Predicting Math Performance Using Natural Language Processing Tools

Scott Crossley
Georgia State University
25 Park Place, Ste 1500
Atlanta, GA 30303
01+404-413-5179
scrossley@gsu.edu

Ran Liu
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
01+412-449-9168
ranliu@cmu.edu

Danielle McNamara
Arizona State University
PO Box 872111
Tempe, AZ 85287
01+480-727-5690
dsmcnamara1@gmail.com

ABSTRACT
A number of studies have demonstrated links between linguistic knowledge and performance in math. Studies examining these links in first language speakers of English have traditionally relied on correlational analyses between linguistic knowledge tests and standardized math tests. For second language (L2) speakers, the majority of studies have compared math performance between proficient and non-proficient speakers of English. In this study, we take a novel approach and examine the linguistic features of student language while they are engaged in collaborative problem solving within an on-line math tutoring system. We transcribe the students' speech and use natural language processing tools to extract linguistic information related to text cohesion, lexical sophistication, and sentiment. Our criterion variables are individuals' pretest and posttest math performance scores. In addition to examining relations between linguistic features of student language production and math scores, we also control for a number of non-linguistic factors including gender, age, grade, school, and content focus (procedural versus conceptual). Linear mixed effect modeling indicates that non-linguistic factors are not predictive of math scores. However, linguistic features related to cohesion affect and lexical proficiency explained approximately 30% of the variance (R^2 = .303) in the math scores.

Categories and Subject Descriptors
K.3.1 [**Computer Uses in Education**]: Computer-assisted Instruction (CAI); J.5 [**Computer Applications: Arts and Humanities**]: Linguistics

General Terms
Algorithms, Measurement, Performance

Keywords
On-line tutoring systems, educational data mining, natural language processing, sentiment analysis, predictive analytics

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
LAK '17, March 13 - 17, 2017, Vancouver, BC, Canada
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-4870-6/17/03…$15.00
DOI: http://dx.doi.org/10.1145/3027385.3027399

1. INTRODUCTION
It has long been argued that there are strong links between language skills and the ability to engage with math concepts and problems. For instance, success in math is argued to be partially based on the development of language that affords children the ability to participate in math instruction in the classroom as well as "engage quantitatively with the world outside the classroom." [1]. Similarly, strong math skills are presumed to interact with language ability because math literacy is not just about knowing numbers and symbols, but also understanding the words surrounding those numbers and symbols. Thus, strong overlap is thought to exist between math and print literacy [2].

Notably, it may not only be language skills that are related to math ability, but also a number of other cognitive predictors that are developed before formal education. Along with linguistic skills, cognitive skills such as a spatial attention and quantitative ability may be related to early math skills in young children [3]. Nonetheless, linguistic skills may be one of the more important factors. For instance, Cummins [4] identified language difficulties in second language (L2) speakers as a key obstacle in solving math problems and linked difficulty in transferring cognitive operations across math and language domains as a barrier to success in math.

One problem with previous studies linking math success and linguistic factors is that the studies have generally relied on correlational analyses among standardized tests of math and linguistic knowledge. For instance, several studies have examined links between tests of language proficiency (e.g., syntax, knowledge of language ambiguity, verbal ability, and phonological skills) and success on tests of math knowledge including algebraic notation, procedural arithmetic, and arithmetic word problems [1, 5]. Other studies have compared success on standardized math tests between first language (L1) speakers of English and second language speakers of English, who have lower linguistic ability [6, 7, 8]. To our knowledge, no studies have examined the relationship between language complexity and language affect in student discourse to their success on math assessments.

The purpose of this study is to fill that gap by examining the language used by students engaged in collaborative math problem solving in an on-line tutoring system. To do so, we transcribed recordings of student discourse during math problem solving and analyzed the language produced for a number of linguistic features related to text cohesion, lexical sophistication, and sentiment that were derived from natural language processing (NLP) tools. In this study, we examined the extent to which the

derived linguistic features are predictive of students' pretest and posttest math scores. We also examine a number of non-linguistic factors that are potentially predictive of math success including age, gender, school, and content focus (procedural versus conceptual). Our goal is to directly investigate links between linguistic production and math success.

1.1 Math and Language Connections

A number of studies have examined links between math skills and language abilities in first language (L1) speakers of English. These studies generally indicate that there are strong links between language proficiency and math ability. For instance, Macgregor and Price [5] examined the relations between three cognitive indicators of language proficiency (metalinguistic awareness of symbols, syntax, and language ambiguity) and algebraic notation. Their data came from pencil-and-paper tests taken by 1500 students aged between 11 and 15 whose length of algebra instruction varied between 1 and 4 years. The majority of students who scored high on language tests also scored high on the algebra test. A follow-up study included a more difficult algebra test that led to greater variance in high and low scores math scores and indicated a stronger relationship between language ability and algebraic notation. The authors concluded that limited metalinguistic awareness seemed to negatively affect success in algebra learning.

In a similar study, Vukovic and Lesaux [1] investigated links between linguistic skills (i.e., general verbal ability and phonological skills), symbolic number skills and arithmetic knowledge (procedural arithmetic and arithmetic word problems). They also included working memory and visual–spatial ability as control variables. Their data came from 287 third graders enrolled at five different schools using the same curriculum for math education. A path model analysis indicated that the linguist skills differed in their degree of relation with arithmetic knowledge. While phonological skills were found to be directly related to arithmetic knowledge, general verbal ability was indirectly related through symbolic number skills. They concluded that "general verbal ability is involved in how children reason numerically whereas phonological skills are involved in executing arithmetic problems." (p.90).

Hernandez [9] investigated links between math and language skills indirectly, by examining relationships between reading ability and math achievement levels. He hypothesized that there was a positive correlation between reading skills and math scores. To test this hypothesis, he analyzed 652 ninth-grade students' scores from the reading and math sections of the *Texas Assessment of Knowledge and Skills.* Correlations between the reading scores and the math scores were computed for texts taken in sixth, seventh, and eighth grades. The results revealed significant positive correlations (with small and medium effect sizes) between reading ability and math achievement. Hernandez suggested that students' reading skills should be taken into account in order to provide more effective math instruction, especially for poor readers. Such instruction could include reading strategy training and collaboration between reading and math teachers.

However, not all studies have found significant links between language skills and math knowledge. LeFevre et al. [3] conducted a longitudinal study that followed children's math progress. The study focused on a year-long data collection from 182 children ages 4 to 8 (37 in preschool and 145 in kindergarten), including linguistics skills (receptive vocabulary and phonological awareness) and non-linguistic skills such as quantitative knowledge, spatial attention, early numeracy skills (number naming and nonlinguistic arithmetic). The outcome measures in the study included standardized and research-based tests of math knowledge. Path modeling resulted in three paths which showed that linguistic skills were significantly related to number naming, that quantitative abilities were related to processing numerical magnitudes, and that spatial attention was related to a variety of numerical and math tests. The last two paths, related to quantitative predictors of arithmetic knowledge, found that non-linguistic features were stronger predictors of math success.

An additional source of evidence for connections between linguistic and math abilities have come from studies comparing L1 and L2 English language speakers. The basic notion behind these studies is that students with lower language skills in English (i.e., second language speakers of English who are less proficient) will have lower math skills in English based classroom. And, it is further assumed that once L2 students reach a threshold of language proficiency, they will have the resources to perform on par with L1 speakers [4].

The assumption that L2 students do not perform as well as L1 students is supported by the US Department of Education [10], which reports that over a five-year period (from 1st to 5th grade), L1 speakers of English report higher math scores than proficient L2 speakers who, in turn, report higher math scores low proficient L2 students.

For the most part, results from research investigating the differences in math skills between L1 and L2 speakers of English concur with US Department of Education report. For instance, Alt et al. [6] investigated relations between math and language achievement among school-age children (ages 7-10) who were grouped into native students (N=21), L2 learners whose first language was Spanish (N=20), and students with specific language impairment (SLI) (N=20). The researchers hypothesized that there would be differences in math skills between the groups due to language proficiency differences. Data were collected using two standardized math tests (one in English and one in Spanish) and three experimental tasks (number comparison, quantity comparison and concept mapping games). The tests and tasks were categorized as either heavy or light processing in terms of language, symbol, and visual working memory. For instance, the math test in English was classified as heavy language processing, heavy symbol processing, and light on visual working memory. The math test in Spanish was classified as light language processing, heavy symbol processing, and light on visual working memory. The concept mapping game was classified as light language processing, light symbol processing, and heavy on visual working memory. The results showed that students with SLI achieved performed significantly worse than native speakers in all tests and tasks. When L1 and L2 speakers were compared, Alt et al. found that L1 students significantly outperformed the L2 students only in language-heavy tests and games. These results led Alt et al. to conclude that language proficiency is a crucial factor in math success for students who have language-related challenges.

Martinello [8] investigated item difficulty differences across math tests between L1 and L2 students in terms of different levels of linguistic complexity and contextual support provided by pictures and schemas. Standardized math test scores for 68,839 fourth-grade students, 3179 of which were non-native speakers of English, were used in the study. The test scores consisted of 39 items that assessed knowledge of number sense and operations, patterns and relations, algebra, geometry, measurement, and

probabilities. For each item in the test, two researchers rated the grammatical and lexical complexity of the item. The results showed that linguistic complexity and the non-linguistic representations that accompanied the items accounted for around 66% of the variation in scores between native and non-native students such that linguistically complex items were found to be more difficult for nonnative speakers. These items included complex grammatical structures and low frequency non-math words that were central to the items and hard to guess from the context. Non-linguistic representations (especially schemas) were found to decrease the difficulty of more linguistically complex items.

Similar findings that support the notion that L2 speakers of English are at a disadvantage in math performance when compared to L1 speakers have been reported in a number of studies [11, 12, 13]. Of course, while language is a predictor of success, it is not the only consideration. Language skills can interact with background differences such as parent education, levels of poverty, and ethnicity [10], courses taken [12], and immigrant status [13]. Nonetheless, correlational studies generally support the threshold hypothesis [4] that proficiency in the language of instruction is necessary for academic achievement in disciplines such as math.

1.2 Current Study

In summary, a number of studies have demonstrated strong links between linguistic knowledge and success in math. Studies examining these links in L1 speakers have traditionally relied on correlational analyses between linguistic knowledge tests and standardized math tests [1, 3, 5]. For L2 speakers, the majority of studies have compared math success between proficient and non-proficient speakers of English [6, 10, 11, 12, 13]. In this study, we take a novel approach and examine the linguistic features of students' language production during math problem solving in an on-line tutoring system. To derive our linguistic features of interest, we transcribe student speech and use a number of natural language processing tools to extract linguistic information related to text cohesion, lexical sophistication, and sentiment. Thus, in contrast to previous studies, our interest is not on linguistic performance as measured by standardized tests, but on linguistic performance as a function of language production during collaborative math learning activities. Our criterion variables are pretest and posttest math performance scores. In addition to examining relations between linguistic features of student language production and math scores, we also control for a number of non-linguistic factors including gender, age, grade, school, and content focus (procedural versus conceptual). Thus, in this study, we address three research questions:

1. Are non-linguistic factors significant predictors of math performance in a collaborative on-line tutoring environment?
2. Are linguistic factors related to lexical sophistication, cohesion, and affect significant predictors of math performance in a collaborative on-line tutoring environment?
3. Are linguistic features stronger predictors of math performance than non-linguistic factors?

METHOD

2.1 Procedure

The data used in this study come from an experiment that compared the effectiveness of collaborative versus individual learning of fraction concepts and procedures from an intelligent tutoring system. Students in the study were randomly assigned to one of two conditions: Collaborative, in which two students worked with a partner through the full tutor curriculum (i.e., collaborative dyad), and Individual, in which students worked by themselves on the entire tutor. Since audio recordings of student dialogue were only collected for the Collaborative condition, the present investigation only applies to students in that condition.

The fractions tutoring system that students used is online software, built using an extension of Cognitive Tutor Authoring Tools [14] designed to support collaborative learning [15]. The fractions tutoring system helps students become better at understanding and using fractions. It covers six sub-topics, including naming, picturing, equivalent, ordering, adding, and subtracting fractions. Its effectiveness has previously been demonstrated in prior classroom deployment studies [16, 17]. These studies showed that students' mistakes decrease as they progress through the tutor; students score higher on a fractions test after using the tutoring system compared to before; and scores remain higher than pre-tutoring a week after they have finished using the tutoring system.

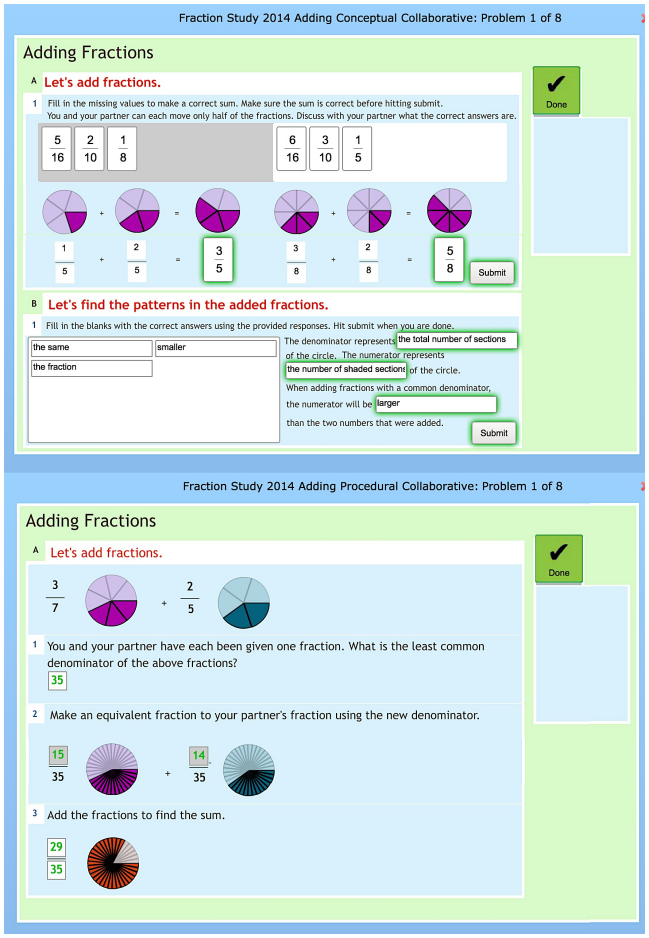

Figure 1. Example problems from the "adding fractions" section of the collaborative fractions tutor.

In this study, the tutoring system was designed to support collaboration between students. Although each student worked on the tutoring system on his or her own computer screen, each student in a pair could control only part of the screen. The students needed to work together to finish the problem (i.e., one student could not do everything). Students worked together at the same time and, ideally, talked about what they were doing, asked for help from their partner, defended a position or explained why

they thought something was the correct answer, and built off of each other's contributions.

All collaborative dyads randomly received a problem set focused on either procedural or conceptual knowledge building. The procedural versus conceptual comparison had been included to investigate whether there were any interactions between collaborative learning and type of knowledge acquired. Figure 1 shows one example of each type of problem, conceptual and procedural. The top and bottom panels show example problems from the conceptual and procedural knowledge conditions, respectively. The figure depicts correctly completed screens; student-input fields are marked with either green text or borders.

The study took place over five consecutive days. On the first day, students individually took a pretest to establish their baseline fractions knowledge. In the following three days, students in the Collaborative condition worked through the tutoring system with a partner. On the last day, students individually took a posttest that also tested fractions knowledge, with content similar to the pretest.

1.3 Participants

A total of 104 fourth and fifth graders participated in the Collaborative condition of the study. There were 19 fifth graders from one classroom of one school and 50 fourth graders and 35 fifth graders from a second school. Of these, only a subset of students completed the full study (pretest, posttest, and three days of tutoring system use), had the same partner during the entire study (no absences for either individual), and consented to audio recording of their dialogue. For consistency purposes, we only analyzed data from students who fit all of these criteria. Thus, our analyses were done on this subset of 36 students (14 fifth graders from the first school, and 16 fourth graders and 6 fifth graders from the other school). There were 15 males and 21 females in the analysis subset. Student pairs were determined by the teachers. Teachers were asked to pair each student with a partner that they would get along with, and who was at a similar knowledge level.

1.4 Transcriptions

A professional transcriber transcribed each of the speech samples collected from the participants. The transcriptions contained the speaker's words, some metalinguistic data (singing, laughing, sighing) and filler words (e.g., ummm, ahhhh). Disfluencies that were linguistic in nature (e.g., false starts, word repetition, repairs) were also retained. If any portion of the audio was not transcribable, the words were annotated either with an underscore or the flag "INAUDIBLE" depending on the transcriptionist. The files were cleaned so that metalinguistic data, filler words, untranscribale portions were removed prior to analysis.

1.5 Linguistic Variables

The transcripts were separated by learner and then cleaned to remove all non-linguistic information including metadata and non-linguistic vocalizations such as coughs and laughs. Each transcript was run through a number of natural language processing tools including the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) [18], the Tool for the Automatic Analysis of Cohesion (TAACO) [19] and the SEntiment ANalysis and Cognition Engine (SEANCE) [20]. The selected tools reported on language features related to lexical sophistication, text cohesion, and sentiment analysis respectively. The tools are discussed in greater detail below.

1.5.1 TAALES

TAALES is a computational tool that is freely available and easy to use, works on most operating systems (Windows, Mac, Linux), allows for batch processing of text files, and incorporates over 150 classic and recently developed indices of lexical sophistication. These indices measure word frequency, lexical range, n-gram frequency and proportion, academic words and phrases, word information, lexical and phrasal sophistication, and age of exposure. Each of these are discussed briefly below. For more detailed accounts of TAALES please see Kyle and Crossley [18].

Word frequency indices. TAALES calculates a number of word frequency indices with frequency counts retrieved from Thondike-Lorge [21], Kucera-Francis [22], Brown [23], and SUBTLexus databases [24]. In addition, TAALES derives frequency counts from the British National Corpus (BNC) [25]. TAALES calculates scores for all words (AW), content words (CW), and function words (FW).

Range indices. In addition to frequency information, TAALES includes a number of range indices which calculate how many texts within a corpus a word appears (i.e., specificity). Range indices are calculated for the spoken (574 texts) and written (3,083 texts) subsets of the BNC, SUBTLEXus (8,388 texts) and Kucera-Francis (500 texts).

N-gram frequency and proportion indices. TAALES calculates bigram and trigram frequencies and proportion scores (i.e., the proportion of n-grams in a text that are common in a reference corpus) from both the written (80 million words) and spoken subcorpora (10 million words) of the BNC.

Academic list indices. TAALES includes word and n-gram level academic lists. These indices are calculated from the Academic Word List (AWL) [26] and the Academic Formula List (AFL) [27].

Word information indices. Word information scores are derived from the MRC Psycholinguistic Database [28, 29, 30]. Word information scores are calculated for word familiarity, concreteness, imageability, meaningfulness, and age of acquisition.

1.5.2 TAACO

TAACO (Crossley et al., in press-b) incorporates over 150 classic and recently developed indices related to text cohesion. For a number of indices, the tool incorporates a part of speech (POS) tagger from the Natural Language Tool Kit [31] and synonym sets from the WordNet lexical database [32]. The POS tagger affords the opportunity to look at content words (i.e., nouns, verbs, adjectives, adverbs) as well as function words (i.e., determiners, propositions). TAACO provides linguistic counts for both sentence and paragraph markers of cohesion and incorporates WordNet synonym sets. Specifically, TAACO calculates type token ratio (TTR) indices (for all words, content words, function words, and n-grams), sentence overlap indices that assess local cohesion for all words, content words, function words, POS tags, and synonyms, paragraph overlap indices that assess global cohesion for all words, content words, function words, POS tags, and synonyms, and a variety of connective indices such as logical connectives (e.g., *moreover, nevertheless*), causal connectives (*because, consequently, only if*), sentence linking connectives (e.g., *nonetheless, therefore, however*), and order connectives (e.g., *first, before, after*).

1.5.4 SEANCE

SEANCE is a sentiment analysis tools that relies on a number of pre-existing sentiment, social positioning, and cognition dictionaries. SEANCE contains a number of pre-developed word vectors developed to measure sentiment, cognition, and social order. These vectors are taken from freely available source databases such as SenticNet [33, 34] and EmoLex [35, 36]. In some cases, the vectors are populated by a small number of words and should be used only on larger texts that provide greater linguistic coverage in order to avoid non-normal distributions of data as found in the Lasswell dictionary lists [37] and the Geneva Affect Label Coder (GALC) [38] lists. For many of these vectors, SEANCE also provides a negation feature (i.e., a contextual valence shifter [39]) that ignores positive terms that are negated (e.g., not happy). The negation feature, which is based on Hutto and Gilbert [40], checks for negation words in the 3 words preceding a target word. SEANCE also includes the Stanford part of speech (POS) tagger [41] as implemented in Stanford CoreNLP [42]. The POS tagger allows for POS tagged specific indices for nouns, verbs, and adjectives.

1.6 Statistical Analysis

We first conducted a paired samples *t*-test to examine if there were differences between pretest and posttest scores for the data. We then conducted linear mixed effect (LME) models to answer our three research questions. The purpose of the LME was to determine if linguistic features in the students' language output along with other fixed effects could be used to predict the students' pretest and posttest math scores. Thus, the LME model modeled the pretest and posttest results in terms of random factors (i.e., repeated variance explained by the students as they moved through the intervention longitudinally) and fixed or between factors (e.g., the linguistic features in their transcripts, gender, age, school). Such an approach allows us to examine math growth over time for individual learners using random factors as well as investigate if individual differences related to the learner such as demographic information, age, and linguistic ability predict math development. Lastly, the approach allows us to also examine if different classroom interventions influence math scores (i.e., procedural versus conceptual approaches to teaching math in the classroom).

Prior to the LME analysis, we first checked that the linguistic variables were normally distributed as well as controlled for multicollinearity between all the linguistic variables ($r > .700$). We used R [43] for our statistical analysis and the package *lme4* [44] to construct linear mixed effects models (LME). We also used the package *lmerTest* [45] to analyze the LME output and derive p-values for individual fixed effects. Final model selection and interpretation was based on *t* and *p* values for fixed effects and visual inspection of residuals distribution. To obtain a measure of effect sizes, we computed correlations between fitted and predicted residual values, resulting in an R^2 value for both the fixed factors and the fixed factors combined with the random factor (i.e., the repeated participant data from the pretest and the posttest). We first developed a baseline model that included gender, grade, condition, and school as fixed effects and participants as random effect. We next developed a full model that included gender, grade, condition, and school as fixed effects along with linguistic features and participants as random effect.

2. RESULTS

2.1 Math Gains

A paired *t*-test examining differences between the pretest and the posttests scores indicated significant differences between the pretest (M= .469, SD=.170) and the posttest (M= .603, SD= .185); $t(35)= 5.988$, $p < .001$.

2.2 Baseline Model

A baseline model considering all fixed effects aside from linguistic revealed no significant effects on math scores. Table 1 displays the coefficients, standard error, *t* values, and *p* values for each of the non-linguistic fixed effects. Inspection of residuals suggested the model was not influenced by homoscedasticity. The non-linguistic variables explained around 2% of the variance ($R^2 = .016$) while the fixed and random variables together explained around 55% of the variance ($R^2 = .553$). Thus, the majority of change found in the pretest and posttest was due to time.

2.3 Full Model

A full model was developed that including the nested baseline model and linguistic fixed effects. The model included five linguistic features related to cohesion (*sentence linking connectives* and *adjacent overlap of adjectives*), affect (*respect terms*), and lexical proficiency (*number of function word types* and *verb hypernymy*). None of the variables showed suppression effects. The model indicated that a greater number of sentence linking connectives (e.g., *nonetheless, therefore, however*), function word types (e.g., prepositions, connectives, and articles), and overlap of adjectives predicted higher math scores. Conversely, more respect terms and greater use of more specific words (i.e., greater hypernymy scores) related to lower math scores. Table 2 displays the coefficients, standard error, *t* values, and *p* values for each of the fixed effects ordered by strength of *t* value. A log likelihood comparisons found a significant difference between the baseline and full models, ($\chi2(2) = 42.486$, $p < .001$), indicating that the inclusion of linguistic features contributed to a better model fit. Together, the fixed factors including the linguistic and non-linguistic variables explained around 30% of the variance ($R^2 = .303$) while the fixed and random variables combined to explain around 82% of the variance ($R^2 = .823$).

Table 1. Baseline model for predicting math scores

Fixed Effect	Coefficient	Std. Error	*t*	*p*
(Intercept)	0.564	0.059	9.543	< .001
Gender (male)	-0.039	0.061	-0.650	0.521
Grade (5)	-0.029	0.082	-0.350	0.729
Condition (procedural)	-0.024	0.060	-0.397	0.694
School	0.038	0.086	0.436	0.666

Table 2. Full model for predicting math scores

Fixed Effect	Coefficient	Std. Error	*t*	*p*
(Intercept)	0.557	0.055	10.106	< .001
Gender (male is contrast)	0.007	0.057	0.121	0.905
Grade (5th grade is contrast)	-0.021	0.077	-0.284	0.778
Condition (procedural content is contrast)	-0.036	0.057	-0.639	0.527
School	0.032	0.080	0.401	0.691
Sentence linking connective	0.059	0.018	3.246	< .001
Number of function word types	0.044	0.0193	2.273	< .050
Respect words	-0.032	0.013	-2.518	< .050
Adjacent overlap of adjectives	0.039	0.015	2.549	< .050
Verb hypernymy	-0.038	0.017	-2.265	< .050

3. DISCUSSION

Previous studies that have investigated links between language use and math performance have reported strong links between the two indicating that language skills are an important prequisite for effectively engaging with math concepts and problems. These previous studies have traditionally relied on analyzing links between language proficiency tests and/or surveys and standardized math scores. Similar studies have also examined differences in math performance between L1 and L2 speakers of English to test threshold hypotheses predicated on the notion that less proficient speakers of English will have more difficulty in math classes taught in English.

Our study takes a novel approach to understanding links between math performance and language use by examining the actual language produced during math problem solving and examining if features of this language are predictive of math performance in standardized tests. Beyond language features, this study also examined a number of non-linguistic student factors including gender, age, grade, school, and content focus (procedural versus conceptual). The findings indicate that the non-linguistic factors were not significant predictors of math performance. However, time between the pretest and posttest was a strong predictor of performance. In addition, linguistic features were significant predictors of math performance. We address each of these below.

That non-linguistic features were not significant predictors of math performance has important implications for understanding math performance. Specifically, male students performed no better than female students and 4th grade students performed no better than 5th grade students. In addition, no differences were reported for students from two different schools. These findings provide evidence that learning within a math tutoring system may not favor one gender over another nor grade or school. Lastly, the two types of knowledge conditions (conceptual or procedural) showed no difference in performance indicating equivalence between the two. However, performance did increase between the pretest and the posttest according to the paired samples *t*-test indicating that significant learning occurred as a result of interacting with the online math tutor. Much of this increase can be attributed to the effects of repeated measures across time (i.e., the random effects). These random effects explained above 50% of the variances in the math performance scores.

The full model LME model demonstrated that a number of linguistic features were significant predictors of math performance. Specifically, a greater number of sentence linking connectives and function words were predictive of math performance. These findings indicate that math performance is likely linked with the production of more complex syntactic structures such as those found in coordinated sentences and sentences with more structural components (i.e., function words). Lexically, math performance is associated with the production of more abstract words (i.e., words with greater hypernymy scores). Intuitively this makes sense because math solutions are based on abstract thinking and language use. In addition, a greater overlap of adjectives between sentences is a strong predictor of math performance likely indicating that the repetition of math adjectives such as *greater than* and *less than* may be related to math performance. Lastly, our analysis demonstrated that math performance was related to the use of fewer words related to respect. This finding may seem counter-intuitive, but performance within a math tutoring system that requires collaboration and timed completion of problems may favor curt and direct discourse between participants that may be interpreted as less respectful. In total, the linguistic factors explained about 28% of the variance in the math performance data over and above the 2% explained by the non-linguistic factors. When both fixed and random factors were included in the model, over 80% of the variance in the math performance was predicted.

To provide examples of the linguistic features above, we extracted excerpts from a student dyad on the last day of the study. The first student (Student 137) in the dyad had the highest posttest score (97%) and showed a 15% gain from the pretest. The second student (Student 128) scored low on the posttest (44%) and had the third lowest score on the pretest (22%). The student's posttest score was almost double that of his pretest score though, showing strong gains in learning.

Table 3: Text excerpts from students that completed and did not complete the EDM MOOC

Sentence linking examples
STUD_137: Yeah it is, but it could be 80/80. It could – why didn't they just have a card with a one? … STUD_137: It's equal. You're rushing. And just look at the numbers and then you put one in.
Function words and Respect example
STUD_128: How did you get them all wrong? STUD_137: You got two of them wrong. STUD_128: You got them all wrong. STUD_137: Wait. If the cement is green. STUD_128: Why did you get them all wrong? STUD_137: You did. STUD_128: Why did you get them all wrong?
Adjective overlap example
STUD_137: No, two-ninths is greater than one-ninth. STUD_128: Two-ninths is greater. … STUD_137: It's greater than. It was either greater than or equal to. STUD_128: It was greater than.
Hypernymy example
STUD_137: And that one you have too. You should be able to figure that one out. STUD_128: Three, four, five. Three-fifths. STUD_137: Dude, when are you going to – what are you doing? Oh. Wait, hold on. STUD_128: You have it. STUD_137: Oh, they're equal. STUD_128: You had it the whole time.

Linguistically, the excerpts provide illustrations for the trends reported in the statistical analysis. For instance, Student 137 links many sentences together with connectors such as *but* and *and*. In addition, both students tend to use a greater number of function words such as *how, did, you, all, of, if,* and *the.* The use of a greater number of function words indicates the use of stronger sentence-based structural elements, which may be important in discussing more abstract ideas. In terms of abstract words, the excerpts show that these two students use a number of abstract words that are less specific such as *that, one, have, able, go, are, do, you, it, they,* and *time*. The two students also demonstrate a directness with one another that could be viewed as disrespectful from an outsider's perspective. For instance, the two students seem comfortable accusing one another of getting the answers wrong. Lastly, in terms of argument overlap, the excerpts provide instances of students repeating adjectives related to math solving problems (i.e., *greater*). In total, the excerpts provide illustrations of what the natural language processing tools are likely capturing in their estimations of math performance. These excerpts help provide contextualized details for the math discourse in the data.

4. CONCLUSION

The findings from this study have practical implications for understanding math performance and math instruction. Specifically, the findings provide support for the notion that language proficiency is strongly linked to math performance such that more complex language and a greater overlap of adjectives equates to higher performance. Similarly, discourse that contains fewer terms related to respect equates to higher performance. From an instructional perspective, the findings also indicate that collaborative, on-line tutoring instruction can lead to improved math performance. These findings could inform math pedagogy practices by providing support for language instruction within the math classroom. For instance, it may be the case that providing students with a solid math vocabulary foundation would improve students' math success by providing them with the means to discuss complex math problems in a collaborative environment. It is likely that focusing both on abstract math principles and on the language needed to communicate these principles would push students over the language threshold needed for success in the math classroom. Additionally, in terms of respect, it is likely that students that show less deference and are more likely to challenge ideas are more successful in the math classroom.

Future studies can build on the results presented here by sampling larger populations of students that come from more diverse backgrounds and more varied grade levels. Such a study would build on the relatively low sample size found in this paper and provide greater evidence for the importance of linguistic proficiency and math success. Of interest in future replications of this research would be to examine if the results reported here persist with older students, in educational settings outside of an intelligent tutoring environment, and with different math topics. Such research would help extend the current study past the single context on which it focused and provide evidence that linguistic proficiency is an important indicator or math success in a variety of learning contexts.

5. ACKNOWLEDGMENTS

This research was supported in part by the Institute for Education Sciences and National Science Foundation (IES R305A080589, IES R305G20018-02, and DRL- 1417997). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES or the NSF.

6. REFERENCES

[1] Vukovic, R. K., & Lesaux, N. K. (2013). The relationship between linguistic skills and arithmetic knowledge. *Learning and Individual Differences*, *23*, 87-91.

[2] Adams, T. L. (2003). Reading math: More than words can say. *The Reading Teacher*, *56*(8), 786-795.

[3] LeFevre, J. A., Fast, L., Skwarchuk, S. L., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010).

Pathways to math: Longitudinal predictors of performance. *Child development*, *81*(6), 1753-1767.

[4] Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49*, 222-251.

[5] MacGregor, M., & Price, E. (1999). An exploration of aspects of language proficiency and algebra learning. *Journal for Research in Math Education*, 449-467.

[6] Alt, M., Arizmendi, G. D., & Beal, C. R. (2014). The relationship between math and language: Academic implications for children with specific language impairment and English language learners. *Language, speech, and hearing services in schools*, *45*(3), 220-233.

[7] Hampden-Thompson, G., Mulligan, G., Kinukawa, A., & Halle, T. (2008). *Math Achievement of Language-Minority Students During the Elementary Years*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

[8] Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational assessment*, *14*(3-4), 160-179.

[9] Hernandez, F. (2013). *The Relationship Between Reading and Math Achievement of Middle School Students as Measured by the Texas Assessment of Knowledge and Skills* (Doctoral dissertation).

[10] Hampden-Thompson, G., Mulligan, G., Kinukawa, A., & Halle, T. (2008). *Math Achievement of Language-Minority Students During the Elementary Years*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

[11] Ardasheva, Y., Tretter, T., Kinny, M. (2012). English Language Learners and Academic Achievement: Revisiting the Threshold Hypothesis. *Language Learning, 62*(3), 769-812.

[12] Mosqueda, E., & Maldonado, S. I. (2013). The effects of English language proficiency and curricular pathways: Latina/os' math achievement in secondary schools. *Equity & Excellence in Education*, *46*(2), 202-219.

[13] Wang, J., & Goldschmidt, P. (1999). Opportunity to learn, language proficiency, and immigrant status effects on math achievement. *The Journal of Educational Research*, *93*(2), 101-111.

[14] Aleven, V., McLaren, B.M., Sewall, J., & Koedinger, K.R. (2009). A New Paradigm for Intelligent Tutoring Systems: Example-Tracing Tutors. *International Journal of Artificial Intelligence in Education*, 19(2), 105-154.

[15] Olsen, J. K., Belenky, D. M., Aleven, V., Rummel, N., & Ringenberg, M. Authoring collaborative intelligent tutoring systems. . In Lane, H. C., Yacef, K., Mostow, J., & Pavlik, P. (Eds.). *Proceedings of the Artificial Intelligence in Education* (AIED) *Conference*. Heidelberg, Germany: Springer.

[16] Rau, M. A., Aleven, V., & Rummel, N. (2009, July). Intelligent Tutoring Systems with Multiple Representations and Self-Explanation Prompts Support Learning of Fractions. In *Proceedings of the Artificial Intelligence in Education* (AIED) *Conference.* (pp. 441-448). Heidelberg, Germany: Springer.

[17] Rau, M. A., Aleven, V., Rummel, N., & Rohrbach, S. (2012). Sense making alone doesn't do it: Fluency matters too! ITS support for robust learning with multiple representations. In *International Conference on Intelligent Tutoring Systems* (pp. 174-184). Springer Berlin Heidelberg.

[18] Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757-786. doi:10.1002/tesq.194

[19] Crossley, S. A., Kyle, K., & McNamara, D. S. (in press). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*.

[20] Crossley, S. A., Kyle, K., & McNamara, D. S. (in press). Sentiment Analysis and Social Cognition Engine (SEANCE): An Automatic Tool for Sentiment, Social Cognition, and Social Order Analysis. *Behavior Research Methods*. (Thorndike & Lorge, 1944)

[21] Thorndike, E. L., & Lorge, I. (1944). *The teacher's wordbook of 30,000 words.* New York: Columbia University, Teachers College: Bureau of Publications.

[22] Kučera, H., & Francis, N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

[23] Brown, G. D. (1984). A frequency count of 190,000 words in theLondon-Lund Corpus of English Conversation. *Behavior Research Methods, Instruments, & Computers, 16*(6), 502-532.

[24] Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977-990. doi:10.3758/brm.41.4.977

[25] *The British National Corpus*, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

[26] Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

[27] Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied LInguistics, 31*(4), 487-512.

[28] Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904-911.Coltheart, 1981

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

[30] Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*(4), 978–990.

[31] Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media, Inc.

[32] Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41.

[33] Cambria, E., Grassi, M., Hussain, A., & Havasi, C. (2012). Sentic computing for social media marketing. *Multimedia tools and applications, 59*(2), 557-577.

[34] Cambria, E., Speer, R., Havasi, C., & Hussain, A. (2010). *SenticNet: A Publicly Available Semantic Resource for Opinion Mining.* Paper presented at the AAAI fall symposium: commonsense knowledge.

[35] Mohammad, S. M., & Turney, P. D. (2010). *Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon.* Paper presented at the Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text.

[36] Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence, 29*(3), 436–465.

[37] Lasswell, H. D., & Namenwirth, J. Z. (1969). *The Lasswell Value Dictionary*. New Haven: Yale University Press.

[38] Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social science information, 44*(4), 695-729

[39] Polanyi, L., & Zaenen, A. (2006). Contextual Valence Shifters. In J. G. Shanahan, Y. Qu, & J. Wiebe (Eds.), *Computing Attitude and Affect in Text: Theory and Applications* (pp. 1-10). Dordrecht: Springer Netherlands.

[40] Hutto, C. J., & Gilbert, E. (2014). *Vader: A parsimonious rule-based model for sentiment analysis of social media text.* Paper presented at the 8th Int. AAAI Conf. on Weblogs and Social Media, Ann Arbor, MI.

[41] Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). *Feature-rich part-of-speech tagging with a cyclic dependency network.* Paper presented at the Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1.

[42] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). *The Stanford CoreNLP Natural Language Processing Toolkit.* Paper presented at the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MA.

[43] R Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013: ISBN 3-900051-07-0.

[44] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

[45] Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package 'lmerTest'. *R package version*, 2.0-29.