DETERMINING THE DIFFICULTY OF MATH PROBLEMS FROM SURFACE LINGUISTIC FEATURES

by

Xin Ju

Submitted in Partial Fulfillment

Of the requirements for a Degree with Honors

(Mathematical Sciences)

Helen Hardin Honors Program

University of Memphis

May 2020

ABSTRACT

Ju, Xin. Bachelor of Science in Mathematics. University of Memphis, December 2019. Determining the Difficulty of Math Problems From Surface Linguistic Features. Major Professor, Alistair Windsor.

Investigating sources of difficulty in mathematical word-problem solving is crucial to improving student performance on word problems. The goal of this ecological study was to examine results of mathematics competition word problems in order to identify linguistic features contributing to problem difficulty, and to consequently provide insight into areas of mathematics that students find the most challenging. Data used for this study consisted of a set of math competition word problems, the percentage of students who answered each question correctly, and expert classification of problems based on strategy. Questions were cleaned, represented as vectors of weighted term frequencies, and clustered. Multiple logistic regression analyses were conducted on each of the following sets of problem difficulty predictors: linguistic clusters ($R^2 = 0.22$), expert classifications ($R^2 = 0.28$), and weighted features ($R^2 = 0.32$). Results from our best model indicated that students tended to experience the most difficulty with problems involving comparison terms, area, and units of measurement.

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

Math tends to be a difficult subject for students today, as evidenced by underwhelming improvement of student test scores during the past few years. According to recent reports from the National Assessment of Educational Progress (NAEP), average math scores of participating students showed a significant drop in 2015, the first in 25 years (NW, Suite 800Washington, & Inquiries, n.d.). Most concerningly, no significant improvements were reported by 2017 (“NAEP Mathematics: National Achievement-Level Results,” n.d.-a, n.d.-b). Average math proficiency for 4th and 8th graders in the U.S. remains consistently low, with most recent results indicating only 40 percent of 4th graders and 34 percent of 8th graders performing at or above *Proficient* level (“NAEP Mathematics: National Achievement-Level Results,” n.d.-a, n.d.-b).

It then comes as no surprise that students today experience particular difficulty when dealing with math word-problems, a genre of questions which require not just mathematical knowledge, but also the ability to accurately and appropriately apply it to representations of real-world situations (Corte, Verschaffel, & Greer, 2000). In fact, math word-problems have been described as belonging to “the most difficult and complex problem types that pupils encounter during their elementary-level mathematical development” (Daroczy, Wolska, Meurers, & Nuerk, 2015, p.1). This belief has been emphasized by numerous other studies (Carpenter & And Others, 1980; Cummins, Kintsch, Reusser, & Weimer, 1988; Geary, 1994; Hegarty, Mayer, & Green, 1992). Cummins et al. (1988) state that “word problems are notoriously difficult to solve” (p.405), and found that performance on questions presented in numeric format was significantly higher than that on questions presented as word-problems for first-grade students. In addition, as reported by Carpenter et al. (1980), children across the nation "perform 10 to 30% worse on arithmetic word problems than on comparable problems presented in numeric format" (Cummins et al., 1988, p.405).

This glaring issue has drawn the attention of educators and researchers alike, and many possible theories for it have been offered. One widely accepted explanation, or contributor, to word-problem difficulty described by Schoenfeld (1991) speculates that many students have disconnected the process of solving word-problems from the real world, and that this mindset is a result of, rather than alleviated by their schooling. Students appear to approach word-problems with a "calculational orientation" (Hoogland, de Koning, Bakker, Pepin, & Gravemeijer, 2018, p. 123), while falsely assuming "all problems to be reasonable" and "to have answers derivable from the data given in them" (Schoenfeld, 1991, p.320). Thus they often haphazardly apply arithmetic operations and procedures to figures presented in the questions without thought as to whether or not those methods are appropriate in the presented situations, resulting in nonsensical answers (Schoenfeld, 1991).

For instance, Schoenfeld (1991) presents the following problem, first posed to first and second grade students by Kurt Reusser in his 1986 research: "There are 26 sheep and 10 goats on a ship. How old is the captain?" (p.315). It may seem fairly obvious that neither the number of sheep nor the number of goats aboard the ship should have any relation to the age of the captain; yet 76 out of the 97 student participants were able to "solve" the problem by adding 26 and 10, resulting in a nonsensical answer (Schoenfeld, 1991). Schoenfeld (1991) describes this phenomenon as “suspension of sense-making” (p.316) where students “suspend the requirement that the problem statements make sense” (p.316).

Although sources for word-problem difficulty are still under investigation by researchers, the crucial role that word-problems play in today’s mathematics curriculum is unquestionable. Various sources suggest that math word-problems help to bridge the gap between, as well as to assess, students’ computational skills and their ability to apply them to real-world situations (Corte et al., 2000; Geary, 1994; Hoogland et al., 2018). As a result, they are strongly emphasized by standardized tests across the nation (like the NAEP) and are deemed to be the “best predictor of adult employment and wages” (Wang, Fuchs, & Fuchs, 2016, p.79), making them critical to students’ classroom and occupational success.

However, improving student performance on math word-problems requires that we first identify and address specific factors contributing to student difficulties. Much of the early research on student word-problem solving involved isolating certain features of word-problems and observing their impact on problem difficulty (Geary, 1994, p.96). Although “not directly related to arithmetic” (Daroczy et al., 2015, p.1), the category of linguistic features, or factors, is considered to be relevant to the study of word-problems (Acosta-Tello, 2010, p.16), and have been shown to contribute to problem difficulty in previous studies (Daroczy et al., 2015; Geary, 1994, p.96-104; Jerman & Mirman, 1974; Nathan & Koedinger, 2000, p.169; Powell, Fuchs, Fuchs, Cirino, & Fletcher, 2009). Past research pertaining to the effects of linguistic features on word-problem difficulty is well-documented, and will be discussed in greater detail in the literature review.

No prior work, however, was found specifically examining the role of surface linguistic features as predictors of word-problem difficulty using an ecological approach. Surface linguistic features can be roughly defined as “any feature of text that can be easily extracted without a serious analysis of the text” (Lee, Song, & Rim, 2007, p.111). Examples may include length of the text, number of sentences, or average word length; but in this case, we are referring only to the words present in the math problems themselves.

To fill this gap in research, the present study seeks to identify surface linguistic features influencing student performance from math competition word-problems not specifically designed for analysis purposes. We then examine each feature’s relationship to word-problem difficulty, in the hopes that they may provide insight into the areas of mathematics that elementary and middle school students are struggling with the most. Furthermore, we aim to evaluate the accuracy of these surface linguistic features as predictors of student performance with the goal of creating an effective model for predicting problem-difficulty in advance.

LITERATURE REVIEW

Much of the research on children's word-problem solving difficulties can be broadly classified into two groups: (1) studies focusing on "identifying the features of the problems themselves that made word problems more difficult than number problems," and (2) those focusing on "how the child understands and conceptually represents word problems and how this affects the child's problem-solving processes" (Geary, 1994, p.96). The assumption underlying the first approach is that the structure of word-problems can, to a large degree, be used to determine the difficulty level of the questions (Lepik, 1990, p.83). Studies from category one can be further broken down into (a) those examining linguistic features, which describe the textual information provided by the problem, and (b) those examining mathematical, or computational, features, which describe standard solutions to the problems (Daroczy et al., 2015, p. 2; Lepik, 1990, p.83).

The current study's subject pertains more closely to the former feature type, so this review will primarily focus on previous research examining the effects of linguistic features on word-problem difficulty. In the following sections, we seek to define a few types of the commonly observed linguistic features. Word-problem readability will be considered as a potential predictor for student problem-solving accuracy, and text analysis, as it relates to the current study, will be discussed.

Linguistic Sources of Word-Problem Difficulty

The primary and most obvious distinguishing factor between simple arithmetic problems and math word-problems is the "inclusion of linguistic information" (Wang et al., 2016, p.3). Therefore, it makes sense to hypothesize that many linguistic features of word-problems could have a marked influence on problem difficulty. Prior studies supporting this claim have investigated numerous linguistic features and their effects on student performance, several of them also taking into account students' individual and social factors, like learning or developmental disabilities (Daroczy et al., 2015, p.2).

In their 2015 review, Daroczy and colleagues compiled a comprehensive overview of previous studies addressing linguistic and numerical sources of word-problem difficulty. According to their report, linguistic features contributing to word-problem difficulty fall into one of two subcategories: (1) structural features or (2) semantic features (Daroczy et al., 2015, p.2). Features corresponding to these subcategories will be discussed in detail in the subsequent sections of this literature review.

Structural Linguistic Features

Structural linguistic features can be roughly categorized into (a) "quantitative properties" of the question's text (e.g., number of characters, words, sentence length, proportion of complex words); (b) vocabulary "(e.g., polysemous words, prepositional phrases, passive voice, clause structure)"; and (c) question wording or placing (Daroczy et al., 2015, p.2). These features have been thoroughly investigated by earlier studies, several of which approached them as predictor variables for problem difficulty (M. E. Jerman & Mirman, 1974; M. Jerman & Rees, 1972; Lepik, 1990).

In a nearly comprehensive study of linguistic and computational variables impacting word-problem difficulty, Jerman and Mirman (1974) attempted to identify a set of features which would improve a linear regression model's predictions for proportion of students who solve a set of arithmetic word-problems correctly. Features investigated included 73 linguistic factors grouped into 7 categories: (1) Length, (2) Parts of Speech,

(3) Words, (4) Numbers, (5) Sentences, (6) Parts of Sentences, and (7) Punctuation, Symbols and characters (M. E. Jerman & Mirman, 1974). A close examination of these features reveals that all relate to quantitative properties of the text and vocabulary.

Two sets of exercises, prepared at different levels of computational difficulty, were administered to students of mixed abilities from grades 4-9; and analyses were performed using stepwise linear regression (M. E. Jerman & Mirman, 1974). No set of features were reported as both significant and accounting for a large proportion of the variance in the response for all grade levels (M. E. Jerman & Mirman, 1974). However, Jerman and Mirman (1974) discovered that two linguistic variables consistently showed up as statistically significant predictors of problem difficulty, namely the number of mathematical terms in the problem (e.g., “diameter”, “speed”, “average”, “inches”) (p.320) for grades 4-6 and the number of words found between the first and last numbers in the problem for grades 7-9.

Lepik (1990) performed a study with a similar objective, but considered far fewer linguistic and computational features (31 in total). All linguistic features examined fall under quantitative properties of the text. A set of word-problems were generated to imitate standard textbook questions and administered to 8th grade students (Lepik, 1990). Lepik (1990) considered two measures of problem difficulty: the proportion of students who approached the question with the correct strategy and the average solving time for the question. Surprisingly, results from Lepik’s correlation analysis indicated poor predictive performance from all linguistic features considered. In addition, a shocking negative correlation was observed between longer question length and average solving time (Lepik, 1990, p.89). However, it is possible that such unexpected findings could be attributed to Lepik's unique measures for student performance, as results from prior studies typically reported length-related variables as significant predictors for problem-solving accuracy (M. E. Jerman & Mirman, 1974; M. Jerman & Rees, 1972).

Other studies examining structural linguistic features and their impact on word-problem difficulty did so in the context of students' reading comprehension difficulties. In her 2008 study, Martiniello investigated linguistic word-problem features that differentially affected the performance of English-language learners (ELLs) as opposed to non-ELLs. Questions collected from the MCAS statewide mathematics assessment were administered to a sample of 4th grade ELLs Spanish students "using think-aloud protocols" (Martiniello, 2008, p.334). Expert reviews and text analyses were then used to determine each question's linguistic complexity (Martiniello, 2008, p.334).

The basis of her study rested on the idea that "linguistic features of natural language that create comprehension difficulties for ELLs relate to complex vocabulary (lexical complexity) and sentence structure (syntactic complexity)" (Abedi & Lord, 2001; Abedi, Lord, & Hofstetter, 1998; Abedi et al., 1997; Butler, Bailey, Stevens, Huang, & Lord, 2004; Spanos, Rhodes, Dale, & Crandall, 1988; as cited in Martiniello, 2008, p.336). Lexical features found to pose disproportionate difficulty to ELLs included unfamiliar terms, polysemous words, and references to mainstream American culture (Martiniello, 2008, pp.357-358) Syntactic features contributing to difficulties for ELLs include presence of multiple clauses, long noun phrases, and "limited syntactic transparency in the text" (Martiniello, 2008, p.357).

Even minor changes to the wording of questions have been shown to significantly influence student performance (Cummins et al., 1988). Cummins and colleagues (1988), who examined the structural recall accuracy of children before and after solving word-problems, found that certain “linguistic forms” (i.e., “SOME,” “How many more X’s then Y’s”, and certain uses of “altogether”) (p.435) resulted in greater comprehension difficulties. Their findings are consistent with the linguistic development view of student difficulties in solving mathematical word-problems.

According to Cummins et al. (1988) this theory “holds that certain word problems are difficult to solve because they employ linguistic forms that do not readily map onto children’s existing conceptual knowledge structures” (p.407). This suggests that students who exhibit word-problem solving difficulties are not necessarily deficient on conceptual knowledge required to solve the problems, but rather may be misinterpreting the questions.

Semantic Linguistic Features

Some semantic linguistic features of word-problems include (a) “verbal cues”, (b) “phrasing in cue words”, (c) “conceptual rewording”, and (d) presence of distracting information (Daroczy et al., 2015, p.2). Verbal cues refer to words or phrases within the question that hint at operations required to reach the solution, and typically take the form of relational statements such as “more than” for addition or “less than” for subtraction (Daroczy et al., 2015). However, the relational terms present in the questions are not always consistent with the operations required to solve them, resulting in difficulties for inexperienced problem-solvers. This phenomenon is commonly referred to as the consistency effect, and its impact on student performance has been well-documented (Hegarty, Mayer, & Green, 1992; van der Schoot, Bakker Arkema, Horsley, & van Lieshout, 2009).

The consistency effect was investigated by Hegarty et al. (1992) and van der Schoot et al. (2009). Both implemented the eye movement approach, in which student's eye fixations were monitored as they worked through questions (Hegarty et al., 1992; van der Schoot et al., 2009). In both cases, two-step word-problems were written and distributed to participants (Hegarty et al., 1992; van der Schoot et al., 2009). Although the two studies used very different student populations, (Hegarty et al. used a small sample of college undergraduates while van der Schoot et al. used 5th to 6th grade students of differing problem-solving abilities), both reached the same conclusion: the presence of inconsistent verbal cues resulted in higher error rates and longer response times (Hegarty et al., 1992; van der Schoot et al., 2009). According to van der Schoot et al. (2009), the consistency effect was "more pronounced for less successful than more successful problem solvers" (p.59).

Small changes in the phrasing of verbal cues have also been shown to significantly impact word-problem difficulty (Daroczy et al., 2015, p.5). Conceptual rewording, which highlights semantic relations to make connections between objects referred to in the problems more explicit, has been shown to improve problem-solving accuracy (Vicente, Orrantia, & Verschaffel, 2007), and can be done by rewording ambiguous verbal cues found in the questions (LeBlanc & Weber-Russell, 1996).

For example, the question, "David and Kathy have 8 soda cans altogether. David has 5 soda cans. How many soda cans does Kathy have?", can be reworded "David and Kathy have 8 soda cans altogether. David has 5 of them. The rest of them are

Kathy's. How many soda cans does Kathy have?" (LeBlanc & Weber-Russell, 1996, p.381). Vicente et al. (2007), who tested easy and difficult two-step word-problems, with and without conceptual rewording, on students from grades 3-5, concluded that conceptual rewording improved student performance, especially with difficult problems requiring two steps.

Another semantic feature found to impact word-problem difficulty is "presence of numeric or linguistic distractors" (Daroczy et al., 2015, p.5). Muth (1992) generated and administered word-problems similar to those found on the NAEP to 8th grade students. Results indicated that word-problems containing extraneous information and extra steps lowered problem-solving accuracy (Muth, 1992).

Word-Problem Readability as a Predictor of Difficulty

Readability formulas, which assess text comprehensibility, are typically calculated from various linguistic features such as word length, sentence length, or word difficulty (Walkington, Clinton, Ritter, & Nathan, 2015). Thus, the readability, or reading level, of math questions merits consideration as a potential source of word-problem difficulty. However, while previous studies examining the effects of word-problem readability levels on student performance have produced mixed results, they generally provide insufficient evidence supporting a link between traditional readability measures and student achievement in mathematics.

Thompson (1967) as well as Linville (1970) published reports claiming to demonstrate a link between lower readability levels and higher student scores (as cited in Paul, Nibbelink, & Hoover, 1986). Thompson created two sets of word-problems testing parallel mathematical concepts, distinguished from each other only by their readability levels (as cited in Paul, Nibbelink, & Hoover, 1986). The readability of the easier set of problems was assessed using the Spache Formula, while the readability of the more difficulty set was assessed using the Dale-Chall Formula (as cited in Paul, Nibbelink, & Hoover, 1986). Linville, too, generated problem sets of different reading levels, but did so by varying the complexity of syntax and vocabulary in the questions (as cited in Paul, Nibbelink, & Hoover, 1986).

Although both conclude that students performed better on questions with low readability levels, Paul et al. (1986) suggest that these studies failed to control other variables contributing to word-problem difficulty, such as “sequence of presenting information” and “the amount of extraneous numerical information” (p.163). In their own study which implemented both “vocabulary control and sentence control” (p. 163), Paul and colleagues (1986) found no evidence suggesting that readability levels impact problem difficulty, or that traditional readability measures accurately reflect the grade-level readability of word-problems.

While Paul et al. (1986) do not express direct criticism of traditional readability metrics, emphasizing that these measures were not expressly designed for evaluating mathematical texts, there are studies questioning even the validity of readability metrics in assessing the readability of non-mathematical passages (Bailin & Grafstein, 2001; Bruce, Rubin, & Starr, 1981).

For example, Bruce, Rubin, and Starr (1981) attribute the failure of traditional readability metrics to several fundamental flaws, one of them being a “discrepancy between the characteristics of texts which readability formulas measure and those which we know to influence text comprehensibility” (p.4). Commonly used readability formulas include the Flesch-Kincaid readability tests (Flesch, 1948), the Fry readability metric (Fry, 1968), and the Spache (Spache, 1953) and Dale-Chall (Dale & Chall, 1948) formulas, all of which rely heavily on a few quantitative linguistic features of the text such as sentence length, word length, word syllables, and percent of unfamiliar words.

All of these readability metrics make the problematic assumption that these basic properties of text account for both syntactical and vocabulary difficulty, when in fact they do not (Bailin & Grafstein, 2001; Bruce et al., 1981). According to Bruce et al., (1981) the “degree of discourse cohesion, number of inferences required, number of items to remember, complexity of ideas, rhetorical structure, dialect, and background knowledge required” (p.4) are all factors readability formulas fail to take into consideration.

While the question of whether or not readability levels are valid predictors of problem difficulty requires further investigation, evidence appears to suggest that most traditional readability formulas are too simplistic to capture “the cohesion of text” (Walkington, Clinton, Ritter, & Nathan, 2015, p.1054), and do not accurately estimate the reading levels of mathematical word-problems. Therefore, readability is unlikely to be a reliable predictor of word-problem difficulty and will not be considered in the current study.

Text Analysis

With the advent of new technologies, large amounts of digital text have now been made available to the public, giving rise to a new field of research: text analysis (Gentzkow, Kelly, & Taddy, 2017; Grimmer & Stewart, 2013). Text analysis, a term often used interchangeably with text mining, refers to the process of transforming text into a quantitative, structured form suitable for analysis; and provides researchers with a systematic and cost-efficient method of analyzing and extracting information from large-scale textual data (Gentzkow et al., 2017; Grimmer & Stewart, 2013).

Data derived from text can often serve as an excellent “complement to the more structured kinds of data traditionally used in research” (Gentzkow et al., 2017, p.2). Previous studies from the political science and economics fields have demonstrated successful use of text as data (Gentzkow et al., 2017; Grimmer & Stewart, 2013). For example, electronic text collections used for text analysis include documented political speeches, news reports, social media postings, and “text from financial news” (Gentzkow et al., 2017, p.2).

However, text analysis does have its limitations, most of them deriving from the innate “complexity of language” (Grimmer & Stewart, 2013, p.2). Grimmer and Stewart (2013) emphasize that text analysis methods can never fully replace “careful and close reading of texts” (p.2), as they are based on “necessarily incorrect models of language” (p.28) that often fail to recognize syntactic ambiguities. Nevertheless, text analysis methods hold great potential when carefully validated by researchers, and when used to augment, rather than substitute for, human insight (Grimmer & Stewart, 2013).

Due to the nature of the data available to us, text analysis techniques are required to transform our word-problems into numerical representations suited for statistical analysis. Rather than construct math questions around specific features to be investigated, as was done in so many previous studies, we seek to identify surface linguistic features contributing to word-problem difficulty from unstructured text using various text analysis methods. Inferences can then be made about the areas of math students today find particularly challenging using results from our statistical analyses.

METHODOLOGY

Purpose

The goal of the current study is to analyze data collected from a set of math competition word-problems in order to identify surface linguistic features contributing to lower student performance. We follow an approach similar to that taken by earlier studies examining features impacting problem solving accuracy (M. E. Jerman & Mirman, 1974; M. Jerman & Rees, 1972; Lepik, 1990), and consider the linguistic features we identify as predictors of word-problem difficulty.

However, we differentiate our work from previous research by performing an observational study, in which we make inferences on data gathered from an outside source, rather than handpicking or adjusting word-problems in order to observe the effects of certain features on student performance. Due to the nature of our data, we implement text analysis methods with the goal of transforming text into numerical data suited for statistical modeling. Through our research, we aim to answer the following questions:

- “Which of our models provides the most reliable predictions for student performance?”,
- “How do surface linguistic features compare with expert classified problem types in terms of predictive accuracy?”,
- “Which features correspond to the lowest student performance?”, and
- “What can they tell us about the areas of math that students today find particularly challenging?”.

MOEMS Competition Overview

Math Olympiads for Elementary and Middle Schools (MOEMS) is an international math competition designed to help young students develop their mathematical problem-solving ability (“MOEMS program description,” n.d.). Contests are held annually with one exam distributed per month from November through March (“MOEMS Contest Dates,” n.d.). Each contest contains a total of five questions and are split into two divisions: Division E and Division M, which represent the Elementary (grades 4-6) and Middle School (grades 6-8) grade levels respectively (“MOEMS Contest Dates,” n.d.). It is from math competition textbooks compiled by MOEMS board members that we obtained our data.

Student Participants

Though not much is known about the student population participating in the competition, the contest is open to all students from grades 4-8 attending schools, home-schools, or institutes (“MOEMS program description,” n.d.). Student teams represent all 50 U.S. states as well as 30 different countries, and the MOEMS participant pool has grown steadily since the competition’s first opening in 1979 to include nearly 170,000 student participants from 6,000 teams in a single year (“MOEMS program description,” n.d.).

While it is recommended that all students compete within their respective grade divisions, students below grade 4 are also permitted to participate (“MOEMS program description,” n.d.). Individual factors such as each students’ grade, preparation, mathematical ability, and reading comprehension level are unknown, but students likely participate on a voluntary basis

Testing Conditions

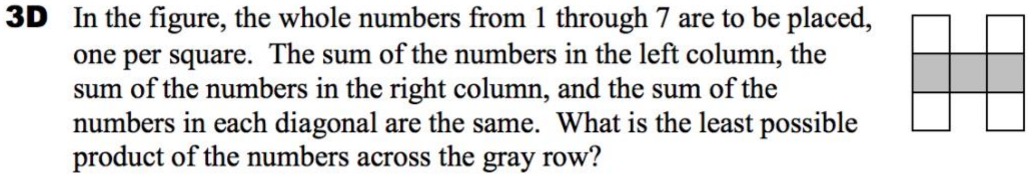
The MOEMS program description informs us that contests are held within each students' respective participating school or institution. Each contest word-problem has a time-limit that must be adhered to, and "calculators are not permitted" during the exams ("MOEMS program description," n.d.). Definitions are provided to students for all "advanced" concepts ("MOEMS program description," n.d.).

Data Collection

We gathered our data from the following textbooks containing word-problems taken from MOEMS contests held from 1979 through 2013:

- *Math Olympiad Contest Problems for Elementary and Middle Schools, Vol. 1* by Dr. G. Lenchner;
- *Math Olympiad Contest Problems, Volume 2* edited by Richard Kalman; and
- *MOEMS Contest Problems, Volume 3* edited by Richard Kalman & Nicholas J. Restivo.

All contest questions in the textbooks were "reviewed by a select committee of mathematicians and teachers for ambiguity, language, and level of difficulty" (Lenchner, 1990/1997, p. 6).



3D In the figure, the whole numbers from 1 through 7 are to be placed, one per square. The sum of the numbers in the left column, the sum of the numbers in the right column, and the sum of the numbers in each diagonal are the same. What is the least possible product of the numbers across the gray row?

Figure 1. Sample 2018 MOEMS Contest Problem (3D) from Division E, Contest 3. Taken From: Sample

Contest. (n.d.). Retrieved April 4, 2019, from https://www.moems.org/sample.htm. Copyright © 2017 by Mathematical Olympiads for Elementary and Middle Schools, Inc.

Content of interest to our study included: (a) the 1,225 contest word-problems provided by the three textbooks, (b) the percent of students who answered each question correctly (determined from reports by individual schools and institutes), and (c) a set of expert classified “Problem Types” (see Table 1), as determined by the writers of the textbooks.

Table 1. *List of Compiled MOEMS Contest Problem Types.*

1. Addition Patterns	2. Algebraic Thinking	3. Area	4. Arithmetic Sequences
5. Arithmetic Series	6. Averages (Arithmetic Means)	7. Book Pages	8. Blindfold
9. Circles	10. Clock Problems	11. Coin Problems	12. Combinations
13. Consecutive Numbers	14. Consecutive Even Numbers	15.Consecutive Odd Numbers	16. Cryptarithms
17. Divisibility	18. Divisibility Combinations	19. Fractions, Decimals, Percents	20. Factors
21. Flashing Lights	22. Magic Squares	23.Motion Problems	24. Multiples
25. Multiplication Patterns	26. Cubes And Rectangular Solids	27. Palimage	28. Perimeter
29.Postage Stamps	30. Prime Factorization	31. Prime Numbers	32. Remainders
33. Sequence Of Partial Sums	34. Sequences And Series	35.Square And Cube Numbers	36. Terminal Zeros
37. Tower Problems	38. Tree Diagram	39. Unit Fractions	40. Venn Diagrams
41. Volume	42. Work	43. Working Backwards	44. Age Problems
45. Angles	46. Business Problems	47. Calendar Problems	48. Certainty Problems
49.Cycling Numbers	50. Digit Problems	51. Distributive Property	52. Draw A Diagram
53. Exponents	54. Factorials	55. Graphs	56. Logic

Table 1. *List of Compiled MOEMS Contest Problem Types (continued).*

57. Border Problem	58. Clover Problem	59. Fence-Post Problem	60. Funny Numbers
61. Three Intersecting Figures	62. Traffic Flow	63. Turnover Card Problem	64. Twinners
65. Up-And-Down Numbers	66. Number Sense	67. Order Of Operations	68. Organizing Data
69. Palindromes	70. Parity (Odd Vs. Even Numbers)	71. Paths	72. Patterns
73. Probability	74. Process Of Elimination	75. Ratios And Proportions	76. Rectangles And Squares
77.Signed Numbers	78. Statistics	79. Tables	80. Triangles
81. Triangular Numbers	82. Arithmetic Operations And Properties	83. Binary Numbers	84. Congruent Figures
85. Fibonacci Numbers	86. Target Problems	87. Asterisk Array Problem	88. Clock-Angle Problem
89. Ducks Problem	90. Interesting Date Problem	91. Math-Olympiad Problem	92. Number Recycling Machine Problem
93. Triangle Inequality Problem	94. Quiz Game Problem	95. Triangle Inequality Problem	96. Wandering Pet Problem

Question and Problem Type data were scanned from the textbook pages and converted to machine-readable text using ABBYY FineReader 14 OCR (Optical Character Recognition) technology, which also enabled us to distinguish between useful figures and text. Information we retained from the scanned pages included, Problem Types, problem identifiers, the suggested time-limits, the percentage of correct responses, and the full text of each question. Any extraneous information not useful to our analysis was discarded.

Due to the OCR software's inability to recognize special mathematical characters (e.g., π, √, ≥) , equations, or displayed figures, we replaced them with descriptive tag abbreviations (see Table 2) to identify their location within the text. This allowed us to somewhat preserve the content of questions whose meaning derived mainly from displayed figures and equations. Our collected data was then restructured, converted to CSV form, and imported into the R software environment as for further cleaning and analysis.

Table 2. *List of all 29 tag names with their abbreviations.*

TAG ABBREVIATION	TAG NAME
[SYM]	[SYMBOL]
[DE]	[DISPLAYED EQUATION]
[IE]	[INLINE EQUATION]
[FG]	[FIGURE GEOMETRIC]
[DX]	[DISPLAYED EXPRESSION]
[IX]	[INLINE EXPRESSION]
[FT]	[FIGURE TABLE]
[FP]	[FIGURE PATTERN]
[FMS]	[FIGURE MAGIC SQUARE]
[FMD]	[FIGURE MISSING DIGITS]
[FMDS]	[FIGURE MISSING DIGITS SUM]
[FMDD]	[FIGURE MISSING DIGITS DIVISION]
[FMDM]	[FIGURE MISSING DIGITS MIULTIPLICATION]
[FMDF]	[FIGURE MISSING DIGITS DIFFERENCE]
[FC]	[FIGURE CUBES]
[FF]	[FIGURE FLOW]
[IM]	[INLINE MIXED NUMBER]
[IF]	[INLINE FRACTION]
[IEX]	[INLINE EXPONENT]
[LS]	[LINE SEGMENT]
[FC]	[FIGURE CLOCK]
[FV]	[FIGURE VENN]
[FS]	[FIGURE SEATING]
[FM]	[FIGURE MAP]
[ANG]	[ANGLE]
[DI]	[DISPLAYED IMAGE]
[FGR]	[FIGURE GRAPH]
[RAY]	[RAY]

Preprocessing

We performed the following preprocessing steps on our data using the gsub() function ("grep function | R Documentation," n.d.) in order to ensure that it was suitable for analysis: removal of unwanted characters such as HTML character entities, expansion of mathematical symbols and abbreviations to their word equivalents (e.g., "$" to "dollar" and "ft." to "feet"), making spelling corrections, and stripping whitespaces. Tokenization, which involves breaking up text into tokens (words) (Allahyari et al., 2017), was performed later using the dfm() function ("dfm function | R Documentation," n.d.) during feature generation, which will be discussed in the next section. Stemming and lemmatization were not performed, as the affixes of many mathematical terms contained information crucial to identifying their semantic meanings.

Feature Generation

Feature generation techniques were performed to define a suitable set of surface linguistic features for statistical analysis. The quanteda package, which provides tools for the "management, processing and quantitative analysis of textual data" ("quanteda package | R Documentation," n.d.), was used for the entirety of this step. Here, not only do we establish features we wish to use for analysis, we also filter out unimportant terms to reduce dimensionality of the data.

Due to image removal during our data collection process, our corpus of word-problems contained several questions with very few terms and limited content (e.g., "Find the value of ."). As a result, we used quanteda's corpus_subset() ("corpus_subset function | R Documentation," n.d.) function to remove all questions containing fewer than 6 terms. This cut-off threshold was determined to be most ideal, as it eliminated questions with little to no content, while retaining questions with few words but enough content to be useful to our analysis.

Next, we built a quanteda dictionary (“dictionary,” n.d.) in order to narrow down and define the features to be used during analysis. Our dictionary’s original list of terms, or features, included our abbreviated tags (see Table 2) as well as all words occurring 10 times or more within the corpus. Frequently occurring but relatively uninformative terms like “number” and “value”, as well as any English stop-words remaining in our dictionary were removed.

A quanteda dictionary allows for the application of multiple terms to the same dictionary key (“dictionary,” n.d.), a capability we used to effectively identify multi-word phrases (e.g., “square unit”, “arithmetic mean”) and categorize associated terms (e.g., “second", "seconds", "minute", "minutes", "hour" are all generalized under the feature “time_unit”). This, to a certain extent, improves the accuracy of our feature space, reduces dimensionality, and accounts for polysemous words indicative of very different problem types. For instance, the word “square”, on its own, may refer to a geometric shape; but takes on a different meaning in the phrase “square root.”

The Vector Space Model (VSM) representation for documents was implemented to create a numeric representation for our corpus of word-problems. From here on, the collection of texts (our entire set of word-problems) will be referred to as the corpus, the “units of analysis” (Grimmer & Stewart, 2013, p.6) (each individual word-problem) will be referred to as documents, and our dictionary-defined terms will be referred to as features. In the VSM representation, each document is converted from a vector of strings to a vector of numeric values indicating the relative weight, or importance, of each feature to the document (Allahyari et al., 2017). We used the dfm() function (“dfm function | R Documentation,” n.d.) to automatically tokenize our text, and to create a document-feature matrix (DFM) that makes use of our dictionary-defined features. After a final reduction of our feature-space by removing documents with low-frequency terms, we applied the term frequency – inverse document frequency (TF-IDF) term weighting scheme, to our DFM using quanteda’s tfidf() function (“tfidf function | R Documentation,” n.d.).

TF-IDF is a technique used to quantify the predictive capacity of each term, and has been described as an effective approach for filtering uninformative words from text , that “improves on simply excluding words that occur frequently” (Gentzkow et al., 2017). Through its two components: TF and IDF, it provides a score that considers both the frequency of a term to its document as well as its frequency throughout all corpus documents. Quanteda’s tfidf() function (“tfidf function | R Documentation,” n.d.) allowed us to compute normalized term frequency with inverse document frequency using the following formula:

$$TF\text{-}IDF = \frac{T_{ij}}{\Sigma_j TF_{ij}} \times \ log_{base}\left(s + \frac{N}{k + DF_j}\right) \tag{1}$$

, where smoothing constant $s = 0$, predefined constant $k = 0$, and the logarithm $base = 10$ (“dfm_weight function | R Documentation,” n.d.; “docfreq function | R Documentation,” n.d.).

Our feature generation step leaves us with a DFM consisting of 154 surface linguistic features (see Appendix A) and a remainder of 1,046 documents (contest word-problems) from our original collection of 1,225 questions. These 154 weighted features compose our first set of predictor variables for word-problem difficulty, and will be further examined during regression analysis.

Data Analysis

Our first task in the analysis step was to generate linguistic clusters of the documents to be later used as an additional set of predictors for word-problem difficulty. Cluster analysis is an unsupervised learning technique that identifies and sorts together "groups of similar documents in a collection of documents" (Allahyari et al., 2017, p.6). The main purpose of this step was to see if we could create a reliable set of predictor variables from linguistic clusters, constructed based on underlying properties of the text rather than on predefined categories.

Clustering algorithms require distance matrices as input, which display pairwise distance scores between documents. Distance scores, in short, measure the "closeness", or dissimilarity, between documents (Gentzkow et al., 2017; Grimmer & Stewart, 2013); and can be computed using various metrics, two of the most popular being the Euclidean distance metric and the Manhattan distance metric (Madhulatha, 2012). Although both metrics were tested, we chose to use the Manhattan distance metric, as it is considered to be better suited for "high dimensional data mining applications" (Aggarwal, Hinneburg, & Keim, 2001). This metric computes the absolute distance between vectors (documents) ("dist function | R Documentation," n.d.) using the formula

$$d = \sum_{i=1}^{n} |x_i - y_i| \tag{2}$$

, where d represents the distance between points $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$ in n-space (Madhulatha, 2012, p.719).

Generating a distance matrix allowed us to cluster our documents using the K-medoids algorithm. The K-medoids algorithm is a popular partitioning method that iteratively selects medoids, or representative center points, to minimize their dissimilarities to other data points within the same clusters (Arora, Deepali, & Varshney, 2016). Although the K-means algorithm appears to be more commonly used, the K-medoids method has been shown to be better "in terms of execution time" (Arora, Deepali, & Varshney, 2016, p.507), sensitivity to outliers, noise reduction; and results in less overlap between clusters (Arora et al., 2016). An algorithm for K-medoids partitioning is as follows:

Input: K_y: the number of clusters, D_y: a data set containing n objects.

Output: A set of k_y clusters.

Algorithm:

- Randomly select k_y as the Medoids for k data points.
- Find the closest Medoids by calculating the distance between data points n and Medoids k and map data objects to that.
- For each Medoids m and each data point o associated to m do the following:

Swap m and o to compute the total cost of the configuration than

Select the Medoids o with the lowest cost of the configuration.

- If there is no change in the assignments repeat steps 2 and 3 alternatively (Arora et al., 2016, p.509).

Using the pamk() function (“pamk function | R Documentation,” n.d.), we were able to perform Partitioning Around Medoids (PAM) clustering on our documents, with the optimum number of clusters (k) estimated using average silhouette width. Our clustering step resulted in a set of 98 linguistic clusters of documents, grouped based on features of the text, to be examined later for their relations to word-problem difficulty.

At this stage, we have obtained three possible sets of predictor variables for word-problem difficulty, as it relates to student performance. These sets of predictors will be referred to as follows:

- Weighted Features (see Appendix A), which refer to the TF-IDF weighted features from the feature generation step;
- Linguistic Clusters, which refer to our 98 clusters of contest questions; and
- Problem Types, which refer to the expert classified problem types found in the MOEMS textbooks (see Table 1).

We ran multiple logistic regression analyses separately on each of these three sets of predictors using the glm() function (“glm function | R Documentation,” n.d.) with quasibinomial error distribution and logit link function. This allowed us to preserve probability (values between 0 and 1) and model a binary response variable with multiple independent predictors. The response in each case was the observed proportion of correct student answers to each question. Each of our three logistic regression models predicts the likelihood of a given student answering each question correctly.

Results

Statistically significant (p < 0.05) predictors from our Weighted Features and Linguistic Clusters models are listed in Tables 3 and 4, respectively. While 137 out of our 154 weighted linguistic features (see Appendix B) and 28 out of our 98 linguistic clusters (see Table 4) were reported to be statistically significant, none of our 98 expert classified problem types were significant below 0.1 level. Statistically significant features from our weighted features model include variables "comparison" (p < 0.01) , "area" (p < 0.001) , "length_unit" (p < 0.01), and "area_units" (p < 0.01), listed in order of greatest to lowest difficulty. Further details regarding our statistically significant linguistic clusters (labeled by cluster number), however, are not provided, as many of them are difficult to characterize.

Predictions for proportion of correct student responses were made for each model and compared to observed values in order to evaluate the predictive accuracy of each set of predictors. Each model demonstrated a positive linear relationship between observed and predicted values for the response: Weighted Features model ($r = 0.56, R^2 = 0.32$), Linguistic Clusters model ($r = 0.47, R^2 = 0.22$), Problem Types model ($r = 0.53, R^2 = 0.28$). Based on these scores, we conclude that our most reliable set of predictors for word-problem difficulty are our weighted features, which interestingly performed even better than our expert classified problem types. Thus, we make our inferences regarding areas of math in which students are experiencing difficulty from the results of our weighted features model.

Table 3. *List of Statistically Significant Linguistic Clusters by Order of Coefficient Size.*

Cluster number	Regression Coefficient	Standard Error	t value	Pr(>|t|)
76 ***	-1.54535	0.540624	-2.85846	0.004350
31 **	-1.40382	0.449256	-3.12476	0.001834
50 *	-1.39659	0.648318	-2.15417	0.031478
34 *	-1.29897	0.629208	-2.06446	0.039246
27 ***	-1.27699	0.320857	-3.97992	7.42E-05
62 *	-1.12061	0.443860	-2.52468	0.011742
11 **	-1.10859	0.416042	-2.66462	0.007838
12 ***	-1.06868	0.308013	-3.46959	0.000545
83 **	-1.05341	0.390087	-2.70044	0.007048
63 *	-1.02826	0.434139	-2.36849	0.018060
38 ***	-1.02426	0.258559	-3.96144	8.01E-05
58 **	-0.99251	0.324416	-3.05939	0.002280
19 **	-0.99201	0.368451	-2.69239	0.007219
75 *	-0.97071	0.403225	-2.40736	0.016258
71 ***	-0.94438	0.267058	-3.53625	0.000425
91 **	-0.91653	0.338240	-2.70971	0.006856
13 **	-0.90437	0.319370	-2.83171	0.004728
59 **	-0.89364	0.327184	-2.73130	0.006426
20 *	-0.88255	0.347359	-2.54073	0.011221
44 **	-0.86995	0.310206	-2.80442	0.005144
65 *	-0.85509	0.345635	-2.47398	0.013536
22 *	-0.84267	0.392748	-2.14558	0.032160
79 *	-0.83206	0.372802	-2.23191	0.025854
35 *	-0.78839	0.354449	-2.22428	0.026364
23 *	-0.76582	0.368343	-2.07911	0.037876
54 *	-0.71395	0.286460	-2.49231	0.012861
51 *	-0.68407	0.276976	-2.46978	0.013695
33 *	-0.67819	0.266730	-2.54262	0.011160

Note. Significance codes: *** p < 0.001 ** p < 0.01 * p < 0.05

Table 4. *List of Top 5 Statistically Significant TF-IDF Weighted Features by Order of Coefficient Size.*

Feature	Regression Coefficient	Standard Error	t value	Pr(>|t|)
comparison **	-8.56066	2.928106	-2.92362	0.003547

Table 4. *List of Top 5 Statistically Significant TF-IDF Weighted Features by Order of Coefficient Size (continued).*

area ***	-7.77352	2.347475	-3.31144	0.000965
length_unit **	-7.66779	2.357979	-3.25185	0.00119
area_units **	-7.15393	2.362285	-3.0284	0.002529
FG **	-7.06517	2.348576	-3.00828	0.002701

Note. Significance codes: *** p < 0.001   ** p < 0.01   * p < 0.05

Discussion

An examination of our weighted features model reveals that the statistically significant feature corresponding to the lowest proportions of correct responses (indicated by a strongly negative regression coefficient, as seen in Table 4) is the variable “comparison.” Put simply, our model indicates that a high TF-IDF weighting of any comparison term in any question is expected to lower the proportion of students likely to answer the question correctly. Terms associated with the “comparison” feature include "increase", "decrease", “more" ,"less" , “fewer”, “most”, “least”, and “greatest” (a comprehensive list of terms can be found in Appendix A); all of which, based on our results, may be expected to contribute to word-problem difficulty. A few sample problems, retrieved from the MOEMS textbooks, in which “comparison” terms have been weighted highly are displayed in Table 5.

This result is unsurprising, as word-problems featuring relational terms (e.g., “more”, “less”) inconsistent with the required operations often result in difficulties for students (Hegarty et al., 1992; Lewis & Mayer, 1987; van der Schoot et al., 2009). It is possible that the performance of MOEMS participants were affected by such inconsistencies in the questions.

A few other significant predictors for problem difficulty corresponding to low student performance scores included "area" and units of area (e.g., "area", "square unit", "square meter", "square inch"), units of length (e.g., "inch", "feet", "meter", "mile"), and "FG" (indicative of a geometric figure displayed in the question). Sample questions involving these features are shown in Table 6.

Table 5. *Sample Problems with High TF-IDF weight for "comparison" Feature.*

Question	Feature TF-IDF Weight					Percent Correct
	comparison	area	length unit	area units	FG	
"In an election, Ethan got 5 fewer votes than Christopher, who got 3 votes more than Olivia, who got 4 fewer votes than Ava. How many more votes did Ava get than Ethan?"	0.622	0.00	0.00	0.00	0.00	0.52
"Noelle collects autographs of famous mathematicians. Exactly one statement below is true: * Noelle owns at least 77 autographs. * Noelle owns at least 62 autographs. * Noelle owns at least 45 autographs. What is the greatest number of autographs Noelle can own?"	0.622	0.00	0.00	0.00	0.00	0.20
"Bay Street has from 2 through 15 houses, numbered 1,2, 3, and so on. Mr. Sullivan lives in one of the houses. The sum of all the house numbers less than his equals the sum of all the house numbers greater than his. How many houses are there on Bay Street?"	0.414	0.00	0.00	0.00	0.00	0.32

Note. Feature weights have been rounded. Percents are represented as proportions.

While we do not claim to know the underlying factors responsible for student difficulties with questions involving these features, some previous research suggests that word-problems requiring unit conversions tend to be more difficult (Loftus & Suppes, 1972). Other possible factors contributing to student difficulties with questions involving area, units of measurement, and geometric figures are beyond the scope of the current study, and should be further investigated, as little prior work on the topic has been identified.

Table 6. *Sample Problems with High TF-IDF weight for "comparison", "area", "length_unit", "area_units", and "FG" Features.*

Question	Feature TF-IDF Weight					Percent Correct
	comparison	area	length unit	area units	FG	
"A circle and a triangle overlap as shown. The area of the circle is three times the area of the triangle. If the common region is removed, then the area of the rest of the circle would be 14 square centimeters more than the area of the rest of the triangle. What is the area of the complete triangle, in sq cm?"	0.00	0.284	0.00	0.00	0.123	0.52
"Amy can mow 600 square yards of grass in _hours. At this rate, how many minutes would it take her to mow 600 square feet?"	0.00	0.00	0.776	0.00	0.00	0.54
"A tractor wheel is 88 inches in circumference. How many complete turns will the wheel make in rolling one mile on the ground? (1 mile = 5,280 feet)"	0.00	0.00	0.097	0.257	0.00	0.06
"An acute angle is an angle whose measure is between 0→∞ and 90→∞. Using the rays in the diagram, how many different acute angles can be formed?"	0.00	0.00	0.00	0.00	0.288	0.18

Note. Figures originally displayed to supplement questions are not shown here due to copyright laws. Feature weights have been rounded. Percents are represented as proportions.

CONCLUSION

In the present study, we have established a precedent for identifying surface linguistic features of word-problems contributing to problem difficulty, for the purpose of creating a predictive model for student performance, using text analysis methods. Our main goal was to find potential determinants of problem difficulty that would be indicative of areas of math students today are having trouble with.

While the reasons behind why certain predictor variables (discussed in the last chapter) caused difficulties for students are beyond the scope of the present study, we believe our weighted linguistic features provide valuable insight into problem types and mathematical topics that students find particularly challenging; and can therefore help to inform the mathematical teaching practice as well as future work on features contributing to word-problem difficulty. In the remainder of this chapter, we discuss some of the limitations of the present study as well as directions for future research.

Limitations of the Current Study

Unfortunately, due to limitations related to our dataset, our knowledge of the participating student population is incredibly limited. Individual factors of participating students, which were typically recognized and controlled in prior research to account for differences in performance, were not considered in the present study. It is important to recognize that such factors, which include the students' mathematical skill, extent of preparation, or reading ability, are likely to have interfered, at least to a small degree, to the students' performance.

What can be inferred from the data provided to us, however, are the respective divisions of each problem; and by extension, the grade-level categories of participating students (Division E for grades 4-6 and Division M for grades 6-8). For simplicity's sake, these division levels were disregarded during analysis. However, it may be interesting to observe how student performance for different areas of math differed based on division level. It should also be noted that although the textbooks provided us with the percent of correct student responses for each question, we cannot ascribe precision to these percentages as they are self-reported. The exact number of students who attempted each question is also unknown.

Lastly, the results of our models appear to indicate that TF-IDF weighted features perform more reliably as predictors of student performance than expert classifications. However, this cannot be stated in absolute terms, as only a fraction of the total contest questions incorporated in our analysis were sorted into problem type categories by the textbook writers.

Future Work

Due to the limitations discussed above, we cannot state that the results presented in the previous chapter are strongly conclusive. Further research, preferably with more transparent demographics for the student body, is necessary to verify our conclusions. It may also be worthwhile to investigate the applicability of our predictive model to other different sets of word-problems, to see if they produce similar results.

Through the present study, we have identified features of mathematical word-problems which may contribute to problem difficulty. Future research should seek to examine why these particular features cause problems for students, with the objective of producing new and improved mathematics word-problem solving strategies. In addition, it is our hope that the features we've identified may be considered concurrently with other features defined in previous research to help educators build word-problems of specified difficulty levels.

REFERENCES

Acosta-Tello, E. (n.d.). *Making Mathematics Word Problems Reliable Measures of Student Mathematics Abilities*. 12.

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. In J. Van den Bussche & V. Vianu (Eds.), *Database Theory — ICDT 2001* (pp. 420–434). Springer Berlin Heidelberg.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *ArXiv:1707.02919 [Cs]*. Retrieved from http://arxiv.org/abs/1707.02919

Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-Means and K-Medoids Algorithm For Big Data. *Procedia Computer Science*, *78*, 507–512. https://doi.org/10.1016/j.procs.2016.02.095

Bailin, A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. *Language & Communication*, *21*(3), 285–301. https://doi.org/10.1016/S0271-5309(01)00005-2

Bruce, B., Rubin, A., & Starr, K. (1981). Why readability formulas fail. *IEEE Transactions on Professional Communication*, *PC-24*(1), 50–52. https://doi.org/10.1109/TPC.1981.6447826

Carpenter, T. P., & And Others. (1980). Solving Verbal Problems: Results and Implications from National Assessment. *Arithmetic Teacher*, *28*(1), 8–12.

corpus_subset function | R Documentation. (n.d.). Retrieved April 22, 2019, from https://www.rdocumentation.org/packages/quanteda/versions/1.4.3/topics/corpus_subset

Corte, E. D., Verschaffel, L., & Greer, B. (2000). Connecting mathematics problem solving to the real world. *Proceedings of the International Conference on Mathematics Education into the 21st Century: Mathematics for Living*, 66–73.

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

Dale, E., & Chall, J. S. (1948). A Formula for Predicting Readability: Instructions. *Educational Research Bulletin*, *27*(2), 37–54. Retrieved from JSTOR.

Daroczy, G., Wolska, M., Meurers, W. D., & Nuerk, H.-C. (2015). Word problems: a review of linguistic and numerical factors contributing to their difficulty. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00348

dfm function | R Documentation. (n.d.). Retrieved April 21, 2019, from https://www.rdocumentation.org/packages/quanteda/versions/1.4.1/topics/dfm

dfm_weight function | R Documentation. (n.d.). Retrieved April 22, 2019, from https://www.rdocumentation.org/packages/quanteda/versions/1.3.4/topics/dfm_weight

dictionary: Create a dictionary in quanteda: Quantitative Analysis of Textual Data. (n.d.). Retrieved April 22, 2019, from https://rdrr.io/cran/quanteda/man/dictionary.html

dist function | R Documentation. (n.d.). Retrieved April 22, 2019, from https://www.rdocumentation.org/packages/stats/versions/3.5.3/topics/dist docfreq function | R Documentation. (n.d.). Retrieved April 22, 2019, from https://www.rdocumentation.org/packages/quanteda/versions/1.3.4/topics/docfreq

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

Fry, E. (1968). A Readability Formula That Saves Time. *Journal of Reading*, *11*(7), 513–578. Retrieved from JSTOR.

Geary, D. C. (1994). *Children's mathematical development: Research and practical applications*. https://doi.org/10.1037/10163-000

Gentzkow, M., Kelly, B. T., & Taddy, M. (2017). *Text as Data* (Working Paper No. 23276). https://doi.org/10.3386/w23276

glm function | R Documentation. (n.d.). Retrieved April 24, 2019, from https://www.rdocumentation.org/packages/stats/versions/3.5.3/topics/glm

grep function | R Documentation. (n.d.). Retrieved April 21, 2019, from https://www.rdocumentation.org/packages/base/versions/3.5.3/topics/grep

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Hegarty, M., Mayer, R. E., & Green, C. E. (1992). Comprehension of arithmetic word problems: Evidence from students' eye fixations. *Journal of Educational Psychology*, *84*(1), 76–84. https://doi.org/10.1037/0022-0663.84.1.76

Hegarty, M., Mayer, R. E., & Green, C. E. (n.d.). *Comprehension of Arithmetic Word Problems: Evidence From Students' Eye Fixations*. 9.

Hoogland, K., de Koning, J., Bakker, A., Pepin, B. E. U., & Gravemeijer, K. (2018). Changing representation in contextual mathematical problems from descriptive to depictive: The effect on students' performance. *Studies in Educational Evaluation*, *58*, 122–131. https://doi.org/10.1016/j.stueduc.2018.06.004

Huang, A. Y.-Q. (2008). *Similarity Measures for Text Document Clustering*.

Jerman, M. E., & Mirman, S. (1974). Linguistic and Computational Variables in Problem Solving in Elementary Mathematics. *Educational Studies in Mathematics*, *5*(3), 317–362. Retrieved from JSTOR.

Jerman, M., & Rees, R. (1972). Predicting the Relative Difficulty of Verbal Arithmetic Problems. *Educational Studies in Mathematics*, *4*(3), 306–323. Retrieved from JSTOR.

LeBlanc, M. D., & Weber-Russell, S. (1996). Text integration and mathematical connections: A computer model of arithmetic word problem solving. *Cognitive Science*, *20*(3), 357–407. https://doi.org/10.1016/S0364-0213(99)80010-X

Lee, J.-T., Song, Y.-I., & Rim, H.-C. (2007). Predicting the Quality of Answers Using Surface Linguistic Features. *Sixth International Conference on Advanced Language Processing and Web Information Technology (ALPIT 2007)*, 111–116. https://doi.org/10.1109/ALPIT.2007.40

Lepik, M. (1990). Algebraic Word Problems: Role of Linguistic and Structural Variables. *Educational Studies in Mathematics*, *21*(1), 83–90. Retrieved from JSTOR.

Madhulatha, T. S. (2012). An Overview on Clustering Methods. *ArXiv:1205.1117 [Cs]*. Retrieved from http://arxiv.org/abs/1205.1117

Martiniello, M. (2008). Language and the Performance of English-Language Learners in Math Word Problems. *Harvard Educational Review*, *78*(2), 333–368. https://doi.org/10.17763/haer.78.2.70783570r1111t32

MOEMS Contest Dates. (n.d.). Retrieved April 21, 2019, from http://www.moems.org/contests.htm

MOEMS program description. (n.d.). Retrieved April 21, 2019, from http://www.moems.org/program.htm

Muth, K. D. (1992). Extraneous information and extra steps in arithmetic word problems. *Contemporary Educational Psychology*, *17*(3), 278–285. https://doi.org/10.1016/0361-476X(92)90066-8

NAEP Mathematics: National Achievement-Level Results. (n.d.). Retrieved March 12, 2019, from https://www.nationsreportcard.gov/math_2017/nation/achievement/?grade=4

Nathan, M. J., & Koedinger, K. R. (2000). Teachers' and Researchers' Beliefs about the Development of Algebraic Reasoning. *Journal for Research in Mathematics Education*, *31*(2), 168–190. https://doi.org/10.2307/749750

NW, 1615 L. St, Suite 800Washington, & Inquiries, D. 20036USA202-419-4300 | M.-857-8562 | F.-419-4372 | M. (n.d.). U.S. academic achievement lags that of many other countries. Retrieved March 12, 2019, from Pew Research Center website: http://www.pewresearch.org/fact-tank/2017/02/15/u-s-students-internationally-math-science/

pamk function | R Documentation. (n.d.). Retrieved April 23, 2019, from https://www.rdocumentation.org/packages/fpc/versions/2.1-11.1/topics/pamk

Paul, D. J., Nibbelink, W. H., & Hoover, H. D. (1986). The Effects of Adjusting Readability on the Difficulty of Mathematics Story Problems. *Journal for Research in Mathematics Education*, *17*(3), 163–171. https://doi.org/10.2307/749299

Powell, S. R., Fuchs, L. S., Fuchs, D., Cirino, P. T., & Fletcher, J. M. (2009). Do Word-Problem Features Differentially Affect Problem Difficulty as a Function of Students' Mathematics Difficulty With and Without Reading Difficulty? *Journal of Learning Disabilities*, *42*(2), 99–110. https://doi.org/10.1177/0022219408326211

quanteda package | R Documentation. (n.d.). Retrieved April 22, 2019, from https://www.rdocumentation.org/packages/quanteda/versions/0.9.2-0

Sample Contest. (n.d.). Retrieved April 4, 2019, from https://www.moems.org/sample.htm

Schoenfeld, A. H. (1991). On mathematics as sense-making: An informal attack on the unfortunate divorce of formal and informal mathematics. In *Informal reasoning and education* (pp. 311–343). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Spache, G. (1953). A New Readability Formula for Primary-Grade Reading Materials. *The Elementary School Journal*, *53*(7), 410–413. https://doi.org/10.1086/458513

tfidf function | R Documentation. (n.d.). Retrieved April 22, 2019, from https://www.rdocumentation.org/packages/quanteda/versions/0.9.4/topics/tfidf

van der Schoot, M., Bakker Arkema, A. H., Horsley, T. M., & van Lieshout, E. C. D. M. (2009). The consistency effect depends on markedness in less successful but not successful problem solvers: An eye movement study in primary school children. *Contemporary Educational Psychology*, *34*(1), 58–66. https://doi.org/10.1016/j.cedpsych.2008.07.002

Vicente, S., Orrantia, J., & Verschaffel, L. (2007). Influence of situational and conceptual rewording on word problem solving. *The British Journal of Educational Psychology*, *77*(Pt 4), 829–848. https://doi.org/10.1348/000709907X178200

Walkington, C., Clinton, V., Ritter, S. N., & Nathan, M. J. (2015). How readability and topic incidence relate to performance on mathematics story problems in computer-based curricula. *Journal of Educational Psychology*, *107*(4), 1051–1074. https://doi.org/10.1037/edu0000036

Wang, A. Y., Fuchs, L. S., & Fuchs, D. (2016). Cognitive and Linguistic Predictors of Mathematical Word Problems With and Without Irrelevant Information. *Learning and Individual Differences*, *52*, 79–87. https://doi.org/10.1016/j.lindif.2016.10.015

APPENDIX A

Table A. *Dictionary of Features for Weighted Features Model.*

	Feature	Dictionary Associated Terms
1	XSYM	“XSYM”
2	XDE	“XDE”
3	XIE	“XIE”
4	XFG	“XFG”
5	XDX	"XDX"
6	XIX	"XIX"
7	XFT	"XFT"
8	XFP	"XFP"
9	XFMS	"XFMS"
10	XFMDS	"XFMDS"
11	XFMDM	"XFMDM"
12	XFC	"XFC"
13	XIF	"XIF"
14	XIEX	"XIEX"
15	XLS	"XLS"
16	XFM	"XFM"
17	place_value	"units digit", "ones", "ones digit", "tens digit", "tens", "tens column", "tens columns", "hundreds digit", "hundreds", "hundreds column", "hundreds columns", "thousands digit", "thousands", "thousands column", "thousands columns"
18	weight	"weight", "weights", "weigh", "weighs"
19	parity	"odd", "even"
20	comparison_age	"older", "younger", "oldest", "youngest"
21	comparison_length	"shortest", "shorter", "longer", "longest", "widest", "wider", "highest", "higher", "deepest", "deeper", "nearest", "nearer", "furthest", "farther"
22	comparison_size	"smaller", "smallest", "larger", "largest"
23	comparison	"increase", "increases", "increased", "decrease", "decreases", "decreased", "more", "less", "fewer", "most", "greater", "greatest", "least", "lower", "fewest"
24	dimension	"wide", "width", "long", "length", "high", "height", "deep", "depth", "distance", "distances"
25	faces	"face", "faces", "side", "sides", "bottom", "top"
26	opposite	"opposite", "opposites"
27	straightline	"straight line", "straight lines", "line segment", "line segments"
28	chosen	"chose", "chosen", "choose", "chooses"
29	simplest	"simplest"
30	overlap	"overlap", "overlaps"
31	travel	"travel", "travels"
32	obtain	"obtain", "obtained"
33	identical	"identical"
34	stand	"stand", "stands"
35	join	"join", "joins", "joined"
36	lose	“lose”, "loses", "lost"
37	equal	“equal”, "equals", "equally"

APPENDIX A (continued)

Table A. *Dictionary of Features for Weighted Features Model (continued).*

38	sign	"positive", "negative", "sign", "signs"
39	show	"show", "shows"
40	original	“original”, "originals"
41	write	"write", "writes", "written", "wrote"
42	give	“give”, "gives", "given", "gave"
43	cover	“cover”, "covers"
44	spend	"spend", "spent"
45	sell	"sell", "sold"
46	paint	“paint”, "paints", "painted"
47	contain	"contain", "contains", "contained"
48	arrange	"arrange", "arranges", "arranged"
49	express	“express”, "expresses", "expressed"
50	without	"without"
51	inclusive	"inclusive"
52	half	"half"
53	entire	"entire"
54	start	"start", "starting"
55	buy	"buy", "buys"
56	note	"note", "notes"
57	work	"work", "works"
58	probability	"probability", "probabilities", "chance", "probable", "likely"
59	top	"top", "tops"
60	multiplication	"multiplication", "multiply", "multiplied", "multiplies", "times", "product", "products"
61	division	"division", "divide", "divided", "divides", "quotient", "quotients"
62	divisible	"divisible", "divisor", "factor"
63	addition	"addition", "add", "added", "adds", "plus"
64	congruent	"congruent"
65	consecutive	"consecutive"
66	money	"money"
67	correct	"correct"
68	together	"together"
69	possible	"possible"
70	value	"value", "values"
71	area	“area”, "areas"
72	total	"total", "totals"
73	rectangle	"rectangle", "rectangles"
74	figure	"figure", "figures"
75	point	"point", "points"
76	time	"time"
77	cent	"cent", "cents"
78	order	"order", "orders"
79	coin	"coin", "coins"
80	perimeter	"perimeter", "perimeters”
81	length	"length", "lengths"

APPENDIX A (continued)

Table A. *Dictionary of Features for Weighted Features Model (continued).*

82	length_unit	"meter","meters","centimeter","centimeters","inch","inches", "millimeter", "millimeters", "foot", "feet", "mile", "miles"
83	clock	"clock", "clocks", "standard clock", "standard clocks"
84	time_unit	“second”, "seconds", "minute", "minutes", "hour", "hours"
85	calendar_unit	"month", "months", "day", "days", "week", "weeks"
86	year	"year", "years"
87	circle	"circle", "circles", "circular", "diameter", "radius", "circumference"
88	box	"box", "boxes"
89	side	"side", "sides"
90	letter	"letter", "letters"
91	student	"student", "students"
92	sequence	"sequence", "sequences"
93	remainder	"remainder", "remainders"
94	form	"form", "forms", "formed"
95	result	"result", "results"
96	line	"line", "lines"
97	stamp	"stamp", "stamps"
98	example	"example", "examples"
99	term	"term", "terms"
100	group	"group", "groups"
101	row	"row", "rows"
102	diagram	"diagram", "diagrams"
103	region	"region", "regions"
104	triangle	"triangle", "triangles"
105	price	"price", "prices"
106	coin_types	"nickel", "nickels", "penny", "pennies", "dime", "dimes", "quarter", "quarters"
107	set	"set", "sets"
108	answer	“answer”, "answers"
109	game	“game”, "games"
110	child	“child”, "children"
111	list	“list”, "lists"
112	multiple	“multiple”, "multiples"
113	rate	“rate”, "rates", "speed", "speeds", "per hour", "per minute", "per second"
114	cut	"cut", "cuts"
115	school	"school", "schools"
116	book	"book", "books"
117	paper	"paper", "papers"
118	store	"store", "stores"
119	page	"page", "pages"
120	age	“age”, "ages"
121	amount	"amount", "amounts"
122	team	“team”, "teams"
123	trip	"trip", "trips"

APPENDIX A (continued)

Table A. *Dictionary of Features for Weighted Features Model (continued).*

124	class	"class", "classes"
125	integer	"integer", "integers"
126	score	"score", "scores"
127	factor	"factor", "factors"
128	column	"column", "columns"
129	car	“car, "cars"
130	color	"color", "colors"
131	fraction	"fraction, "fractions"
132	case	"case", "cases"
133	way	"way", "ways"
134	girl	"girl, "girls"
135	piece	"piece, "pieces"
136	midpoint	"midpoint", "midpoints"
137	size	"size", "sizes"
138	corner	"corner", "corners"
139	segment	"segment", "segments"
140	center	"center", "centers"
141	edge	"edge", "edges"
142	place	"place", "places"
143	countingnumbers	"counting number", "counting numbers"
144	linesegment	"line segment", "line segments"
145	dollars	"dollar", "dollars", "bill", "bills", "dollar bill", "dollar bills"
146	squares	"square", "squares"
147	cubes	"cube", "cubes"
148	missingdigits	"blank", "blanks", "missing digit", "missing digits", "missing entry", "missing entries", "blank space", "blank spaces"
149	perfectsquares	"perfect square", "perfect squares", "square number", "square numbers"
150	wholenumbers	"whole number", "whole numbers"
151	primenumbers	"prime number", "prime numbers", "prime", "primes"
152	area_units	"square unit", "square units", "square yard", "square yards", "square meter", "square meters", "square centimeter", "square centimeters", "square foot","square feet", "square inch", "square inches"
153	average	"arithmetic mean", "mean", "average", "averages"
154	pagenumbers	"page number", "page numbers"

APPENDIX B

Table B. *List of 137 Statistically Significant Weighted Features.*

Feature	Estimate	Standard Error	t value	Pr(>|t|)	
comparison	-8.56066	2.928106	-2.92362	0.003547	**
area	-7.77352	2.347475	-3.31144	0.000965	***
length_unit	-7.66779	2.357979	-3.25185	0.00119	**
area_units	-7.15393	2.362285	-3.0284	0.002529	**
XFG	-7.06517	2.348576	-3.00828	0.002701	**
faces	-6.84461	2.055646	-3.32967	0.000905	***
squares	-6.62331	2.274425	-2.91208	0.00368	**
perimeter	-6.60139	1.778202	-3.7124	0.000218	***
value	-6.47553	1.987103	-3.25878	0.001161	**
fraction	-6.31317	1.277191	-4.94301	9.19E-07	***
cubes	-6.15859	1.468199	-4.19466	3.01E-05	***
dimension	-5.98196	1.884494	-3.17431	0.001553	**
countingnumbers	-5.95478	1.674143	-3.55691	0.000395	***
possible	-5.66795	1.482483	-3.82328	0.000141	***
rate	-5.65141	1.414361	-3.99574	6.98E-05	***
way	-5.60338	1.292156	-4.33646	1.61E-05	***
multiplication	-5.48979	1.959249	-2.80198	0.005189	**
time_unit	-5.22999	1.694537	-3.08638	0.002089	**
total	-5.20215	1.768257	-2.94197	0.003346	**
letter	-5.08382	1.432769	-3.54825	0.000408	***
length	-5.05044	1.708779	-2.95558	0.003203	**
segment	-5.03589	1.664739	-3.02503	0.002557	**
form	-4.95331	1.540861	-3.21464	0.001353	**
rectangle	-4.93875	1.603244	-3.08048	0.00213	**
inclusive	-4.9362	1.216774	-4.05679	5.41E-05	***
amount	-4.92261	1.302651	-3.77892	0.000168	***
set	-4.90082	1.250873	-3.91792	9.61E-05	***
dollars	-4.7541	1.697348	-2.8009	0.005206	**
figure	-4.71197	1.439255	-3.2739	0.001102	**
give	-4.55424	1.330179	-3.42378	0.000646	***
time	-4.5473	1.484309	-3.06358	0.002253	**
XIEX	-4.49453	1.035054	-4.34231	1.57E-05	***
row	-4.47219	1.244651	-3.59313	0.000345	***
entire	-4.43621	1.477403	-3.00271	0.002751	**
show	-4.43605	1.291074	-3.43594	0.000618	***
cent	-4.43463	1.278704	-3.46807	0.000549	***
cut	-4.41092	1.361734	-3.23919	0.001243	**
obtain	-4.39415	1.204066	-3.64942	0.000278	***

APPENDIX B (continued)

Table B. *List of 137 Statistically Significant Weighted Features.*

primenumbers	-4.39059	1.222304	-3.59206	0.000346	***
average	-4.36154	1.346755	-3.23856	0.001246	**
region	-4.34828	1.535409	-2.832	0.00473	**
wholenumbers	-4.32356	1.602685	-2.6977	0.007114	**
circle	-4.32238	1.432148	-3.01811	0.002616	**
point	-4.31913	1.542897	-2.79936	0.005231	**
comparison_size	-4.29911	1.647128	-2.61006	0.009204	**
diagram	-4.23157	1.312191	-3.22481	0.001306	**
side	-4.16744	1.96161	-2.1245	0.033903	*
contain	-4.16657	1.472152	-2.83026	0.004756	**
coin	-4.16397	1.169009	-3.56197	0.000388	***
calendar_unit	-4.16061	1.42403	-2.92171	0.003569	**
comparison_length	-4.13452	1.569304	-2.63462	0.008569	**
size	-4.13262	1.164399	-3.54914	0.000407	***
start	-4.11071	1.298215	-3.16643	0.001596	**
factor	-4.09712	1.182231	-3.46558	0.000554	***
without	-4.09018	1.276911	-3.20318	0.001407	**
example	-4.08419	1.397142	-2.92325	0.003552	**
page	-4.05664	1.366395	-2.96887	0.003069	**
remainder	-4.05514	1.327674	-3.05432	0.002323	**
line	-4.03818	1.444943	-2.7947	0.005306	**
division	-4.0267	1.572227	-2.56114	0.010596	*
work	-4.0259	1.200274	-3.35415	0.00083	***
term	-4.0109	1.274267	-3.14761	0.001701	**
XIX	-4.00912	1.402086	-2.8594	0.004343	**
equal	-3.99854	1.3604	-2.93924	0.003375	**
corner	-3.98555	1.295437	-3.07661	0.002158	**
missingdigits	-3.95686	1.061566	-3.72738	0.000206	***
perfectsquares	-3.95078	1.23604	-3.19632	0.001441	**
correct	-3.87211	1.270253	-3.0483	0.002369	**
together	-3.84069	1.301357	-2.95129	0.003247	**
original	-3.8193	1.14603	-3.33264	0.000896	***
list	-3.77688	1.174285	-3.21632	0.001345	**
store	-3.74907	1.193091	-3.14232	0.001732	**
probability	-3.74095	1.257633	-2.97459	0.003013	**
box	-3.73837	1.310169	-2.85335	0.004426	**
consecutive	-3.73782	1.280139	-2.91985	0.00359	**
sell	-3.7054	1.070458	-3.46151	0.000563	***
divisible	-3.68493	1.224189	-3.0101	0.002685	**

APPENDIX B (continued)

Table B. *List of 137 Statistically Significant Weighted Features.*

parity	-3.67183	1.161449	-3.16142	0.001623	**
triangle	-3.66574	1.335187	-2.74549	0.006164	**
result	-3.65073	1.257384	-2.90344	0.003782	**
XIF	-3.63189	1.228351	-2.95672	0.003191	**
paper	-3.62484	1.225149	-2.95869	0.003171	**
game	-3.60754	1.069532	-3.37301	0.000776	***
simplest	-3.60421	1.288529	-2.79715	0.005267	**
team	-3.58828	1.082918	-3.31353	0.000958	***
book	-3.53402	1.132791	-3.11974	0.001868	**
XDX	-3.51642	1.253904	-2.80437	0.005151	**
group	-3.51355	1.060836	-3.31206	0.000963	***
weight	-3.49958	1.018847	-3.43485	0.00062	***
join	-3.48131	1.243503	-2.7996	0.005227	**
order	-3.47578	1.425466	-2.43835	0.014949	*
trip	-3.41107	1.189771	-2.867	0.004241	**
case	-3.31843	1.002063	-3.3116	0.000965	***
spend	-3.28528	1.434483	-2.29022	0.022241	*
multiple	-3.28362	1.162398	-2.82487	0.004836	**
paint	-3.28159	1.215741	-2.69925	0.007081	**
arrange	-3.2287	1.203454	-2.68286	0.007435	**
school	-3.22715	1.099789	-2.93434	0.003428	**
write	-3.20394	1.238797	-2.58633	0.009858	**
money	-3.19	1.147451	-2.78008	0.005549	**
class	-3.18658	1.021131	-3.12063	0.001863	**
sequence	-3.14839	1.073845	-2.93189	0.003455	**
year	-3.1462	1.18404	-2.65717	0.008021	**
student	-3.14219	1.098285	-2.861	0.004322	**
XFM	-3.13153	1.080553	-2.89808	0.003847	**
XIE	-3.12448	1.261113	-2.47756	0.013413	*
price	-3.11633	1.094055	-2.84842	0.004495	**
XSYM	-3.10967	1.317677	-2.35997	0.018491	*
column	-3.09103	1.305229	-2.36819	0.018088	*
car	-3.083	0.98305	-3.13616	0.001768	**
XFMDS	-3.06831	1.201007	-2.55478	0.01079	*
answer	-3.06307	1.170616	-2.61663	0.00903	**
XFMDM	-3.04925	1.214731	-2.51023	0.012241	*
lose	-3.04914	1.310708	-2.32633	0.020224	*
color	-3.02826	0.959321	-3.15667	0.001649	**
coin_types	-2.99227	1.202278	-2.48883	0.012998	*

APPENDIX B (continued)

Table B. *List of 137 Statistically Significant Weighted Features.*

child	-2.98787	1.038783	-2.87632	0.004119	**
integer	-2.98441	1.12807	-2.64559	0.008298	**
travel	-2.97391	1.352729	-2.19845	0.028173	*
XLS	-2.97259	1.169681	-2.54137	0.01121	*
piece	-2.91961	1.013607	-2.88042	0.004067	**
chosen	-2.90825	1.146168	-2.53737	0.011338	*
place	-2.88961	1.106444	-2.61162	0.009163	**
stand	-2.80022	1.095444	-2.55624	0.010746	*
clock	-2.78582	1.102901	-2.5259	0.011712	*
place_value	-2.76846	1.137145	-2.43457	0.015104	*
cover	-2.73427	1.173349	-2.33031	0.020012	*
sign	-2.7162	1.090517	-2.49074	0.012929	*
age	-2.62566	1.04647	-2.50907	0.012281	*
stamp	-2.60741	0.963844	-2.70523	0.006956	**
girl	-2.54541	1.006007	-2.53022	0.01157	*
note	-2.48595	1.23254	-2.01693	0.044002	*
comparison_age	-2.47891	1.020603	-2.42887	0.015343	*
XDE	-2.4388	1.058869	-2.30321	0.021496	*
buy	-2.43854	1.141608	-2.13606	0.032946	*
score	-2.21886	1.099219	-2.01858	0.04383	*
(Intercept)	5.058634	1.80487	2.80277	0.005176	**

Note. Significance codes: *** p < 0.001 ** p < 0.01 * p < 0.05