and scaled it as a continuous variable ranging from 0 to 1, with half of the score coming from the books rating (rated on a 1-4 or 1-5 scale, with each rating level representing a different range of books) and half coming from the computer rating (a simple yes/no regarding whether students had a computer in home). On the 1996 NAEP, there was no computer question, so number of books was used alone to compute Luxury. On the TIMSS, there was no number of books measure for 1995 Grade 4, and no computer measure for 1999 Grade 8, so only one measure was used to compute Luxury. This variable was missing for a small percentage of students on TIMSS (1.2%) and NAEP (2.1%).

Mathematics Attitudes. For each year on the TIMSS, a measure was collected where the student responded to the text "I enjoy learning mathematics" on a 1-4 scale. This value was missing for a small percentage (2.96%) of participants (see Table 4). These values were normalized. Measures for attitudes are not included for the NAEP as they were not available in all years.

Plausible Values/ Mathematics Achievement. All NAEP and TIMSS datasets include five "plausible values" that estimate students' overall mathematics achievement based on their responses to the mathematics items answered. There were 5 such values because, as mentioned previously, students typically received different items, and any estimates of achievement contain measurement error. Plausible values use multiple imputations to show a likely distribution of a students' proficiency, and cannot simply be averaged (see Von Davier, Gonzalez, & Mislevy, 2009). We used these plausible values as a proxy for mathematics achievement. There were no missing PVs on the TIMSS, and a very small (0.3%) percentage of missing sets of PVs on the NAEP. These values were normalized.

Problem Variables Used in Analyses

We also used descriptive variables to control for characteristics of individual problems (see Table 5) that did not directly relate to readability.

Problem Type. On the NAEP, problems were organized into 3 types: multiple choice, short constructed response, and extended constructed response. Extended constructed response items were few (30 total items), and differed substantially from the other two problem types – they

would ask students to write extended mathematical explanations. These were omitted from the analysis. On the TIMSS, there were constructed response and multiple choice items. While most of the constructed response items asked only for short "fill in the blank" answers, several asked for extended written explanations (15 total items) – these were also omitted.

Problem Difficulty. On the NAEP, each problem had a Difficulty rating of Easy, Medium, or Hard given on the NAEP website. On the TIMSS, there was no such rating given explicitly, so Easy problems were defined as those with an international performance level of 67% or greater, medium had 34-66% correct, and high had 33% correct or less. Note that as these are *international* performance levels, we would not necessarily expect them to be directly indicative of U.S. students' actual performance levels. Students in different countries have different opportunities to learn various mathematics concepts due to differences in curriculum, policy, and instruction. Also, as the TIMSS is offered in different languages, the language of the assessment itself may impact item difficulty. Thus, this is not a completely redundant measure when predicting U.S. item performance, and can be thought of as a broad measure of the difficulty of the mathematics problem independent of a country's particular context.

Problem Complexity. For 2005 and forward, NAEP explicitly rated the Complexity of their problems as Low, Moderate, or High. Prior to 2005, NAEP rated problems as Procedural Knowledge (which we mapped to Low), Conceptual Understanding (which we mapped to Moderate) and Problem Solving (which we mapped to High). The Moderate and High problems were later collapsed such that Complexity only had two levels – Low and High. NAEP differentiates complexity from difficulty: "Mathematical complexity is not necessarily related to item difficulty, which is based on actual student performance. Mathematical complexity should also be independent of curriculum, meaning it is determined assuming that students are familiar with the mathematical content of the item" (Neidorf, Binkley, Gattis, & Nohara, 2006, p. 30). The system for the TIMSS was more complex – they had 12 different "Cognitive Domain" categories, which evolved over the years the test was administered. However, we mapped these categories into Low (Knowing, Knowing Facts and Procedures, Performing Routine Procedures,

Solving Routine Problems, Using Routine Procedures) and High (Applying[2], Communicating and Reasoning, Reasoning, Investigating and Solving Problems, Solving Problems, Using Complex Procedures, Using Concepts) Complexity.

Grade Level. All problems were either Grade 4 or Grade 8. Grade 12 data were not considered. Grade level could be considered a proxy for age and thus a student variable rather than a problem variable. However, here this did not make sense because Grade 4 and Grade 8 had completely distinct problem sets, with different mathematical and reading demands, different uses of visuals, symbols, equations, etc. Thus, comparing differential effects of readability based on age using these problems would not be a sensible comparison.

Content Domain. Both NAEP and TIMSS items were organized into mathematical content domains. The NAEP was organized into the 5 categories we use here, while the TIMSS had categories that were similar but shifted year-by-year in their wording. The TIMSS domains were grouped according to the five NAEP categories: Algebra (which included Patterns and Relationships), Data (which included Data Representation, Probability and Chance), Geometry, Measurement, and Number (which included Fractions and Proportions).

Model Fitting Techniques

Mixed-effects logistic regression models were used (Snijders & Bosker, 1999) and fit separately for the NAEP and TIMSS. We included random intercept terms for Problem ID and which year's data the observation was drawn from (Tables 6-7). We did not fit a random effect for student ID, as the Plausible Value variable was a similar, and perhaps better, measure of student performance on the mathematics items on the test. The plausible value provided by NAEP/TIMSS had the advantage of being calculated based on students' performance on all mathematics items, rather than only based on their performance on released items included in the analysis. The models for the NAEP included 1,511,700 observations of 705,600 students solving

[2] The "Applying" category on the TIMSS had problems of disparate complexity, and the decision was made to call category "high complexity." However, in practice, the cognitive complexity of the individual problems varied. Regression models were run both ways, and results were similar.

565 mathematics problems. The models for the TIMSS included 720,010 observations of 40,570 students solving 428 problems (all samples rounded to the nearest 10).

We also included all student and problem level variables as fixed effect predictors. The outcome variable was whether the student got the problem correct or incorrect, coded as a 0/1. For short constructed response items where partial credit was a possibility, we did not award any partial credit. Overall, on the TIMSS, students had an accuracy rate of 55.69%, while on the NAEP they had an accuracy rate of 51.39% correct. We first fit a null model, which included only our student and problem variables (Model 1). We then fitted a model that included our 4 readability predictors and real-world context (Model 2), and a subsequent model that allowed these predictors to interact with student background characteristics (Model 3). We then fit a model that did not include the plausible value measures of mathematics achievement (Model 4) to examine how omitting this variable changed the results. Indeed, including a predictor for mathematics achievement to model performance on a problem, when that achievement measure was in part computed from performance on that problem, could be considered somewhat circular. However, given the well-known achievement gaps between diverse student groups, fitting the model with these plausible values seemed important as well. For the TIMSS, Model 4 was fit with Student ID as a random effect, as we were using the entire dataset. However, in the NAEP, there were too few observations per student (because we were using a 10% subset) for a random effect for Student ID to be computationally feasible. We also fit some additional models (described later) to explore alternative hypotheses about results that arose.

Models were implemented using the binomial family in the *glmer()* function in the lme4 package (Bates, Maechler, Bolker, & Walker, 2014) in Revolution R Enterprise (Revolution Analytics, 2014). Fitting models with plausible values involves special techniques for model computation. There are 5 plausible values, so each model is fit 5 separate times, once using each value. These models are then compiled by averaging the regression coefficients to obtain overall regression coefficients. Computation of the error term of these new coefficients was computed by combining the sampling variance and imputation variance, as outlined in Mislevy, Beaton,

Kaplan, and Sheeham (1992). We also implemented *p*-value corrections, given the large number of significance tests we were conducting. We used the False Discovery Rate (FDR) procedure, described in Benjamini and Hochberg (1995), also described in the NAEP technical documentation (http://nces.ed.gov/nationsreportcard/tdw/). The FDR is a multiple comparison procedure that holds the significance for a set of comparisons at a certain level (here $\alpha = 0.05$) by controlling for the expected proportion of errors from among the rejected null hypotheses.

As the models were logistic, *d*-type (standardized mean difference) effect sizes were estimated using the procedure outlined in Chinn (2000), where the coefficient is divided by 1.81. While this is straightforward for indicator variables or variables that range from 0 to 1, it is more difficult to interpret for variables that take a continuous range. For our normalized variables (3 of our readability variables, mathematics attitudes and plausible values) we took the difference between +1.5 and -1.5 standard deviations to compute this effect size measure.

In Cohen (1988), effect sizes of 0.2, 0.5, and 0.8 are considered small, medium, and large, respectively. However, as Hill, Bloom, Black, and Lipsey (2008) argue, it is important to consider benchmarks for effect size in the unique context of our study. They show that in educational research, effect sizes found using broad standardized mathematics tests often range from 0.1 to 0.3. Prior research on mathematics story problems (Walkington et al., 2015) suggests that the effects for readability measures are likely to be small, even when the effects of several measures are grouped together. When using text analysis tools, individual measure effect sizes near to $d = 0.1$ have been considered reasonable for practical significance (e.g., Newman et al., 2008). The combined effect of several readability measures that vary together may be more considerable and practically meaningful in the context of problem design.

Results

Findings for Null Models

Models 1 and 2 in Table 8 (NAEP) and Table 9 (TIMSS) are two null models that do not have interaction terms between readability factors and student background characteristics. Model 1 includes only student and problem variables, and shows similar results across the NAEP and

TIMSS. Unsurprisingly, medium and high difficulty problems have significantly lower accuracy than low difficulty problems, and high complexity problems have significantly lower accuracy than low complexity problems. On both tests, males score significantly higher than females, owning more luxury items is associated with significantly higher accuracy, and higher plausible values (mathematics achievement) are associated with significantly greater accuracy. On the NAEP, short constructed response problems are associated with significantly lower accuracy than multiple choice problems, African-American and Hispanic students score significantly lower than Caucasian students while students of other race/ethnicities score significantly higher, and Grade 8 has significantly lower accuracy than Grade 4.

Model 2 adds in readability predictors, with models from both datasets showing that the presence of second-person pronouns is associated with significantly lower accuracy (d = -.18 for NAEP and d = -.42 for TIMSS). The NAEP models additionally show that higher word counts are associated with significantly lower accuracy (d = -.17), more concrete words are associated with significantly higher accuracy (d = .12), and a greater pronoun density is associated with significantly higher accuracy (d = .13). The effect of a "real world" context is not significant in either model, but is directionally positive in NAEP and directionally negative in TIMSS.

Research Question 1: Differential effects of Readability Based on Student Characteristics

Model 3 in Table 8 gives the findings for the interaction of readability and student characteristics for the NAEP. In terms of gender differences, positive associations between real world contexts and accuracy and between pronouns and accuracy are higher for males than females (d = .03 and d = .02). However, these effect sizes are negligible. In addition, negative associations between word count and accuracy and second-person pronouns and accuracy are stronger for males than females, but again effect sizes are small (d = -.07, d = -.04). In terms of racial differences, positive associations between pronouns and accuracy are higher for African-American students than Caucasian students (d = .07). However, positive associations between concrete words and accuracy are lower for African-American students than Caucasian students, although the effect size is negligible (d = -.03). Finally, we see that negative associations

between second-person pronouns and accuracy are stronger for African-American students (d = -.11) or Hispanic students (d = -.08) than Caucasian students. Thus, the difficult readability factor of second-person pronouns might be differentially negatively associated with performance for these groups. For language status, we see that negative associations between word count and accuracy are higher for students with a language other than English spoken at home than other students, although the effect size is negligible (d = -.03). Also, negative associations between second-person pronouns and accuracy are stronger as students have fewer luxury items, although the effect size is very small (d = -.05).

The strongest effects for differential findings for readability characteristics are for mathematics achievement (measured by plausible values). Negative associations between word count and accuracy, and second-person pronouns and accuracy are less strong for higher achievement students (d = .19, d = 0.54). In addition, positive associations between concrete words and accuracy are lower for higher achievement students (d = -.11). Finally, positive associations between presence of a real-world context and accuracy are higher for higher achievement students, although the effect size is very small (d = .05). This suggests that stronger students have less negative associations between their performance and readability factors that make a problem difficult to read, while weaker students have stronger positive associations between their performance and readability factors that facilitate easy reading.

Model 3 in Table 9 gives the findings for the interaction of readability and student characteristics for the TIMSS. In terms of gender, negative associations between word count and accuracy are stronger for males than females; however, this has a negligible effect size (d = -.03). In addition, negative associations between second-person pronouns and accuracy are stronger for males than females, although this too has a very small effect size (d = -.07). As in the NAEP data, there is evidence in the TIMSS that males are more negatively impacted than females by textual factors that make reading the problem more difficult, although the effects are very small. In terms of birth status, negative associations between second-person pronouns and accuracy are stronger for students not born in the U.S. This suggests that difficult readability factors may have

a stronger association to performance for non-U.S. born students than other students, which has some similarity to the language finding in the NAEP. However, the effect size is again very small (d = -0.07). For mathematics enjoyment, the positive association between concrete words and accuracy is significantly greater for students who enjoy mathematics more (d = .08), and the negative association between word count and accuracy are significantly less strong for students who enjoy mathematics more (d = .01). Although the latter result has a negligible effect size, there is some evidence that enjoying mathematics allows for stronger positive associations between performance and textual characteristics that make a problem easier to read, even when controlling for other background variables like mathematics achievement.

Finally, in terms of mathematics achievement (i.e., plausible values), we see that negative associations between word count and accuracy are less strong for higher achievement students, and this has a considerable effect size (d = 0.34). Further, negative associations between second-person pronouns and accuracy are less strong for higher achievement students (d = 0.30). There is also some evidence that positive associations between pronoun density and accuracy are lower for higher achievement students, but this has a negligible effect size (d = -.04). Overall, having higher mathematics achievement is associated with student performance having a weaker association with difficult readability factors, whereas students with lower mathematics achievement may be differentially negatively impacted by these factors.

Results for the NAEP and the TIMSS on Research Question 1 are summarized in Table 10. Note that, surprisingly, there were few significant interactions for SES (luxury items), language spoken at home, or birth status. This may be because controls for mathematics achievement were masking these differences. We explore this issue next.

Research Question 2: Differential Effects of Readability without Mathematics Achievement

For the analysis for Research Question 2, we omitted the Plausible Value (mathematics achievement) variable from the models. This is because there are well-known achievement gaps between students of different race, genders, and SES, and language backgrounds. The inclusion of this predictor may mask the differential effects of readability these students actually

experience, as much of the inequity is likely explained by falling behind in mathematics knowledge. The results for this analysis are in Model 4 in Tables 8 (NAEP) and 9 (TIMSS). Here we focus on results that are new or different between these models and the previous models.

In the NAEP models, the biggest difference when removing the mathematics achievement controls was the interaction effects that appeared for luxury items (our proxy for SES). Results showed that for students who own fewer luxury items, there is a stronger positive association between accuracy and more concrete words (d = .09) and the same held for more pronouns (*d* = .06), compared to those who owned more luxury items. Students who owned more luxury items had a weaker negative association between second-person pronouns and performance (*d* = 0.16). In terms of findings related to race, the new models did not detect a significant interaction between the racial category of African-American and word concreteness, but did find that African-American students had a stronger negative association between increased word count and decreased performance (*d* = -.06), compared to Caucasian students. The opposite trend was true for students of "Other" races/ethnicities, compared to Caucasian students (*d* = .07). Finally, the interaction between male students and second-person pronouns was no longer significant. Overall, when controls for mathematics achievement are removed, we see stronger differential readability impacts for luxury items (SES), and some racial differences.

In the TIMSS models, we see a similar trend that luxury items (our proxy for SES) are now more important for differential readability effects. The negative association between word count and accuracy is less strong for students with more luxury items students (*d* = .10). Similarly, the negative association between word count and accuracy is stronger for students born outside of the U.S., although the effect size us very small (*d* = -.05). This suggests that the inclusion of mathematics achievement may have been masking some of the difficulties that students with fewer luxury items (low SES) and students born outside the U.S. experience with difficult-to-read word problems. In addition, two more interactions that were very close to significance in Model 3 now show as significant in Model 4 – the positive association between pronoun density and accuracy is stronger for students who enjoy mathematics more (*d* = .05),

and the association between presence of a real-world context and accuracy is significantly more positive for males than females, although the effect size is negligible (d = .02).

Additional Analyses

Exploring Word Count and Second-Person Pronouns

The strongest and most consistent effects in the above models tended to be for word count and presence of second-person pronouns. One hypothesis is that problems that are more mathematically difficult tend to have more words, and tend to use second-person pronouns, thus the interactions we are detecting are actually between problem difficulty and student characteristics, not problem readability and student characteristics. While it is not clear whether there is a purely "mathematical" difficulty for a problem that is completely distinct from the "reading" difficulty of the problem, or whether it would even be sensible to try to separate these components, we present some additional explorations in this section.

We examined the problems that tended to be especially long or short and the problems that tended to have second-person pronouns. Examples of such problems were shown in Table 1. Problems that used second-person pronouns sometimes put the reader into the problem (e.g., the problem was a story about something "you" did). However, more often, these problems were using second-person pronouns to give the reader specific technical instructions (e.g., round your answer) or more general instructions to show their work or steps. Problems that had low word counts tended to give brief information for a simple calculation or definition, or had a small amount of text because they referred the reader to a visual. Problems that were long tended to describe a complex mathematical situation or give a significant amount of procedural instruction to the reader. Note that the presence of a real-world context was controlled for in all models, thus the findings were not being driven by an increased tendency of longer problems or problems with second-person pronouns to be in a real-world context.

This suggests that word count and second-person pronouns may be associated with mathematical aspects of the problem's difficulty and complexity. In addition, these readability factors may be more likely to occur when the question is short answer (rather than multiple

choice). Thus, we next did analyses that controlled for interactions between all student background characteristics and the variables that were included for problem difficulty, problem complexity, and problem type (multiple choice or short constructed response). It is important to note that categorizations of difficulty and complexity were not necessarily made based on mathematical structure alone, and also likely involved aspects of the text's readability; indeed, the link between mathematical difficulty and reading difficulty is impossible to untangle in word problems that are not artificially experimentally manipulated.

Models Controlling for Interactions between Student Characteristics and Problem Difficulty

When the NAEP models were fit with the additional student background by problem difficulty interactions, we found slightly different results than in Models 3 and 4. The interaction effects that were lost included the interaction between gender and presence of second-person pronouns and between luxury items and second-person pronouns. A new significant interaction between presence of a real-world context and African-American students was detected, but had a negligible effect size (d = -.02). In addition, new interactions were detected between luxury and real world contexts and luxury and word count, but also had negligible effect sizes (d = .03 and d = .04). Finally, there was a new, significant interaction between mathematics achievement and pronoun density (d = 0.14), which suggested a stronger positive association between pronoun density and performance for higher achievement students than lower achievement students. Overall, for the NAEP, we see limited evidence that some of the interactions detected for second-person pronouns were interactions with problem difficulty. Second-person pronoun interactions remained significant for mathematics achievement (d = .21) and racial/ethnic categories (d = -.10 for African-American students and d = -.08 for Hispanic students).

When TIMSS models were fit with the additional student background by problem difficulty interactions, we again found similar results to those found in Models 3 and 4. The interaction effects that were lost included the interaction between mathematics achievement and pronouns and the interaction between mathematics achievement and second-person pronouns (ps > 0.5). In turn, the model showed a new, significant interaction between word concreteness and

mathematics achievement, suggesting that the positive association between word concreteness and accuracy was less strong for high achieving students (d = -.10). Overall, there is little evidence in the TIMSS data to suggest that interactions detected between readability and student background characteristics were interactions between problem difficulty and student background.

Discussion

We conducted analyses of nearly 20 years of mathematics achievement data from the 4th and 8th grade NAEP and TIMSS, and explored whether the readability characteristics of mathematics word problems (length, pronouns, word difficulty, and real world context) tend to differentially impact different student groups. The key student demographic characteristics we examined were gender, race, mathematics achievement, mathematics attitudes, language(s) spoken at home, birth status, and luxury items (SES). Here we focus on results that were replicated across both datasets, and/or results that had effect sizes nearing practical significance.

Gender

Although there was some suggestion in both datasets that males tend to benefit more from readability characteristics that make problems easier to read (i.e., pronouns that are not second-person pronouns) and tend to be harmed more by readability characteristics that make problems harder to read (i.e., more words and second-person pronouns), effect sizes were all very small (ds < .08). These differential effects for readability based on gender may stem from reading achievement differences favoring females, which have been documented in numerous studies (Chatterji, 2006; Department for Education and Skills, 2007; Halpern et al., 2007; Logan & Johnston, 2010; Mullis et al., 2012; Twist et al., 2007).

Race

Differential effects of readability based on racial/ethnic background were only examined for the NAEP. However, this analysis suggested that African-American students and Hispanic students may tend to suffer more from readability characteristics that make a mathematics word problem less readable – specifically, second-person pronouns. Here, effect sizes were bordering practical significance. An analysis of problems that contained second-person pronouns suggested

that these problems often give the students instructions (e.g., show your work) that take them outside of the mathematical scenario and ask them to consider the manner in which they are approaching the problem and presenting their solution.

Language and Country of Birth

Although language(s) spoken at home (from NAEP) and country of birth (from TIMSS) are clearly very different background measures, in both cases we found few statistically or practically significant differential effects for readability characteristics of word problems. This is surprising given previous work on language and story problems (e.g., Abedi et al., 2000; Martiniello, 2008). It is important to note that neither of these measures were technically a measure of ELL status, and that the TIMSS measure was likely only indirectly related to language background.

Socioeconomic Status

When the models included controls for mathematics achievement, we found few differential effects based on owning luxury items (our proxy for SES). The strong relationship between SES and mathematics achievement is well-known. However, when mathematics achievement was removed as a variable, we saw effects in the expected direction in both datasets. Students with fewer luxury items seemed to benefit more from readability characteristics that made problems easier to read (i.e., more pronouns and more concrete words). Correspondingly, they were negatively impacted more by readability characteristics that made problems harder to read (i.e., more words and second-person pronouns). This makes sense given that students of lower SES struggle more in reading and mathematics than students of higher SES (Noble et al., 2006; Perry & McConney, 2010; Sirin, 2005).

Mathematics Attitudes

We only had mathematics attitudes data in the TIMSS, which was the smaller dataset. However, students with more positive mathematics attitudes tended to benefit more from concrete words. It is important to note that these analyses controlled for mathematics achievement. Previous findings have indicated that positive attitudes can increase focus and

attention towards material (Pekrun, Goetz, Titz, & Perry, 2002) and are positively associated with students making connections within a text and between a text and their background knowledge (Clinton & van den Broek, 2012). Positive attitudes may thus allow students to engage more deeply with the problem situation, which may allow the benefits of concrete words for situation model construction to be realized.

Mathematics Achievement

We saw the strongest differential effects for readability for students with different mathematics achievement levels. Students who were weaker in mathematics tended to benefit more from factors that made problems easier to read (i.e., concrete words and pronouns). These students tended to suffer more from factors that made problems harder to read (i.e., more words and second-person pronouns). These results make sense given that learners with high mathematics knowledge can better compensate for fewer supports in the environment (Mayer, 2001), and that mathematics and reading achievement are closely related (Wu, 2010).

Limitations

There are a number of methodological limitations to this analysis. First, all analyses are strictly correlational, thus only associations can be examined, and no conclusive evidence for causation can be gleaned. Studies of this type should be supplemented by experimental research where readability characteristics of word problems are systematically varied, and performance is examined on mathematically-matched problems that vary only with respect to a readability characteristic. However, it is difficult to do this type of experimental study on a scale that allows for a sensible examination of how readability characteristics differentially impact different student groups, and very large sample sizes would be needed to get enough students with each combination of demographic characteristics. This is the strength of the NAEP and TIMSS data – although correlational, we have the sample size needed to make sensible comparisons between student demographic groups across a variety of demographic variables.

A second limitation was the demographic measures that were available. We did not have a measure of reading ability, we only had racial background in one dataset, we did not have a

measure of ELL status that was available in both datasets, and SES was approximated based on books and computer in home variables. A more robust set of demographic variables certainly would have strengthened the analysis. However, having enough mathematics word problems such that we had appropriate variance on all of our problem-level and readability predictors was essential, and another important strength of using these large-scale datasets without omitting years where less or different demographic data were collected. Because of problem sample size considerations, we were also unable to look at how the impact of readability may have changed over time, for particular groups of years; this would be an interesting direction for future work if problem banks of sufficient size become available.

A third limitation was the problem-level variables that were available in the NAEP and TIMSS. Specifically, there was no sensible way to separate problem "mathematical" difficulty from "reading" difficulty. In prior work (Walkington et al., 2015), we analyzed a corpus of problems that all covered the same narrow mathematical concept – linear functions. This allowed us to put in careful controls for aspects of the problem's mathematical structure, but at the expense of greatly hindering the generalizability of any results we detected. Both types of analyses are needed to understand how readability of mathematics word problems is related to student performance.

A fourth limitation is the narrow scope of readability variables examined. Both to avoid Type I errors and for methodological reasons, only a select number of variables were examined. Unfortunately, this meant many sentence-level variables that may be especially important in the short texts used in these analyses were not considered, such as the number of words before the main verb and the number of phrases (McNamara et al., 2013). It is possible that the variable of text length used in this study encompassed some of these important sentence-level measures as more words before the main verb and a greater number of phrases would likely lead to a greater number of words. Moreover, many of the cohesion metrics in Coh-Metrix require multiple paragraphs and at least 300 words to be valid (McNamara et al., 2014). In addition, future work in variables known to be important for readability in short texts, such as the number of

propositions and the length of propositions (Miller & Kintsch, 1980), would contribute to better connecting fields of text comprehension and mathematical problem solving. As discussed earlier, our analyses and results are ultimately limited by how to measure and operationalize "readability," and the inherent limitations therein.

Implications

Findings support previous work indicating the importance of reading comprehension in solving story problems (Hecht et al., 2001; Lerkkanen et al., 2005). Reading comprehension is a complex, multidimensional construct (Kintsch, 1998), and traditional measures of reading difficulty provide only a coarse estimate of the amount of information a reader needs to process, but do not capture the nuance of language (Graesser, McNamara, Louwerse, & Cai, 2004). Readability tools like Coh-Metrix provide fine-grained measures that can specify different features of readability. Knowing these specific features allows for more targeted interventions to improve the readability of mathematics items.

Our analyses suggest that readability characteristics of mathematics word problems are differentially impacting different student groups. Groups of students that have historically stronger performance in mathematics or reading – including Caucasian students and higher SES students – tend to see less impact on their performance on mathematics test items when the reading difficulty of the mathematics problem is greater. This is also true for females who have stronger academic performance in reading, which suggests that more developed mathematical knowledge may allow these students to compensate for difficult textual and linguistic structures, facilitating situation and problem model construction. On the other hand, students from groups who have historically performed less well in mathematics and reading – including African-American and Hispanic students and lower SES students – tend to benefit more when problems are concrete, concise, and understandable. This is also true for males, who have weaker academic performance in reading. Such characteristics may support the challenging task of constructing a situation model and coordinating it with a problem model. These types of readability modifications may be critical to fairly teaching and evaluating students from diverse groups.

Explicit considerations of reading difficulty when evaluating students' mathematical knowledge via assessments may be crucial to using assessments to create meaningful targets and interventions to improve mathematics instruction.

Although we evaluated pre-selected readability categories here, it seems likely that effects would be similar for other factors that are known to make texts more or less easy to read for students – such as other measures of word difficulty and measures of similarity and overlap between ideas in sentences. It was also interesting to find that the effect of a real-world context was not pronounced as a main effect and only produced a few, very weak interactions that suggested these contexts may benefit male students, Caucasian students, and higher SES students. Although such contexts have been hailed as making mathematical ideas more accessible compared to mathematical abstractions by allowing for the support of a situation model (Walkington et al., 2012), if not designed and written with diverse students in mind they may serve to further alienate these students.

Given that readability matters for mathematics word problems, and readability issues differentially impact different student groups, how can mathematics instruction be designed to support all students? One idea is that teachers can focus on a few critical readability factors when posing problems that cover mathematics concepts they know will be challenging and new for students. Results here suggest that shortening the problem text and including concrete words and scenarios may be particularly important to support struggling learners.

In addition, several interventions to support students in reading mathematics word problems have been proposed. In *schema-based instruction* (Fuchs et al., 2004), students learn to identify different types or classes or word problems before choosing strategies to solve them. *Solve It!* is another such intervention where "Students are taught how to read the problem for understanding, paraphrase by putting the problem into their own words, visualize the problem by drawing a picture or making a mental image, set up a plan for solving the problem, estimate the answer, and compute and verify the solution" (Montague, Warner, & Morgan, 2000, p. 111). These interventions are intended to move students away from direct translation and keyword-

type strategies, and encourage careful reading of the problem situation and the formulation of a sensible situation model, a key element of the theory of successful word problem solving we discussed. To form accurate situation models, students need significant experiences making sense of meaningful mathematical scenarios, discussing alternative interpretations, reading critically, defending claims with evidence, and drawing out misconceptions. This kind of engagement, detailed in a recent discussion of *Anticipation Guides* for reading mathematical texts (Adams, Pegg, & Case, 2015), can help deepen students mathematical understanding while simultaneously developing their reading skills. In these guides, students determine whether statements are true or false using evidence they recorded from the text, by reading interactively and drawing upon their mathematical knowledge.

Our results also raise questions about how reading and mathematics instruction should be intertwined. Mathematics teachers could develop an understanding of what factors make reading mathematics problems difficult, and implement strategies to assist students in overcoming these difficulties. Particularly, discipline-specific reading behaviors that are endemic to the genre of mathematics story problems could be developed and cultivated (see Shanahan & Shanahan, 2008). The Common Core Standards for English Language Arts, while perhaps not familiar to all mathematics teachers, include discipline-specific reading standards for technical texts, like "Determine the meaning of symbols, key terms, and other domain-specific words and phrases as they are used in a specific scientific or technical context…" (CCSS, 2010). The departmentalized nature of secondary mathematics instruction may make the divide between reading and mathematics particularly wide, and new models for collaboration and joint planning could be explored. Having students take on the role as the author and creator of mathematics story problems may be another promising approach to help students understand the discursive structure of story problems. Indeed, Walkington (in press) reports an intervention at a diverse urban school with a high proportion of ELLs, where students posed their own algebra story problems related to their interests, and then showed improved post-test performance on standard story problems compared to a control group. Additionally, embedding compelling mathematical

tasks that are motivated by complex reading passages, such as model-eliciting activities (Lesh & Harel, 2003), into the curriculum with appropriate support for the reading demands can allow students to gain exposure to and experience with difficult reading in the mathematics classroom.

A recent article in the *New York Times* (Hartocollis, 2016) titled "New, Reading-Heavy SAT Has Students Worried" discussed the College Board's newly released SAT. The test has raised concerns among some that it may contain lengthier and more difficult reading passages and mathematics word problems. These changes are thought to differentially impact some of the most vulnerable student populations on a test that has compelling implications for college and careers. It is critical to continue to investigate and understand how the verbal structure of mathematics word problems may impact the performance and access of different student groups.

Acknowledgements

This research was supported by a grant from the American Educational Research Association which receives funds for its "AERA Grants Program" from the National Science Foundation under Grant #DRL-0941014. Opinions reflect those of the author and do not necessarily reflect those of the granting agencies. Special thanks to Paul Yovanoff for his assistance, and to Alyssa Holland, Valentin Ortiz, and Brannon Bradshaw for their help with data entry. We thank Nia Dowell, a doctoral student at the Institute for Intelligent Systems in the University of Memphis, for her assistance processing the text files. We also acknowledge the FedEx Institute for Technology for funding the Coh-Metrix Text Analysis Service. Finally, we acknowledge the assistance of the IES Data Security office for their help with accessing the NAEP student-level dataset.

References

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). NAEP Mathematics Performance and Test Accommodations: Interactions with Student Language Background. CSE Technical Report.

Abedi, J., Leon, S., & Mirocha, J. (2001). Validity of Standardized Achievement Tests for English Language Learners. Paper presented at the American Educational Research Association Conference, Seattle, WA. Retrieved from: http://eric.ed.gov/?id=ED455292

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16–26.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*, 219-234.

Abedi, J., Lord, C., & Hofstetter, C. (2001). Impact of Selected Background Variables on Students' NAEP Mathematics Performance. U.S. Department of Education, National Center for Education Statistics. NCES Working Paper Series 2001-11.

Adams, A. E., Pegg, J., & Case, M. (2015). Anticipation Guides: Reading for Mathematics Understanding. *Mathematics Teacher*, *108*, 498-504.

Arnold, J. E., Losongco, A., Wasow, T., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 28-55. doi: 10.2307/417392

August, D., & Shanahan, T. (Eds.). (2008). Developing reading and writing in second-language learners: Lessons from the report of the National Literacy Panel on Language-Minority Children and Youth. New York: Routledge; Newark, DE: Center for Applied Linguistics and the International Reading Association.

Bailin, A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. *Language & Communication*, *21*(3), 285-301.

Ball, A. F. (1996). Expository writing patterns of African American students. *The English*

Journal, *85*(1), 27-36.

Barth, A. E., Tolar, T. D., Fletcher, J. M., & Francis, D. (2014). The effects of student and text characteristics on the oral reading fluency of middle-grade students. *Journal of Educational Psychology*, *106*, 162-180.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7.

Bauer, L. M., & Altarriba, J. (2008). An investigation of sex differences in word ratings across concrete, abstract, and emotion words. *The Psychological Record*, *58*, 465–474.

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review, 24*, 63-88.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Bernhardt, E. B., & Kamil, M. L. (1995). Interpreting relationships between L1 and L2 reading: Consolidating the linguistic threshold and the linguistic interdependence hypotheses. *Applied Linguistics*, *16*(1), 15-34.

Bertram, B. & Newman, S. (1981). *Why readability formulas fail*. Reading Education Report No. 28. Illinois University, Urbana. Center for the Study of Reading.

Carpenter, T., Fennema, E., Franke, M., Levi, L., & Empson, S. (1999). *Children's Mathematics: Cognitively Guided Instruction.* Heinemann.

Carpenter, T., & Moser, J. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal for Research in Mathematics Education*, *15*(3), 179-202.

Charity, A. H., Scarborough, H. S., & Griffin, D. M. (2004). Familiarity with school English in African American children and its relation to early reading achievement. *Child development*, *75*, 1340-1356.

Chatterji, M. (2006). Reading achievement gaps, correlates, and moderators of early reading achievement: Evidence from the Early Childhood Longitudinal Study (ECLS)

kindergarten to first grade sample. *Journal of Educational Psychology*, *98*, 489-507.

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in medicine*, *19*(22), 3127-3131. Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in medicine*, *19*, 3127-3131.

Clinton, V., Seipel, B., van den Broek, P., McMaster, K.L., Kendeou, P., Carlson, S., & Rapp, D.N. (2014). Gender differences in inference generation by fourth-grade students. *Journal of Research in Reading 37*, 356-374.

Clinton, V., & van den Broek, P. (2012). Interest, inferences, and learning from texts. *Learning and Individual Differences, 22*, 650-663.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Coltheart, M. (1981). *MRC Psycholinguistic Database User Manual: Version 1.*

Common Core State Standards (CCSS) Initiative (2010). Common Core State Standards for English Language Arts. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers.

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.

Crawford, L., Tindal, G., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment*, *7*(4), 303-323.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology*, *20*, 405-438.

De Groot, A., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, *50*(1), 1-56.

Delpit, L., & Dowdy, J. K. (Eds.). (2002). *The skin that we speak: Thoughts on language and*

culture in the classroom. The New Press.

Department for Education and Skills (2007). Gender and education: The evidence on pupils in England (Reference No. 00389-2007BKT-EN). Nottingham: DfES Publications

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.

Frankenstein (2009). Realistic contexts, mathematics assessment, and social class. In B. Greer, L. Verschaffel, W. Van Dooren, & S. Mukhopadhyay (Eds.) *Word and Worlds: Modelling Verbal Descriptions of Situations*. Rotterdam, Netherlands: Sense Publishers.

Fry, R. (2007). How Far behind in Mathematics and Reading Are English Language Learners? Report. *Pew Hispanic Center*.

Fuchs, L. S., Fuchs, D., Prentice, K., Hamlett, C. L., Finelli, R., & Courey, S. J. (2004). Enhancing Mathematical Problem Solving Among Third-Grade Students with Schema-Based Instruction. *Journal of Educational Psychology*, *96*, 635-647.

Fuentes, P. (1998). Reading comprehension in mathematics. *The Clearing House, 72*(2), 81-88.

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.

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*. Rotterdam, the Netherlands: Sense Publishers.

Halpern, D.F., Benbow, C.P., Geary, D.C., Gur, R.C., Hyde, J.S. & Gernsbacher, M.A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*(1), 1–51.

Hartocollis, A. (2016, February 9). New, reading-heavy SAT has student worried. *The New York Times*, pp. A1.

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.

Hegarty, M., Mayer, R. E., & Monk, C. A. (1995). Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers. *Journal of Educational Psychology*, *87*, 18-32.

Helwig, R., Rozek-Tedesco, M. A., Tindal, G., Heath, B., & Almond, P. J. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *The Journal of Educational Research*, *93*(2), 113-125.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172-177.

Inoue, N. (2005). The realistic reasons behind unrealistic solutions: The role of interpretive activity in word problem solving. *Learning and Instruction*, *15*(1), 69-83.

Jiban, C. L., & Deno, S. L. (2007). Using Mathematics and Reading Curriculum-Based Measurements to Predict State Mathematics Test Performance Are Simple One-Minute Measures Technically Adequate?. *Assessment for Effective Intervention*, *32*(2), 78-89.

Jin, Y. S. (1990). Effects of concreteness on cross-language priming in lexical decisions. *Perceptual and Motor Skills*, *70*, 1139-1154.

Kazemi, E. (2002). Exploring test performance in mathematics: The questions children's answers raise. *The Journal of Mathematical Behavior*, *21*(2), 203-224.

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. *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: Cambridge University Press.

Kim, S., Jiang, Y., & Song, J. (2015). The effects of interest and utility value on mathematics engagement and achievement. In A. Renninger, M. Nieswandt, & S. Hidi (Eds,) *Interest in Mathematics and Science Learning* (pp. 63-78), American Educational Research Association.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge university press.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review*, *92*(1), 109-129.

Kirkwood, K. J., & Wolfe, R. G. (1980). *Matching Students and Reading Materials: A Cloze-Procedure Method for Assessing the Reading Ability of Students and the Readability of Textual Materials*. Toronto, Canada: Ontario Government Bookstore.

Koedinger, K., & Nathan, M. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *Journal of the Learning Sciences, 13*(2), 129-164.

Ladsen-Billings (1995). Making mathematics meaningful in multicultural contexts. In W. Secada (Ed.), *For Equity in Mathematics Education* (pp. 126-145). Cambridge University Press.

Ladsen-Billings, G. (1997). It doesn't add up: African American students' mathematics achievement. *Journal for Research in Mathematics Education*, *28*, 697-708.

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*(2), 121-137.

Lesh, R., & Harel, G. (2003). Problem solving, modeling, and local conceptual development. Mathematical Thinking and Learning, *5*(2/3), 157-190.

Lesaux, N. K., & Kieffer, M. J. (2010). Exploring sources of reading comprehension difficulties among language minority learners and their classmates in early adolescence. *American Educational Research Journal*, *47*, 596-632.

Logan, S. & Johnston, R. (2010). Investigating gender differences in reading. *Educational Review*, *62*(2), 175– 187.

Loveless, T, Fennel, F., Williams, V., Ball, D., & Banfield, M. (2008). Chapter 9: Report of the Subcommittee on the National Survey of Algebra I Teachers. In *Foundations for Success: Report of the National Mathematics Advisory Panel*.

Mayer, R. (2001). *Multimedia Learning*. Cambridge University Press.

Martiniello, M. (2008). Language and the performance of English-language learners in mathematics word problems. *Harvard Educational Review*, *78*, 333-368.

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

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

Miller, J.R. & Kintsch, W. (1980). Readability and recall of short prose passages: A theoretical analysis. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(4), 335-353.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133-161.

Montague, M., Warger, C., & Morgan, T. H. (2000). Solve it! Strategy instruction to improve mathematical problem solving. *Learning Disabilities Research & Practice*, *15*(2), 110-116.

Mullis, I. V., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 International Results in Reading*. International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, the Netherlands.

Mullis, I. V., Martin, M. O., & Foy, P. (2013). The impact of reading ability on TIMSS mathematics and science achievement at the fourth grade: An analysis by item reading demands. In Martin, M. O., & Mullis, I. V. S. (Eds.) T*IMSS and PIRLS 2011:*

Relationships among reading, mathematics, and science achievement at the fourth grade - Implications for early learning, (pp. 67-108). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and IEA.

Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, The Netherlands.

Nasir, N., Hand, V., & Taylor, E. V. (2008). Culture and mathematics in school: Boundaries between "cultural" and "domain" knowledge in the mathematics classroom and beyond. *Review of Research in Education*, *32*(1), 187-240.

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.

National Center for Educational Statistics, U.S. Department of Education (2013). *NAEP 2012: Trends in Academic Progress*. Retrieved from http://nces.ed.gov/nationsreportcard/subject/publications/main2012/pdf/2013456.pdf

National Center for Educational Statistics, U.S. Department of Education (2009). *The Nation's Report Card: An Overview of Procedures for the NAEP Assessment* (NCES 2009–493) U.S. Department of Education. Institute of Education Sciences. National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Neidorf, T.S., Binkley, M., Gattis, K., and Nohara, D. (2006). *Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments* (NCES 2006-029). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, *45*, 211-236.

Noble, K. G., Farah, M. J., & McCandliss, B. D. (2006). Socioeconomic background modulates cognition–achievement relationships in reading. *Cognitive Development*, *21*, 349-368.

Organisation for Economic Cooperation and Development. (2010). *PISA 2009 results: What students know and can do*. Paris: OECD Publications.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, *37*(2), 91-105.

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*.

Perry, L., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *The Teachers College Record*, *112*(4), 7-8.

Pimm, D. (1987). Speaking mathematically: Communication in mathematics classrooms. New York, NY: Routledge.

Pimm, D. (1994). Spoken mathematical classroom culture: Artifice and artificiality. In S. Lerman (Ed.), Cultural perspectives on the mathematics classroom (pp. 133–147). Norwell, MA: Kluwer Academic Publishers.

Provasnik, S., Kastberg, D., Ferraro, D., Lemanski, N., Roey, S., & Jenkins, F. (2012). Highlights from TIMSS 2011: Mathematics and Science Achievement of US Fourth-and Eighth-Grade Students in an International Context. NCES 2013-009. *National Center for Education Statistics*.

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In G.J. Duncan & R.J. Murnane (Eds.) *Whither opportunity: Rising inequality, schools, and children's life chances*, pp. 91-116. New York, NY: Russell Sage Foundations.

Robinson, J. P., & Lubienski, S. T. (2011). The Development of Gender Achievement Gaps in Mathematics and Reading During Elementary and Middle School Examining Direct

Cognitive Assessments and Teacher Ratings. *American Educational Research Journal*, *48*, 268-302.

Saxe, G. (1988). Candy selling and mathematics learning. *Educational Researcher*, *17*(6), 14-21.

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*(2), 105-126.

Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harvard Educational Review*, *78*(1), 40-59.

Schulz, R.A. (1981). Literature and readability: Bridging the gap in foreign language reading. The Modern Language Journal, 65, 43-53.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417-453.

Snijders, T. A. B., & Bosker, R. (1999). *Introduction to multilevel analysis.* London: Sage.

Sweller, J., van Merrienboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*, 251–296.

Tate, W. F. (1994). Race, retrenchment, and the reform of school mathematics. *The Phi Delta Kappan*, *75*, 477-484.

Taylor, E. (2005). Low income first and second grade students' engagement in currency exchange: The relationship to mathematical development. University of California, Berkeley. *Doctor of Philosophy in Education*.

Twist, L., Schagen, I. & Hodgson, C. (2007). Readers and reading: The National Report for England 2006 PIRLS (Progress in International Reading Literacy Study). Slough: NFER.

Vilenius-Tuohimaa, P. M., Aunola, K., & Nurmi, J. E. (2008). The association between mathematical word problems and reading comprehension. *Educational Psychology*, *28*, 409-426.

Von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI monograph series*, *2*, 9-36.

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*(4), 932-945.

Walkington, C. (in press). Design research on personalized problem-posing. Research report to appear in *Proceedings of the 39th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Indianapolis, IN.

Walkington, C., Clinton, V., Ritter, S., & Nathan, M. J. (2015). How readability and topic incidence relate to performance on mathematics story problems in computer-based curricula. *Journal of Educational Psychology*, *107*(4), 1051-1074.

Walkington, C., Clinton, V., Shivraj, P., & Yovanoff, P. (April, 2015). Association between readability and topic of mathematics word problems and performance on large-scale assessments. Presentation at *2015 Annual Meeting of the American Educational Research Association*, Chicago, IL.

Walkington, C., Cooper, J., & Howell, E. (2013). The effects of visual representations and interest-based personalization on solving percent problems. In Martinez, M. & Castro Superfine, A (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., Sherman, M., & Petrosino, A. (2012). 'Playing the game' of story problems: Coordinating situation-based reasoning with algebraic representation. *Journal of Mathematical Behavior*, *31*(2), 174-195.

Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment*, *14*(3-4), 139-159.

Wu, M. (2010). *Comparing the similarities and differences of PISA 2003 and TIMSS* (OECD Working Papers No. 32). Paris: OECD Publishing. doi: 10.1787/5km4psnm13nx-en

Yore, L.D., Pimm, D., & Tuan, H-L. (2007). The literacy component of mathematical and scientific literacy. International *Journal of Science and Mathematics Education, 5*, 559-

589.

Zamanian, M., & Heydari, P. (2012). Readability of Texts: State of the Art. *Theory and Practice in Language Studies*, *2*, 43. doi: 10.4304/tpls.2.1.43-53

Zygouris-Coe, V. (2012). Disciplinary literacy and the common core state standards. *Topics in Language Disorders, 32*(1), 35-50.

Tables

Table 1. Example Problems from different readability categories

	Example 1	Example 2	Example 3
Problems Using Second-Person Pronouns	On average, thunder is heard in Tororo, Uganda, 251 days each year. What is the probability that thunder will be heard in Tororo on any day? (1 year = 365 days) Give your answer to the nearest percent.	There is only one red marble in each of the bags shown below. Without looking, you are to pick a marble out of one of the bags. Which bag would give you the greatest chance of picking the red marble?	Use 2 of the triangle tiles to make one large black triangle. Then show what you did with your tiles by shading in your triangle below.
Problems with Low Word Count	There are 600 balls in a box, and 1/3 of the balls are red. How many red balls are in the box?	Write the names of shapes A, B, and C in the spaces provided.	On the grid, draw a line parallel to line L.
Problems with High Word Count	In a car rally two checkpoints are 160 km apart. Drivers must travel from one checkpoint to the other in exactly 25 hours to earn maximum points. A driver took 1 hour to travel through a 40 km hilly section at the beginning of the course. What must the average speed, in kilometers per hour, be for the remaining 120 km if the total time between checkpoints is to be 25 hours?	In a game, Mysong and Naoki are making problems. They each have four cards like these. The winner of the game is the person who can make the problem with the largest answer. Write numbers in the squares below to show how you would place the cards to beat both Mysong and Naoki.	On this grid, find the dot with the circle around it. We can describe where this dot is by saying it is at First Number 1, Second Number three. Now find the dot with the triangle around it. Describe where the dot is on the grid in the same way. Fill in the numbers we would use.

Table 2. Descriptive Statistics for Readability Variables on NAEP (left) and TIMSS (right)

	NAEP		**TIMSS**	
Variable	**Average (SD)**	**Min/Max**	**Average (SD)**	**Min/Max**
Word Count	26.82 (16.4)	4/147.	24.06 (13.4)	5/75
Word Concreteness	401.9 (52.6)	255/612	394.8 (58.2)	225/599
Pronoun Density	7.3 (5.2)	0/33	8.9 (7.0)	0/40
	No. Problems Present (%)	**No. Problems Not Present (%)**	**No. Problems Present (%)**	**No. Problems Not Present (%)**
Second-person Pronouns (0/1 factor)	30 (5.3%)	535 (94.7%)	26 (6.1%)	402 (93.9%)
Real World Context (0/1 factor)	312 (55.2%)	253 (44.8%)	186 (43.5%)	242 (56.5%)

Note. In the models word count, word concreteness, and pronoun density were normalized, whereas here we provide unmodified values to allow for comparability to Coh-Metrix documentation.

Table 3. Descriptive Statistics for Student Demographic Variables on NAEP (left) and TIMSS (right)

		NAEP			TIMSS		
Variable	**Levels**	**Number of Observations**	**Number of Students (%)**	**Average % Correct**	**Number of Observations**	**Number of Students (%)**	**Average % Correct**
Sex	Male	765,120	357,730 (49.3%)	52.05	355,450	20,550 (50.6%)	56.5
	Female	745,890	347,560 (50.7%)	50.72	364,570	20,030 (49.4%)	54.9
	Missing	680	310 (.04%)	49.00	0	0	NA
Race	Caucasian	890,930	414,510 (58.7%)	56.76	242,470	13,240 (32.6%)	61.4
	African-American	257,870	120,840 (17.1%)	39.75	62,660	3,400 (8.4%)	46.9
	Hispanic	235,350	110,840 (15.7%)	43.55	101,070	5,470 (13.5%)	50.5
	Other	102,860	47,620 (6.7%)	52.20	36,080	1,940 (4.8%)	61.2
	Missing	24,680	11,790 (1.7%)	50.75	277,740	16,520 (40.7%)	54.5
Language	Speaks only English	830,770	387,500 (54.9%)	52.75	Not avail.	Not avail.	Not avail.
	Speaks another language	642,180	300,060 (42.5%)	50.02	Not avail.	Not avail.	Not avail.
	Missing	38,740	18,040 (2.6%)	45.07	Not avail.	Not avail.	Not avail.
Birth Status	U.S. Born	Not avail.	Not avail.	Not avail.	630,670	35,440 (87.3%)	56.9
	Non-U.S. Born	Not avail.	Not avail.	Not avail.	75,960	4,330 (10.7%)	48.1
	Missing	Not avail.	Not avail.	Not avail.	13,400	810 (2.0%)	44.0

Note. The overall average percent correct for NAEP was 51.39%, and for TIMSS was 55.69%.

Table 4. Descriptive Statistics for Student Demographic Variables that are Continuous on NAEP (left) and TIMSS (right)

	NAEP				TIMSS			
Variable	**No. Observ**	**Avg (SD)**	**No. Missing Observ (%)**	**Avg % Correct - Missing**	**No. Observ**	**Avg (SD)**	**No. Missing Observ (%)**	**Avg % Correct - Missing**
Plausible Values (Normalized)	1,507,010	0 (1)	4,680 (0.31%)	35.7	720,010	0 (1)	0 (0%)	NA
Luxury	1,479,330	0.75 (0.26)	32,360 (2.14%)	46.4	711,270	0.39 (0.27)	8,750 (1.2%)	43.0
Enjoy Mathematics (Normalized)	Not avail.	Not avail.	Not avail.	Not avail.	698,680	0 (1.0)	21,330 (2.96%)	43.5

Note. The overall average percent correct for NAEP was 51.39%, and for TIMSS was 55.69%.

Table 5. Descriptive Statistics for Problem Variables on NAEP (left) and TIMSS (right)

		NAEP			TIMSS		
Variable	**Levels**	**No. of Observ**	**No. of Problems (%)**	**Avg % Correct**	**No. of Observ**	**No. of Problems (%)**	**Avg % Correct**
Difficulty	Easy	633,150	230 (40.7%)	72.1	142,430	82 (19.2%)	79.3
	Medium	500,540	189 (33.5%)	45.4	467,140	276 (64.5%)	55.3
	Hard	378,000	146 (25.8%)	24.6	110,440	70 (16.4%)	26.9
Complexity	Low	803,090	287 (50.8%)	57.9	335,750	201 (47.0%)	59.5
	High	708,600	278 (49.2%)	44.1	384,270	227 (53.0%)	52.4
Type	Multiple Choice	1,062,230	390 (69.0%)	55.8	462,840	279 (65.2%)	60.5
	Constructed Response	449,470	175 (31.0%)	41.1	257,170	149 (34.8%)	47.1
Content Domain	Algebra	287,970	108 (19.1%)	46.9	105,550	63 (14.7%)	53.8
	Data	195,770	74 (13.1%)	54.9	100,940	59 (13.8%)	66.0
	Geometry	275,700	102 (18.1%)	52.2	142,740	89 (20.8%)	54.6
	Measurement	279,060	109 (19.3%)	52.3	78,140	48 (11.2%)	44.9
	Number	473,180	172 (30.4%)	51.7	292,650	169 (39.5%)	56.2

Table 6. Descriptive Statistic for Problem/Student Variables on NAEP

Variable	**Levels**	**Number of Observations**	**Number of Students (%)**	**Number of Problems (%)**	**Average % Correct**
Year	1996	11,590	7,470 (1.1%)	44 (7.8%)	39.6
	2003	376,870	162,570 (23.0%)	108 (19.1%)	53.6
	2005	265,630	125,960 (17.9%)	80 (14.2%)	52.8
	2007	359,620	156,690 (22.2%)	102 (18.1%)	50.5
	2009	107,390	57,240 (8.1%)	59 (10.4%)	50.9
	2011	172,730	85,280 (12.1%)	88(15.6%)	49.6
	2013	217,860	110,380 (15.6%)	84 (14.9%)	49.5
Grade	4	768,750	368,460 (52.2%)	264 (46.7%)	51.7
	8	742,940	337,150 (47.8%)	301(53.3%)	51.1
Overall		***1,511,690***	***705,600***	***565***	***51.4***

Table 7. Descriptive Statistics for Problem/Student Variables on TIMSS

Variable	**Levels**	**Number of Observations**	**Number of Students (%)**	**Number of Problems (%)**	**Average % Correct**
Year	1995	118,710	6,980 (17.2%)	63 (14.7%)	49.7
	1999	144,980	8,780 (21.6%)	70 (16.4%)	57.5
	2003	228,260	12,480 (30.8%)	156 (36.4%)	57.3
	2011	228,070	12,340 (30.4%)	139 (32.5%)	56.0
Grade	4	355,550	20,060 (49.4%)	205 (47.9%)	58.8
	8	367,580	20,510 (50.6%)	223 (52.1%)	52.7
Overall		***720,010***	***40,570***	***428***	***55.7***

Table 8. Mixed Effects Logistic Regression Results for NAEP dataset

	Model 1 – Null	**Model 2 - Readability**	**Model 3 – Readability Interactions**	**Model 4 – No Plausible Values**
Random Intercepts	**Var.**	**Var.**	**Var.**	**Var.**
Problem	0.3767	0.3565	0.3720	0.2771
Year	0.0182	0.0170	0.0171	0.0040
Fixed Effect	**Est. (SE) Sig**	**Est. (SE) Sig**	**Est. (SE) Sig**	**Est. (SE) Sig**
(Intercept)	2.038 (0.0901)*	1.887 (0.0968)*	1.9134 (0.0984)*	0.5251 (0.0584)*
Grade-4	(ref.)	(ref.)	(ref.)	(ref.)
Grade-8	-1.383 (0.0525)*	-1.306 (0.0535)*	-1.3142 (0.0542)*	0.0126 (0.0434)
Type-MultipleChoice	(ref.)	(ref.)	(ref.)	(ref.)
Type-ConstructedResponse	-0.590 (0.0583)*	-0.471 (0.0608)*	-0.4786 (0.0624)*	-0.4428 (0.0494)*
Difficulty-Easy	(ref.)	(ref.)	(ref.)	(ref.)
Difficulty-Medium	-1.420 (0.0602)*	-1.364 (0.0594)*	-1.3584 (0.0607)*	-1.1744 (0.0443)*
Difficulty-Hard	-2.520 (0.0677)*	-2.446 (0.0678)*	-2.4568 (0.0687)*	-2.1288 (0.0522)*
Complexity-Low	(ref.)	(ref.)	(ref.)	(ref.)
Complexity-High	-0.209 (0.0586)*	-0.153 (0.0586)*	-0.1439 (0.0597)*	-0.1465 (0.0448)*
Domain-Algebra	(ref.)	(ref.)	(ref.)	(ref.)
Domain-Data	0.054 (0.0871)	0.057 (0.0886)	0.0528 (0.0913)	0.0212 (0.0634)
Domain-Geometry	0.104 (0.0806)	0.102 (0.0807)	0.0897 (0.0829)	0.0666 (0.0577)
Domain-Measurement	0.006 (0.08)	-0.085 (0.0805)	-0.1002 (0.0827)	-0.1094 (0.0564)
Domain-Number	0.062 (0.0724)	0.005 (0.0717)	-0.0009 (0.0737)	-0.0296 (0.0501)
Sex-Female	(ref.)	(ref.)	(ref.)	(ref.)
Sex-Male	0.031 (0.0043)*	0.031 (0.0043)*	0.0049 (0.007)	0.0712 (0.0063)*
Race-Caucasian	(ref.)	(ref.)	(ref.)	(ref.)
Race-AfricanAmerican	-0.027 (0.006)*	-0.027 (0.006)*	-0.0064 (0.0099)	-0.6788 (0.0088)*
Race-Hispanic	-0.022 (0.0068)*	-0.022 (0.0068)*	-0.0219 (0.0114)	-0.4467 (0.0101)*
Race-Other	0.021 (0.0094)*	0.021 (0.0094)	0.0253 (0.0149)	-0.084 (0.0131)*
Lang-EnglishOnly	(ref.)	(ref.)	(ref.)	(ref.)
Lang-OtherLang	0.0002 (0.0047)	0.0002 (0.0047)	-0.002 (0.0079)	0.0079 (0.0071)
LuxuryItems	0.045 (0.009)*	0.045 (0.0090)*	0.0268 (0.0145)	1.013 (0.0127)*
MathAchievement	1.230 (0.0036)*	1.230 (0.0036)*	1.2009 (0.0054)*	
RealWorldContext		0.122 (0.0638)	0.0789 (0.0668)	0.1141 (0.0502)
WordCount		-0.104 (0.0325)*	-0.1156 (0.0345)*	-0.1636 (0.0275)*
WordConcreteness		0.069 (0.0284)*	0.0904 (0.03)*	0.1097 (0.0254)*
PronounDensity		0.080 (0.0283)*	0.0827 (0.03)*	0.0934 (0.0246)*
2ndPersonPronoun		-0.328 (0.1144)*	-0.3948 (0.1256)*	-0.4268 (0.0703)*
Sex-Male: RealWorldContext			0.0499 (0.0094)*	0.0369 (0.0086)*
Sex-Male: WordCount			-0.0399 (0.0049)*	-0.0262 (0.0044)*
Sex-Male: WordConcreteness			-0.0076 (0.0045)	-0.0067 (0.0041)

Sex-Male: PronounDensity	-0.0134 (0.0044)*	-0.0109 (0.0041)*
Sex-Male: 2ndPersonPronoun	-0.0667 (0.0207)*	-0.032 (0.0181)
Race-AfricanAmerican :RealWorldContext	-0.025 (0.0136)	-0.018 (0.012)
Race-Hispanic: RealWorldContext	0.01 (0.0157)	0.0038 (0.0139)
Race-Other: RealWorldContext	-0.0022 (0.0201)	-0.0211 (0.018)
Race-AfricanAmerican: WordCount	-0.0146 (0.0073)	-0.0362 (0.0064)*
Race-Hispanic: WordCount	-0.011 (0.0083)	-0.0109 (0.0072)
Race-Other: WordCount	0.0045 (0.0105)	0.0393 (0.0092)*
Race-AfricanAmerican: WordConcreteness	-0.0184 (0.0064)*	-0.0004 (0.0057)
Race-Hispanic: WordConcreteness	0.0083 (0.0074)	0.0128 (0.0066)
Race-Other: WordConcreteness	-0.0061 (0.0097)	-0.0092 (0.0087)
Race-AfricanAmerican: PronounDensity	0.0402 (0.0065)*	0.04 (0.0057)*
Race-Hispanic: PronounDensity	0.012 (0.0071)	0.0089 (0.0065)
Race-Other: PronounDensity	0.0081 (0.0095)	0.0043 (0.0086)
Race-AfricanAmerican: 2ndPersonPronoun	-0.2071 (0.0321)*	-0.3088 (0.0272)*
Race-Hispanic: 2ndPersonPronoun	-0.1501 (0.0348)*	-0.206 (0.03)*
Race-Other: 2ndPersonPronoun	-0.0805 (0.0433)	-0.0623 (0.0363)
Lang-OtherLang: RealWorldContext	0.0006 (0.011)	0.0036 (0.0098)
Lang-OtherLang: WordCount	-0.0171 (0.0057)*	-0.0188 (0.005)*
Lang-OtherLang: WordConcreteness	-0.0085 (0.0051)	-0.003 (0.0047)
Lang-OtherLang: PronounDensity	-0.0083 (0.005)	-0.004 (0.0046)
Lang-OtherLang: 2ndPersonPronoun	0.0284 (0.0235)	0.027 (0.0207)
LuxuryItems: RealWorldContext	0.0335 (0.0199)	-0.0208 (0.0172)
LuxuryItems: WordCount	0.0288 (0.0112)*	0.1285 (0.0096)*
LuxuryItems: WordConcreteness	-0.0163 (0.0094)	-0.0543 (0.0083)*
LuxuryItems: PronounDensity	-0.0096 (0.0092)	-0.035 (0.0082)*
LuxuryItems: 2ndPersonPronoun	0.0653 (0.0466)	0.2831 (0.037)*
MathAchievement: RealWorldContext	0.0292 (0.0074)*	
MathAchievement: WordCount	0.0382 (0.0041)*	
MathAchievement: WordConcreteness	-0.0213 (0.004)*	
MathAchievement: PronounDensity	0.0067 (0.004)	
MathAchievement: 2ndPersonPronoun	0.3239 (0.0193)*	

Note. (ref.) denotes the reference category to which all comparisons are made. Each column gives the estimated coefficient from the logistic regression in log-odds format, along with its standard error and significance. All significance levels are denoted by a single "*", as our method for p-value corrections does not allow for direct interpretation of the magnitude of the p-value. Instead, this method simply gives a binary significant/not significant ruling. The following variables have been normalized: Mathematics Achievement, Word Concreteness, Pronoun Density, and Word Count. The "Mathematics Achievement" rows are calculated using the Plausible Values for estimated mathematics ability on the NAEP provided with the dataset.

Table 9. Mixed Effects Logistic Regression Results for TIMSS dataset

	Model 1 – Null	**Model 2 - Readability**	**Model 3 – Readability Interactions**	**Model 4 – No Plausible Values**
Random Intercepts	**Var.**	**Var.**	**Var.**	**Var.**
Problem	0.47668	0.4465	0.4500	0.4564
Year	0.03073	0.0263	0.0274	0.1399
Student ID				1.1768
Fixed Effect	**Est. (SE) Sig**	**Est. (SE) Sig**	**Est. (SE) Sig**	**Est. (SE) Sig**
(Intercept)	1.871 (0.1479)*	1.911 (0.1449)*	1.8944 (0.1487)*	1.6481 (0.2299)*
Grade-4	(ref.)	(ref.)	(ref.)	(ref.)
Grade-8	0.016 (0.0875)	-0.009 (0.0855)	-0.0092 (0.0863)	-0.2907 (0.0868)*
Type-MultipleChoice	(ref.)	(ref.)	(ref.)	(ref.)
Type-ConstructedResponse	-0.129 (0.0776)	-0.053 (0.0788)	-0.0546 (0.0794)	-0.0505 (0.0809)
Difficulty-Easy	(ref.)	(ref.)	(ref.)	(ref.)
Difficulty-Medium	-1.433 (0.0894)*	-1.376 (0.0875)*	-1.3617 (0.0884)*	-1.3662 (0.0913)*
Difficulty-Hard	-3.003 (0.1251)*	-2.862 (0.1242)*	-2.8747 (0.126)*	-2.9014 (0.1307)*
Complexity-Low	(ref.)	(ref.)	(ref.)	(ref.)
Complexity-High	-0.234 (0.0697)*	-0.231 (0.0693)*	-0.2308 (0.0696)*	-0.2238 (0.0707)*
Domain-Algebra	(ref.)	(ref.)	(ref.)	(ref.)
Domain-Data	0.335 (0.1228)*	0.385 (0.1258)*	0.4012 (0.1273)*	0.4159 (0.1336)*
Domain-Geometry	-0.178 (0.1129)	-0.233 (0.1116)	-0.2349 (0.1125)	-0.2102 (0.1179)
Domain-Measurement	-0.68 (0.1333)*	-0.715 (0.1299)*	-0.7137 (0.1309)*	-0.7106 (0.1353)*
Domain-Number	-0.053 (0.102)	-0.067 (0.0989)	-0.072 (0.1008)	-0.0510 (0.1051)
Sex-Female	(ref.)	(ref.)	(ref.)	(ref.)
Sex-Male	0.027 (0.0111)*	0.027 (0.0111)*	0.0205 (0.0129)	0.1371 (0.0137)*
U.S. Born	(ref.)	(ref.)	(ref.)	(ref.)
Non-U.S.Born	0.024 (0.0167)	0.024 (0.0167)	0.0357 (0.0205)	-0.5196 (0.0228)*
LuxuryItems	0.043 (0.0143)*	0.043 (0.0143)*	0.0684 (0.0201)*	0.9726 (0.0294)*
MathAchievement	1.116 (0.0043)*	1.116 (0.0043)*	1.1159 (0.0062)*	
EnjoyMath	-0.002 (0.0039)	-0.002 (0.0039)	-0.0022 (0.005)	-0.1188 (0.0071)*
RealWorldContext		-0.107 (0.0854)	-0.1056 (0.087)	-0.0984 (0.0891)
WordCount		-0.011 (0.0417)	-0.0128 (0.0426)	-0.0221 (0.0432)
WordConcreteness		0.042 (0.0395)	0.0524 (0.0403)	0.0438 (0.0407)
PronounDensity		0.027 (0.0402)	0.0164 (0.041)	0.0313 (0.0414)
2ndPersonPronoun		-0.762 (0.1431)*	-0.7878 (0.1475)*	-0.6962 (0.1527)*
Sex-Male: RealWorldContext			0.0303 (0.0145)	0.0385 (0.0145)*
Sex-Male: WordCount			-0.0174 (0.0077)*	-0.0224 (0.0077)*
Sex-Male: WordConcreteness			-0.0099 (0.007)	-0.0026 (0.0069)
Sex-Male: PronounDensity			-0.0031 (0.007)	-0.0014 (0.0069)
Sex-Male: 2ndPersonPronoun			-0.1173 (0.0264)*	-0.1284 (0.0262)*
Non-U.S.Born: RealWorldContext			-0.0112 (0.0246)	0.0030 (0.0239)
Non-U.S.Born: WordCount			-0.0099 (0.0134)	-0.0320 (0.0130)*

Non-U.S.Born: WordConcreteness	-0.0206 (0.0112)	-0.0205 (0.0113)
Non-U.S.Born: PronounDensity	-0.007 (0.0117)	-0.0034 (0.0112)
Non-U.S.Born: 2ndPersonPronoun	-0.1316 (0.0477)*	-0.2295 (0.0444)*
LuxuryItems: RealWorldContext	-0.0635 (0.0334)	-0.0447 (0.0319)
LuxuryItems: WordCount	0.0355 (0.0177)	0.0628 (0.0173)*
LuxuryItems: WordConcreteness	-0.0137 (0.0144)	-0.0188 (0.0141)
LuxuryItems: PronounDensity	0.0222 (0.0149)	-0.0054 (0.0145)
LuxuryItems: 2ndPersonPronoun	0.1055 (0.0551)	0.0098 (0.0543)
EnjoyMath: RealWorldContext	-0.0003 (0.0075)	-0.0020 (0.0075)
EnjoyMath: WordCount	0.0118 (0.0039)*	0.0081 (0.0039)
EnjoyMath: WordConcreteness	0.0162 (0.0036)*	0.0183 (0.0036)*
EnjoyMath: PronounDensity	0.0075 (0.0036)	0.0097 (0.0036)*
EnjoyMath: 2ndPersonPronoun	0.0146 (0.0147)	0.0133 (0.0143)
MathAchievement: RealWorldContext	-0.0189 (0.0117)	
MathAchievement: WordCount	0.0687 (0.0052)*	
MathAchievement: WordConcreteness	-0.0074 (0.0043)	
MathAchievement: PronounDensity	-0.026 (0.0047)*	
MathAchievement: 2ndPersonPronoun	0.1837 (0.0179)*	

Note. (ref.) denotes the reference category to which all comparisons are made. Each column gives the estimated coefficient from the logistic regression in log-odds format, along with its standard error and significance. All significance levels are denoted by a single "*", as our method for p-value corrections does not allow for direct interpretation of the magnitude of the p-value. Instead, this method simply gives a binary significant/not significant ruling. The following variables have been normalized: Mathematics Achievement, Word Concreteness, Pronoun Density, and Word Count. The "Mathematics Achievement" rows are calculated using the Plausible Values for estimated mathematics ability on the TIMSS provided with the dataset.

Table 10. Summary of Results of Interaction between Background Characteristics and Readability Categories from NAEP and TIMSS

Comparison	NAEP	TIMSS
Males compared to Females	1) Positive associations between real world contexts and accuracy are higher for males than females (d = .03) **2) Negative associations between word count and accuracy are more strong for males than females** (d = -.07) 3) Positive associations between pronouns and accuracy are lower for males than females (d = -.02) **4) Negative associations between 2nd person pronouns and accuracy are more strong for males than females** (d = -.04)	**1) Negative associations between word count and accuracy are more strong for males than females** (d = -.03) **2) Negative associations between 2nd person pronouns and accuracy are more strong for males than females** (d = -.07)
African-American students compared to Caucasian students	1) Positive associations between concrete words and accuracy are lower for African-American students than Caucasian students (d = -.03) 2) Positive associations between pronouns and accuracy are higher for African-American students than Caucasian students (d = .07) 3) Negative associations between 2nd person pronouns and accuracy are more strong for African-American students than Caucasian students (d = -.11)	Not available
Hispanic students compared to Caucasian students	Negative associations between 2nd person pronouns and accuracy are higher for Hispanic students than Caucasian students (d = -.08)	Not available
Home Language or Birth Status	Negative associations between word count and accuracy are higher for students with a foreign language spoken at home than other students (d = -.03)	Negative associations between 2nd person pronouns and accuracy are higher for students not born in the U.S. than other students (d = -.07)
Poverty Level (as measured by Luxury items)	Negative associations between 2nd-person pronouns and accuracy are less strong as students have more luxury items (d = .05)	Nothing significant
Mathematics achievement (as measured by Plausible values)	1) Positive associations between presence of a real world context and accuracy are higher for higher achievement students (d = .05) **2) Negative associations between word count and accuracy are less strong for higher achievement students** (d = .19) 3) Positive associations between concrete words and accuracy are lower for higher achievement students (d = -.11) **4) Negative associations between 2nd-person pronouns and accuracy are less strong for higher achievement students (d = 0.54)**	**1) Negative associations between word count and accuracy less strong for higher achievement students** (d = 0.34) 2) Positive associations between pronouns and accuracy are lower for higher achievement students (d = -.04) 3) **Negative associations between 2nd-person pronouns and accuracy are less strong for higher achievement students (d = .30)**
Mathematics Enjoyment (1-4 scale)	Not available	1) Positive associations between concrete words and accuracy are greater for students who enjoy mathematics more (d = .08) 2) Negative associations between word count and accuracy are less strong for students who enjoy mathematics more (d =

.01)

Note. Results that replicate in both datasets are bolded.