Predicting Math Success in an Online Tutoring System Using Language Data and Click-Stream Variables: A Longitudinal Analysis

Scott Crossley
Georgia State University, Applied Linguistics/ESL, Atlanta, GA, USA
scrossley@gsu.edu

Shamya Karumbaiah
The University of Pennsylvania, Philadelphia, PA, USA
shamya@upenn.edu

Jaclyn Ocumpaugh
The University of Pennsylvania, Philadelphia, PA, USA
ojaclyn@upenn.edu

Matthew J. Labrum
Imagine Learning, Provo, UT, USA
matthew.labrum@imaginelearning.com

Ryan S. Baker
The University of Pennsylvania, Philadelphia, PA, USA
rybaker@upenn.edu

Abstract

Previous studies have demonstrated strong links between students' linguistic knowledge, their affective language patterns and their success in math. Other studies have shown that demographic and click-stream variables in online learning environments are important predictors of math success. This study builds on this research in two ways. First, it combines linguistics and click-stream variables along with demographic information to increase prediction rates for math success. Second, it examines how random variance, as found in repeated participant data, can explain math success beyond linguistic, demographic, and click-stream variables. The findings indicate that linguistic, demographic, and click-stream factors explained about 14% of the variance in math scores. These variables mixed with random factors explained about 44% of the variance.

2012 ACM Subject Classification Applied computing → Computer-assisted instruction; Applied computing → Mathematics and statistics; Computing methodologies → Natural language processing

Keywords and phrases Natural language processing, math education, online tutoring systems, text analytics, click-stream variables

Digital Object Identifier 10.4230/OASIcs.LDK.2019.25

Funding This research was supported in part by NSF 1623730. Opinions, conclusions, or recommendations do not necessarily reflect the views of the NSF.

1 Introduction

Students need a number of cognitive skills including spatial attention and quantitative ability to be successful within a math classroom [28]. In addition, recent research has shown strong links between students' language production and math success. This research demonstrates that students that are more proficient in math are generally also more proficient language users. There are several potential reasons for links between math and language domains, both of which rely on the ability to interpret and manipulate abstract symbolic systems [40]. One

2nd Conference on Language, Data and Knowledge (LDK 2019).
Editors: Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski; Article No. 25; pp. 25:1–25:13

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

key reason is that language skills help students learn knowledge and math operations from text books and tutoring systems, as well as from other people. More generally, students with greater language proficiency are better able to engage individually and collaboratively with math concepts and solve math problems because math is not purely based on numbers and abstract symbols but also on real world problems that involve the words surrounding math numbers and symbols [2]. Thus, language skills help students to participate constructively and collaboratively in math discourse and engage with and solve math problems both inside and outside of the classroom [30, 41].

Some previous studies that have examined links between math success and language have relied on correlational analyses between standardized tests of language and math. As an example, previous studies have analyzed association between language proficiency tests that assess syntax, lexical, and phonological skills and math scores on standardized tests that assess arithmetic and algebra and found strong links [29, 41]. Another area of inquiry between language proficiency and math skills has been to compare success rates on standardized math tests between native and non-native speakers of English. These studies often find that non-native speakers of English perform lower on math assessments [3, 20, 31] although see [1] for counter argument. A final approach to examining math and language links is to assess links between the complexity of language produced by students and their success on math assessments. Such studies generally find that students that produce more complex language features score higher in math, possibly because students' ability to switch from conversational language to the conventions required in mathematics requires high level metalinguistic skills [19].

The current study builds on previous studies that have focused on links between the language produced by students and their math success, by combining fine-grained click-stream variables and simple demographic data (i.e., grade and gender) with language features in student production to predict math success. We also assess math performance over time to better control for variance associated with participants. To do so, we use natural language processing (NLP) tools to assess language production in e-mail messages sent by elementary students within an online tutoring system over the course of a year. We then examine the students' behaviors within the tutoring system in terms of actions completed, entries into various elements of the system, and time spent within these elements. The goals of the study are to combine these data to increase prediction rates within the system. Additionally, we examine performance over time to assess the degree to which random variance found in repeated participant data can explain math success beyond linguistics, demographic, and click-stream data.

1.1 Relationships between Language and Math Skills

The body of research demonstrating connections between proficiency in language and math skills continues to grow, becoming more robust as researchers explore the potential underlying causes. Early studies focused on links between scores on math and language tests. For instance, MacGregor and Price [29] found that students who scored high on an algebra test also scored well on language tests. Using a more difficult algebra test produced a stronger relationship between algebraic notation and language ability. Similarly, Hernandez [22] found significant positive correlations between reading and math scores in standardized tests. Vukovic and Lesaux [41] also reported links between language and math skills, but additionally found that language skills differed in their degree of relation with math knowledge. For example, general verbal ability was indirectly related with symbolic number skills while phonological skills were directly related to arithmetic knowledge. Lastly, LeFevre et al. [28] reported that language ability was positively related to number naming.

More recent studies have begun to examine links between the language features found in students' language production and their success in math learning using NLP tools. These studies have focused on elementary and college level students. In an early study of elementary students, Crossley et al. [12] examined language features found in transcribed student speech during collaborative math projects and found that language features related to cohesion, affect, and lexical proficiency explained a significant amount of variance in students' math scores. More mathematically proficient students produced more cohesive language that was comprised of more lexically sophisticated words. In another study, Crossley and Kostyuk [11] examined links between the language features of elementary students' language production while e-mailing a virtual pedagogical agent in an online math tutoring system and math success within the system. They found that students who expressed more certainty in their writing and followed standardized language patterns scored higher in math assessments. In a more recent study, Crossley et al. [13] used linguistic features found in student e-mails within an online math tutoring system to predict math success. They found that lexical features and syntactic complexity indices were significant predictors of math success such that more successful students used words that were found across a variety of registers and used more sophisticated words. In addition, higher scoring math students produced fewer complex sentences.

Studies assessing links between language features and math success for college level students have reported similar findings. For instance, Crossley et al. [10] examined college students' forum posts in an online tutoring system that was part of a blended math class (i.e., a class with both online and traditional face-to-face instruction). They investigated relationships between language features in these posts and final scores in the class, finding that success in the class was predicted by language features related to affect, syntactic complexity, and text cohesion. Specifically, more complex syntactic structures and fewer explicit cohesion devices were associated with higher course performance. The linguistic model also indicated that less self-centered students and students using words related to tool use were more successful. In a similar study using the same data set, Crossley et al. [16] examined how linguistic features derived from cohesion network analyses could predict math success. The models from this study indicated that students who encouraged greater language collaboration within forum posts (i.e., those students that precipitated discussion among other students) received higher final scores in the class.

In general, these studies demonstrate that linguistic features from students' language can predict math performance across grade levels (from elementary to college level students) in different types of learning environments (collaborative online tutors, traditional online tutoring systems, and blended math courses). Overall, younger students that are more proficient at math produce more cohesive language that includes more sophisticated vocabulary. In addition, younger students that are more proficient at math are better at following expected language patterns and produce less complex syntactic structures. In contrast, older students who receive higher grades in math class produce more syntactically complex structures that are less cohesive. These students also encourage greater collaboration through their language use. The differences between older and younger students' language production is likely related to different stages of language acquisition.

1.2 Click-stream Data and Student Success in Math

There is growing research that demonstrates the strength of using student interaction data [34] in online learning environments (i.e., click-stream data) to predict short- to long-term learning, engagement and interest in mathematics. Data on fine-grained aspects of student

behavior provides opportunities to explore how patterns of interaction relate to outcomes. For instance, Beal et al. [8] reported that students' use of interactive multimedia hints in an online tutor for SAT-Math problems were predictive of learning gains in the system. To extract richer information from the raw student log data, researchers have also extensively used student interaction data for a discovery with models approach [23]. For example, student interactions in a math tutor were used to build a predictive model of students' careless errors [38], and those models were connected with predictive models of affective states to study the relationship between affective states and carelessness [39].

Log data from math tutors have also been used to predict student scores on end-of-year state accountability exams, resulting in better prediction than paper-pencil benchmark tests and standardized tests [4, 18, 21]. These models become better still when supplemented with data on student strategies [36]. Xie et al. [42] showed how learning strategies defined by interaction data (e.g., learning from errors, switching to a new topic, and reviewing previously mastered topics) predicted end of semester assessments.

Other research has found that student behavior in math tutors in middle school year are predictive of long term success. For example, San Pedro et al. [37] found that student carelessness, and intentional misuse to complete problems without learning in a middle school math tutor are associated with lower probability of college attendance and STEM major. Similarly, Ocumpaugh et al. [34] conducted a longitudinal study of the relationship between middle school math performance and interaction-based affect detectors with student's vocational self-efficacy and interest. They found that both self-efficacy and interest in high school were negatively correlated with confusion during middle school, but that both were positively correlated with carelessness.

Recent studies have also examined click-stream data and math success, usually in conjunction with NLP tools. For instance, Crossley and Kostyuk [11] reported that elementary students who met more objectives within an online math tutoring system and those that sent fewer messages to a pedagogical agent, performed better on math problems. Crossley et al. [10, 16] reported that college level students that received higher final scores in a blended math class spent more time in a forum that allowed postings between students and teachers and visited the online learning platform more often.

1.3 Current Study

As discussed above, a number of studies have demonstrated strong links between students' linguistic knowledge, their affective language patterns and their success in math. In addition, studies have shown that click-stream variables are important predictors of success in online learning systems. This study builds on this previous research in two ways. First, it combines linguistics and click-stream variables along with demographic information to increase prediction rates for math success. Second, it examines how random variance, as found in repeated participant data, can explain math success beyond linguistic, demographic, and click-stream variables.

To derive our language features of interest, we analyzed the language produced by students sending email messages to a virtual pedagogical agent within an online math tutoring system. We analyzed the language using several Natural Language Processing (NLP) tools in order to extract language information related to text cohesion, lexical sophistication, and sentiment. Our click-stream data was extracted from the online tutoring data and focused on actions within the system, entries into various modes of the system, and temporal data related to time spent in those modes. Demographic data included grade and gender. We collected data from students in two consecutive semesters (fall and spring) allowing us to track performance over time. Thus, in this study, we address the following research question:

Are linguistic and click-stream factors along with participant variance over time significant predictors of math performance in an online tutoring environment over two semesters of study?

2 Method

2.1 Reasoning Mind

We collected data from Reasoning Mind's *Foundations* product, which is a blended learning mathematics program used in grades 2–5. *Foundations* students learn math in an engaging, animated world at their own pace, while teachers use the system's real-time data to provide one-on-one and small-group interventions [32]. The algorithms and pedagogical logic underlying Foundations (previously called Genie 2) are described in detail by Khatchatryan et al. [24].

The main study mode in *Foundations*, called *Guided Study*, consists of a sequenced curriculum divided into objectives, each of which introduces a new topic (e.g., the distributive property) using interactive explanations, presents problems of increasing difficulty on the topic, and reviews previously studied topics. Within *Guided Study*, every student completes problems addressing the basic knowledge and skills required in the objective. These basic problems (known as A-level problems) typically require only a single step to solve and are the lowest of three possible difficulty levels. Students who do well on A-level problems may also proceed to problems of higher difficulty that require two or three steps to solve (e.g., B-level and C-level problems) within the objective. They may also access the higher-level problems in an independent study mode called *Wall of Mastery*. Other modes in *Foundations* allow students to play math games against classmates, tackle challenging problems and puzzles, and use points earned by solving math problems to buy virtual prizes.

Foundations uses animated characters to provide a backstory to the mathematics being learned and to deliver emotional support. The main character is the Genie, a pedagogical agent who encourages students throughout their work in the system. Students are also able to send emails to the Genie. These messages are answered in character by part-time Reasoning Mind employees who reference an extensive biography of the Genie and project a consistent, warm, and encouraging persona, model a positive attitude toward learning, and emphasize the importance of practice and challenging work for success. The Genie email system is a popular component of the system, having received 129,879 messages from 38,940 different students in the 2016–17 academic year.

2.2 Participants

The students sampled in this study were selected from the 34,602 students who used *Foundations* in the 2016–17 academic year. The students were from 462 different schools located in 99 different districts, most of which were located in Texas. We included those students that had attempted A-level problems in both the fall and spring semester. As an additional requirement, these students needed to have written at least 50 words within the Genie email system (the minimum number of words needed to develop a linguistic profile for the students). From the available student data, 1,036 students met these criteria.

2.3 Genie Email Corpus

Our language sample for this analysis consisted of messages sent from the students to the Genie. Because many messages contained only a few words, we aggregated all emails sent by each student to create a representation of an individual student's linguistic activity.

We then implemented data cleaning procedures to reduce the amount of noise in the data. First, all the data was cleaned of non-ASCII characters that could interfere with the NLP tools. Second, all texts were automatically spell-checked and corrected using an open-source Python spelling correction library, in addition to several Python text-cleaning scripts that we developed. Furthermore, several measures were taken to clean the texts, including removing random, non-math symbols such as "#", "@", and "&", as well as omitting repeating words, excessively long words, words with repeating characters, such as "wooorrrddd", and mixed-type words, such as "$word$" (with the exceptions of currencies, percentages, timestamps, and ordinals). Next, all non-dictionary, invalid words were removed from the data. This was accomplished by first checking each word against synonym sets (synsets) in WordNet, and if a match could not be found, then checking if it consisted of all consonants (always invalid), or if any pair of characters (digraph) in the word were invalid in the English language. Words that met either of these two conditions were removed. Lastly, all texts were cleaned of repeating, non-overlapping groups of words, such as "this word this word this word." Only word groups of lengths two, three, and four were removed by this approach.

2.4 Natural Language Processing Tools

We used several NLP tools to assess the linguistic features in the aggregated posts of sufficient length. These included the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) [26], the Tool for the Automatic Analysis of Cohesion (TAACO) [14], the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) [27], and the SEntiment ANalysis and Cognition Engine (SEANCE) [15]. In addition, we developed specific indices related to topics commonly discussed with the Genie email system using Latent Dirichlet Allocation (LDA). Thus, the selected NLP features consisted of language variables related to lexical sophistication, text cohesion, syntactic complexity sentiment analysis, and topic similarity respectively. The features are discussed in greater detail below.

TAALES. TAALES [26] is a computational tool that is freely available and easy to use, works on most operating systems, affords batch processing of text files, and incorporates over 100s of classic and newly 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. For many indices, TAALES calculates scores for all words (AW), content words (CW), and function words (FW). For instance, for word frequency, TAALES reports frequency counts retrieved the SUBTLexus databases [9].

TAALES also reports on a number of word information and psycholinguistic scores derived from the University of South Florida (USF) norms [33], and the English Lexicon Project (ELP) [5] among others. The USF norms are used to calculate the number of associations per word while the ELP is used to calculate many lexical features including the number of orthographic neighbors a word has (i.e., how many words are spelled similarly). Lastly, TAALES reports on type token ratios (TTR) that reference how many unique words are found within a sample.

TAACO. TAACO [14] incorporates a number of classic and recently developed indices related to text cohesion. TAACO has features for content and function words and provides linguistic counts for both sentence and paragraph markers of cohesion. The tool incorporates WordNet synonym sets, latent semantic analysis, and word2vec features.

Specifically, TAACO calculates sentence and paragraph overlap indices and a variety of connective indices.

TAASSC. TAASSC [27] measures large and fined grained clausal and phrasal indices of syntactic complexity and usage-based frequency/contingency indices of syntactic sophistication. TAASSC includes a number of pre-developed fine-grained indices or clausal complexity and phrasal complexity. In addition, TAASCC reports on features related to verb argument constructions (VACs) including the frequency of VACs and the attested constructions in reference corpora taken from the Corpus of Contemporary American English (COCA) [17] to include sub-corpora such as academic writing, magazines, and fiction.

SEANCE. SEANCE [15] 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 to measure sentiment, cognition, and social order. These vectors are taken from freely available source databases. For many of these vectors, SEANCE also provides a negation feature (i.e., a contextual valence shifter) that ignores positive terms that are negated (e.g., not happy). SEANCE also includes a part of speech (POS) tagger.

2.5 Click-Stream Variables

Reasoning Mind extensively logs the interaction of the students in the system at the action-level. Actions include logging in, entering a mode, seeing a problem, submitting an answer, and reviewing theory. The type of action a student can take depends on the mode they are in. For instance, in *City Landscape*, students can switch modes such as *Guided Study* and *Wall of Mastery*, where students practice math problems. The *Game Room* also provides an opportunity for students to learn math through games. In contrast, no math learning happens in the *City Landscape* mode, which is simply the landing page in Reasoning Mind from where the student navigates to other modes. Students can also spend time in the *Shopping Mall* purchasing items to decorate their *My Place*, which might have an impact on their overall engagement with the system.

To explore the student interaction patterns in the tutor, we engineered features that captured the distribution of a student's effort in the various phases of the tutor. Along with quantifying the kind of actions a student performs, these features also measure their persistence in an activity. For instance, a higher number of visits to *City Landscape* mode would denote that the student is less persistent in focusing on a single activity.

For each student, we extracted features based on the actions in the 27 modes (e.g., *City Landscape*), the actions within the 11 module types of *Guided Study* mode (e.g., *Introduction*, *Theory*, *Problems*, *Homework*) and 6 content types (e.g., *Problem A*, *Problem B*). For each of these, we mined the log data to extract three kinds of features – 1) number of entries to the mode (e.g., number of entries to *City Landscape* mode); 2) number of actions performed in the mode/module/content type (e.g., number of actions performed in *Problems* module type); 3) total time-spent in the mode/content type (e.g., total time spent solving *Problem A* content type). Thus, the feature *Time in City Landscape* refers to the total sum of time spent by a student across all logins in the *City Landscape* mode. Similarly, the feature *Number of entries to Guided Study* is calculated by counting the number of times a student enters the *Guided Study* mode and summing the counts across all their logins in a semester. The feature value varies drastically across the students in this dataset. For instance, *Number of entries to Guided Study* has a mean of 125.96 and a standard deviation of 116. In contrast, *Time in*

Guided Study has a mean of 18.86 hours and a standard deviation of 15.25 hours per semester. In addition, we calculated normalized features measuring number of hints, number of virtual prizes purchased, and problem accuracies. In total, we mined 110 click-stream features.

2.6 Statistical Analysis

Prior to analysis, all numeric scores were standardized. We used linear mixed effects (LME) models in R [35] using the *lme4* package [7] to develop models of math scores over time (i.e., across the fall and spring semesters). We opted to use LME models because they offer statistical advantages over traditional repeated measures analyses of variance (RM ANOVAs). Specifically, LMEs account for both pooled and individual variance among participants as opposed to only pooled group variance by including subjects as random effects (i.e., assigning a unique intercept for each participant), resulting in more accurate estimates based on individual participant variation. The purpose of the model was to test whether any of the independent variables (e.g., grade level, linguistics and affect features, and click-stream variables) significantly predicted math success. Accordingly, in the model, we entered math success as the dependent variable, with grade level, gender, linguistics and affect features, and click-stream variables as fixed effects (i.e., predictor variables). No interactions were conducted between fixed factors. The baseline grade level was second grade. Grade levels were balanced at around 250–300 students in grades 3–5. There were fewer students in first grade (∼150) and sixth grade (∼15).

To help prevent over-fitting, we removed several variables prior to analysis. First, we conducted correlations between the dependent variables and the independent variables. Any independent variable that did not demonstrate at least a small relationship with the dependent variable ($r \geq 0.100$) was removed from the analysis. Next, we checked for multicollinearity between the remaining independent variables using variance inflation factors (VIF) with a threshold set to 5 (i.e., high multicollinearity). All variables showing VIF above 5 were removed from the analysis and the remaining variables were used in the LME analysis. This variable pruning left us with five click-stream variables ad seven linguistic variables. For each dependent variable, an initial LME model was run with all independent variables. After an initial model was constructed, we used a stepwise variable selection technique (backwards) to eliminate non-significant effects. The results of the stepwise model were used as the final models for the analyses.

We used several other packages to aid in our construction and interpretation of our models. We used *lmerTest* [25] to derive p-values from the models and to perform automatic backward elimination of variables in the LME models, and the *MuMIn* package [6] to obtain two measures of variance explained: a marginal R^2 measuring the variance explained by the fixed effects only, and a conditional R^2 measuring the variance explained by the fixed and random effects combined.

3 Results

An LME model predicting math success as the dependent variable reported significant main effects for a number of click-stream and linguistic features. In general, the click-stream effects indicate that students that were more successful at level A math problems spent less time in the main page of the system (i.e., the *City Landscape*), entered *Guided Study* more, and purchased more items. The linguistic effects demonstrated that students that were more successful at level A math problems produced more sophisticated language (i.e., words with fewer associations, fewer orthographic neighbors and lower range scores), used a greater

Table 1 LME model predicting math success.

Fixed effect	Estimate	Percent of estimate	Std. Error	t	p
(intercept)	0.086		0.064	1.335	0.182
6th grade	-0.512	0.265	0.213	-2.405	0.016
3rd grade	-0.291	0.151	0.080	-3.654	0.000
5th grade	-0.274	0.142	0.083	-3.299	0.001
4th grade	0.219	0.114	0.080	2.758	0.006
Time in City Landscape	-0.137	0.071	0.023	-5.911	0.000
Number of entries into Guided Study	0.094	0.049	0.025	3.765	0.000
Word associations (USF) CW	-0.068	0.035	0.022	-3.097	0.002
Number of items purchased	0.061	0.031	0.023	2.629	0.009
Word range (SUBTLEXus) AW	-0.060	0.031	0.023	-2.577	0.010
Attested constructions (Magazine)	0.060	0.031	0.024	2.490	0.013
Moving Avg. Type Token Ratio (MATTR)	0.057	0.029	0.022	2.541	0.011
Semantic overlap between sentences (word2vec)	0.052	0.027	0.025	2.080	0.038
Orthographic neighbors (CW)	-0.047	0.024	0.023	-2.019	0.044

diversity of words, used more common syntactic constructions, and produced language that was more cohesive. Time was not a significant predictor. The model reported a marginal R^2 of 0.139 and a conditional R^2 of 0.438. Table 1 displays the estimates, percent of estimate, standard errors, t-values, and p-values for the fixed effects for this model.

4 Discussion

This study builds on previous work that examines links between math success and language production by examining how language features and click-stream variables combine to explain student success in an online tutoring system. Unlike previous studies, the current study included student growth over time as a variable. In addition, the click-stream variables examined in this study were finer-grained than in previous studies, allowing us to better understand how student behaviors within the system help explain math success in conjunction with language features. Overall, our fixed factors explained about 14% of the variance in math scores (marginal variance) while a mix of both fixed and random factors explained about 44% of the variance (conditional variance).

In general, the results indicate that grade level was the strongest predictor of math success such that there was a decline in the percentage of level A problems correctly answered among students in higher grades, although fourth-grade students showed a notable departure from that trend. Post-hoc analyses (not reported here) indicated that trends were not linear with fourth graders performing better than third, fifth and sixth graders. We hypothesize that these results may be indicative of a developmental milestone or a change in curriculum expectations (e.g., with competence generally increasing, but fifth and sixth graders receiving more challenging material), but more research is needed.

Beyond grade level, the next strongest predictors were related to click-stream variables. Specifically, the more time that students spent in the *City Landscape*, the lower they performed on A-level math problems. This is not surprising, as no math learning happens in *City*

Landscape. In fact, spending more time in *City Landscape* may suggest the student has lower persistence because they are constantly trying to switch modes in the tutor. In comparison, the more entries they made to *Guided Study*, which is the main instructional and study mode in *Foundations*, the better they performed. We also observe that high student performance was correlated with more purchases in the shopping mall; this is because the points used to make purchases are earned through better mathematics performance. It may be an interesting area of future work to see if differences in the items students purchase relate to differences in math success.

In terms of language production, more successful students produced words that were more sophisticated. For example, more successful students used words that had fewer associations, were found in fewer texts (i.e., a lower range score), and had fewer orthographic neighbors (i.e., words that are spelled similarly). All of these indices indicate that more successful students had more depth of lexical knowledge. Not only were the words they produced more complex, these students also used a greater variety of words (i.e., lexical diversity) indicating that they had larger productive vocabularies (i.e., breadth of lexical knowledge). Beyond lexical knowledge, more successful students also produced a greater number of verb argument constructions indicating a greater range of syntactic structures and produced language that was more cohesive in terms of semantic similarity between sentences. These data indicate that students who are more successful at solving math problems are more proficient language users, in line with previous findings [10, 11, 12, 13, 16].

It is interesting to note that time was not a significant predictor of math success in our LME model. Thus, there is no evidence of improved performance between the fall and spring semester in terms of A-level problems. This is likely related to the curriculum design, which scaffolds student learning and arranges content to become increasingly difficult as a student masters easier content. In addition, gender was not a significant predictor of math success; girls and boys performed similarly on level A problems.

5 Conclusion

The work presented here provides additional evidence that links language production to math success. In general, there seems to be strong evidence that students who are more successful in math produce language that is more sophisticated in terms of words and more complex in terms of syntax. In addition, more successful math students also produce language that is more cohesive and follows language conventions found in adult language corpora.

The study also finds that student choice of activities is associated with their degree of success. Specifically, this study looked at the virtual locations within RM City, which represent different activities within the Foundations curriculum. We found that students who were struggling were more likely to be switching from one activity to the next (through the *City Landscape* mode). while high performing students were more likely to engage in other types of activities, like making purchases for their *My Space*) and spending more time in modes related to math problems.

While we were able to capture many features of student behavior within the system and student language production, there is of course room for further feature engineering. For example, it may be possible to better capture the variation in student behavior in the system through creating additional temporal features. Linguistically, features related to intention and meaning should be deployed as well to increase our knowledge of how language and math skills interact.

Lastly, of interest is the amount of variance explained by the random factors (i.e., the conditional variance). While the fixed factors explained about 14% of the variance in the math score, the majority of the variance (∼30%) was explained by the random factor of participant. This finding may indicate that much of math success is not in behaviors within the system, grade level, or language production but likely rather resides in the individual differences of students. These results suggest that it may be useful in future research to look into which individual differences (e.g., ELL status, geographic location, ethnicity, socio-economic status) may best explain math success.

References

1 Jamal Abedi and Carol Lord. The Language Factor in Mathematics Tests. *Applied Measurement in Education*, 14(3):219–234, 2001.

2 Thomasenia Lott Adams. Reading mathematics: More than words can say. *The Reading Teacher*, 56(8):786–795, 2003.

3 Mary Alt, Genesis D Arizmendi, and Carole R Beal. The relationship between mathematics 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, 2014.

4 Nathaniel Anozie and Brian W Junker. Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. Technical report, Educational Data Mining: Papers from the AAAI Workshop. Menlo Park, CA: AAAI Press, 2006.

5 David A Balota, Melvin J Yap, Keith A Hutchison, Michael J Cortese, Brett Kessler, Bjorn Loftis, James H Neely, Douglas L Nelson, Greg B Simpson, and Rebecca Treiman. The English lexicon project. *Behavior research methods*, 39(3):445–459, 2007.

6 Kamil Barton. *MuMIn: Multi-Model Inference*, 2018. URL: `https://CRAN.R-project.org/package=MuMIn`.

7 Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

8 Carole R Beal, Rena Walles, Ivon Arroyo, and Beverly P Woolf. On-line tutoring for math achievement testing: A controlled evaluation. *Journal of Interactive Online Learning*, 6(1):43–55, 2007.

9 Marc Brysbaert and Boris New. Subtlexus: American word frequencies. `http://subtlexus.lexique.org`, 2009.

10 Scott Crossley, Tiffany Barnes, Collin Lynch, and Danielle S McNamara. Linking Language to Math Success in an On-Line Course. In *Proceedings of the 10th International Conference on Educational Data Mining*, pages 180–185, Wuhan, China, 2017.

11 Scott Crossley and Victor Kostyuk. Letting the Genie out of the Lamp: Using Natural Language Processing tools to predict math performance. In *International Conference on Language, Data and Knowledge*, pages 330–342. Springer, 2017.

12 Scott Crossley, Ran Liu, and Danielle McNamara. Predicting math performance using natural language processing tools. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 339–347. ACM, 2017.

13 Scott Crossley, Jaclyn Ocumpaugh, Matthew Labrum, Franklin Bradfield, Mihai Dascalu, and Ryan S Baker. Modeling Math Identity and Math Success through Sentiment Analysis and Linguistic Features. In *International Educational Data Mining Society*. ERIC, 2018.

14 Scott A. Crossley, Kristopher Kyle, and Mihai Dascalu. The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, 51(1):14–27, 2019.

15 Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior research methods*, 49(3):803–821, 2017.

16 Scott A Crossley, Maria-Dorinela Sirbu, Mihai Dascalu, Tiffany Barnes, Collin F Lynch, and Danielle S McNamara. Modeling Math Success Using Cohesion Network Analysis. In *International Conference on Artificial Intelligence in Education*, pages 63–67. Springer, 2018.

17 Mark Davies. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, 14(2):159–190, 2009.

18 Mingyu Feng, Neil T Heffernan, and Kenneth R Koedinger. Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required. In *International conference on intelligent tutoring systems*, pages 31–40. Springer, 2006.

19 Pier Luigi Ferrari. Mathematical Language and Advanced Mathematics Learning. *International Group for the Psychology of Mathematics Education*, 2004.

20 Gillian Hampden-Thompson, Gail Mulligan, Akemi Kinukawa, and Tamara Halle. Mathematics Achievement of Language-Minority Students During the Elementary Years. Research report, The University of York, Washington, DC, 2008.

21 Neil T Heffernan, Kenneth R Koedinger, Brian W Junker, and Steven Ritter. Using Web-based cognitive assessment systems for predicting student performance on state exams. *Research proposal to the Institute of Educational Statistics, US Department of Education. Department of Computer Science at Worcester Polytechnic Institute, Massachusetts*, 2001.

22 Federico Hernandez. *The Relationship Between Reading and Mathematics Achievement of Middle School Students as Measured by the Texas Assessment of Knowledge and Skills*. PhD thesis, University of Houston, 2013.

23 Arnon Hershkovitz, Ryan Shaun Joazeiro de Baker, Janice Gobert, Michael Wixon, and Michael Sao Pedro. Discovery with models: A case study on carelessness in computer-based science inquiry. *American Behavioral Scientist*, 57(10):1480–1499, 2013.

24 George A Khachatryan, Andrey V Romashov, Alexander R Khachatryan, Steven J Gaudino, Julia M Khachatryan, Konstantin R Guarian, and Nataliya V Yufa. Reasoning Mind Genie 2: An intelligent tutoring system as a vehicle for international transfer of instructional methods in mathematics. *International Journal of Artificial Intelligence in Education*, 24(3):333–382, 2014.

25 Alexandra Kuznetsova, Per B Brockhoff, and Rune Haubo Bojesen Christensen. lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26, 2017.

26 Kristopher Kyle, Scott Crossley, and Cynthia Berger. The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior research methods*, 50(3):1030–1046, 2018.

27 Kristopher Kyle and Scott A Crossley. Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. *The Modern Language Journal*, 102(2):333–349, 2018.

28 Jo-Anne LeFevre, Lisa Fast, Sheri-Lynn Skwarchuk, Brenda L Smith-Chant, Jeffrey Bisanz, Deepthi Kamawar, and Marcie Penner-Wilger. Pathways to mathematics: Longitudinal predictors of performance. *Child development*, 81(6):1753–1767, 2010.

29 Mollie MacGregor and Elizabeth Price. An exploration of aspects of language proficiency and algebra learning. *Journal for Research in Mathematics Education*, 30:449–467, 1999.

30 Maria Martiniello. Language and the performance of English-language learners in math word problems. *Harvard Educational Review*, 78(2):333–368, 2008.

31 Maria Martiniello. Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational assessment*, 14(3-4):160–179, 2009.

32 William L Miller, Ryan S Baker, Matthew J Labrum, Karen Petsche, Yu-Han Liu, and Angela Z Wagner. Automated detection of proactive remediation by teachers in Reasoning Mind classrooms. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, pages 290–294. ACM, 2015.

33 Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. The University of South Florida word association, rhyme, and word fragment norms, 1998.

34 Jaclyn Ocumpaugh, Maria Ofelia San Pedro, Huei-yi Lai, Ryan S Baker, and Fred Borgen. Middle school engagement with mathematics software and later interest and self-efficacy for STEM careers. *Journal of Science Education and Technology*, 25(6):877–887, 2016.

35 R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL: `https://www.R-project.org/`.

36 Steven Ritter, Ambarish Joshi, Stephen Fancsali, and Tristan Nixon. Predicting standardized test scores from Cognitive Tutor interactions. In *Proceedings of the International Conference on Educational Data Mining*, 2013.

37 Maria Ofelia San Pedro, Jaclyn Ocumpaugh, Ryan S Baker, and Neil T Heffernan. Predicting STEM and Non-STEM College Major Enrollment from Middle School Interaction with Mathematics Educational Software. In *Proceedings of the International Conference on Educational Data Mining*, pages 276–279, 2014.

38 Maria Ofelia Clarissa Z San Pedro, Ryan SJ d Baker, and Ma Mercedes T Rodrigo. Detecting carelessness through contextual estimation of slip probabilities among students using an intelligent tutor for mathematics. In *International Conference on Artificial Intelligence in Education*, pages 304–311. Springer, 2011.

39 Maria Ofelia Z San Pedro, Ryan SJ d Baker, and Ma Mercedes T Rodrigo. Carelessness and affect in an intelligent tutoring system for mathematics. *International Journal of Artificial Intelligence in Education*, 24(2):189–210, 2014.

40 David Tall. Thinking Through Three Worlds of Mathematics. *International Group for the Psychology of Mathematics Education*, 2004.

41 Rose K Vukovic and Nonie K Lesaux. The relationship between linguistic skills and arithmetic knowledge. *Learning and Individual Differences*, 23:87–91, 2013.

42 Jun Xie, Alfred Essa, Shirin Mojarad, Ryan S Baker, Keith Shubeck, and Xiangen Hu. Student learning strategies and behaviors to predict success in an online adaptive mathematics tutoring system. In *Proceedings of the International Conference on Educational Data Mining*, pages 460–465, 2017.