Modeling Math Identity and Math Success through Sentiment Analysis and Linguistic Features

Scott Crossley
Georgia State University
Atlanta, GA 30303
scrossley@gsu.edu

Jaclyn Ocumpaugh
University of Pennsylvania
Philadelphia, PA 19127
jlocumpaugh@gmail.com

Matthew Labrum
Reasoning Mind
Houston, TX 77507
mlabrum@reasoningmind.org

Franklin Bradfield
Georgia Tech Research Institute
Atlanta, GA 30318
Franklin.Bradfield@gtri.gatech.edu

Mihai Dascalu
University of Bucharest
Bucharest, Romania 060042
mihai.dascalu@cs.pub.ro

Ryan S. Baker
University of Pennsylvania
Philadelphia, PA 19127
ryanshaunbaker@gmail.com

ABSTRACT

A number of studies have demonstrated strong links between students' language features (as found in spoken and written production) and their math performance. However, no studies have examined links between the students' language features and measures of their Math Identity. This project extends prior studies that use natural language processing (NLP) features to examine student language features and math performance, replicating their analyses. The study then uses NLP features to model students' Math Identity. Specifically, the study compares performance on basic math skills within an online math tutoring system to both student language (as captured in emails to a virtual pedagogical agent) and to survey measures of Math Identity (math self concept, interest, and value). Language features were analyzed by a number of NLP tools that extracted information related to text cohesion, lexical sophistication, and sentiment. The findings indicate weak to medium relationships between math scores and Math Identity and language features were able to predict a significant amount of the variance in each Math Identity variable and in math scores. The potential for these measures to inform interventions for students with lower Math Identity is discussed.

Keywords

Natural language processing (NLP), math, math identity, student success, on-line learning

1. INTRODUCTION

Educational Data Mining (EDM) has, among its many applications, been employed to better understand student-level differences that are important to personalization efforts in educational settings [1, 2]. These include efforts to better understand constructs like student engagement (e.g., [3]), self-efficacy [4], and self-concept [5]. Many of these studies have relied upon sensors (e.g. posture sensors, vocal recognition, heartbeat, video, sweat/skin conductance, EEG), which can sometimes make it challenging to implement interventions *in situ*. Research using student interaction data has become more common even when modeling highly qualitative constructs like student engagement (c.f., [3]), but to date, much of these efforts have focused on temporally short variables (e.g., state-based variables like behaviors and affect), rather than on trait-based variables such as identity, which are larger in scope and duration.

Work in related research areas has shown results that suggest that trait-based variables may be a promising area for investigation. Within the EDM community, there is now a growing body of research on identity-related constructs, such as motivation and self-regulated learning strategies (cf. [6]). Meanwhile, the related field of Natural Language Processing (NLP) has demonstrated relationships between language use and personality characteristics (cf.,[7, 8]). Detecting a construct like identity, which underlies motivation and goals [9], could further advance efforts toward personalized learning within educational setting, including the development of effective intervention strategies.

Identity, broadly, refers to a person's sense of who they are and the development of an identity permits people to make predictions about their abilities to navigate different aspects of their life (cf. [9]). While identity is the focus of this study, we do not attempt to investigate all aspects of student identity, but instead focus specifically on how they identify with math. Math Identity is often described as "the association between math and the self" [10], a definition that might be paraphrased as the degree to which one considers oneself to be a math person. We do so within the context of Reasoning Mind, a blended learning curriculum that offers significant metacognitive support to K-6th grade students through an on-line learning platform [11]

Specifically, we use language features produced in within-system emails to predict three aspects of Math Identity in self-reported survey data: math self-concept, math interest, and math value. These constructs have been used to understand social influences on mathematic achievement in previous studies of identity (e.g., [12]). In addition, we examine links between math success in the system and the three Math Identity scales. We also use language features in the language produced by students to model math success, math value, math self-concept, and math interest. Our goal is to examine the potential for linguistic predictors within student data to identify math success and identity. If successful, such linguistic predictors could be used to better identify students in need of intervention.

2. Language and Math Ability

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, [13] 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, [14] reported links between language and math skills, but also 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.

Other research has focused on more indirect links between math and language skills, such as reading ability. For example, Hernandez [15] found significant positive correlations between reading and math scores in standardized tests. Based on these findings, Hernandez recommended that reading skills and reading strategies should be factored into math instructions to increase math ability, especially for poor readers. In another study, LeFevre et al. [16] reported that language ability was positively related to number naming, but that non-language abilities such as quantitative skills and spatial attention were stronger predictors of math ability than language abilities.

A number of 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. For instance, Crossley et al. [17] examined linguistic and non-linguistic features of elementary student discourse while students were engaged in collaborative problem solving within an on-line math tutoring system. NLP tools that reported on affect, text cohesion, and lexical sophistication were used to extract linguistic information from transcribed student speech. These language features along with a variety of non-linguistic features such as gender, age, grade, and school were used to predict pre- and post-test math scores. The results showed that language features related to cohesion, affect, and lexical proficiency explained around 30% of the variance in students' math scores, while the selected non-language features were not significant predictors. A second study by Crossley and colleagues examined students' forum posts in an online tutoring system. Using these posts, Crossley et al. [18] investigated relationships between math success, click-stream data within the system, and language features reported by NLP tools for students in a university level blended math class (i.e., a class with both on-line and traditional face to face instruction). The study found that math success was best predicted by a non-language feature (days on the system) and language features related to affect (egotism), syntactic complexity and text cohesion. Specifically, more complex syntactic structures and fewer explicit cohesion devices equated to 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 addition, the results indicated that students that are more active in on-line discussion forums are more likely to be successful. In a final study, Crossley and Kostyuk [19] examined links between the language features of young students' language production (grades 2nd through 5th) while e-mailing a virtual pedagogical agent in an online math tutoring system, and success within that system. Using NLP tools that reported language features related to affect, lexical sophistication, and text cohesion, Crossley and Kostyuk found that students who expressed more certainty in their writing and followed standardized language patterns scored higher in math assessments. In addition, students from higher grades who met more objectives, received more messages from teachers, and sent fewer messages to the agent, performed better on math problems.

Overall, these studies demonstrate that features from students' language productions can be used to predict math success (i.e., performance) in a variety of domains and across a number of ages and proficiency levels. In general, older students who produce more complex language, which is more positive and less self-centered, tend to have stronger math skills. For younger students, adherence to expected language patterns relates to higher math performance. However, to our knowledge, no research has attempted to extend this approach to predicting larger student identity features that are trait-based such as Math Identity.

3. Math Identity

Math Identity, or the degree to which one considers oneself a "math person," has become an area of interest among social scientists hoping to better understand what drives students to enter Science, Technology, Engineering, and Math (STEM) fields (cf. [20]). However, broader issues of self-definition (identity) are not new to educational research, especially when considering long-term development. For example, Bandura's research [21] on self-efficacy discusses the role of self-attributional processes (including a wide range of self-definitions studied by Bem, [22] many of which are directly related to educational identities. In this research, a student's cognitive appraisal (self-evaluation of ability) is thought to be susceptible to a form of confirmation bias where the student ignores demonstrable achievements and improvements when they contrast with a previously established self-definition [21]. Bandura's observations on the role of self-definitions in the development of self-efficacy are highly compatible with other research paradigms, which describe identity as an anchor that people use to understand their own interests and abilities [23]. This may explain Bandura's findings that students who show improvement that is contrary to self-appraisals often attribute their performance to environmental factors rather than to their own persistence [21].

Constructs considered to be a core part of one's identity are long thought to start developing in adolescence ([24]. There is some support that Math Identity should be included in this timeframe with research suggesting that it develops early in life. For instance, [25] showed that students who start in a non-STEM degree program rarely transfer into a STEM program (despite the high frequency of major changes more generally). Similarly, within the EDM community, student engagement indicators in middle school online mathematics tutors have been shown to correlate with college enrollment more generally [26], and with STEM-major enrollment more specifically [27]. Math Identity is most often studied through ethnographic studies (e.g., [28]), implicit association tests (e.g., [29, 10]), and surveys (e.g., [30, 31]).

In this study, we operationalize Math Identity as math self-concept, math interest, and math value. We defined these constructs using self-report scales adapted from Ryan & Ryan [12], who examined how these constructs performed during conditions likely to trigger stereotype threat effects. While these are well-established constructs in research on the effects of social evaluations of mathematics, they are not unique to research on identity. In addition to their appearance in Bandura's work, they appear in Eccles' [32] expectancy value theory, where self-efficacy (among a variety of other factors) is hypothesized to influence both intrinsic value (interest) and utility value (the usefulness of the task). We discuss each of these briefly below.

3.1.1 Math Self-Concept

Research in self-concept overlaps considerably with two related constructs—identity and self-efficacy—because all three are related to the mental schema a person uses when calculating their ability to negotiate different challenges in their lives. In general, social-psychologists are more likely to refer to the concept of identity when discussing issues related to social processes, while they are more likely to use the term self-concept when discussing internal mental processes ([9]).

In education research, self-concept and self-efficacy are often used to discuss domain-specific evaluations (e.g., self-concept in mathematics), and they are sometimes used synonymously. However, there are education researchers who draw a distinction between these two constructs, limiting the term self-efficacy to self-evaluations of specific tasks, often specifying that it must be measured directly after the task has been completed [33, 34]. For example, they might use a Likert scale administered after each math problem to measure self-efficacy by asking a student to indicate his/her confidence that each problem had been completed correctly.

In this research tradition, self-concept is a broader measure of ability within the domain, where its meaning more closely approaches its use among social-psychologists, who tend to define it as a theory of self (e.g., [35]) which often operates below the level of consciousness, guiding people's interpretations and expectations of external events (cf. [9]). For example, in a situation where a student failed a task in a domain for which they have high self-concept, they might be more willing to retry than someone with low self-concept. Alternatively, they might interpret the task as flawed since their performance did not match the expectations created by their self-concept.

Like researchers who study educational outcomes, social psychologists tend to believe that people develop self-concept from experience, so that those with more shallow or limited experiences are likely to be more susceptible to changes in self-concept [35]. For example, academic self-concept tends to be positively correlated with achievement indices, [36], but there appears to be some reciprocity in this relationship. High self-concept can make students more likely to persist through difficult mathematics instruction, leading to improved academic outcomes. However, repeated failure could theoretically lower self-concept, particularly if a student did not have other mastery experiences in mathematics to serve as a sort of buffer.

3.1.2 Interest in Mathematics

Motivational research defines interest as the propensity to engage with a particular subject over time through both affective and cognitive components [37]. Studies on the relationship of interest to other constructs such as self-concept have repeatedly found that self-concept drives intrinsic interest in a given subject [38, 39], with theorists suggesting that as self-efficacy increases, it becomes safe for the ego to become invested in a particular topic [40].

Researchers have identified a number of simple strategies that appear to increase interest in the classroom, such as creating more challenging tasks for students or adding variety to the ways in which a student is asked to perform a task. However, others caution that some of these strategies may only improve situational interest (e.g., [37]), suggesting that intrinsic interest (which they refer to as individual interest) is almost always self-driven, possibly because it seems to be fed by increased self-efficacy. Others researchers have found that interest is highly susceptible to contextual effects that vary from student to student (cf. [39]). Researchers in Career Theory (e.g., [41]) have found that interest, like self-efficacy, is directly responsive to performance success and failure.

Interest is an important complement to self-concept when defining Math Identity, since its development is known to improve self-regulatory strategies [37]. Students with a stronger sense of interest in a subject are more likely to persist when confronted with frustrating challenges [42, 37; 43], so that strengthening skills in mathematics is a self-feeding cycle. Eccles' [32] discussion of identity development mentions this cycle and state that enjoyable or pleasant experiences with a subject are likely necessary to develop the persistence needed to become an expert in that subject.

3.1.3 Value of Mathematics

Math value is the degree to which a student thinks that math is or will be useful to their life. Like self-concept and interest, value (utility) has been linked to motivation in a number of different research traditions. Among social psychologists, research has shown that value is influenced by self-concept, and, in turn, that value positively influences the kind of goal-setting practices that lead to increased effort [44]. However, research also finds that (perhaps more than self-concept or interest), parents can have a substantial effect on math value [44, 45], which suggests the construct could also be more susceptible to other social pressures or interventions. Cumulatively, these findings suggest that value is often the last component of Math Identity to develop unless external influences (e.g., parents) are involved.

4. Current Study

A number of studies have demonstrated strong links between students' linguistic knowledge and affect (as found in language production), and their success in math. However, to our knowledge, no studies have examined the links between the linguistic features in student language production and variables related to Math Identity. In the current study, we attempt to replicate previous studies that have investigated how linguistic features and affective aspects of students' language production can predict success. More importantly, we also derive models of math identify based on student survey responses related to math value, interest, and self-concept. 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 a number of NLP tools in order to extract language information related to text cohesion, lexical sophistication, and sentiment. While our primary interest is in using NLP features to predict variables related to math value, interest, and self-concept, we are also interested in studying the links between NLP features and accuracy scores on beginning level math problems within the online tutoring system. Thus, in this study, we address two research questions:

1. Are linguistic features significant predictors of self-reported student traits related to math value, interest, and self-concept?
2. Are linguistic factors significant predictors of math performance in an on-line tutoring environment?

5. METHOD

5.1 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 [46]. The algorithms and pedagogical logic underlying *Foundations* (previously called *Genie 2*) are described in detail by Khachatryan et al. [11].

The main study mode in *Foundations, 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 (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.

5.2 Participants

The students sampled in this study came from a large sample of *Foundations* students in the 2016-17 academic year, who had written messages for the Genie in the email system. The dates sampled were from August 1, 2016 to June 17, 2017. There were a total of 34,602 such students. The students were from 462 different schools located in 99 different districts, most of which were located in Texas. This analysis samples students in 4th-5th grades because their writing skills are developed enough to be captured by NLP tools. We also included only those students that had completed the post-test survey (discussed in the next sub-section) and those students that had attempted A Level problems. This subset of the data consisted of 970 students.

5.3 Survey Data

The measures used in the present study consisted of three 4-point scales adapted from [47] and administered at the start/end of the 2016/2017 school year. The first was *mathematics self-concept*, which comprised five items that captured the degree to which the student see themselves as a “math person” (e.g., “I have always been good at math”). The second was *interest in mathematics*, which consisted of three items that capture intrinsic curiosity or enjoyment of mathematics (e.g., “How much do you like math?”). The last scale measured *value of mathematics* and consisted of five items that captured the degree to which students find math to be useful (e.g., “How important is it to you to get good grades in math class?”). The Cronbach α of these scales were 0.72, 0.69, and 0.72, respectively.

5.4 Final Corpus

Our language sample for this analysis consisted of messages sent from the students to the Genie. Because many messages contained few words, we aggregated all e-mails 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 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 two of these 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.

Finally, we removed data from students who had produced fewer than 150 words in writing to the Genie (calculated after cleaning). This cut-off ensures that students produced a large enough language sample to provide a clear representation of their linguistic ability including bag-of-word assumptions for Latent Dirichlet Allocation (LDA) analyses. This left us with data from 351 students for analyses.

5.5 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) [48], the Tool for the Automatic Analysis of Cohesion (TAACO) [49], the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) [50], and the SEntiment ANalysis and Cognition Engine (SEANCE) [51]. In addition, we developed specific indices related to topics commonly discussed with the Genie e-mail 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.

5.5.1 TAALES

TAALES reports on a number of indices related to basic lexical information (e.g., the number of tokens, and types), lexical frequency, lexical range, lexical registers, word information features (e.g., concreteness, meaningfulness, polysemy [the number of senses a word has]), and psycholinguistic variables. For instance, the tool uses the Kucera-Francis corpus to compute the number of registers (e.g., humor academic, or fiction registers) that words occur in (a measure of register specificity). The tool also reports on a number of phonological, orthographic, and phonographic neighborhood effects that calculate how many near neighbors based on sound or spelling that a word has. TAALES also reports on variables that measure how long a word takes to name, how accurately words are pronounced, and how many senses a word contains (i.e., polysemy).

5.5.2 TAACO

TAACO incorporates a variety of classic and recently developed indices related to text cohesion. For a number of indices, the tool incorporates the Stanford part of speech (POS) tagger [52] and synonym sets from the WordNet lexical database [53]. 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, sentence overlap indices that assess local cohesion, paragraph overlap indices that assess global cohesion, and a variety of connective indices such as logical connectives (e.g., *also, next, so*) and sentence linking connectives (e.g., *but, if, then*).

5.5.3 TAASSC

TAASSC measures large and fined grained clausal and phrasal indices of syntactic complexity and usage-based frequency/contingency indices of syntactic sophistication. TAASSC includes indices measured by Lu's [54] Syntactic Complexity Analyzer (SCA) and a number of pre-developed fine-grained indices or clausal complexity and phrasal complexity, The SCA measures are classic measures of syntax based on t-unit analyses [19] where t-units are defined as a dominant and subordinate clause. For instance, SCA measures the number of complex t-units in a text (i.e., T-units that includes both an independent and a dependent clause). The fine-grained clausal indices calculate the average number of particular structures per clause and dependents per clause. The fine-grained phrasal indices measure noun phrase types and phrasal dependent types.

5.5.4 SEANCE

SEANCE is a sentiment analysis tool that relies on a number of pre-existing sentiment, social positioning, and cognition dictionaries. SEANCE contains a number of pre-developed word vectors that 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. Examples of affective variables reports by SEANCE include positive and negative polarity metrics, terms related to arousal (as compared to calmness), and respect terms. Cognition examples include words related to socially defined ways of doing work, acts and methods to accomplish goals, time and space, and quantity.

5.5.5 Latent Dirichlet Allocation (LDA) features

We developed measures of domain topicality for the messages found in the corpus using LDA. LDA is a computational modeling technique used to infer underlying topics through a generative probabilistic process. We conducted an LDA analysis on the entire corpus of student messages to the Genie and fit 200 topics to the data - the optimal number of topics was inferred using Hierarchical Dirichlet processes [55]. Using these latent topics, each word is perceived as a probability distribution across all topics; if irrelevant for a topic, the corresponding weight is 0, whereas more relevant topics for a given word have higher probabilities. These word weights were then used to create topic distributions for each student in order to identify how strongly student language overlapped with topics covered in the entire Genie message corpus.

5.6 Statistical Analysis

We first calculated correlations between the students' accuracy on A Level problems and their survey scores for Math Identity (self concept, interest, and value). These relationships allow us to better understand how basic math skills interacted with student survey responses for Math Identity.

We followed this up by calculating linear models to assess the degree to which linguistic features in the students' emails to the Genie, along with other behaviors (e.g., question/note posted, questions answered, site visits) were predictive of students' math skills and their self-reported Math Identity. As part of this analysis, we first checked that all variables were normally distributed. For the linguistic variables, we tested only those variables that showed at least a small effect size ($r > .100$) with the response variable. We also controlled for multicollinearity between all the linguistic and non-linguistic variables ($r \geq .700$) such that if two or more variables were highly similar, all but one of the variables (the one with the strongest relationship with the response variable) were removed from the analysis.

We cross-validated our results by dividing data into training and test sets based on a 67/33 split. We used stepwise linear models on the training set to find the best fitting models for each analysis. After model selection, coefficients were checked for suppression and visual inspection of residuals distribution for non-standardized variables was conducted. To obtain a measure of effect sizes, we computed correlations between the fitted and observed values, resulting in an overall R^2 value for the fixed factors in the training set. The model from the training set was used to derive an r and R^2 value for the test data.

6. RESULTS

6.1 Correlations

Pearson correlations were conducted among the response variables to assess links between Math Identity and math scores. The results, reported in Table 1, indicate that all three Math Identity variables were positively and significantly correlated with performance on A level math problems. Medium effects were found for self-concept. Weak effects were found for interest and value. None of the Math Identity variables were strongly associated with one another (i.e., $r < .500$), although correlations with interest approached that threshold for both self-concept ($r = .489$) and value ($r = .491$).

Table 1. Correlations between response variables

Variable	Self-concept	Interest	Value
A level score	0.341**	0.205**	0.145*
Self Concept		0.489**	0.309**
Interest			0.491**

Note * $p < .010$, **p < .001

6.2 Linear Model for Self-Concept

A linear model to predict students' self-concept including linguistic, affect, and click-stream variables yielded a significant model, $F(5, 242) = 2.861$, $p < .001$, $r = .356$, $r^2 = .127$ (see Table 2 for details). Two linguistic variables: *Phonographic neighbors, function words* and *word naming accuracy, function words* were significant predictors as were three affective variables: *Methods and goals words, words related to work*, and *polarity verbs*. No click-stream variables were significant predictors. The combination of the five variables accounted for 13% of the variance in the students' self-concept scores. Cross-validating the model on the test set yielded $r = .371$, $r^2 = .138$, demonstrating that the combination of the five variables accounted for 14% of the variance in the student samples comprising the test set.

Table 2. Linguistic model for predicting self-concept scores

Fixed Effect	Coefficient	Std. Error	t
(Intercept)	61.518	21.309	2.887**
Phonographic Neighbors: Function words	-0.284	0.081	-3.512***
Acts and methods terms to accomplish goals	9.441	3.113	3.033**
Words related to work	-6.609	2.342	-2.822**
Polarity verbs	0.247	0.087	2.857**
Word naming accuracy: Function words	-57.807	21.413	-2.700**

Note * $p < .050$, ** $p < .010$, ***p < .001

6.3 Linear Model for Interest

A linear model using linguistic and click-stream variables to predict students' interest yielded a significant model, $F(4, 218) = 4.943$, $p < .001$, $r = .419$, $r^2 = .176$ (see Table 3 for details).. Four affective variables were significant predictors in the model: *Hu Liu negative terms, power words, arousal ratings, and words related to methods and goal*. No click-stream variables were significant predictors. The combination of the four variables accounted for 17% of the variance in the students' interest scores. Using the model from the training set on the samples in the test set yielded $r = .360$, $r^2 = .130$, demonstrating that the combination of the four variables accounted for 13% of the variance in the student samples comprising the test set.

Table 3. Linguistic model for predicting interest scores

Fixed Effect	Coefficient	Std. Error	t
(Intercept)	3.523	0.137	25.708***
Hu Liu negative terms	-0.928	0.201	-4.612***
Power words	-8.440	3.335	-2.531**
Arousal ratings	-9.407	3.336	-2.820**
Acts and methods terms to accomplish goals	8.056	2.951	2.730**

Note * $p < .050$, ** $p < .010$, ***p < .001

6.4 Linear Model for Value

A linear model to predict students' math value using linguistic and click-stream variables yielded a significant model, $F(3, 217) = 7.843$, $p < .001$, $r = .313$, $r^2 = .098$ (see Table 4 for details).. Three variables were significant predictors in the model: polarity verbs component score (verbs related to polarity, aptitude, and pleasantness), time and space terms, and words related to respect. No click-stream variables were significant predictors. The combination of the three affect variables accounted for 10% of the variance in the students' math value scores. Using the model from the training set on the samples in the test set yielded $r = .303$, $r^2 = .091$, demonstrating that the combination of the five variables accounted for 9% of the variance in the student samples comprising the test set.

Table 4. Linguistic model for predicting value scores

Fixed Effect	Coefficient	Std. Error	t
(Intercept)	3.301	0.082	40.254**
Polarity verbs	0.15	0.048	3.107**
Time/space terms	2.932	1.048	2.799**
Respect words	4.776	2.119	2.254*

Note * $p < .050$, ** $p < .010$, ***p < .001

6.5 Linear Model for Math Success

A linear model to predict math success including linguistic and click-stream variables yielded a significant model, $F(5, 217) = 9.130$, $p < .001$, $r = .417$, $r^2 = .174$ (see Table 5 for details).. Five linguistic variables were significant predictors in the model: *Kucera-Francis categories, phonological neighbors distances, complex t-units, polysemy (adverbs), and quantitative terms*. No click-stream variables were significant predictors. The combination of the five variables accounted for 17% of the variance in the students A level math scores. Using the model from the training set on the samples in the test set yielded $r = .378$, $r^2 = .143$, indicating that the combination of the five variables accounted for 14% of the variance in the student samples comprising the test set.

Table 5. Linguistic model for predicting math scores

Fixed Effect	Coefficient	Std. Error	T
(Intercept)	33.544	15.331	3.508***
Kucera-Francis categories	2.721	0.776	2.12*
Phonological neighbor Levenshtein distances	15.225	7.18	-2.701**
Complex T-units	-5.256	1.946	-3.019**
Polysemy (adverbs)	-1.212	0.401	2.348**
Quantitative terms	62.983	26.82	3.508**

Note * $p < .050$, ** $p < .010$, ***p < .001

7. DISCUSSION AND CONCLUSION

Investigating the degree to which students identify with math (e.g., their Math Identity) can provide important information related to student-level differences which in turn could allow for personalization efforts within educational settings. The purpose of this study was to examine links between students' self-reported Math Identity (e.g., math self-concept, value, and interest) and language features found in student e-mails within an on-line math tutoring system. The study also examined links between student math scores and self-reported Math Identity and between math scores and language features. Overall, we find weak to medium relationships between Math Identity variables and math scores. Additionally, language features were able to explain a significant amount of variance for each Math Identity variable and for student math scores. These findings are discussed below along with implications for better understanding Math Identity and developing pedagogical interventions within Reasoning Mind's *Foundation* system.

Our first analysis examined links between A level math scores within the *Foundations* system and student's self-reported Math Identity variables (self concept, interest and value). All of the Math Identity variables were positively correlated with each other as well as with the math-performance metric, although this effect was stronger for self-concept than for interest or value. The correlation matrix in Table 1 provides evidence that the Math Identity variables self-reported by the students were related to math ability within the system.

Our next goal was to investigate if linguistic models could be developed for each of the Math Identity variables. Specifically, we were interested in examining links between the words and language structures produced by the student in their e-mails to the Genie and their self-ratings of self-concept, interest, and value. Our model of student ratings for self-concept explained 14% of the variance in the test set (r = .371). The model was informed by five language features. Three sentiment and cognition features were reported by SEANCE while two features related to lexical sophistication were reported by TAALES. Polarity verbs were again positively related to a math identify variable indicating that students who used more positive verbs reported higher math self-concept. Additionally, students who produced more words related to accomplishing goals (e.g., *build, make,* and *formulate*) reported higher self-concept. Conversely, words related to ways of doing work were negatively associated with self-concept. This may be an effect of the word *grade,* which is included in this category and was common in the e-mails (i.e., students worried about low grades). Two lexical indices for function words were also negatively predictive of self-concept scores: phonographic neighbors and word naming accuracy. These findings suggest that students with higher self-concept produced function words that had fewer neighbors and lower word naming accuracy. In both cases, the results indicate that students producing more sophisticated function words had greater self-concept.

Our model for math interest explained 13% of the variance in the test set (r = .360) and included only sentiment and cognition variables reported by SEANCE. These variables indicate that students with greater math interest used fewer negative terms, fewer words related to arousal (i.e., more words related to calmness), and more words related to acts and methods to accomplish goals, which was also a predictor of self-concept scores. Lastly, words related to power yielded a negative co-efficient with math interest scores. This finding suggests that students that use power words (e.g., *force* and *command*) have lower interest in math.

With respect to students' ratings of their math value, language features were able to predict about 9% of the variance in student test set ratings. (r = .303). Three features were positive predictors of value: polarity verbs, time/space terms, and respect terms. All variables were reported by SEANCE and were related to either sentiment or cognition. The results show that students that reported higher math value produced language in their e-mails that included more positive verbs and showed greater respect through the use of terms such as *honor, admire,* and *respect.* In addition, these students produced more words related to time and space. Time words include prepositions such as *across* and *above* but also space verbs that may be related to math concepts including *circle, curve,* and *distance*.

Finally, we developed a model to predict math success (i.e., scores on A Level problems). This model explained 14% of the variance in math scores (r = .378) using lexical features, a measure of syntactic complexity, and a measure of cognition. The three lexical indices included the number of registers in which a word occurs, phonological neighbors based on Levenshtein distances (i.e., words that words that require more substitutions, insertions, or deletion operations to transform that word into its closest phonologic neighbors), and the polysemy value of adverbs. The first index suggests that students with high math scores produced words that were found across a variety of registers. The second and third indices indicate that students with higher math scores produced more sophisticated language (i.e., adverbs with fewer senses and words that required more operations to find a phonological neighbor). Students with higher math scores also produced fewer complex sentences (sentences with an independent and dependent clause) and used more quantitative words.

Overall, the findings suggest that language variables related to sentiment and cognition can explain a significant amount of the variance in a number of self-reported survey variables related to math self-concept, interest, and value. These variables have the potential to not only better explain the constructs of Math Identity, but also have the potential to be useful for student interventions.

The findings from this study indicate that students who produce more positive language e-mails within the *Foundations* system are more likely to have a positive Math Identity. Conversely, those that use more negative language are more likely to have lower Math Identity. However, it is not just positive and negative terms that are related to Math Identity. Students with stronger Math Identity use more respectful language, less power-related language, and language that is more calm. Lastly, students with stronger Math Identity were more likely to use more sophisticated words or words related to accomplishing goals.

The findings from this study also suggest little overlap between the language features that predict Math Identity and those that predict math success even though we see links between our Math Identity variables and math success within the system. While there are some similarities between self-concept scores and math scores with respect to phonological neighbors, these features differ in their parts of speech (content versus function words). In general, most predictors of math success are related to linguistic features (lexical, syntactic, and cohesion features) while predictors of Math Identity are related to sentiment and cognition features. In total, these sentiment and cognition features provide a profile of students within the system that have high math interest.

Using the models reported here, a number of different interventions could be developed for students identified as likely having low math interest. These interventions could be as simple as having the Genie send an e-mail to students that provides statistics on their successes within the system, their perseverance in answering problems, or simply the number of problems they have attempted or accurately solved over a specific time period. Students could also be asked to correspond with the Genie using metacognitive strategies related to self-assessment and goal-setting activities, as this corresponds with both the interest models we developed here and with long-standing interventions designed to support self-efficacy and interest (cf. [21]). Interventions such as these may assist students in more critically thinking about themselves in relation to math and in better understanding their math knowledge and acquisition.

While the Math Identity profiles developed should be strong enough to drive interventions, the models report only medium effect sizes. Thus, much variance remains to be identified within the existing survey data. Some of that variance may emerge in language features that are not yet captured by NLP tools, while other variance may be related to demographic or other click-stream data available within the system such as the number of messages sent and received by the students within the e-mail system, hours spent on-line within the tutoring system, and number of objectives met within the system. Thus, the findings here should be seen as preliminary with implications for future development.

8. ACKNOWLEDGEMENTS

The authors are indebted to Victor Kostyuk for his help in organizing the data here. In addition, the authors than Stefan Slater for helping with final touches. This research was supported in part by the National Science Foundation (DRL- 1418378). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

9. REFERENCES

[1] Corbett, A.T., McLaughlin, M.S., and Scarpinatto, K.C. 2000. Modeling student knowledge: cognitive tutors in high school and college. *User Modeling and User-Adapted Interaction* 10 (Jun. 2000), 81-108. DOI= https://doi.org/10.1023/A:102650562

[2] Romero, C., Ventura, S., Delgado, J. A., and De Bra, P. 2007. Personalized links recommendation based on data mining in adaptive educational hypermedia systems. In *European Conference on Technology Enhanced Learning*, EC-TEL 2007. Springer, Berlin, Heidelberg, 292-306.

[3] Baker, R.. and Ocumpaugh, J. 2014. Interaction-Based Affect Detection in Educational Software. In *The Oxford Handbook of Affective Computing,* R. Calvo, S. D'Mello, J. Gratch, A. Kappas, Eds. Oxford U. Press, Oxford, UK.

[4] Mcquiggan, S. W., Mott, B. W., and Lester, J. C. 2008. Modeling self-efficacy in intelligent tutoring systems: An inductive approach. *User Modeling and User-Adapted Interaction* 18, 1-2, 81-123. DOI= https://doi.org/10.1007/s11257-007-9040-y

[5] Cooper, D. G., Arroyo, I., Woolf, B. P., Muldner, K., Burleson, W., and Christopherson, R. 2009. Sensors model student self concept in the classroom. In *International Conference on User Modeling, Adaptation, and Personalization*, 30-41.

[6] Winne, P. H. and Baker, R. S. 2013. The potentials of educational data mining for researching metacognition, motivation and self-regulated learning. *Journal of Educational Data Mining* 51 (May 2013), 1-8.

[7] Pennebaker, J. W. and King, L. A. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology* 77, 6 (Dec. 1999), 1296-1312.

[8] Pennebaker, J. W. and Graybeal, A. 2001. Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science* 103 (Jun. 2001), 90-93. DOI= https://doi.org/10.1111/1467-8721.00123

[9] Schlenker, B. and Weigold, M. 1989. Goals and the self-identification process: Constructing desired identities. In *Goal Concepts in Personality and Social Psychology,* Pervin, Ed.. Lawrence Erlbaum Assoc., Hillside, NJ, 243-89.

[10] Nosek, B., Banaji, M., and Greenwald, A. 2002. Math= male, me= female, therefore math≠ me. *Journal of Personality and Social Psycholog*y 83, 1 (Jul. 2002), 44-59.

[11] Khachatryan, G., Romashov, A., Khachatryan, A., Gaudino, S., Khachatryan, J., Guarian, K., and Yufa, N. 2014. Reasoning Mind Genie 2: An intelligent tutoring system as a vehicle for international transfer of instructional methods in mathematics. *International J. A.I. Ed*. 243, 3 (Sep. 2014), 333-382. DOI= https://doi.org/10.1007/s40593-014-0019-7

[12] Ryan, K. and Ryan, A. 2005. Psychological processes underlying stereotype threat and standardized math test performance. *Edu'nal Psychologist* 40, 1 (Jun. 2010), 53-63. DOI= https://doi.org/10.1207/s15326985ep4001_4

[13] MacGregor, M. and Price, E. 1999. An exploration of aspects of language proficiency and algebra learning. *Journal for Research in Math Education* 30, 4 (Jul. 1999), 449–467. doi: 10.2307/749709

[14] Vukovic, R. K. and Lesaux, N.K. 2013. The relationship between linguistic skills and arithmetic knowledge. *Learning and Individual Differences* 23 (Feb. 2013), 87-91. DOI= https://doi.org/10.1016/j.lindif.2012.10.007

[15] 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 Thesis.

[16] LeFevre J., Fast, L., Skwarchuk, S., Smith-Chant, B., Bisanz, J., Kamawar, D., and Penner-Wilger, M. 2010. Pathways to math: Longitudinal predictors of performance. *Child Development* 81, 6 (Nov. 2010), 1753–1767. DOI= 10.1111/j.1467-8624.2010.01508.x

[17] Crossley, S. A., Liu, R., and McNamara, D. S. 2017. Predicting math performance using natural language processing tools. *Proceedings of the 7th International Learning Analytics and Knowledge LAK Conference*. LAK'17. ACM, New York, NY, 339-347.

[18] Crossley, S.A., Barnes, T., Lynch, C., and McNamara, D.S. 2017. Linking language to math success in a blended course. In *Proceedings of the 10th International Conference on Educational Data Mining* (Wuhan, China), Hu, X., Barnes, T., Hershkovitz, A., and Paquette, L. Eds. 180-185

[19] Crossley, S. A. and Kostyuk, V. 2017. Letting the Genie out of the Lamp: Using Natural Language Processing tools to predict math performance. In *Language, Data, and Knowledge LDK 2017,* Gracia J., Bond F., McCrae J., Buitelaar P., Chiarcos C., and Hellmann S. Eds. In. *Lecture Notes in Computer Science*, vol 10318. Springer, Cham, Switzerland.

[20] Syed, M., Azmitia, M., and Cooper, C. 2011. Identity and academic success among underrepresented ethnic minorities: An interdisciplinary review and integration. *Journal of Social Issues* 67, 3 (Sep. 2011), 442-468. DOI= 10.1111/j.1540-4560.2011.01709.x

[21] Bandura, A. 1977. Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review* 84, 2, 191-215. DOI= http://dx.doi.org/10.1037/0033-295X.84.2.191

[22] Bem, S. 1974. The measurement of psychological androgyny. *J. of Consulting and Clinical psychology* 42, 2, 155-162. DOI= http://dx.doi.org/10.1037/h0036215

[23] Hitlin, S. 2003. Values as the core of personal identity: Drawing links between two theories of self. *Social Psychology Quarterly* 66, 2 (Jun. 2003), 118-137. DOI= 10.2307/1519843

[24] Erikson, E. 1968. *Youth: Identity and Crisis*. Norton & Company, *New York, NY*.

[25] Syed, M., Azmitia, M., & Cooper, C. R. 2011. Identity and academic success among underrepresented ethnic minorities: An interdisciplinary review and integration. *Journal of Social Issues, 67*(3), 442-468.

[26] San Pedro, M.O.Z., Baker, R.S.J.d., Bowers, A., Heffernan, N. 2013. Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. In *Proc. 6th International Conf. on Educational Data Mining*, 177-184.

[27] San Pedro, M.O.Z., Ocumpaugh, J., Baker, R., Heffernan, N. 2014. Predicting STEM and Non-STEM College Major Enrollment from Middle School Interaction with Mathematics Educational Software. In *Proc. 7th International Conf. on Educational Data Mining*, 276-279.

[28] Landers, M. 2013. Buying in and checking out: Identity development and meaning making in the practice of mathematics homework. *Qualitative Research in Ed.* 2, 2, 130-160.

[29] Cadinu, M. and Galdi, S. 2012. Gender differences in implicit gender self-categorization lead to stronger gender self-stereotyping by women than by men. *European Journal of Social Psychology* 4, 25 (Apr. 2012), 546-551. DOI= 10.1002/ejsp.1881

[30] Keller, J. 2007. Stereotype threat in classroom settings: The interactive effect of domain identification, task difficulty and stereotype threat on female students' maths performance. *British J. of Ed. Psych.* 77, 2 (Jun/ 2017), 323-338. DOI= 10.1348/000709906X113662

[31] Steele, C. 1997. A threat in the air. How stereotypes shape intellectual identity and performance. *Am. Psychologist* 52, 6, 613-629.

[32] Eccles, J. 2009. Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. *Educational Psychologist* 44, 2 (Apr. 2009), 78-89. DOI= https://doi.org/10.1080/00461520902832368

[33] Bong, M. and Skaalvik, E. M. 2003. Academic self-concept and self-efficacy: How different are they really?. *Educational Psychology Review* 15, 1 (Mar. 2003), 1-40. DOI= https://doi.org/10.1023/A:1021302408382

[34] Pajares, F. and Miller, M. D. 1994. Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal Of Educational Psychology* 86, 2, 193-203.

[35] Epstein, S. 1973. The self-concept revisited: Or a theory of a theory. *American Psychologist* 28, 5, 404-416. DOI= http://dx.doi.org/10.1037/h0034679

[36] Shavelson, R. and Bolus, R. 1982. Self concept: The interplay of theory and methods. *J. Educational Psychology* 74, 1, 3-17. DOI= http://dx.doi.org/10.1037/0022-0663.74.1.3

[37] Hidi, S. and Renninger, K. 2006. The four-phase model of interest development. *Ed. Psychologist* 41, 2 (Jun. 2010), 111-127. DOI= https://doi.org/10.1207/s15326985ep4102_4

[38] Bandura, A. and Schunk, D. 1981. Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *J. of Personality and Social Psych.* 41, 3, 586-598. DOI= http://dx.doi.org/10.1037/0022-3514.41.3.586

[39] Sansone, C., Weir, C., Harpster, L., and Morgan, C. 1992. Once a boring task always a boring task? Interest as a self-regulatory mechanism. *J. of Personality and Social Psych.* 63, 3 (Sep. 1992), 379-390.

[40] Roberts, B. and DelVecchio, W. 2000. The rank-order consistency of personality from childhood to old age: A quantitative rev. of longitudinal studies. *Psych. Bulletin* 126, 1 (Jan. 2000), 3-25.

[41] Campbell, N. and Hackett, G. 1986. The effects of mathematics task performance on math self-efficacy and task interest. *J. of Vocational Behavior* 28, 2, 149-162. DOI= https://doi.org/10.1016/0001-8791(86)90048-5

[42] Fink, R. P. 1998. Interest, gender, and literacy development in successful dyslexics. In L. Hoffmann, A. Krapp, K. A. Renninger, & J. Baumert (Eds.), *Interest and learning: Proceedings of the Seeon Conference on interest and gender* (pp. 402–407). Kiel, Germany: IPN.

[43] Prenzel, M. 1992. The selective persistence of interest. In *The Role of Interest in Learning and Development,* Renninger, K. A.. S. Hidi, and A. Krapp, Eds. Lawrence Erlbaum, Hillsdale, NJ, 71–98.

[44] Chouinard, R., Karsenti, T., and Roy, N. 2007. Relations among competence beliefs, utility value, achievement goals, and effort in mathematics. *British Journal of Educational Psychology* 77, 3 (Sep. 2007), 501-517. DOI= 10.1348/000709906X133589

[45] Harackiewicz, J., Rozek, C., Hulleman, C., and Hyde, J. 2012. Helping parents to motivate adolescents in mathematics and science: An experimental test of a utility-value intervention. *Psychological Science* 23, 8 (Jul. 2012), 899-906. DOI= https://doi.org/10.1177/0956797611435530

[46] Miller, W., Baker, R., Labrum, M., Petsche, K., Liu, Y-H., and Wagner, A. 2015. Automated Detection of Proactive Remediation by Teachers in Reasoning Mind Classrooms. In *Proc. 5th International Learning Analytics and Knowledge Conf*. (Poughkeepsie, NY, March 16 - 20, 2015). AMC, New York, NY, 290-294.

[47] Mingle, L. (2013). *Threats to success in mathematics: examining the combined effects of choking under pressure and stereotype threat* (Doctoral dissertation, University of Illinois at Urbana-Champaign).

[48] Kyle, K., Crossley, S. A., and Berger, C. 2017. The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior research methods,* 1-17. DOI= https://doi.org/10.3758/s13428-017-0924-4

[49] Crossley, S.A., Kyle, K. and McNamara D.S. 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods* 28, 4 (Sep. 2015), 1227-1237. DOI= https://doi.org/10.3758/s13428-015-0651-7.

[50] Kyle, K. and Crossley, S. A. 2017. Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing* 34, 4 (Sep. 2017), 513–535. DOI= https://doi.org/10.1177/0265532217712554

[51] .Crossley, S.A., Kyle, K. and McNamara D.S. 2016. Sentiment Analysis and Social Cognition Engine (SEANCE):

An automatic tool for sentiment, social cognition, and social order analysis. *Behavior Research Methods* 49, 3 (Jun. 2017), 803-821. DOI= https://doi.org/10.3758/s13428-016-0743-z.

[52] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. ACL, Baltimore, MA, 55–60

[53] Miller, G. A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38, 11 (Nov. 1995), 39–41. DOI= 10.1145/219717.219748

[54] Lu, X. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15, 4, 474-496.

[55] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. 2006. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association* 101, 1566–1581.