Automatic Analysis of Language Use in K-16 STEM Education and Impact on Student Performance

Farah Nadeem

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Mari Ostendorf, Chair

Eve Riskin

Hannaneh Hajishirzi

Program Authorized to Offer Degree:
Electrical and Computer Engineering

University of Washington

Abstract

Automatic Analysis of Language Use in K-16 STEM Education and Impact on Student Performance

Farah Nadeem

Chair of the Supervisory Committee:
System Design Methodologies Professor Mari Ostendorf
Electrical & Computer Engineering

There is a growing community of research focusing on educational applications of natural language processing (NLP). The applications tend to focus on analysis of student writing for scoring and feedback, and analysis of language learning. There has been less focus on analysis of language use in educational content, like assessment questions and textbooks, which is largely an expert driven process. This work examines this space, presenting automated tools for analysis of language use in K-16 science, technology, engineering and mathematics (STEM) education, and demonstrates the utility of automatically extracted features in studying student performance. This work also serves to bridge research in educational measurement and machine learning, providing a machine learning framework for analysis of factors that contribute to the difficulty of science assessment items.

Within the broader umbrella of language use, this work focuses on two aspects: language difficulty (or linguistic complexity), and gender representation. Linguistic complexity has been studied from both the expert driven educational perspective and in the context of machine learning and NLP based tools. For the latter, models have shown a high agreement with expert annotation for longer documents, however have not been shown to work well for shorter, informational texts. This work presents a discourse aware hierarchical neural model for classification of linguistic complexity quantified as grade level, demonstrated to work accurately for shorter texts, achieving state-of-the-art performance. Unlike most existing NLP based methods, the performance of our model is also validated for the downstream task of predicting student performance, where we find an impact both

for K-12 and college level STEM assessments. The model for classification also generalizes to other text classification problems.

Educational measurement research for prediction of difficulty of assessments questions is important in the context of assessment design and analysis of student learning. To understand the relative importance of factors impacting difficulty, many past studies have relied on use of linear models for predicting item difficulty given item characteristics. Some more recent work has looked at non-linear tree-based ensemble methods, but without analysis to identify important item characteristics. In our work with linear methods, we provide specific examples showing that the commonly used assumptions of feature independence and linear relationship between features and difficulty do not hold in practice. We also use non-linear ensemble models for the prediction problem, but unlike previous work, present a robust analysis of model performance, and apply recently introduced methods of feature interpretation to analyze aspects that contribute to question difficulty. Our results demonstrate that some item characteristics, including linguistic complexity, have a non-linear impact on item difficulty.

Analysis of how gender roles are depicted in content, including assessment questions, is also a growing area of research in the educational space. This is important since negative stereotypes can impact both student performance and retention of students in STEM. Expert annotation for this task is very time consuming and can be prohibitively expensive for large text collections. Our work presents NLP based methods to automate this process for STEM textbooks and middle school assessment items. Specifically, we extract gendered mention counts, more nuanced aspects of roles, agency and authority of gendered characters, and activity characteristics. Using these features, we develop tools for analysis of content and assessments for gender biases, showing that biases exist both in terms of the frequency with which masculine and feminine characters appear in the texts, as well as in terms of the activities, roles, agency and authority of these mentions.

Together, these results show the utility of NLP tools for analysis of language use in educational content, providing downstream validation with analysis of student performance. Our findings demonstrate that NLP-based analysis tools can identify sources of difficulty even in expert-curated educational content.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am incredibly grateful to my advisor Professor Mari Ostendorf for her mentorship and support. I would also like to thank my committee members and collaborators, Eve Riskin, Hannaneh Hajishirzi, Min Li, Paula Heron, Peter Shaffer, Dongsheng Dong, Sheh Lit Chang and Gabriella Gorsky, who contributed their time and expertise for this work.

My colleagues, friends and family have played a huge role in my journey, providing support, encouragement, help and mentorship, including Trang Tran, Kevin Lybarger, Sara Ng, Kevin Everson, Ellen Wu, Roy Lu, Sitong Zhou, Micheal Lee, Aaron Jaech, Ji He, Hao Fang, Vicky Zayats, Yi Luan, Hao Cheng, Elizabeth Oestreich, Niveditha Kalavakonda, Glenna Chang, Megan Kennedy, Maarten Sap, Tarah Helliwell, Sandamali Devadithya, Irina Tolkova, Abid Ahmad, Asma Ahmed, Saadiya Nadeem, Daniella Suarez, Giuliana Conti, Shugla Kakar, Stephanie Swanson, Brenda Larson, Patrick Heneghan, Meghann Gerber, Natacha Kune, Bree Callahan and Lincoln Johnson.

For the work on linguistic complexity classification I also thank Dr. Meurers, Professor University of Tubingen, and Dr. Vajjala-Balakrishna, National Research Council Canada, for sharing the WeeBit training corpus, their trained readability assessment model and the Common Core test corpus.

DEDICATION

To my son Mustafa, who inspires me everyday to be my best self.

Chapter 1

INTRODUCTION

Performance in science, technology, engineering and math (STEM) education in K-12 impacts post-secondary achievement and career outcomes. National level data shows achievement gaps in performance based on race or ethnicity, socioeconomic status (SES), gender and English language learner (ELL) status (White and Rotermund, 2019; Hussar et al., 2020; Maarouf, 2019), and that historically underrepresented groups still make a small part of the STEM workforce (Khan et al., 2020; Funk and Parker, 2018). The primary indicators of achievement include performance on standardized assessments and attainment (e.g. graduation and employment). Achievement gaps in STEM are a concern both for K-12 and 2 and 4-year colleges. Several factors contribute to this gap, including community and family support, aspects of school resources and environment, teaching methodology, educational content, and assessment methods. Within these factors, our work focuses on educational content and assessments.

Assessments, which are one of the primary tools for measuring achievement (Taylor and Nolen, 2005; National Research Council, 2001), and educational content like textbooks, which are crucial for K-12 and college education, are ideally equitable tools for learning and teaching. However, they can often carry unintended biases that impact students as well as the validity of assessment scores. Various characteristics of assessment questions have a complex impact on difficulty level and measurement bias for different student groups. This in turn affects students' understanding and performance. This is also true for other educational content like textbooks and articles.

Different elements of language use in educational content can cause unintended biases. The aspect of linguistic complexity is a particular consideration for ELLs, which make up just over 10% (or five million) of all students in K-12 (Hussar et al., 2020). A study looking at State level assessments for 5th grade showed that students who are English language learners were more likely to answer incorrectly, despite understanding the underlying concepts being tested in the questions (Noble et al., 2012). Another consideration is the presence of negative stereotypes about students'

identity groups, which can negatively impact their performance. For gender identity this includes prescriptive gender stereotypes (Prentice and Carranza, 2002) and unconscious gender-stereotypical cues (Cheryan et al., 2009). Cues about ethnic groups can also be detrimental to students, propagating microaggressions on a macro level (Allen et al., 2013). There is an increased focus on examining educational materials for biases, and studies have shown that some textbooks perpetuate stereotypes (Deckman et al., 2018; Lee and Chin, 2019).

Research in (Santelices and Wilson, 2010) has indicated that SAT questions show negative biases towards African American students, validating similar findings presented earlier in (Freedle, 2003). A study on SAT scores from 2003 comparing the impact of family income for African American and White students found an impact on both groups, greater for African American students (Dixon-Román et al., 2013). Ethnic biases have also been explored in physics (Henderson and Stewart, 2018), with Caucasian students outperforming African American students. Gender-based performance differences have been explored by the work in (Hazari et al., 2007), showing women perform worse than men on introductory physics when controlling for academic preparedness. Research studies generally indicate that the impact of student demographics are intertwined with other factors. Investigating the extent to which performance differs by demographic, and how this effect is dependent on other factors, is useful for informing the scaffolding students need coming into the classroom to be successful.

It is important to identify biases and quantify language difficulty for educational resources, however current methods for doing so are largely expert driven, where annotation is time consuming. Additionally, open source educational content for K-12 is increasingly available, which gives teachers the flexibility to adapt the content to suit their students (de los Arcos et al., 2016). However, K-12 classrooms are often constrained in terms of resources (Leachman et al., 2017), and it is difficult to assess this content for being suitable for students. It can often be prohibitively expensive to get expert annotation for language use that can negatively impact students for this set of resources.

The broad aim of this work is to automate the analysis of language use in educational texts by creating machine learning (ML) and natural language processing (NLP) tools, and provide methods to study the impact of automatically extracted text features on student performance. The goal of these tools is to help educators, including teachers and assessment writers, and give better feedback to students, with the ability to scale to large data sets without the need for extensive expert annota-

tion. We focus on two aspects of language use, linguistic complexity and gender representation, for educational texts and assessments. Pejorative stereotypes based on ethnicity and socio-economic status are also important to identify in this context; however, automatic analysis of socio-cultural aspects of language use is less well developed than gendered language use, and will not be a focus of our work.

1.1 Problem Statement

Our work addresses two key computational problems:

- automatic analysis of language use in STEM educational texts (profiling), both content and assessments, in terms of linguistic complexity and gender representation, and

- methodologies for identifying the impact of the automatically extracted text features for STEM assessments on student performance, and interaction effects with other predictors, including item format and cognitive demand, as well as student demographics.

Our methods are developed on K-12 STEM texts and assessed on middle school STEM and college physics assessments.

For STEM environments, linguistic complexity of assessments and texts can potentially impact student performance regardless of subject knowledge, particularly for English language learners. To study this impact, language difficulty needs to be quantified. Several methods have been proposed, which are either based on expert annotation, or automatically extracted. Expert annotation is typically expensive and time consuming. For automated measures, shallow count-based features do not perform well for informational/STEM texts (Sheehan et al., 2013), and feature based systems generally do not perform well for short texts (Nadeem and Ostendorf, 2018). Our work aims to provide more accurate linguistic complexity scores for STEM texts, including short items. Automatic profiling of gendered language in the educational domain remains a largely unexplored area. Like linguistic complexity, there have been expert annotation efforts, however these do not provide the ability to scale to large corpora. Within the domain of educational applications of NLP, there is research on analysis of student writing, typically for language learning, now expanding to STEM (Cahill et al., 2020; Lee et al., 2019; Riordan et al., 2020). However these systems are trained to

analyze student writing, not educational resources. Specifically, there has been very little work looking at automated analysis of language of STEM educational resources, both content and assessment items.

The impact of cognitive demand (e.g. skills needed to solve a problem and content knowledge) and linguistic characteristics of assessment questions (items) on students' performance, particularly for diverse groups, is difficult to predict. The existing methodologies for analysis of factors that contribute to the difficulty of assessment questions often rely on assumptions of feature independence and linearity. Given the factors used in the prediction models, these assumptions often do not hold in practice, which we hypothesize has led to mixed findings about factors that contribute to difficulty of assessments. Recent work has examined non-linear models, including ensemble methods, for difficulty analysis (Sinharay, 2016). However, there has not been research on interpretable feature analysis for these models to identify what is contributing to item difficulty, which is very important in the educational context. Our work aims to provide a methodology that benefits from the added power of non-linear models but also provides robust interpretation of features and feature interactions when working with small item collections. In addition, a goal is to provide methods that apply to aggregate student performance as well as individual student responses where it is possible to consider demographic factors.

1.2 Contributions

Broadly, our work makes contributions in automated analysis of language use in the educational context, focusing on K-16 STEM. We highlight the utility of this analysis in examining student performance, providing downstream validation of automated measures, and contextualizing existing educational measurement research on analysis of assessment difficulty. While most existing research on educational applications of NLP looks at automated student response scoring and aspects of language acquisition, our work presents analysis of content created for teaching and assessing students, which has been largely unexplored.

Specifically, we focus on the aspects of linguistic complexity and gender representation in language use. Our work presents a novel discourse-aware neural architecture for classification of linguistic complexity that achieves state-of-the-art performance for short texts and generalizes to other text classification problems. The ability to automatically and accurately predict linguistic complex-

ity is important for identifying appropriate content for students with diverse language proficiency, e.g. in the case of English language learners. Additionally, accurate results on short texts allows the application of the classifier to assessment questions. In addition to algorithmic advances, a key contribution of this work is the validation of the automated system predictions in analysis of question difficulty, providing insight into how linguistic complexity interacts with other features of assessment questions in contributing to overall difficulty.

In analysis of difficulty of assessment questions, our work introduces machine learning methodologies to the established problem of item difficulty analysis in educational measurement research, linking the terminology and analysis approaches for the two fields. In this context we make three key contributions: i) demonstrating the utility of using hold-out testing in reporting results for small datasets, ii) highlighting limitations of existing work in terms of model assumptions and feature analysis, and iii) applying tools to interpret more powerful tree-based ensemble methods presented in (Lundberg et al., 2020). While these methods are not new to machine learning, they have not been adopted in educational measurement research, and we argue that the prior work analyzing fit to the training data has resulted in misleading findings. Our work provides more generalizable methods of exploring feature importance. We build on this analysis for introductory calculus-based physics, and show that linguistic complexity, where predictive of student performance, shows an impact similar to that identified for middle school science.

For gender representation, we build a system to profile text for gendered language use based on automatically extracting counts of feminine, masculine and gender neutral mentions presented in a text, as well as the agency and authority of associated verbs. In addition to distributional analysis, we develop a tool that can be used by educators to automatically identify gender biases in texts. We identify examples of highly biased texts, which underscores the importance of assessing resources available on the internet, including educational content, as well as datasets used for training ML/NLP systems. By providing a bias profiling tool, our work makes an important contribution towards assessing educational corpora for biases that can negatively impact students. This is significant in the context of increasing adoption of open source educational material in K-12, since it is generally expensive and infeasible to carry out expert annotation of large existing datasets. It can also be used to identify unintended sources of gender bias in curated assessments, which is important because negative stereotypes can impact student performance, creating unintended biases in

measurement of learners' proficiency.

1.3 Thesis Overview

The remainder of the thesis is organized as follows: Chapter 2 covers relevant issues and prior work in educational measurement, including impact of gender representation and student demographics in education. This is followed by an overview of NLP techniques including text classification and automated parts-of-speech (POS) tagging systems relevant to our work. Next we look at some of the main themes for current NLP applications in education. More specific background for each of the experiments is detailed in the relevant chapters.

Chapter 3 introduces neural models for linguistic complexity classification, introducing a novel hierarchical neural network with bidirectional context with attention (BCA) that incorporates cross-sentence information sharing. We train the model to predict grade level of text, achieving state-of-the-art performance on short texts, including assessment items, specifically for informational texts such as science and mathematics. Classification results for shorter texts and assessment questions are presented and compared with the current state-of-the-art feature-based system. In subsequent chapters we show a non-linear relationship between linguistic complexity of questions and student performance, providing external validation for complexity analysis, which has been largely unexplored for existing automated measures of complexity analysis.

Analysis of item difficulty and impact of linguistic complexity is described in Chapter 4, with an overview of existing literature on STEM item difficulty analysis. To contextualize existing research on analysis of item difficulty, this work provides new results and insights into prior work by conducting experiments with a variety of models and model selection criteria in studying the utility of different factors for predicting item difficulty for a set of middle school science items. The automatically identified linguistic complexity of individual items, in addition to other item characteristics (cognitive and response type), is used for analysis of student response data to study the impact of item characteristics on aggregate student performance. We investigate the generalizability of prior research methods in identifying features that are predictive of item difficulty. The analysis also provides insights into mixed and mostly negative findings in prior literature about the impact of linguistic complexity on student performance in STEM, showing that linguistic complexity is significant when non-linearity and interaction with other item features are incorporated in the model.

Analysis of assessments and student performance for introductory college-level calculus-based physics is presented in Chapter 5, with results for tests broken down by gender and ethnicity, as well as correlation of question difficulty with predicted linguistic complexity. We look at answering the question of whether student demographics are predictive of performance, and find that this is true at the start of the course, however the impact decreases over the course of instruction. We further examine which factors are most predictive of performance, including gender, underrepresented minority status and linguistic complexity of questions, finding non-linear effects of linguistic complexity.

Analysis of gendered language in educational texts in presented in Chapter 6, with a focus on quantitative analysis for open source educational corpora and assessment items. Using standard part-of-speech tagging and parsing tools, gendered lexica, and verb framing, several features are extracted to characterize gendered language use. These features are then used in distributional analyses and automatic gender prediction to profile text from K-12 STEM textbooks and middle school math and science assessments, highlighting bias in widely used resources and carefully curated assessments.

Chapter 7 concludes the thesis, with a summary, discussion of broader impact of the methods, and directions for future work.

Chapter 2

BACKGROUND

This chapter gives a broad overview of relevant research in education, NLP tools and methods, educational applications that leverage NLP, and text corpora used that are used for this thesis. The educational aspect includes research on what contributes to difficulty of assessment questions, focusing on commonly used analysis methods and limitations, followed by a focused discussion on linguistic complexity since it is a focus of this work. In the domain of NLP, we overview existing work on automatic text classification methods, which we build on for our work on automatic profiling and present methods for automated annotation of linguistic structure, which we use in our bias profiling analysis. In the space of NLP applications for education, we focus on automated student response scoring (the predominant area of research in automated language analysis in education), and readability analysis. We wrap up with educational corpora used in this work, including datasets compiled for readability analysis, text simplification, and question answering, and a corpus of open source K-12 textbooks, as well as sets of assessment items, some with associated student performance and demographic data.

2.1 Educational Measurement

The key focus of research in the space of educational measurement is to provide methods to create tests that reliably measure student progress, and to identify factors that impact validity with the goal of enhancing student learning (Brookhart and McMillan, 2019; Pidgeon and Yates, 2018). This section provides a brief overview of existing research relevant to our work in this domain.

2.1.1 Item difficulty analysis

The ability to accurately identify how item characteristics influence performance is required for assessment writing, and for diagnosing student learning to guide teaching strategies. Analysis can focus on both aggregate and individual student performance. The effects of item characteristics on

student test performance has been the subject of research in different subjects (Little and Jones, 2010; Reisslein et al., 2010), ranging from problem solving (Council et al., 1992; Marshall, 1995), mathematics (Hickendorff, 2013), statistics (Quilici and Mayer, 1996; Scheines et al., 2007), economics (Cronin et al., 2009), and physics (Chi et al., 1981; McDermott et al., 1987). Item characteristics that are studied include the skills and content knowledge required to answer the question (cognitive demand), the response type and presence of illustrations (format), and linguistic complexity. In looking at student related factors, prior research has identified aspects like past grades and classroom performance, demographics, and socio-economic information as predictive of individual performance (Saa et al., 2019).

Most research on item difficulty using aggregate student response data uses linear regression models for analysis (Crisp et al., 2008; Crisp and Grayson, 2013; El Masri et al., 2017). The approach relies on the assumptions of feature independence and linear relationship between performance and predictive features. In addition, analyses are based on fitting a model to the full dataset, where the number of items is typically less than 200 and model fit criteria are likely to be biased. We present solutions to these challenges in our work, applying non-linear models to item difficulty prediction and using two hold-out methods for reporting results. We further use techniques from machine learning to analyze the contribution of individual features to overall difficulty for these models, as well as pairwise feature interaction.

For analysis of student related factors, in addition to linear models, existing research has explored decision trees and Naive Bayes classifiers (Saa et al., 2019). Item response theory (IRT) models, which are latent trait models, are also extensively used in analysis of items and student proficiency. Latent implies that the trait or class is not directly observable and is inferred from observed performance on assessment. The basic IRT model relates the probability of a correct response to student proficiency (latent trait) and item parameter(s). Item parameters define aspects of the item which make it difficult or easy to score well on it (Joachims, 1998). For analysis of performance differences based on student demographics, IRT can be used to estimate the parameters of an item for each student sub-population, which allows us to examine whether the parameters differ significantly across student groups, i.e. if the item exhibit differential item functioning (DIF). For tests that examine multiple skills, multi-dimensional IRT (MIRT) models are used for DIF analysis, including the Lord's wald test (Lee and Suh, 2018). A challenge with using this approach is that imbalance

in the number of respondents that belong to the focal group and reference group can lead to higher rates of error in detection of DIF. For our analyses of college level physics, the student demographics, e.g. female students and underrepresented minority students, make up a smaller fraction of the respondent group, which can lead to errors in detecting DIF using MIRT based approaches. Another challenge can be a small test size with less than 25 items, which is also true for our physics data. We present an alternate analysis method that does not build on MIRT, applying a non-parametric model to look at both test level and individual question level analysis, using feature analysis to flag the impact of student demographics.

2.1.2 Impact of linguistic complexity on student performance

A key focus of our work is linguistic complexity, which is considered a potential source of measurement error in STEM assessments. Unnecessary linguistic complexity may interfere with students' comprehension, leading to inaccurate score interpretations and uses. Previous work has shown that the impact of linguistic complexity varies across students. Some features may impact students regardless of their backgrounds, e.g. difficult vocabulary (Shaftel et al., 2006), while other features may be more biased against certain groups.

Ambiguous wording, difficult vocabulary, and syntactic complexity with longer sentences may contribute to the item difficulty, thus creating unnecessary comprehension difficulties for ELLs (Abedi and Lord, 2001). Wolf and Leon (2009) stressed the impact of academic vocabulary when it interacts with content difficulty on ELLs' performance on mathematics and science items. Specifically, when an item requires relatively easy content knowledge, the number of general academic vocabulary involved can disadvantage a language learner. While vocabulary can have a negative impact on language learners, length does not seem to. The language of science items may distract students from what is assessed (Li et al., 2017) and this language may allow for alternative interpretations (Kachchaf et al., 2016).

Prior research has shown mixed findings on the impact of linguistic complexity, often finding that language difficulty does not significantly contribute to item difficulty. A key shortcoming of these methods is the assumption of feature independence. As we show in our work, linguistic complexity is often correlated with other features of a question, and it is not possible to discern this

interaction when the model assumes feature independence. Another challenge is that there is no agreed upon measure for language difficulty. Most research studies present their own measure, such as length of text (which does not capture linguistic complexity), individual features like average age-of-acquisition and average length of sentences, commercial readability scores (Flesch-Kincaid, Coh-Metrics), or subjective difficulty levels such as high, medium and low, that are not explicitly identified with particular text characteristics. This makes it infeasible to draw conclusions that generalize across studies. Reading level prediction systems such as WeeBit (Vajjala and Meurers, 2012) give a more holistic score, combining several different features (including some used in education research) through machine learning. However they do not work well on short STEM texts, so we develop a new approach.

2.2 Natural Language Processing Methods

This section outlines general NLP techniques used in our work. We look at text classification methods, which we build on for linguistic complexity analysis and activity classification. We then describe current methods for automatically extracting linguistic structure in text, which we use for extracting gendered language.

2.2.1 Automatic Text Classification

Text classification systems label text sequences such as documents, paragraphs or sentences with one or more of a predefined list of tags. In our work this is used to label texts with linguistic complexity, and for math items for the type of activities associated with the context. Text classification is a well-studied problem with many different solutions available, including naive Bayes, support vector machines, decision trees and k-nearest neighbors, summarized in (Ikonomakis et al., 2005).

Currently, the most effective methods leverage neural networks. Unlike prior work, neural models of language represent a word in continuous vector space (referred to as word embeddings) using a mapping learned from a large text collection. After mapping the word sequence to a vector sequence, different neural architectures can be applied, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Convolutional neural networks (CNNs) have been shown to work well for text classification, particularly shorter sequences (Kim, 2014). Recurrent neural networks (RNNs) are adept at learning text representations, as demonstrated by language modeling

(Mikolov et al., 2010) and text classification tasks (Yogatama et al., 2017). Additional RNN structures have been proposed for improved representation, including tree LSTMs (Tai et al., 2015) and a hierarchical RNN (Yang et al., 2016). Hierarchical models have been shown to better represent document structure (Yang et al., 2016). However each sentence is encoded independently, with no cross-sentence information sharing, which fails to account for local discourse. Our work presents a variant of this neural network that incorporates cross-sentence information, and provides better performance than the basic hierarchical network for the task of linguistic complexity quantification.

Attention mechanisms were introduced to improve neural machine translation tasks (Bahdanau et al., 2015), and have also been shown to improve the performance of text classification (Yang et al., 2016). Attention weights are used to generate a single vector from a sequence of vectors. This attention weight is based on a score computed between the target hidden state h_t and a subset of the source hidden states h_s. In machine translation, attention is computed over the source sequence when predicting the words in the target sequence. For text classification, attention weights are learned that target the final classification decision. This approach is referred to as "self attention" in (Lin et al., 2017), but will be referred to as "task attention" for our work. The hierarchical RNN in (Yang et al., 2016) uses task attention mechanisms at both word and sentence levels. We propose extensions of the hierarchical RNN that leverage attention in different ways, including combining the concept of context attention from machine translation with task attention to capture interdependence of adjoining sentences in a document.

For supervised learning, annotated data are used to both train and evaluate models. Having more labeled data allows use of more complex models and leads to better performance, if the annotated training data is matched in character to the test data. When such data is limited, leveraging unlabeled external data has been shown to improve text classification accuracy. One popular method of leveraging external data is to use unsupervised (or self-supervised) learning to represent words and word sequences with a learning objective based on predicting words based on the neighboring context (Mikolov et al., 2013; Pennington et al., 2014). We use the pretrained word representations from (Pennington et al., 2014) as input to our linguistic complexity prediction model.

The most recent set of models that provide state-of-the art performance on a variety of NLP tasks, including classification, are transformers, which are designed to handle sequential data without relying on recurrence like RNNs. Transformers give contextualized word embeddings, which

capture word semantics in different contexts. Thus the same word can have different embeddings based on the context it appears in. One of the most widely used models is bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019). While transformers achieve state-of-the-art performance on several classification tasks, they require both extensive computing resources and data to train. For our work, we fine-tune smaller pretrained BERT models for classification of activities in math story problem contexts.

2.2.2 Annotating linguistic structure

Automatic part-of-speech (POS) tagging is labelling words with their parts of speech, such as `noun`, `verb`, `adjective` etc., based on their context. POS tags are not static, since the context can change the part-of-speech the word is used as. POS tagging is an established NLP problem, with systems achieving a high accuracy (Manning et al., 2014). We use the spaCy POS tagging implementation (Honnibal and Montani, 2017),[1] which uses a set of 12 universal POS tags presented in (Petrov et al., 2012), including `noun`, `verb`, `adjective`, `adverb`, `pronoun` etc. POS tagging allows us to extract nouns and pronouns that might be indicative of people in a text. It is also used in the downstream task of parsing to extract syntactic structure, discussed below.

Named entity recognition (NER) is automatically extracting entities from text, and categorizing them into predefined groups like people, places, object, date, value etc. Named entities can be represented by one word, e.g. `Seattle` (place), or by a span of words, e.g. `Mary Montagu` (person). Both feature-based and neural network based NER systems are able to achieve high accuracy on this task (Yadav and Bethard, 2018). Our work uses the NER extraction implementation from spaCy, which has been evaluated and shown to perform well across several datasets (Jiang et al., 2016). Named entity recognition allows for extraction of names that indicate people in a text.

Additionally, we can automatically extract the grammatical structure of a sentence using parsing, which provides the relationship between words and phrases. There are two types of parsing, constituency and dependency, based on the set of rules (or grammar) used to create the tree (Jurafsky and Martin, 2008). Constituency parsing depends on constituency or phrase-structure grammars, while dependency parsing uses dependency grammars to parse the tree. Phrase structure allows us

[1] https://spacy.io/

to organize words into nested groups, which can be combined into sentences. The constituency parsing of a sentence provides the constituents of a sentence, including verb phrase (`VP`), noun phrase (`NP`), or prepositional phrase (`PP`), and the words that make up these phrases. On the other hand, dependency structures indicate how words modify or are modified by other words. A dependency parse of a sentence provides the relationships between words and their root words, e.g. `subject` of `verb`. For a person, we can use dependency parsing to identify aspects like whether the mention appears as a `subject` or `object` in the text, and whether it appears in a group with other mentions (`conjunction`). Both types of parsing are established NLP tasks, with various tools available. For the work presented in Chapter 3, we use both constituency and dependency parsing, using the tools presented in Stanford Core NLP (Manning et al., 2014).[2] The parser implementations are based on the work presented in (Zhu et al., 2013) for constituency parsing, and on (Chen and Manning, 2014) for dependency parsing. We also use dependency parsing for our work in Chapter 6, using the spaCy implementation based on the transition-based parsing system proposed in (Honnibal and Johnson, 2015).

2.3 Educational Applications of NLP

Research on educational applications for NLP spans a diverse array of topics, with a few of the key areas being automatic grading of student writing, readability assessment, grammatical error detection and correction, and computer aided language teaching. This section provides a brief overview of two applications, student response scoring and readability analysis which are most relevant to our work.

2.3.1 Student response scoring

Existing response scoring systems rely on feature-based and/or neural models. A number of systems have been developed for student response scoring, including essay scoring, constructed response scoring, content evaluation, as well as automatically generating feedback for students. These systems are important both for online learning platforms and online testing, as well as for resource-constrained classrooms.

[2] https://stanfordnlp.github.io/CoreNLP/

Automated essay scoring (AES) is the more established application, with commercially deployed systems like E-Rater® (Attali and Burstein, 2006). More recent systems often rely on neural methods, such as hierarchical recurrent neural networks (Nadeem et al., 2019), LSTM-based classifiers (Farag et al., 2018; Wang et al., 2018; Cummins and Rei, 2018), combination LSTMs and CNNs (Taghipour and Ng, 2016; Zhang and Litman, 2018), transformer-based methods (Mayfield and Black, 2020), as well as feature-based systems (Klebanov et al., 2016; Nguyen and Litman, 2018). Systems that use ensembling and/or combine neural and feature-based approaches (Liu et al., 2019; Taghipour and Ng, 2016) show better performance on smaller data sets than complex neural-models which benefit from large training sets. where it is difficult to train more complex neural models. For settings where the AES training corpora are small, ensembling is a promising direction, as in the case of the Automated Student Assessment Prize (ASAP) dataset.[3]

Automated scoring in the domain of STEM is also gaining interest. There has been recent work on grading complex mathematical responses (Cahill et al., 2020), science constructed response scoring (Riordan et al., 2020), and feedback for writing scientific arguments in real time (Lee et al., 2019). The system described in (Riordan et al., 2020) is currently being used by educators in an online learning system in the implementation of the next generation science standards (NGSS) to provide feedback to educators about student performance. The system C-Rater (Leacock and Chodorow, 2003) assigns full or partial credit to short constructed student responses for mathematics. When tested for 4th and 8th grade National Assessment of Educational Progress (NAEP) assessments, it showed an agreement of over 80% with human raters.

The key focus of existing work in this space is on scoring students' writing, which is a very useful application in the context of resource-constrained classrooms. It is related to our work in that it involves a text classification task in an educational context, so methods in this field may be of interest for our work and vice versa. It differs in that it is applied to student written text vs. the text of assessment items and content used in our work.

---

[3]https://www.kaggle.com/c/asap-aes

2.3.2 Readability analysis

There has been substantial work in developing measure of readability level (which we equate with linguistic complexity) that go beyond shallow count based features like the Flesch-Kincaid (Kincaid et al., 1975) and Coleman-Liau index (Coleman and Liau, 1975). The most common approach has been to leverage features that can be automatically extracted using NLP systems, including lexical, syntactic and discourse features, and use these as input to models for predicting reading level. Some systems include support vector machine classifiers (Schwarm and Ostendorf, 2005; Petersen and Ostendorf, 2009), and logistic regression (Feng et al., 2010). The work in (Vajjala and Meurers, 2012) introduced features from second language acquisition to improve performance on this task, which was expanded to include morphological, semantic and psycho-linguistic features (Vajjala and Meurers, 2012).

Most models are trained and tested on full length documents, with the exception of the feature-based model presented in (Vajjala and Meurers, 2012). When tested on middle school science assessment questions, however, this system did not perform well. A closer look revealed that, due to the short length of questions, the feature vector was sparse, which impacted the performance of the classifier. Since a key focus of our work is on analyzing assessment questions, we develop a novel text classification model that is designed to work well for STEM assessment questions.

Subsequent to the work described in this thesis, other neural deep learning models have been applied to this task. Researchers in (Deutsch et al., 2020) present readability assessment models, comparing feature-based, neural networks, and a combination of the two. Results indicate that for smaller training datasets, linguistic features help. For larger data sets, performance of pretrained models like transformers did not improve by the addition of linguistic features. This suggests that pretrained models might be implicitly learning the features used in traditional models. The work in (Martinc et al., 2019) presented results for automated readability analysis comparing transformers, LSTMs and hierarchical networks. It showed that neural models achieve results comparable to or better than feature-based models, but the performance varies across corpora, with BERT performing the best on one dataset, but doing worse than hierarchical networks on two other corpora.

2.4 Educational Datasets

We use two types of data in this work: textbook and other content sources, and items associated with student performance data. Some educational corpora have been compiled for training linguistic complexity classification models, including WeeBit (Vajjala and Meurers, 2012) which has texts for 5 different reading levels, and OneStopEnglish (Vajjala and Lučić, 2018) with variants of articles with the same content for three different reading levels. NewsELA (Staff) is a website for supporting English language learners in the classroom (Drinkwater, 2016), which has made available a corpus of news articles for research. The corpus contains versions of a news article modified for five reading levels. The aim of our work is to provide a model which quantifies the linguistic complexity as grade level. The above corpora provide a good starting point, however they are not as fine-grained as needed for grade-level prediction. To overcome this issue, we compile a data set of open source textbooks available online (Michigan; Siyavula; CK12) spanning grades Kindergarten through 12.[4]

Another key source of data for our study is STEM assessment items. For standardized published assessments, we use Program for International Student Assessment (PISA) science assessment items for 15-year-old students, and NAEP science and mathematics assessment items for Grade 8 students. For a limited number of PISA and NAEP data, we also have the aggregate student success rate (percentage of students in standardized testing who attempted the question correctly). We also use pretest data from college-level calculus-based physics, where we have three pretests, individual student performance and associated student demographics including gender, ethnicity and under-represented minority status.

Since the number of publicly available items with response data is limited, we use additional sources of assessments mainly collections compiled by AI2 researchers including items from the CK-12 Foundation (Kembhavi et al., 2017) and from Regents Examination (New York State Education Department NYSED), a collection of short answer science items used in SemEval 2013 (Dzikovska et al., 2016). We also use a large dataset of Algebra problems (AQuA) compiled for automatic question answering (Ling et al., 2017).

---

[4]Details of the textbook dataset can be found in Chapter 3.

Chapter 3

ESTIMATING LINGUISTIC COMPLEXITY FOR SCIENCE TEXTS

This chapter presents our work on automated models for linguistic complexity classification and results for texts of varying length and assessment questions.[1] A typical classroom presents a diverse set of students in terms of their reading comprehension skills, particularly in the case of English language learners (ELLs). Supporting these students often requires educators to estimate accessibility of instructional texts. To address this need, several automated systems have been developed to estimate text difficulty, including readability metrics like Lexile (Stenner et al., 1988), the end-to-end system TextEvaluator (Sheehan et al., 2013), and linear models (Vajjala and Meurers, 2014; Petersen and Ostendorf, 2009; Schwarm and Ostendorf, 2005). These systems leverage knowledge-based features to train regression or classification models. Most systems are trained on literary and generic texts, since analysis of text difficulty is usually tied to language teaching. Existing approaches for automated text complexity analysis pose two issues: 1) systems using knowledge-based features typically work better for longer texts (Vajjala and Meurers, 2014), and 2) complexity estimates are less accurate for informational texts such as science (Sheehan et al., 2013). In the context of science, technology and engineering (STEM) education, both problems are significant. Teachers in these areas have less expertise in identifying appropriate reading material for students as opposed to language teachers, and shorter texts become important when dealing with assessment questions and identifying the most difficult parts of instructional texts to modify for supporting students who are ELLs.

Our work specifically looks at ways to address these two problems. First, we propose recurrent neural network (RNN) architectures for estimating linguistic complexity, using text as input without feature engineering. Second, we specifically train on science and other informational texts, using the grade level of text as a proxy for linguistic complexity and dividing grades k-12 into 6 groups. We explore four different RNN architectures in order to identify aspects of text which contribute more

[1] This work was published in (Nadeem and Ostendorf, 2018)

to complexity, with a novel structure introduced to account for cross-sentence context. Experimental results show that when specifically trained for informational texts, RNNs can accurately predict text difficulty for shorter science texts. The models also generalize to other types of texts, but perform slightly worse than feature-based regression models on a mix of genres for texts longer than 100 words. We use attention with all models, both to improve accuracy, and as a tool to visualize important elements of text contributing to linguistic complexity. The key contributions of the work include new neural network architectures for characterizing documents and experimental results demonstrating good performance for predicting reading level of short science texts.

The rest of the chapter is organized as follows: section 3.1 looks at existing work on automated readability analysis and introduces RNN architectures we build on for this work. Section 3.2 lays out the data sources, section 3.3 covers proposed models, and section 3.4 presents results. Discussion and concluding remarks follow in sections 3.5 and 3.6.

3.1 Background

Studies have shown that language difficulty of instructional materials and assessment questions impacts student performance, particularly for language learners (Hickendorff, 2013; Abedi and Lord, 2001; Abedi, 2006). This has lead to extensive work on readability analysis, some of which is explored here.

Traditional reading metrics including Flesch-Kincaid (Kincaid et al., 1975) and Coleman-Liau index (Coleman and Liau, 1975) are often used to assess a text for difficulty. These metrics utilize surface features such as average length of sentences and words, or word lists (Chall and Dale, 1995). The development of automated text analysis systems has made it possible to leverage additional linguistic features, as well as conventional reading metrics, to estimate text complexity quantified as reading level. NLP tools can be used to extract a variety of lexical, syntactic and discourse features from text, which can then be used with traditional features as input to models for predicting reading level. Some of the models include statistical language models (Collins-Thompson and Callan, 2004), support vector machine classifiers (Schwarm and Ostendorf, 2005; Petersen and Ostendorf, 2009), and logistic regression (Feng et al., 2010). Text coherence has also been explored as a predictor of difficulty level in (Graesser et al., 2004), with an extended feature set that includes syntactic complexity and discourse in addition to coherence (Graesser et al., 2011).

A study conducted in (Nelson et al., 2012) indicates that metrics that incorporate a large set of linguistic features perform better at predicting text difficulty level; the metrics were specifically tested on the Common Core Standards (CCS) texts.[2] Features from second language acquisition complexity measures were used in a linear regression model in the WeeBit system (Vajjala and Meurers, 2012) to improve readability assessment. This feature set was further extended to include morphological, semantic and psycholinguistic features to build a readability analyzer for shorter texts (Vajjala and Meurers, 2014). A tool specifically built for text complexity analysis for teaching and assessing is the TextEvaluator™. While knowledge-based features offer interpretability, a drawback is that if the text being analyzed is short, the feature vector is sparse, and prediction accuracy drops (Vajjala and Meurers, 2014). This is particularly true for assessment questions, which are shorter than the samples most models are trained on.

Generally, for any text classification task, the type of text used for training the model is important in terms of how well it performs; training on more representative text tends to improve performance. The work in (Sheehan et al., 2013) shows that traditional readability measures under-estimate the reading level of literary texts, and overestimate that of informational texts, such as history, science and mathematics articles. This is due, in part, to the vocabulary specific to the genre. Science texts have longer words, though they may be easier to infer from context. Literary texts, on the other hand, might have simpler words, but more complicated sentence structure. The work demonstrated that more accurate grade level estimates can be obtained by two stage classification: i) classify the text as either literary, informational, or mixed, and then ii) use a genre-dependent analyzer to estimate the level. In an analysis on how well a model trained on news and informational articles generalizes to the categories in CCS, the work in (Vajjala and Meurers, 2014) shows better performance on informational genre than literary texts. Training on more representative text, however, requires genre-specific annotated data.

3.2 Data

For our work we consider grade level as a proxy for linguistic complexity. Within a grade level, there is variability across different genres, which students are expected to learn. Since there is no

[2] http://www.corestandards.org/

publicly available data set for estimating grade level and text difficulty aimed at informational texts, we created a corpus using online science, history and social studies textbooks, as mentioned in. Chapter 2. The textbooks are written for either specific grades, or for a grade range, e.g. grades 6-8. There are a total of 44 science textbooks and 11 history and social studies textbooks, distributed evenly across grades K-12. Given the distribution of textbooks for each grade level, we decide to classify into one of six grade bands: K-1, 2-3, 4-5, 6-8, 9-10 and 11-12. Because of our interest in working with short texts, we split the books into paragraphs, using end line as the delimiter.[3] In addition to the textbooks, we also used the WeeBit corpus (Vajjala and Meurers, 2012) for training, again split into paragraphs.

We have three different sources of test data: i) the CCS appendix B texts, ii) a subset of the online texts that we collected,[4] and iii) a collection of science assessment items. The CCS appendix B data is of interest because it has been extensively used for evaluating linguistic complexity models, e.g. in Sheehan et al. (2013); Vajjala and Meurers (2014). It includes both informational and literary texts. We use document-level samples from the CCS data for comparison to prior work, and paragraph-level samples to provide a more direct comparison to the information test data we created.

For the informational texts, we selected chapters from multiple open source texts. Since we had so few texts at the K-1 level, the test data only included texts from higher grade levels, as shown in table 3.1. The paragraphs in these chapters were randomly assigned to test and validation sets. Table 3.2 shows the number of textbooks in the train and test corpus by grade and subject.

To assess the models on stand alone texts, we assembled a corpora of science assessment questions from Khot et al. (2015); Clark et al. (2018), AI2 Science Questions Mercury,[5] and AI2 Science Questions v2.1 (October 2017).[6] This test set includes 5470 questions for grades 6-8 from sources including standardized state and national tests. The average length of a question is 49 words.

For training, two data configurations were used. When testing on the CCS data and the science assessment questions, there is no concern about overlap between training and test data, so all text

[3] In splitting the text into paragraphs, we are implicitly assuming that all paragraphs have the same linguistic complexity as the textbook, which is probably not the case. Thus, there will be noise in both the training and test data, so some variation in the predicted levels is to be expected.

[4] Available at https://tinyurl.com/yc59hlgj.

[5] http://data.allenai.org/ai2-science-questions-mercury/

[6] http://data.allenai.org/ai2-science-questions/

Grade Level	All chapters	Test set chapters
K-1	25	-
2-3	22	2
4-5	53	9
6-8	165	12
9-10	48	5
11-12	28	3

Table 3.1: Chapter-based test data split

Grades	Science	History & Social Studies	Grades	Science	History & Social Studies
K-1	2	2	7	5	1
2	1	1	8	6	1
3	2	1	9	2	0
4	4	1	10	2	0
5	5	1	11	1	0
6	4	1	12	1	0
6-8	7	0	9-12	2	2

Table 3.2: Open source textbooks

can be used for training. We held out 10% of this data for analysis, and the remaining text is used for the D_1 training configuration. Data statistics are given in table 3.3. About 20% of the training samples (5152) are from WeeBit, spread across grades 2-12. For testing on all three sets, we defined a training configuration D_2 that did not include any text from chapters overlapping with the test

Grade Level	Train Samples	Mean Length
K-1	739	24.42
2-3	723	62.05
4-5	4570	63.82
6-8	15940	74.79
9-10	3051	68.24
11-12	2301	75.28

Table 3.3: Training data (D_1) with mean length of text in words

data, so there training set is somewhat smaller than for D_1, except for grades K-1. The same WeeBit training data was included in both cases.

For the elementary grade levels, we have much less data than for middle school, and for high school, we have substantial training data with coarser labels (grades 9-12). To work around both issues, we first used all training samples to train the RNN to predict one of four labels (grades K-3, 4-5, 6-8 and 9-12). We then used the training data with fine labels to train to predict one of six labels. This approach was more effective than alternating the training.

3.3 Models for Estimating Linguistic Complexity

This section describes the four RNN structures for linguistic complexity estimation, including: a sequential RNN with task attention, a hierarchical attention network, and two proposed extensions of the hierarchical model using multi-head attention and attention over bidirectional context. In all cases, the resulting document vector is used in a final stage of ordinal regression to predict linguistic complexity. All systems are trained in an end-to-end fashion.

3.3.1 Sequential RNN

The basic RNN model we consider is a sequential RNN with task attention, where the entire text in a paragraph or document is taken as a sequence. For a document t_i with words K words w_{ik} $k \in \{1, 2, ..., K\}$, a bidirectional GRU is used to learn representation for each word h_{ik}, using a forward run from w_{i1} to w_{iK}, and a backward run from w_{iK} to w_{i1}.

$$\overrightarrow{h_{ik}} = \overrightarrow{GRU}(w_{ik}) \tag{3.1}$$

$$\overleftarrow{h_{ik}} = \overleftarrow{GRU}(w_{ik}) \tag{3.2}$$

$$h_{ik} = [\overrightarrow{h_{ik}}, \overleftarrow{h_{ik}}] \tag{3.3}$$

Attention is computed over the entire sequence α_{ik}, and used to compute the document representation v_i^{seq}:

$$u_{ik} = tanh(W_s h_{ik} + b_s) \tag{3.4}$$

$$\alpha_{ik} = \frac{exp(u_{ik}^T u_s)}{\sum_{ik} exp(u_{ik}^T u_s)} \tag{3.5}$$

$$v_i^{seq} = \sum_k \alpha_{ik} h_{ik} \tag{3.6}$$

The document vector is used to predict reading level. Since the grade levels are ordered categorical labels, we implement ordinal regression using the proportional odds model (McCullagh, 1980). For the reading level labels $j \in \{1, 2, ..., J\}$, the cumulative probability is modeled as

$$P(y \leq j | v_i^{seq}) = \sigma(\beta_j - w_{ord}^T v_i^{seq}), \tag{3.7}$$

where $\sigma(.)$ is the sigmoid function, and β_j and w_{ord} are estimated during training by minimizing the negative log-likelihood

$$\mathcal{L}_{ord} = \quad -\sum_i \log(\sigma(\beta_{j(i)} - w_{ord}^T v_i^{seq}) - \sigma(\beta_{j(i)-1} - w_{ord}^T v_i^{seq})) \tag{3.8}$$

3.3.2 Hierarchical RNN

While a sequential RNN has the capacity to capture discourse across sentences, it does not capture document structure. Therefore, we also explored the hierarchical attention network for text classification from (Yang et al., 2016). The model builds a vector representation v_i for each document t_i

with L sentences s_l, $l \in \{1, 2, .., L\}$, each with T_l words w_{lt}, $t \in \{1, 2, ..., T_l\}$. The first level of the hierarchy takes words as input and learns a representation for each word h_{lt} using a bidirectional GRU. Task attention at the word level α_{lt} highlights words important for the classification task, and is computed using the word level context vector u_w. The word representations are then averaged using attention weights to form a sentence representation s_l

$$\alpha_{lt} = \frac{exp(u_{lt}^T u_w)}{\sum_t exp(u_{lt}^T u_w)} \tag{3.9}$$

$$s_l = \sum_t \alpha_{lt} h_{lt}, \tag{3.10}$$

where $u_{lt} = tanh(W_w h_{lt} + b_w)$ is a projection of the target hidden state for learning word-level attention. The second level of the hierarchy takes the sentence vectors as input, learns representation h_l for them using a bidirectional GRU. Using a method similar to the word-level attention, a document representation v_i is created using sentence-level task attention α_l which is computed using the sentence level context vector u_s

$$\alpha_l = \frac{exp(u_l^T u_s)}{\sum_l exp(u_l^T u_s)} \tag{3.11}$$

$$v_i = \sum_l \alpha_l h_l, \tag{3.12}$$

where $u_l = tanh(W_s h_l + b_s)$ is analogous to u_{lt} at the sentence level. The word- and sentence-level context vectors, u_w and u_s, as well as W_w, W_s, b_w and b_s, are learned during training.

3.3.3 Multi-Head Attention

Work has shown that having multiple attention heads improves neural machine translation tasks (Vaswani et al., 2017). To capture multiple aspects contributing to text complexity, we learn two sets of word level task attention over the word level GRU output. These two sets of sentence vectors feed into separate sentence-level GRUs to give us two document vectors by averaging using task attention weights at the sentence level. The document vectors are then concatenated to form the document representation. The multi-head attention RNN is shown in figure 3.1.

3.3.4 Hierarchical RNN with Bidirectional Context

The hierarchical model is designed for representing document structure, however, the sentences within a document are encoded independently. To capture information across sentences, we extend

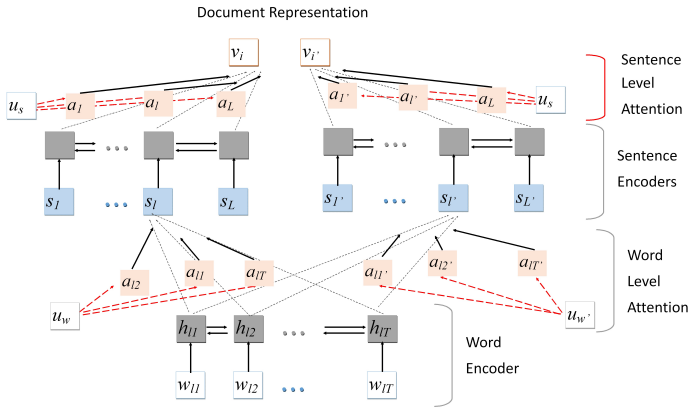


Figure 3.1: RNN with Multi-Head Attention

the concept of context attention used in machine translation, using it to learn context vectors for adjoining sentences. We extend the hierarchical RNN by introducing bi-directional context with attention. Using the word level GRU output, a "look-back" context vector $c_{l-1}(w_{lt})$ is calculated using context attention over the preceding sentence, and a "look-ahead" context vector $c_{l+1}(w_{lt})$ using context attention over the following sentence for each word in the current sentence.

$$\alpha_{(l-1)t}(w_{lt}) = \frac{exp(score(h_{lt}, h_{(l-1)t}))}{\sum_{t'} exp(score(h_{lt}, h_{(l-1)t'}))} \tag{3.13}$$

$$c_{l-1}(w_{lt}) = \sum_{t'} \alpha_{(l-1)t'}(w_{lt}) h_{(l-1)t'} \tag{3.14}$$

$$\alpha_{(l+1)t}(w_{lt}) = \frac{exp(score(h_{lt}, h_{(l+1)t}))}{\sum_{t'} exp(score(h_{lt}, h_{(l+1)t'}))} \tag{3.15}$$

$$c_{l+1}(w_{lt}) = \sum_{t'} \alpha_{(l+1)t'}(w_{lt}) h_{(l+1)t'} \tag{3.16}$$

where $score(h_{lt}, h_{kt}) = h_{lt} W_\alpha h_{kt}^T$ and a single W_α is used for computing the score in both directions. The context vectors are concatenated with the hidden state to form the new hidden state h'_{lt}.

$$h'_{lt} = [c_{l-1}(w_{lt}), h_{lt}, c_{l+1}(w_{lt})] \tag{3.17}$$

The rest of the structure is the same as a hierarchical RNN, using equations 3.9-3.12 with h'_{lt} instead of h_{lt}. Figure 3.2 shows the structure for calculating 'look-back" context.

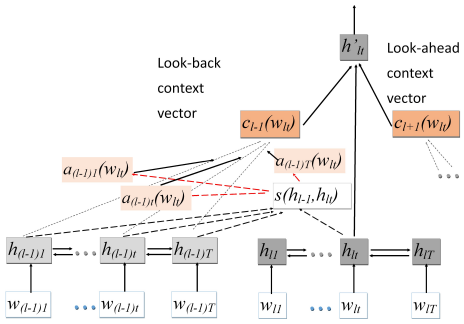


Figure 3.2: RNN with Bidirectional Context and Attention

3.3.5 Implementation Details

The implementation is done via the Tensorflow library (Abadi et al., 2016).[7] All RNNs use GRUs (Cho et al., 2014) with layer normalization (Ba et al., 2016), trained using Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001. Regularization was done via drop out. The validation set was used to do hyper-parameter tuning, with a grid search over drop out rate, number of epochs, and hidden dimension of GRU cells. Good result for all four architectures are obtained with a batch size of 10, a dropout rate of 0.5-0.7, a cell size of 75-250 for the word-level GRU, and a cell size of 40-75 for the sentence-level GRU. For the RNN, we also trained a version with a larger word-level hidden layer cell size of 600.

Pre-trained Glove embeddings[8] are used for all models (Pennington et al., 2014), using a vocabulary size of 65000-75000.[9] The out of vocabulary (OOV) percentage on the CCS test set was 3%, and on the informational test set was 0.5%. All OOV words were mapped to an 'UNK' token. The text was lower-cased, and split into sentences for the hierarchical models using the natural language toolkit (NLTK) (Loper and Bird, 2002).

---

[7]The code and trained models are available at https://github.com/Farahn/Liguistic-Complexity.

[8]http://nlp.stanford.edu/data/glove.840B.300d.zip

[9]In vocabulary words not present in Glove had randomly initialized word embeddings.

Test Set	Model	Samples	WeeBit	RNN	RNN 600	HAN	BCA	MHA
CCS Document	D_1	168	0.69	0.28	0.43	0.47	0.55	0.42
CCS Paragraphs	D_1	1532	0.36	0.30	0.25	0.29	0.32	0.28
CCS Document	D_2	168	0.69	0.34	0.38	0.43	0.48	0.43
CCS Paragraphs	D_2	1532	0.36	0.27	0.26	0.24	0.30	0.29
Informational Paragraphs	D_2	1361	0.22	0.51	0.60	0.60	0.62	0.60

Table 3.4: Results (Spearman Rank Correlation)

3.4 Results and Analysis

We test our models on the two science test sets, as well as on the CCS appendix B document level texts and a paragraph-level version of these texts. We also evaluated the best performing model on the middle school science questions data set. Since both the true reading level and predicted levels are ordered variables, we use Spearman's rank correlation as the evaluation metric to capture the monotonic relation between the predictions and the true levels.

As a baseline, we use the WeeBit linear regression system (Vajjala and Meurers, 2014). The WeeBit system uses knowledge-based features as input to a linear regression model to predict reading level as a number between 1 and 5.5, which maps to text appropriate for readers 7-16 years of age. The feature set includes parts-of-speech (e.g. density of different parts-of-speech), lexical (e.g. measurement of lexical variation), syntactic (e.g. the number of verb phrases), morphological (e.g. ratio of transitive verbs to total words) and psycholinguistic (e.g. age of acquisition) features. There are no features related to discourse, thus it is possible to compute features for sentence-level texts. The system was trained on a subset of the data that our system was trained on, so it is at a disadvantage. We did not have the capability to retrain the system.

3.4.1 Results by Genre

Results for the different models:

- WeeBit
- sequential RNN with self attention (RNN),
- large sequential RNN with self attention (RNN 600),
- hierarchical RNN with attention at the word and sentence level (HAN),
- hierarchical RNN with bidirectional context and attention (BCA), and
- multi-head attention (MHA)

are shown in table 3.4, together with the results for the WeeBit system which has state-of-the-art results on the CCS documents. For the CCS data, both D_1 and D_2 training configurations are used for the neural models; only D_2 is used for the informational test set. For all of these models the hidden layer dimension for the word level was between 125 and 250. We also trained a sequential RNN with a larger hidden layer dimension of 600.

The HAN does better for document level samples than a sequential RNN; the converse is true for paragraph level texts. The RNN with a larger hidden layer dimension performs better for longer texts, while the performance for smaller dimension RNN deteriorates with increasing text length. The BCA model seems to generalize to longer documents and new genres better than the other neural networks.

Figure 3.3 shows the error distribution for BCA(D_1) in terms of distance from true prediction broken down by genre on the 168 CCS documents. The category of informational texts is often over predicted, which we hypothesize is roughly due to specific articles related to the United States history and constitution. The only training data for our models with that subject is in the grades 6-8 and 9-12 categories. The performance for literary and mixed texts, on the other hand, is roughly unbiased; this shows that the model is better at generalizing to non-informational texts, even when there are no literary text samples in the training data.

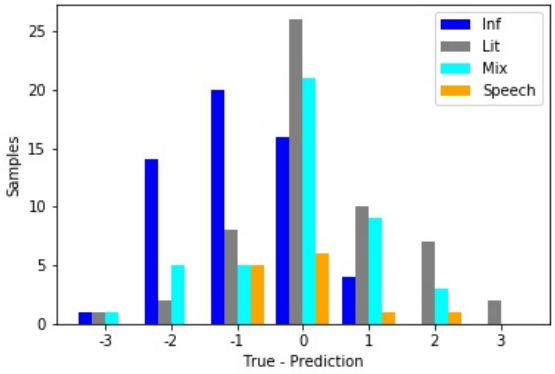


Figure 3.3: Error distribution for the CCS documents BCA(D_1).

3.4.2 Results by Length

Figures 3.4(a) and 3.4(b) show the performance of our models and the WeeBit model as a function of document length, both on the informational paragraphs test set and the CCS paragraph level test set. The results indicate that for shorter texts, particularly under 100 words, neural models tend to do better. Even for a mixture of genres, the model with bidirectional context performs better than the feature-based regression model, as shown in figure 3.4(b).

It is likely that the WeeBit results results on shorter texts would improve if trained on the same training set that is used for the neural models. However, we hypothesize that the feature-based approach is less well suited for shorter documents because the feature vector will be more sparse. Comparing the CCS document- and paragraph-level test sets, the average percentage of features that are zero-valued is 28% for document-level texts and 44% for paragraph-level texts. The most sparse vectors are 40% and 81% for document and paragraph-level texts, respectively.

3.4.3 Results for Science Assessment Questions

Finally, we apply both the baseline WeeBit system and our best model (BCA trained on D_1) to the set of 5470 grade 6-8 science questions. The results are shown in figures 3.5(a) and 3.5(b), where the grade 6-8 category (ages 11-14) corresponds to predicted level 3 for BCA and predicted level 4 for WeeBit. The results indicate that BCA predictions are better aligned with human rankings than

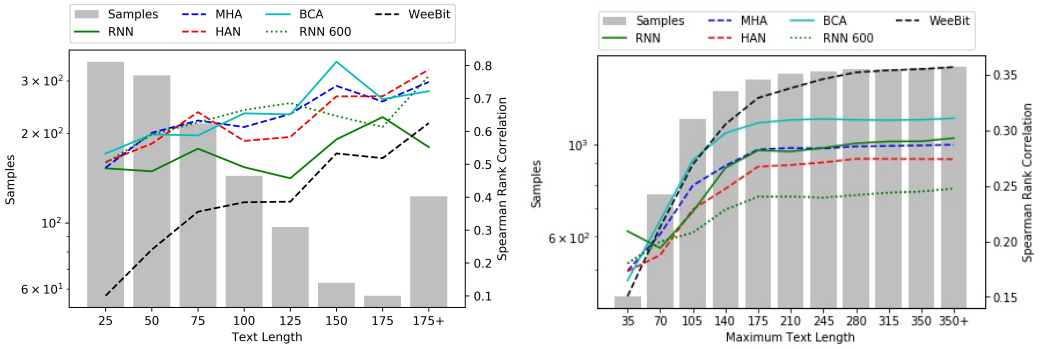


Figure 3.4: Performance vs. text length for informational paragraphs BCA(D_2) (left) and performance vs. maximum text length for CCS paragraphs BCA(D_1) (rigth).

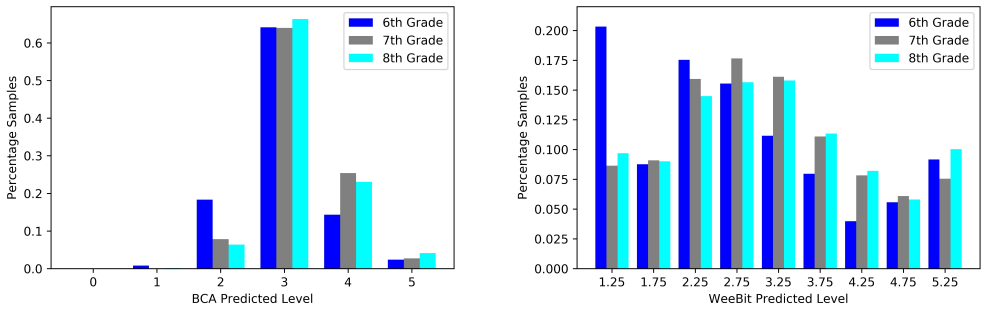


Figure 3.5: BCA predicted levels for middle school science assessment questions (left) and WeeBit predicted levels for middle school science assessment questions (right).

the baseline. As expected, grade 6 questions more likely to be predicted as less difficult than grade 8 questions.

3.4.4 Attention Visualization

Attention can help provide insight into what the model is learning. In the analyses here, all attention values are normalized by dividing by the highest attention value in the sentence/document to account for different sequence lengths. Figure 3.6 shows the word-level attention for the BCA and HAN for

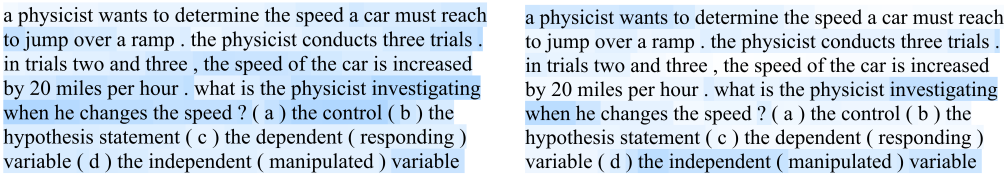


Figure 3.6: Word level attention visualization for BCA (left) and HAN (right) for a middle school science assessment question.

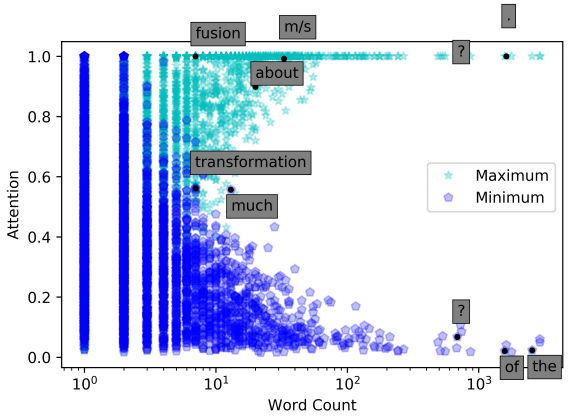


Figure 3.7: Maximum and minimum values of attention as a function of word count for BCA.

a sample text from the science assessment questions test set. (Attention weights in the figure are smoothed to reflect the fact that a word vector from a biLSTM reflects the word's context.) The results show that attention weights are more sparse for HAN than for BCA. At the sentence level (not shown here), the BCA sentence weights tend to be more uniformly distributed, whereas HAN weights are again more selective.

Another aspect of the attention is that a word does not have the same attention level for all occurrences in a document. We look at maximum and minimum values of attention as a function of word frequency for each grade band, shown in figure 3.7 for grade 6-8 science assessment questions.

The pattern is similar for each grade band in the validation and test sets. The minimum attention values assigned to a word drop with increasing word frequency, while the maximum values increase. This suggests that the attention weights are more confident for more frequent words, such as *of*.

Words like *fusion* and *m/s* get high maximum attention values, despite not being as high frequency as words like *of* and *the*. This may indicate that they are likely to contribute to linguistic complexity. The fact that *transformation* has a high minimum is also likely an indicator of its importance. For HAN without bidirectional context, a similar visualization shows that while the trend is similar, the attention weights typically tend to be lower, both for minimum and maximum values.

We find that sentence-end tokens (period, exclamation and question mark) have high average attention weight, ranging from 0.54 to 0.81, while sentence-internal punctuation (comma, colon and semicolon) get slightly lower weights, ranging from 0.20 to 0.47. The trend is similar for all grades. These high attention values might be due to punctuation serving as a proxy for sentence structure. It is interesting to note that the question mark gets higher minimum attention value than period, despite being high frequency. It may be that questions carry information that is particularly relevant to informational text difficulty.

3.5 Discussion

Our work differs from existing models that estimate text difficulty since we do not use engineered features. There are advantages and disadvantages to both approaches, which we briefly discuss here. Models using engineered features based on research on language acquisition offer interpretability and insight into which specific linguistic features are contributing to text difficulty. An additional advantage of using engineered features in a regression or classification model is that less training data is required.

However, given both the evolving theories in language acquisition and the large number of variables that impact second language acquisition, the methodologies used in language acquisition research have certain limitations. For example, the number of variables that can be considered in a study is practically limited, the sample population is often small, and the question of qualitative vs. quantitative methodologies used can influence outcomes (more details in (Larsen-Freeman and Long, 2014; Mitchell et al., 2013)). These limitations can carry into the feature engineering process. Using a model with text as input ensures that these constraints are not inherently part of the model; the performance of the system is not limited by the features provided. Of course, performance is limited by the training data, both in terms of the cost of collection and any biases inherent in the data. In addition, with advances in neural architectures such as attention modeling, there may be

opportunities for identifying specific aspects of texts that are particularly difficult, though research in this direction is still in early stages.

3.6 Conclusion

In summary, this work explored different neural architectures for linguistic complexity analysis, to mitigate issues with accuracy of systems based on engineered features. Experimental results show that it is possible to achieve high accuracy on texts shorter than 100 words using RNNs with attention. Using hierarchical structure improves results, particularly with attention models that leverage bidirectional sentence context. Testing on a mix of genres shows that the best neural model can generalize to subjects beyond what it is trained on, though it performs slightly worse than a feature-based regression model on texts longer than 100 words. More training data from other genres will likely reduce the performance gap. Analysis of attention weights can provide insights into which phrases/sentences are important, both at the aggregate and sample level. Developing new methods for analysis of attention may be useful both for improving model performance and for providing more interpretable results for educators.

Two aspects not considered in this work are explicit representation of syntax and discourse structure. Syntax can be incorporated by concatentating word and dependency embeddings at the token level. Our BCA model was designed to capture cross-sentence coherence and coordination, but it may be useful to extend the hierarchy for longer documents and/or introduce explicit models of the types of discourse features used in Coh-Metrix (Graesser et al., 2004).

Chapter 4

ANALYSIS OF SOURCES OF DIFFICULTY IN SCIENCE ASSESSMENT ITEMS: LESSONS FROM MACHINE LEARNING

The focus of this chapter is the analysis of features that contribute to the difficulty of a question for middle school science assessments. Understanding what makes an item difficult is critical for assessment development. This is particularly true for contextualized items as the inclusion of a context inevitably leads to text and/or non-textual information that test takers need to interact with in their problem solving process. As science assessment items are increasingly situated in real-life scenarios, it becomes more challenging to accurately measure student understanding, as student performance is not only affected by the underlying constructs but also by characteristics of an item such as linguistic demands and item format. Being able to predict item difficulty and pinpoint potential sources of difficulty in assessment items provides several benefits. From the test validation perspective, it can reveal the potential sources of construct-irrelevant variances and thereby evaluate the extent to which items provide precise and accurate interpretation of students' understanding of the assessed constructs. For item writers, it provides valuable guidance that can help them more efficiently produce items that provide an accurate assessment of student knowledge, which is also considered as evidence for the formative stage to evaluate the validity claims for score interpretations (Kane, 2006). From the pedagogical perspective, it offers guidance for classroom assessments in the sense that teachers can select items aligned to specific learning goals and developmentally appropriate for their students so that the assessment results are more readily relevant for teachers to interpret and take action. This work was done in collaboration with Dongsheng Dong and Professor Min Li from the College of Education. My contributions include machine learning specific literature review, linguistic feature extraction, model implementation and feature analysis.

Item difficulty has been indexed in multiple ways. In this work, we consider the task of predicting item difficulty in terms of the percentage of correct responses to the item (p-value), aggregating over a group of examinees, given the text associated with the item. While we know that the proba-

bility that a student answers an item correctly depends on both the item characteristics and student understanding level, in many cases only aggregate performance data is available (e.g. the fraction of students completing the item correctly), in which case individual student parameters cannot be accounted for. In this scenario, as well as when individual student data is available, previous studies have explored a large number of factors that may influence item difficulty (Enright and Sheehan, 2002; Sheehan et al., 2006). However, it is challenging to interpret results across studies when model evaluation criteria have such a large range (e.g., the percentage of variance explained, R^2, varying from 0.10 to more than 0.90), making it difficult to identify specific findings that generalize across multiple studies. This is in part due to the fact that studies look at different sets of item characteristics, different types of items, and different populations of test takers. In addition, we argue here that the methodologies often used are limiting the generalizability of findings. The goals of this work are to provide insights into the findings from previous work by shedding light on issues associated with data analysis methodology, as well as to provide new results related to specific item characteristics.

In this chapter, we employ a machine learning framework to look at the combination of item format (response type), cognitive demands (topics and practices) and linguistic complexity features, drawing on characteristics that have been explored in various forms in several prior studies. Broadly, our approach is to analyze student responses on science assessment items with multiple prediction models, different model selection criteria, and alternative criteria for interpreting importance of item characteristics (input features to the prediction model). We compare methods that have been used in prior measurement research on item difficulty prediction with related approaches that are standard in machine learning. Our work leverages models and the model evaluation criterion used by (Sinharay, 2016), but further provides a detailed comparison to other model evaluation (goodness of fit) methods and a more in-depth study of feature interpretation using multiple methods.

This work makes two main contributions. First, using general machine learning results and an empirical study of predicting difficulty of NAEP items, we outline scenarios where popular linear regression methods for analyzing the impact of item characteristics on item difficulty provide misleading results, and show that tree-based machine learning models and repeated hold-out performance estimation can lead to more robust findings. These results suggest that past findings related to item characteristics may need to be reassessed and that reported goodness of fit results are unre-

liable in many cases. Second, using the new methods, we confirm past findings about item response type, introduce new findings related to science practices, and show that linguistic complexity seems to be non-linearly related to item difficulty.

The remainder of the chapter proceeds as follows. We first provide context for this study with a review of prior work in the education literature and relevant results from machine learning. Next, we detail the experimental methods used, followed by a presentation of results and analyses. The chapter concludes with a summary of the contributions and limitations of the study.

(*Note:* this chapter provides more details on fundamental machine learning concepts that are taken as well known in other chapters. The detailed discussion and experimental analysis of overfitting and measurement error are included because it is intended for an education research audience that is less familiar with machine learning methods and because our work calls into question prior work.)

4.1 Background

In this section, we look at prior research on item difficulty prediction that informs our study, pointing out commonalities and differences in findings associated with these studies. We then cast the problem of modeling the effect of item characteristics on student performance in a machine learning framework and provide links between terminology in the two different research communities, machine learning and educational research. In this context, we explore how machine learning techniques can be used to strengthen the methodology for item difficulty prediction.

4.1.1 Prior Research on Item Difficulty Analysis

Previous studies have explored a large number of factors that may influence item difficulty including cognitive demands, item formats, item topics, linguistic demands and so on. These features are referred to as item response demands by Ferrara and colleagues (Ferrara and Duncan, 2011), which encompass the content, cognitive and linguistic knowledge and skills required to solve or partially solve an item. Based on the item response demand framework, (Ferrara and Steedle, 2018) reviewed 24 studies on item difficulty modeling and summarized item response demands identified as potential predictors of item difficulty. According to (Ferrara and Steedle, 2018), 15 out of 24 studies employed linear regression as the major methodological approach.

As the review by (Ferrara and Steedle, 2018) includes only two studies related to science assessments, we composed a more STEM-focused literature review by keeping the science and mathematical studies from that review and adding more research on science assessments. Table 4.1 summarizes 8 studies on item difficulty modeling that do not incorporate student characteristics, which is the case in our work. As noted in the table, the studies use different parameters for item difficulty. In our study we use p-value, since it is available for more NAEP items and it aligns with the dependent variable used in prior work on NAEP assessment items (Valencia et al., 2017).

Study	**Items**	**Predicted Variable**
Predictors *(italicized when significant)*		
Mesic and Muratovic (2011)	123 physics (66 multiple choice, 57 constructed response)	Rasch item difficulty
item openness, interference effects of intuitive and formal physics, relationships, related relationships, experimental method, mitigating factors (e.g., whether the item can be solved by remembering fragments of knowledge), analytic representation		
Mesic (2011)	123 physics (66 multiple choice, 57 constructed response)	item discrimination power measure
item openness; relationships; related relationships; interference effects of intuitive and formal physics; analytic representation; number of depictors; grade level of item; combined features (grade $\times$ relationships, grade $\times$ related relationships, item openness $\times$ relationships, item openness $\times$ analytic representation)		
Crisp and Grayson (2013)	38 physics (multiple choice)	Rasch item difficulty
question attributes (*total amount of reading*, maximum sentence length, concepts, context, visual resources, importance of options); question processes (e.g., recalling equation or unit, *using physics concepts*, selecting equation or data, *working with symbols*, *calculating*); physics knowledge and understanding (e.g., scientific phenomena, scientific applications); and cognitive demand (e.g., complexity, abstractness, response strategy)		

Rosca (2004)	104 science (multiple choice)	Rasch item difficulty
presence of a figure, Flesch reading level, the *mean number of words in distractors*, number of options, *ratio of number of words in correct option and mean number of words in distractors*, and *cognitive level*		
El Masri et al. (2017)	216 UK science (objective & short constructed response)	2-parameter graded response model threshold parameters
curricular variables (e.g., topic, subtopic, and concept), item type (e.g., *extended construct response*), depth of knowledge, nature of stimulus (i.e., text, *photo*, graph, schematics representation), and language variables (five dimensions from Coh-Metrix software)		
Le Hebel et al. (2017)	103 PISA science (objective & constructed response)	*p*-values
depth of knowledge; necessity of context information; *item format*; and PISA competency		
Turner et al. (Turner et al. (2013))	48 PISA math	item difficulty
communication; *devising strategies*; mathematizing; representation; *using symbolic, formal, and technical language and operations*; and *reasoning and argumentation*		
Morrison and Embretson (2014)	math (number unspecified)	item difficulty
19 attributes in 5 cognitive competencies: *translation (e.g., modifier prop); integration (e.g., translating word equation); solution planning (e.g., number of subgoals); solution execution (e.g., number knowledge);* and *decision processing (e.g., decision processing confirmation)*		

Table 4.1: Existing studies on aggregated item difficulty prediction, listing number and type of items, predicted variable, and predictors used in the study

All of the studies use linear regression to predict item difficulty, not noted in the table for brevity. ((Crisp and Grayson, 2013) also use a linear logistic test model.) Most studies report R^2 and/or adjusted R^2 as an indicator of model fit, using all available data to fit the model. Reported R^2 varies from 0.10 (Ferrara and Steedle, 2018) to 0.999 (Turner and Adams, 2012) across various experimental settings. The number of items used in the studies ranges from 38 to 216. The table also notes the predicted variable and the student/item characteristics explored as predictors, highlighting those characteristics found to be statistically significant for predicting item difficulty, where significance is usually determined by analysis of variance of regression coefficients. We observe that some studies find many characteristics to be significant, whereas others find very few. One would expect that studies based on smaller data sets would tend to have fewer significant characteristics simply due to the data limitations. Thus, it is surprising to see so many significant characteristics in studies with a relatively limited number of items.

Since these studies test different sets of item parameters on different item collections, it is impossible to compare specific findings across studies. Further, there is no one single characteristic explored in all studies. For that reason, we look only at trends in findings for broad categories of features.

All studies look at some type of subject-related cognitive demand, and all but (El Masri et al., 2017) find that at least one cognitive demand factor is significant. However, different categories of cognitive demand are considered, including topic knowledge, question processes, and cognitive level/complexity of the item, and they are represented with different levels of granularity and used in different combinations. Probably the most consistent trend is that characteristics related to question processes are often found significant.

Several studies examine item format (Mesic, 2011; Mesic and Muratovic, 2011; Le Hebel et al., 2017; El Masri et al., 2017), primarily response type, sometimes referred to as openness, but some also look at the use of graphics. Studies distinguish between various forms of constructed response (CR) and multiple choice items, where CR might include short answer, fill in the blank, essay, graphing, or drawing. Most studies found response type to be significant when included with some form of cognitive demand features. Generally, open-ended responses are the most difficult; item format and multiple choice are the easiest.

A few studies look at some form of linguistic complexity,[1] including measures of reading level (Rosca, 2004) and specific variables hypothesized to distinguish between reading levels (Crisp and Grayson, 2013; El Masri et al., 2017). None of these features were found to be statistically significant. El Masri et al. (2017) suggested that the lack of efficacy of using linguistic features for assessment items in their own study might be due to their features (Coh-Metrix variables (Graesser et al., 2011)) being less useful with short passages, which provide limited information in computing these statistics. This effect was also seen in a study on predicting language difficulty in science assessment items (Nadeem and Ostendorf, 2018).

Another body of work involves the use of experimental design, where an item is altered to reflect differences in one or few variables of format features and context while controlling for item features, to examine how these item characteristics as the manipulated variables impact individual student performance. The work in (Höttecke et al., 2018) creates three variants of a question by changing the linguistic complexity to determine whether the language difficulty impacts student performance. In the study by (Song and Bruning, 2016), two different contexts are tested in items about global warming, one based in the US and one based in Korea, to assess whether familiarity with the context impacts students' performance. Similarly, (Solano-Flores et al., 2014) look at providing support to ELLs through visual elements in questions, comparing with performance on non-illustrated questions. The study by (McCullough, 2004) modifies physics questions which have stereotypically male contexts to versions with stereotypically female contexts to look for performance impacts. These studies provide useful insights for guiding future studies, but they typically involve only a small number of items, making it difficult to assess significance of the findings. For example, (Höttecke et al., 2018) use 6 items and find that linguistic complexity has a significant effect in 3 of these, but one of these is associated with a reversal effect.

From our survey of previous work, we see that typically models for item difficulty prediction are trained with very small numbers of samples, and metrics are reported on the data the model is trained on. As we will show empirically, this leads to cases where findings that hold for one data set may not hold for another set of items. This problem may be reflected in Table 4.1 in that different

[1]We consider linguistic complexity as opposed to linguistic demands, since we focus on difficulty associated with language use, as opposed to higher demands posed by longer texts. As such, we do not consider total word count to be a linguistic complexity feature.

features are identified as significant in different studies.

To counter these issues, in this atudy we propose approaches from machine learning to strengthen existing methodologies for model training and evaluation, similar to (Sinharay, 2016) but with empirical analyses to provide additional insights into evaluation measurement error and new methodologies for exploring feature importance and significance testing.

4.1.2 Machine Learning Methodology

The problem of predicting item difficulty (represented here in terms of the item *p*-value) given the text (and non-textual part) of the item can be framed as a machine learning problem, where the task is to learn a function $\hat{p}_i = f(x_i)$ from a collection of examples $\{(x_i, p_i); i = 1, \ldots, N\}$ (the training set), where x_i is a vector of features (explanatory variables[2]) for the i-th item, and $p_i \in [0, 1]$ is the item difficulty (dependent variable). Assuming a particular functional form for $f(\cdot)$, the parameters of the function are learned by minimizing a loss function on the data, a process that is often referred to as "training." Model training corresponds to model fitting, i.e., finding the parameters that best fit the data. The loss function is typically related in some way to the negative version of the goodness of fit criterion.

4.1.2.1 Evaluation, Model Selection and Feature Selection

In machine learning, an important issue is overfitting. As the power of the prediction function is increased (either by adding features or using more complex functions), it may be possible to perfectly fit the training data, but then the model will not generalize well to new data. More specifically, the goodness of fit on the training data is optimistically biased, i.e. the expected performance on the training data is higher than the expected result on independent data.

Because model performance is assessed on a random sample of data, it represents a noisy measurement of the true performance. The noise or measurement error reflects a combination of variance associated with using a random sample and bias associated with the measurement technique, where the bias is the difference between the expected value of that measurement and the true performance. The estimate is unbiased when performance of the learned predictor is measured on data that it is

[2]"Independent variables" is a commonly used term, but these variables are not always statistically independent, and relying on that assumption can sometimes be problematic.

not trained on, which is referred to as the evaluation or test set. However, when the test set is small, then the variance of the estimate will be high, leading to higher measurement error. Variance also increases with more complex predictors. In the studies described previously, data set sizes vary from 18-216, which is problematic for a simple hold-out test set performance measurement. In cases such as this, where labeled training data is limited, other strategies are used to provide a more reliable performance estimate. In one approach, referred to as cross-validation (CV), the data is first split into m subsets (called as folds). One fold of the data is held out, and the model is trained using each of the remaining $m-1$ folds of data, then evaluated on the held out fold. This process is repeated m times, each time with a different fold held out. The final evaluation score is the average of scores on the held-out folds. This allows the use of all the data in both training *and* evaluating the model, at the cost of having to learn multiple models. To further reduce the variance, one can repeat CV with different partitions and average the results. An alternative, referred to as repeated hold-out (RHO), is to randomly select a subset of the data for training, assess performance (goodness of fit) on the remaining samples, repeat a large number of times for different random samples, and average the results. While the RHO method can still have some bias, it is shown empirically to provide performance estimates with lower error (less noisy) in limited data scenarios (Kim, 2009). Our work will rely on RHO esimtates, but we present results for different methods to illustrate the impact of measurement error for predicting item difficulty with a small data set (i.e. less than a few hundred samples, which is typical for publicly available items).

The processes of model selection and feature selection can also result in overfitting, and a variety of strategies have been introduced to counteract this. One approach is to add a penalty function to the loss function. Examples commonly used in educational measurement models are the Akaike information criterion (AIC) and Bayesian information criterion (BIC), where a maximum loglikelihood objective is combined with a penalty term that accounts for model complexity and training set size. Specifically, AIC and BIC are defined as $2k - 2\log(\hat{L})$ and $\log(n)k - 2\log(\hat{L})$, respectively, where $\hat{L}$ is the likelihood of the data given the trained model, n is the number of samples, and k is the number of model parameters. Adjusted R^2 is motivated by the same idea. A problem with AIC, BIC and adjusted R^2 is that they rely on statistical assumptions that do not hold for many item difficulty prediction scenarios, particularly when using more powerful machine learning models. Another approach to avoid over-fitting is to use a parameter regularization or model size penalty,

where the penalty weight is chosen using held-out data or CV. A similar approach can be used for feature selection. While these methods reduce the problem of overfitting, they do not eliminate the problem of bias in the performance estimate. For these reasons, it is important to use these criteria in addition to – and not in place of – evaluating performance on held-out data. A difficulty in interpreting prior work in modeling item difficulty is that most studies do not report results on held-out data.

4.1.2.2 Prediction Models

The problem of predicting item difficulty can be posed in different ways. If it is posed as predicting a continuous variable $p_i \in \mathbb{R}$, it is a regression problem, and the evaluation criterion is typically mean-squared error (MSE) or the normalized version, $R^2 = 1 - MSE/\sigma^2$ (the percentage of variance explained by the model). An equivalent solution can be obtained by using a negative log likelihood loss function with the assumption that $p_i = f(x_i) + v_i$, where v_i has a Gaussian distribution. If the problem is posed as modeling the probability of a binary outcome $p_i \in [0, 1]$ (whether or not the item will be answered correctly) using a logistic function, a standard evaluation criterion would be normalized cross-entropy. Following most prior work on aggregated student performance data, this study will treat item difficulty prediction as a regression problem.

Defining a model requires specifying the form of the prediction function $f(x)$, where x is a d-dimensional vector of features $x = [x^1 x^2 \cdots x^d]$. Because of limited training data, we choose relatively simple forms, including linear functions ($\sum_j b_j x^j$) and different combinations of binary decision functions ("is $x^j = 1$?" or "is $x^j > T$?"), as described in the next section. A key difference between the model forms relates to the potential for making use of any interdependence of features in prediction and in methods for interpreting feature importance.

4.2 Methods

This study follows the general approach of prior work that analyzes the impact of different item characteristics on item difficulty in terms of their utility as features in predicting the percentage of correct responses to an item (p-values). Specifically, we apply this methodology using multiple models, model selection methods and interpretation methods, in order to provide more reliable

findings given the small data set available. In this section we describe the data, features (item characteristics), models, model evaluation methods, and feature interpretation techniques.

4.2.1 Items

Our data set consists of 8th grade released science assessment items from the National Assessment of Educational Progress (NAEP, https://nces.ed.gov/nationsreportcard). We collected 132 items, for the years 2000-2011, each with the associated p-value. For multiple choice items, the p-value is the percentage of students responding correctly; for constructed response items, it is the average score normalized to a [0,1] range. For items used in multiple years, the p-value of the item is averaged over the years the item has been administered.

The associated meta-data for each item includes the response type, content classification (topic), and science practice. For response type, which corresponds to an item format feature, each item was characterized as either being multiple choice, short constructed or extended constructed response. Each item was associated with one of three content topics (Earth and space sciences, life science and physical science) and one of three science practices (identifying science principles, using science principles and using scientific inquiry). Item counts for the different meta-data characteristics are shown in Table 4.2.

Table 4.2: NEAP Data: Distribution of the 132 Items by Content, Practice and Response Type

Content classification		Science practice		Response type	
Earth & space science	41	Identifying science principles	37	Extended response	17
Life science	52	Using science principles	64	Short response	54
Physical science	39	Using scientific inquiry	31	Multiple choice	61

4.2.2 Features

The meta-data labels associated with each item were mapped to multiple binary indicator variables (sometimes referred to as "dummy" variables) for use in the different models for predicting item p-values. For example, the topic label was associated with one indicator each for earth and space science, life science, and physical science. The result is 3 binary format features and 6 binary cognitive demand features (3 topics and 3 practices). Since the three response types are mutually exclusive (as are the topics and science practices), they can be captured with two indicator variables instead of three, as in the work in (Sinharay, 2016). While using 2 vs. 3 indicators does not have a significant effect on model fit, we find that using 3 features is preferable for interpreting the impact on item difficulty, as we will show in empirical analyses. Of course, using 3 indicator features for a 3-way category means that the features are not statistically independent, but we observe other dependencies among the features, so relying on the assumption of independence in linear regression is more generally problematic.

Two linguistic complexity features were used: i) the average age-of-acquisition for all words in an item using word scores pulled from a psycho-linguistic database (Cortese and Khanna, 2008); and ii) the predicted grade level from a pretrained neural network model (NN) from Chapter 3, which is the current state of the art for predicting grade-level for math and science assessment items. These features are extracted with the full text of the items, including answer options when they are present.

4.2.3 Models

This work explores multiple models in order to assess whether more complex models lead to a better fit to the data and/or new insights about the relative importance of different item characteristics.

We fit four types of models for predicting the item p-values: multiple linear regression, decision tree regression (Breiman et al., 1984), random forest regression (Liaw et al., 2002) and gradient boosting regression (Friedman, 2001). All models are implemented using the Scikit-learn software package (Pedregosa et al., 2011).

Multiple linear regression is the most commonly used model in predicting item difficulty. The

p-values are predicted using a linear combination of features:

$$\hat{p} = b_0 + \sum_{j=1}^{d} b_j x^j$$

The linear regression model parameters $\{b_j\}$ are chosen to minimize the mean-squared prediction error on the training set.

Decision tree regression models involve asking a series of questions about the features, following a tree structure, and then assigning a predicted p-value according to the final leaf of the tree that the question answers lead to. The tree structure is learned in a greedy fashion, by finding the partition of the feature that gives the maximum reduction in mean squared error and repeating this until either the maximum specified size is reached or there are no more partitions that improve the MSE. For the continuous-valued linguistic complexity features, questions have the form "is $x^j > T$?" where a different threshold T is learned for each point where the question is asked.

Random forest regression uses an ensemble of decision trees, where different trees are obtained by randomly drawing a subset of samples with replacement from the training data and learning a model using those samples. The final prediction is the average value from all trees learned.

Gradient boosting is also an ensemble method, using a linear combination of regression trees, but the regression trees and their weights are learned in stages that shrink the contribution of successive trees according to a learning rate. For each iteration, a new tree is learned to minimize the sum of losses (mean squared error) given the ensemble at the previous iteration. Often, boosting ensembles explicitly include weak learners, which would mean smaller trees than in a random forest. Here, there are no hard constraints on the tree size, but smaller trees are learned on average due to the stage-wise learning strategy.

When given a large number of features, or continuous-valued features (which can be associated with a large number of binary questions), decision trees are susceptible to overfitting. To counteract this problem, it is common to use some form of pruning. We use 5-fold CV within the training data to select the maximum depth of the tree, minimum number of samples in a leaf and the minimum number of samples for a node to split via grid search. Using the selected parameters, the final tree is retrained on the full training set. The random forest and gradient boosting models require a similar parameter selection step, with the addition of choosing the number of trees in the ensemble. Gradient boosting also requires tuning the learning rate.

4.2.4 Model Evaluation

In assessing *goodness of fit*, it is important to use a criterion that can be applied to all models explored. A popular criterion for evaluating model fit when predicting item difficulty is R^2, which indicates how much variance (σ^2) is explained by the model (higher R^2 is better; negative R^2 indicates that the error is higher than the sample variance). When R^2 is reported on the full data set, it is optimistically biased and sensitive to overfitting, so most studies select the best model and report results based on other criteria (e.g. AIC, BIC or adjusted R^2). Since these criteria are not useful for the tree-based models used in our study, we need an alternative measure to convincingly demonstrate that specific features impact item difficulty. With the constraint of limited data, we measure R^2 using RHO with 1000 runs, in each case having 2/3 of the data for training for each iteration, testing on the remaining 1/3 of the data. To provide insight into measurement error, we also present results using 3-fold CV and fitting on the full data set.

In *model comparison*, the goal is to determine whether the difference between RHO scores for two models is statistically significant. Again, it is important to use a criterion that can be applied to all model comparison pairs. For the RHO estimate, we use a *t*-test on the 1000 R^2 values from the two models and compute the confidence of a difference between the means. For the 3-fold CV scores, we use an item-level bootstrapping test. For two models being compared, we first obtain the model results for all three CV test folds (total of 132 items) for both sets of trained models. We then randomly sample $n = 132$ items with replacement (bootstrapping), and compare the model performance for these two sets. This process is then repeated 10,000 times, and the p-value is the fraction of times the bootstrapping estimate gives a higher score for the model with the lower CV score.

4.2.5 Interpretation

The underlying goal of this work is to determine which item characteristics are important for predicting difficulty of the item. Methods differ depending on the model form, i.e. linear regression vs. tree-based.

For linear regression, we compute the standard error for each coefficient in linear regression, which is an estimate of the standard deviation of the coefficient. The *t*-statistic is computed by

dividing the coefficient value by the standard error, which is compared to the Student's *t* distribution with $n - k$ degrees of freedom to give the *p*-value. In addition, we provide plots showing the distribution of the RHO coefficients, which provides an alternative view of feature importance.

In a decision tree, the regression result is determined by the questions asked in the particular path of a decision tree. By quantifying and accumulating the contributions of a feature at these different points, one can obtain a measure of feature importance. For models based on ensembles of trees (random forest and gradient boosting), the importance scores can of the different trees are combined. Standard machine learning toolkits typically provide feature importance scores based on this heuristic. Because of the popularity of this measure, we present the RHO distribution of feature importance scores, contrasting with linear regression.

The commonly used feature importance score in tree-based models has the limitation that it does not provide insights into whether a characteristic is associated with easier or more difficult items. An alternative analysis is provided by Tree Explainer (Lundberg et al., 2020), which uses the SHapley Additive exPlanation (SHAP) framework for interpreting complex models (Lundberg and Lee, 2017). For each prediction, the contribution of a feature is approximated as the effect of including this feature in the model, allowing the model prediction for that instance to be cast as an additive model, with the ability to handle multicollinearity. Beyond individual feature importance, the framework allows capturing local interaction between pairs of features and how the pairwise interaction contributes to the model prediction. This is particularly important for tree-based models, which, unlike linear models, can leverage feature interaction through successive questions that allow them to perform better when the assumption of independence between input features does not hold.

4.3 Results

In this section, we present results for different models to provide insight into generalizability of methods for measuring goodness of fit, obtain improved model fit with stronger models, and provide analyses to better understand the impact of different item characteristics on student performance.

4.3.1 Measurement Error

This section provides analyses that highlight the problems of overfitting and variation across data samples, which depend on both the performance measurement strategy and the choice of models.

Table 4.3: R^2 for Models Trained Using 3-fold CV with All Features.

(a) Linear Regression

	Fold 1	Fold 2	Fold 3
Train 12	0.37	0.40	*0.04*
Train 23	*0.19*	0.39	0.19
Train 13	0.38	*0.28*	0.19

(b) Decision Tree with Parameter Selection

	Fold 1	Fold 2	Fold 3
Train 12	0.68	0.65	*0.14*
Train 23	*0.13*	0.31	0.40
Train 13	0.49	*0.28*	0.34

(c) Decision Tree Without parameter Selection

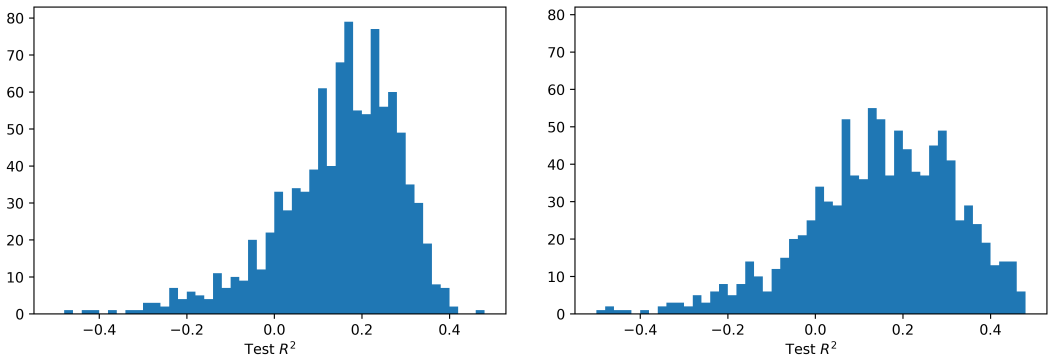
	Fold 1	Fold 2	Fold 3
Train 12	1.00	1.00	*-0.01*
Train 23	*-0.32*	1.00	1.00
Train 13	1.00	*-0.34*	1.00

Note. The performance of the fold used in testing is indicated in italic.

For these experiments, we use linear regression and decision tree models to illustrate random variation in measurements of model fit, since they tend to have low and high variance, respectively.

Table 4.3 shows the model fit as measured by R^2 in each fold of data when training with 3-fold CV using all features (i.e. 3 binary format features, 6 binary cognitive demand features, and 2 linguistic complexity features). One can see that R^2 for a particular fold in a table is always lower when that fold is used in testing than in training. The difference in performance across folds is due to randomness associated with any particular sample. For the decision tree without parameter selection, the R^2 for the training folds can get as high as 1.0, since it is easy to overfit with a decision tree. In this case, the R^2 values for the test folds are very low and sometimes negative, indicating that the model is performing worse than a constant mean predictor.

The variation across folds is also reflected in variation of performance statistics when the test set is defined by randomly selecting (without replacement) a third of the data for testing and then



(a) Linear Regression

(b) Decision Tree with Parameter Selection

Figure 4.1: Histograms for test R^2 for 1000 RHO trials for linear regression and decision tree with parameter selection using all features.

training model parameters using the remaining data. Figure 4.1 shows variation in R^2 for each of the 1000 RHO samples, for linear regression and the decision tree model with parameter selection. Linear regression is a simpler model (lower capacity), so it has less variation. Note that all models have some hold-out test subset that leads to negative R^2. With higher variance models, one has a higher chance of training a model that does not generalize. The CV-Test estimate would correspond to an average of 3 points that could be in the histogram. By averaging 1000 measurements (instead of 3 measurements), the RHO estimate will have lower variance than the 3-fold CV measure in predicting performance on unseen items.

For studies that use linear regression, it is common to use significance tests or criteria such as AIC or BIC to select features and then report R^2 associated with models based on only the significant features. Even with feature selection, reporting R^2 based on the full data set leads to optimistically biased results. For example, when using only the 4 significant features identified by the standard linear regression analysis, R^2 measured on the full data set is 0.28, compared to 0.18 when using an RHO estimate.

4.3.2 Model Comparison

Table 4.4 shows the model fit on the full data set using four different measurements: i) R^2 when all the data is used for both training and testing (All-Train), ii) average R^2 of the training folds using 3-fold CV (CV-Train), iii) average R^2 of the testing folds using 3-fold CV (CV-Test), and iv) average R^2 of 1000 randomly selected subset using hold-out testing (RHO). For any given model, R^2 is highest when training and evaluating on the same data (All-Train and CV-Train). When decision trees are used without CV parameter selection, they are able to fit the training data perfectly but do not generalize to independent data. Using the more reliable CV-Test and RHO methods of measuring goodness of fit, better results are obtained with the more powerful models. The random forest gives the best results, which are significantly better ($p < .001$) than linear regression and decision tree for both CV (via bootstrap) and RHO (via *t*-test). Compared to gradient boosting, random forest is significantly better for RHO ($p < .001$), but not for CV.

Table 4.4: Average R^2 for Models using Different Performance Estimation Scenarios

Model	All-Train	CV-Train	CV-Test	RHO
Linear Regression	0.30	0.32	0.17	0.14
Decision Tree with parameter selection	0.41	0.48	0.18	0.15
Decision Tree	1.00	1.00	-0.22	-0.29
Random Forest with parameter selection	0.70	0.53	0.32	0.21
Gradient Boosting with parameter selection	0.53	0.70	0.30	0.18

4.3.3 Feature Selection

4.3.3.1 Multiple Linear Regression Feature Interpretation

Linear regression models have been used in many studies aiming to identify characteristics of items that are significant for predicting student performance. However, interpreting features based on linear regression coefficients can be problematic when the data does not match the model assumption, e.g., the input variables are not independent. In Table 4.5, we report standard feature analyses associated with linear regression, providing the coefficients b and indicators of significance (denoted by asterisks) for different sets of features. All models are learned on the full data set, but the resulting b coefficients are equivalent to the results of averaging RHO models. The *t*-statistic test is equivalent to using the RHO means and variances. The different feature subsets are chosen to show the impact of feature dependence and dummy variable assignment. Specifically, we compare: i) the full set of features (Model 1); ii) three alternatives for assigning 2 dummy variables to encode the 3-way categorical features (Models 2A-C); and iii) prediction without linguistic features (Model 3), which we hypothesized are correlated with some categorical features. The RHO R^2 estimate is provided for both the full feature set and the reduced subset of features selected based on the significance test.

As expected, the RHO R^2 results are not significantly different for the models using 2 binary indicators vs. 3 for the categorical features when including all features. However, the results in the table point to a problem with using 2 indicators when assessing feature significance. While the contribution from excluded redundant features can be captured in the intercept value, it is impossible to separate out contributions of different factors when there are multiple cases being handled in this way. Further, the choice of which one is left out impacts the significance of other features when there are dependencies across groups, which leads to very different R^2 for the predictors using only the significant features. With Models 2A and 2B, for which the uncoded variables correspond to features with lower significance in Model 1, the multiple choice feature loses significance because of the interdependence between item format and science practices (items associated with "using scientific inquiry" are never "extended response"). In Model 2C, the uncoded variables correspond to the most significant features in Model 1, and their omission gives a result where the topic category has no significance. When two input features are not statistically independent variables, including both can also reduce the statistical significance of the other. For model 3, where the linguistic complexity

Table 4.5: Multiple Linear Regression Analysis Results (*b* Coefficient Values) for Models Trained with Different Subsets of Features.

Model ID	1	2A	2B	2C	3
Intercept	0.261*	0.379*	0.497**	0.726**	0.215**
Linguistic Feature					
Age-of-acquisition	-0.016	-0.016	-0.016	-0.016	–
NN reading level prediction	-0.001	-0.001	-0.001	-0.001	–
Item Format					
Extended Response	0.075	–	–	-0.072	0.056
Short Response	0.041	-0.034	-0.034	-0.106*	0.029
Multiple Choice	0.146**	0.072	0.072	–	0.129**
Topics					
Earth and Space Science	0.047	–	-0.048	-0.072	0.030
Life Science	0.119**	0.072	-0.023	–	0.103**
Physical Science	0.095*	0.048	–	-0.023	0.081**
Science Practices					
Identifying Science Principles	0.199**	0.203**	0.133**	–	0.183**
Using Science Principles	0.066	0.070	–	-0.133*	0.054*
Using Scientific Inquiry	-0.004	–	-0.070	-0.203*	-0.022
RHO R^2 (all features)	0.142	0.139	0.145	0.137	0.174
RHO R^2 (only *,** features)	0.182	0.131	0.131	0.192	0.193

Note. Statistical significance is based on comparing with Student's *t* distribution (*p¡0.05; **p¡0.01).

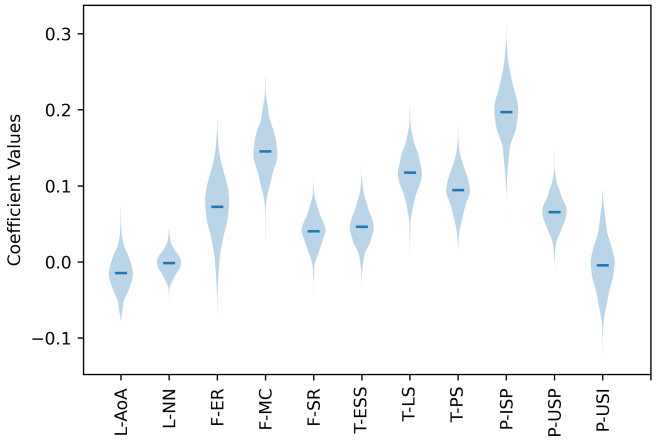


Figure 4.2: Violin plot for linear regression coefficients for 1000 RHO models trained to predict item *p*-values. The horizontal bar for each feature indicates the mean value of the coefficient. *Note:* the constant is not shown.

features are omitted, we see that the topic and science practice features have higher significance than for Model 1, likely due to interdependence between linguistic complexity and topic and science practice features (e.g., more difficult practices tend to have higher predicted reading level).

While problematic for interpreting significance, an advantage of the regression analysis is that it provides insight into which dimensions of a category are more or less difficult. When training the regression function using RHO, each training partition gives different values of b because of random sample differences. Figure 4.2 shows a violin plot that provides a visualization of the distribution of each linear regression coefficient learned from the 1000 random samples. The t-test statistic is roughly a variance normalized distance of the mean from 0, so distributions that are farther from 0 are more significant. The biggest differences are in the science practices category, where identifying science principles is the easiest skill and using scientific inquiry is the most difficult. Among the topics, earth and space science is most difficult, and the multiple choice format is easiest.

4.3.3.2 Feature interpretation in tree-based models

For tree-based models, we present analyses based on the random forest model, since it gave the best performance. Findings with gradient boosting are similar. Figure 4.3 provides violin plots

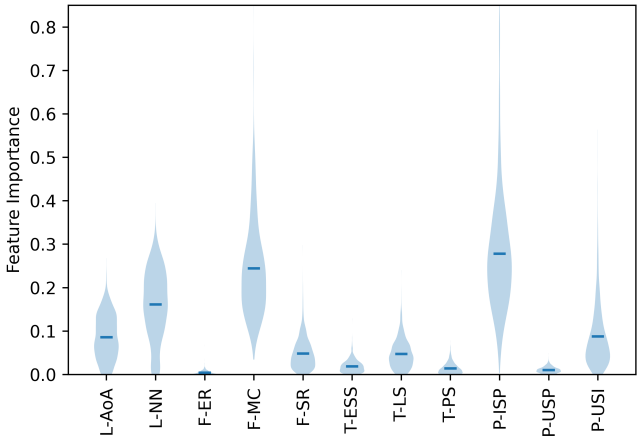
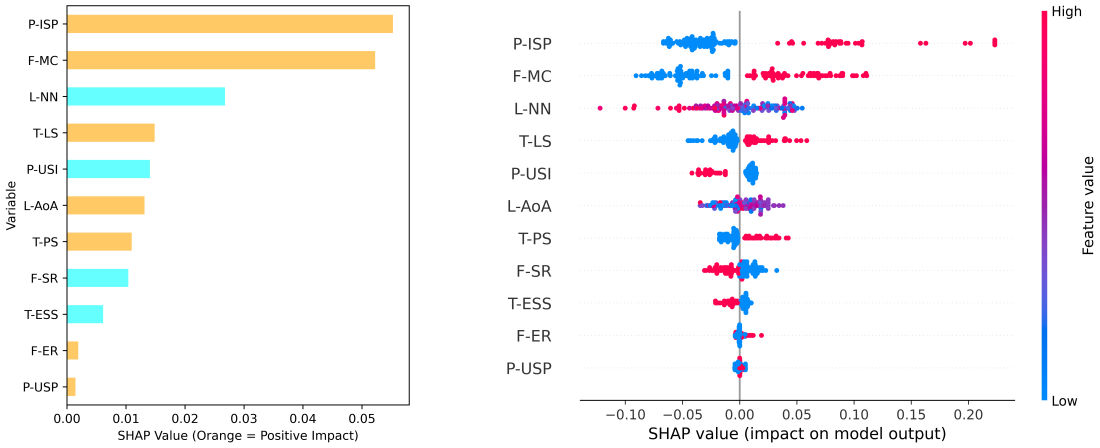


Figure 4.3: Violin plot for average feature importance scores for 1000 RHO for random forest; the line indicates the mean value for each feature.

that show the distribution of feature importance values (provided by the sklearn software package) associated with the 1000 RHO case for the random forest ensemble of regression trees. The tree-based importance agrees with the regression model on the importance of the multiple choice format and identifying science principles practices features. However, the tree-based importance differs from the regression model in assigning more importance to linguistic features and less to topics. This difference could be due to non-linear effects and/or interdependence between features.

For the random forest trained on all data with parameter selection, we also compute the SHAP values, which are visualized in Figure 4.4. The left figure shows average magnitude of the SHAP value for each feature, with the features having positive impact (associated with easier problems) in orange, and those with negative impact (associated with harder problems) in cyan. The right figure illustrates the distribution of SHAP values for individual items (width indicates density of items with a particular SHAP value). Figure 4.5 shows the average SHAP magnitudes for the 1000x RHO setting, where, unlike the feature importance in figure 4.3, we can also see whether the features are associated with increasing or decreasing p-values. The associated sign (whether the feature has a positive or negative impact) is determined by taking the sign of the correlation of the SHAP values with the feature values, computed individually for each of the 1000 runs.

Using the SHAP framework we can examine the main effect of each feature in more detail, as



(a) Average SHAP Magnitudes

(b) SHAP Item Values

Figure 4.4: SHAP analysis of the different features as used in the random forest predictor trained on the full data set with parameter selection.

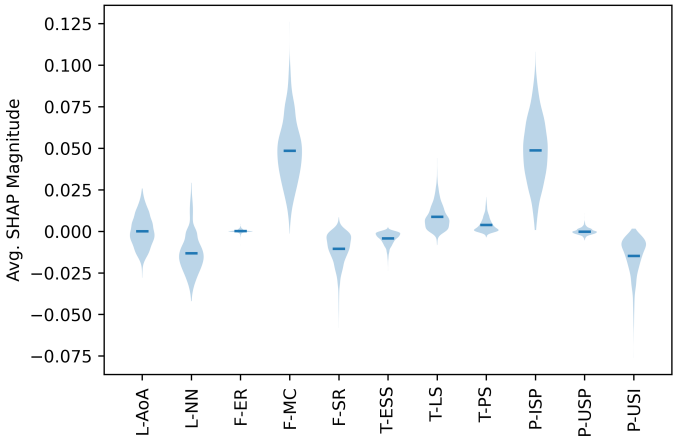
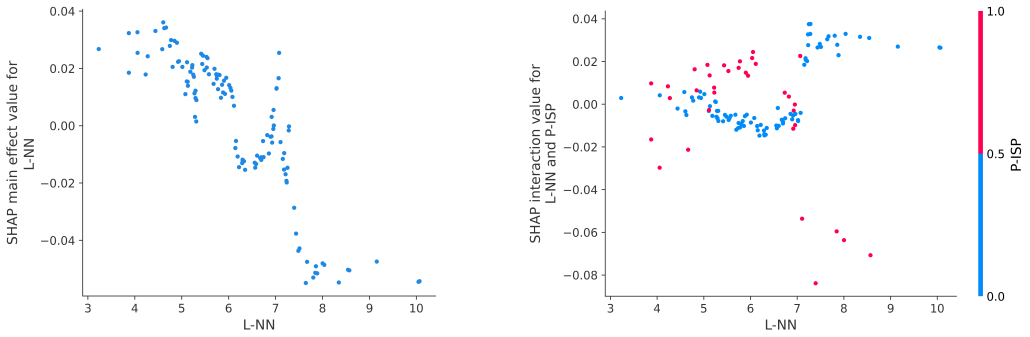


Figure 4.5: Violin plot for average SHAP magnitudes for 1000x RHO for random forest, the line indicates the mean value for each feature.



(a) Main effect of L-NN

(b) Interaction effect of L-NN and P-ISP

Figure 4.6: Main effect of neural network grade level prediction (left) and interaction effect of the grade level prediction and identifying science principles (right) on the item difficulty prediction for SHAP values for the random forest regression model.

well as pairwise interactions of features. To better understand differences between linear regression and random forest feature interpretation, we examine the linguistic feature neural network grade level prediction (L-NN). The SHAP values for the main effect of the L-NN is shown in figure 4.6 (left), and its interaction with the P-ISP feature is shown in figure 4.6 (right). The main effect figure indicates that there is a non-linear relation between L-NN linguistic complexity and item difficulty in that there is little difference between higher grade levels. The interaction effect figure shows a reverse effect of linguistic complexity depending on whether or not the item involves identifying science principles. Higher linguistic complexity only leads to greater difficulty for items involving "identifying science principles."

Looking at pairs of features, we find that the largest effect is from the science practice "identifying science principles" interacting with other variables, specifically the neural network grade level prediction, the format multiple choice, and the topic life sciences. For the interaction of P-ISP with the topic life sciences and multiple choice format, the average interaction effect values are presented in the tables 4.6. The values show that the combination of identifying science principles and multiple choice format contributes to overall item difficulty, while the combination of other question formats and identifying science principles decreases item difficulty. The trend is flipped for the

Table 4.6: Interaction Effects of Important Factors for Item Difficulty Prediction for the SHAP Values for the Random Forest Regression Model.

(a) P-ISP and F-MC

	F-MC	Not F-MC
P-ISP	-0.092	0.078
Not P-ISP	0.020	-0.014

(b) P-ISP and T-LS

	T-LS	Not T-LS
P-ISP	0.016	-0.018
Not P-ISP	-0.007	0.016

topic life sciences, though the effect is weaker.

4.4 Conclusions

In summary, this paper introduced a method for studying the impact of item characteristics on item difficulty, with the goal of identifying trends that are likely to generalize to a new set of items. Our review of prior work on analysis of item characteristics finds that most studies have leveraged methods that are problematic in two respects. First, many studies report goodness of fit using the same data that the model parameters are learned on. When given a small amount of data, as is the case for work on science assessment items, a general problem with learning prediction models is overfitting. When using a large number of features or a complex model, it is possible to fit one data set with high R^2, but have poor generalization to new data. In the measurement literature, standard methods to counteract for overfitting (e.g. adjusted R^2, AIC and BIC) are based on assumptions that do not hold for many models of interest. A more reliable approach for the small datasets used in education research is to conduct repeated trials of hold-out testing, either through cross-validation or random sampling. This approach is widely used in machine learning and grounded in theory; our work simply demonstrates the impact empirically for item difficulty prediction. Second, most work is based on multiple linear regression, which relies on the assumptions that the input features are statistically independent and linearly related to the prediction variable. When these assumptions do not hold (which we show is the case for item characteristics investigated here), the

feature importance analysis is compromised. Tree-based regression functions offer more powerful but still interpretable models.

As an alternative to these commonly used methods, our work supports the use of:

- repeated hold-out (RHO) measurement of R^2 using a large number of iterations (e.g. 1000), with random selection of training samples in each iteration and significance testing of model differences based on sets of RHO measurements for evaluating and comparing models (instead of reporting R^2 fitting to the full data set);

- random forest regression trees with parameter selection, which can capture non-linear effects and feature interactions;

- analysis of feature importance using Tree Explainer (Lundberg et al., 2020).

For the NAEP data set used in our study, the random forest gives an R^2 of 0.21, compared to 0.14 for multiple linear regression, which is statistically significant with $p < .001$ using a t-test. The random forest predictor was also found to be the best model for TIMSS data in (Sinharay, 2016), though no significance tests are reported.

When redundant dummy variables are used in encoding item characteristics, all models we explored identify the same two features as having the greatest impact: the multiple choice format and identifying science principles, both of which are associated with less difficult items. Both linear regression and the random forest show that extended response questions and using scientific inquiry lead to more difficult items, but the effect is smaller because there are fewer items of this type. The item format findings are consistent with prior work on associating item features to item difficulty (e.g. (El Masri et al., 2017; Le Hebel et al., 2017; Mesic, 2011; Mesic and Muratovic, 2011)). However, to our knowledge the findings related to science practice features are new which may be due to the fact that this cognitive framework is used only with NAEP items.

The main effects of the item format and science practice features on item difficulty need nuanced interpretation, partly due to their interaction with other item features, which is further illustrated by our findings based on the Tree Explainer tool. For example, we observe interactions between science practices and both item format and topic on item difficulty. More specifically, items that require

identifying scientific principles in multiple choice format are statistically the easiest whereas those items in the other two formats are found the hardest; however, items that involve using science principles or using scientific inquiry do not yield difference in item difficulty across item formats. We additionally find: i) a significant, but non-linear effect from linguistic complexity as characterized by a neural network predictor of reading level; and ii) interactions between linguistic complexity (reading level) and science practices. (El Masri et al., 2017) hypothesized that negative results associating linguistic complexity with difficulty may be related to the fact that most linguistic complexity features are sparsely represented in assessment items. The neural reading level predictor was designed to address this issue. It was still found to be insignificant in linear regression analysis, but tree-based analysis suggests that this may be due to a non-linear relation to difficulty and interaction with other item characteristics.

Being able to tease out the interaction effects among item features helps revisit and reconcile the inconsistent findings that stem from different research designs. While the item format is reported with significant main effect in research that link item features to item difficulty such as ours, research with experimental design (e.g., (Bridgeman, 1992)) supports the construct equivalence between multiple choice and open-ended formats in some cases as the comparable difficulty of total scores. Such inconsistency stems from the fact that the former type of inquiry may use statistical methods that fail to take into account feature interactions, in contrast to the latter in which one or few item features are manipulated while many others are kept as controlling variables. Therefore, choosing appropriate modeling approaches as we recommend earlier enables us to address the risk of overlooking interaction or non-linear relations.

In this study, we considered only a few easily obtained item characteristics as our main goal is to demonstrate the methodological contributions when analyzing data with a limited sample size. The methodology we highlighted could be applied to other item characteristics. It would also be useful to validate the findings on other sets of assessment items, though there is currently limited public availability of difficulty parameters together with item texts. An important future direction for this inquiry is to extend the methodology to analysis of performance with both item and student characteristics, to assess differential performance as a function of demographics and other student-related variables. Lastly, it is important to note that machine learning is impacted by biases in the data, so factors that are important in one data set may not be in another, particularly if there are

differences in the student population involved in the assessment.

Chapter 5

IMPACT OF LANGUAGE DIFFICULTY AND STUDENT DEMOGRAPHICS ON PERFORMANCE: A STUDY OF CALCULUS-BASED PHYSICS

This chapter further explores student performance on STEM assessments, looking at introductory college-level physics. Our focus is on performance differences among demographic groups, as well as analysis of whether student level demographic factors can predict student performance. Women and minority groups are underrepresented in physical sciences and engineering in 2 and 4-year institutes both in terms of enrollment and degrees awarded (for Science and Statistics, 2019). Studies have looked at performance gaps in terms of gender and ethnicity, but findings are mixed. Multiple studies find that under-represented minorities have a lower conceptual understanding when starting the course, as well as on the post-tests (Henderson and Stewart, 2018; Brewe et al., 2010; Hazari et al., 2007). While (Henderson and Stewart, 2018) shows that ethnicity is a significant predictor of performance, work by (Salehi et al., 2019) indicates that ethnicity is no longer significant when scores on SAT/ACT are accounted for. Additionally, work by (Santelices and Wilson, 2010) found that standardized tests show ethnicity based differential item functioning. Examining the predictors of course outcomes, work in (Hazari et al., 2007) found that while overall differences based on gender are not significant, once the incoming high school academic preparedness was accounted for, female students did significantly worse in introductory physics courses than male students with the same incoming academic proficiency.[1] On the other hand, results from (Salehi et al., 2019) show that when academic preparedness is including in the model, gender is no longer significantly predictive of performance. Existing studies rely on linear models, which we showed in the previous Chapter can give misleading results when there is interdependence among features.

Additionally, when considering course content and assessments, language difficulty can impact students' performance. Studies have examined the impact of language difficulty on question difficulty in K-12 STEM (Rosca, 2004; Crisp and Grayson, 2013; El Masri et al., 2017). Work has

[1]The study only included students who had taken high school physics.

also looked at the impact of language proficiency on student performance in general chemistry (Pyburn et al., 2013) which found that prior language comprehension proficiency shows a correlation with performance on chemistry. The study in (Taibu and Ferrari-Bridgers, 2020) looked at students' anxiety about jargon in physics courses, and showed a week negative correlation with student performance. Work in (Taibu et al., 2017) explores the link between language ambiguities and concept proficiency in the context of the word "weight" for an introductory physics course aimed at preservice elementary and middle school teachers. The work showed that explicitly teaching about language ambiguity after teaching physics concepts without the use of the word "weight" led to significant performance gains for students.

This chapter presents an exploratory analysis of proficiency for different demographic groups including gender, underrepresented status and ethnicity, as well as intersection of these groups for three pretests on an introductory calculus-based physics course taught at the University of Washington from 2005-2016. We explore the utility of student demographics in predicting scores on pretests that students take through the course of instruction, to assess whether demographic factors can be predictive of performance, and if so, whether this changes over the course of instruction.

We explore the impact of both demographics and linguistic complexity, as measured by the automatically extracted measure of linguistic complexity developed in Chapter 3 using non-linear models for predicting student performance explored in Chapter 4. We use techniques presented in (Lundberg et al., 2020) to analyze the relative importance of features for predicting student performance on individual questions.

The results show that for the first pretest, gender is the strongest predictor of student performance, while linguistic complexity is one of the strongest predictors for the third pretest, which is conducted later in the course. When all pretests are taken together, demographics are not predictive of student performance.

The rest of the chapter is organized as follows: section 5.1 presents details on the pretest data and the demographic of the student respondents. Section 5.2 provides our analysis on student performance differences by demographics. Prediction models and results are presented in section 5.3. Discussion and conclusions are presented in section 5.4.

	Pretest 1	Pretest 2	Pretest 3
Topic	Acceleration 1-D	Newton's 2nd & 3rd laws	Dynamics (rigid bodies)
Number of quarters	21	24	13
Total questions	15	14	11
MC questions	9	9	6
Avg. # students/Qtr	244	310	348

Table 5.1: Pretest data

5.1 Data

The pretest data is for an introductory calculus-based physics course on mechanics (Heron, 2015). The three tests are for the topics: i) acceleration in one-dimension, ii) Newton's 2nd and 3rd laws, and iii) dynamics of rigid bodies. The course is taught over a ten week academic quarter, and it is taught every quarter. The data is for fall 2005 - winter 2016, not including summer quarters, except for the pretest on the dynamics of rigid bodies, which is for spring 2011 - winter 2016. The pretests are given after instruction, but before the tutorial sessions. The first pretest, acceleration in one-dimension, is given at the beginning of the quarter, and can be considered as representative of students' high school preparedness for the test topic. (*Note:* We only have results for all 3 pretests in a course for only 3 quarters, and we do not have complete information for students who take all three pretests.)

The details for each pretest are shown in table 5.1. Each pretest has a set of multiple choice (MC) questions, most of which are followed by constructed response questions asking to explain the reasoning behind the selected choice. The scores for the tests are computed using only the multiple choice questions, which are automatically graded. Each correct response has a score of 1, and an incorrect response is assigned a score of zero.

The demographics of the respondents are shown in table 5.2, which include gender, ethnicity and underrepresented minority (URM) status. We see that the demographics of the students are skewed,

(a) URM status and gender

Pretest	**1**	**2**	**3**
Total respondents	5133	7442	4518
URM			
Yes	7.03	7.63	6.71
No	89.07	89.63	91.39
Not indicated	3.90	2.74	1.90
Gender			
Female	29.03	27.41	27.00
Male	70.82	72.39	72.69
Not indicated	0.16	0.20	0.31

(b) Ethnicity

Pretest	**1**	**2**	**3**
Ethnicity			
Afro-Am	1.50	1.42	1.17
Amer-Ind	0.88	1.05	0.97
Haw/Pac	0.72	0.85	0.69
Hispanic	3.94	4.31	3.87
Asian	29.50	29.97	30.10
Caucasian	51.10	45.65	44.05
International	8.47	14.02	17.24
Not indicated	3.90	2.74	1.90

Table 5.2: Percentage of respondents broken down by demographics for the pretests.

with underrepresented students making a small percentage of the group. Students self select as underrepresented, which typically follow ethnicity demographics, with some outliers. For our analysis, we use ethnicity as indicator for URM status, with *Hispanic, Hawaiian/Pacific, African American* and *American Indigenous* students identified as URMs, and *Caucasian, Asian* and *International* as not URMs, and a third group, *Not Indicated*, which includes all students whose ethnicity is *Not Indicated*. We have three groups for gender, *Male, Female* and *Not Indicated*.

5.2 Distributional Analysis

We present analysis of student performance on each of the pretests, and look at differences by demographic. This is done for all data, as well as by quarter. The average score for the three pretests are computed across all respondents, and then computed for students within a demographic group. Results are presented in table 5.3. We tested for significance between a focal group and either the

(a) URM status and gender

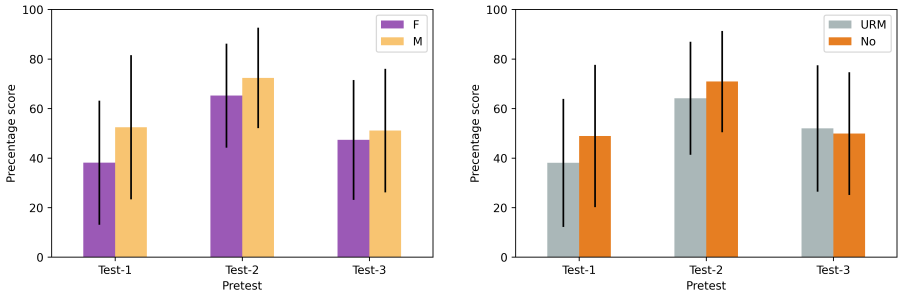
Pretest	1	2	3
Average score	4.35	6.34	3.01
URM			
Yes	3.43*	5.77*	3.12
No	4.42	6.38	3.00
Not indicated	4.74	6.46	3.28
Gender			
Female	3.43*	5.87*	2.84*
Male	4.72	6.51	3.07
Not indicated	4.88	6.47	3.64

(b) Ethnicity

Pretest	1	2	3
Ethnicity			
Afro-Am	2.86**	5.25**	2.83**
Amer-Ind	3.93	6.26	3.57
Haw/Pac	4.00	5.65	2.94
Hispanic	3.43	5.86	3.13
Asian	3.87	5.99	2.89
Caucasian	4.57†	6.63†	3.26†
International	5.27	6.43	2.51
Not indicated	4.74	6.46	3.28

Table 5.3: Average student scores broken down by demographics for the pretests. Maximum scores are 9, 9 and 6 for pretest 1, 2 and 3 respectively. * indicates that the group (URM and F) had significantly worse performance than the control group (not URM and M). ** indicates that the group (Afro-Am) does significantly worse than all other students. † indicates that the group Caucasian does significantly better than all other students. (Significance testing via *t*-test, $p < 0.01$)

control group or all other students' performance using a *t*-test. We see that difference in performance based on gender follow a consistent trend for all three tests, with male students significantly out performing female students. For pretests 1 and 2, we see that URM students perform worse than students who are not URM, however there is no significant difference for pretest 3. For ethnicity, in pretests 1 and 2, African American students perform significantly lower than other students. The difference in aggregate scores for the focal groups (URM and F) compared to the control groups (not URM and M) decreases over the course of instruction, which is shown in figure 5.1. For URM vs. not URM, there is no significant difference for pretest 3.



(a) Gender (b) URM

Figure 5.1: Average pretest scores (percentage) by gender and URM status for pretests 1, 2 and 3, showing that the gap between the focal and control groups decreases from pretest 1 to pretest 3.

Average student scores for the intersection of the demographics gender and ethnicity are shown in table 5.4. We see that the gender differences persist across all ethnicity groups for the first test, and most groups for the second and third tests.

Since the tests were administered across multiple quarters that span several years, we look at each quarter the test was given, specifically for gender (male/female)[2] and URM status (yes/no)[3]. The results for gender are presented in figure 5.2, and those for URM status are presented in figure 5.3. For both sets of figures, we see the most pronounced performance difference for pretest 1, and no visible difference for pretest 3. The results for gender in pretest 1, taken at the start of the quarter, seem to be consistent with the findings in (Salehi et al., 2019), that female students have a lower proficiency in concepts coming into the physics course.

5.3 Prediction of Student Performance

To analyze whether student level demographic features are predictive of students answering a question correctly or incorrectly, we train models to predict student performance. We train binary classification models to predict whether a student will answer a question correctly or not. Our feature

[2]While we have a third category for gender, N*ot Indicated*, there were not enough respondents when broken down by quarter to present meaningful analysis.

[3]Like gender, URM *Not Indicated* did not have enough respondents when broken down by quarter to present meaningful analysis.

(a) Pretest 1-average score 4.35

Gender	Afro-Am	Amer-Ind	Haw/Pac	Hisp	Asian	Caucasn	Intrnl	Not Ind
Female	2.67	2.70	3.00	2.91	3.24	3.36	4.82	3.37
Male	2.98	4.29	4.37	3.66	4.15	5.01	5.51	5.34

(b) Pretest 2-average score 6.34

Gender	Afro-Am	Amer-Ind	Haw/Pac	Hisp	Asian	Caucasn	Intrnl	Not Ind
Female	5.51	5.14	4.85	5.57	5.63	6.01	6.27	5.66
Male	5.13	6.50	6.22	5.97	6.15	6.82	6.50	6.75

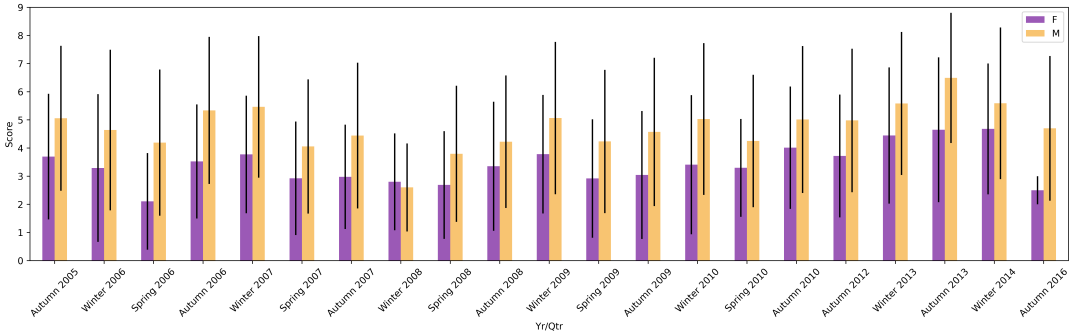
(c) Pretest 3-average score 3.01

Gender	Afro-Am	Amer-Ind	Haw/Pac	Hisp	Asian	Caucasn	Intrnl	Not Ind
Female	2.93	3.30	2.80	2.75	2.70	3.05	2.56	3.51
Male	2.79	3.65	3.06	3.27	2.97	3.32	2.48	3.15

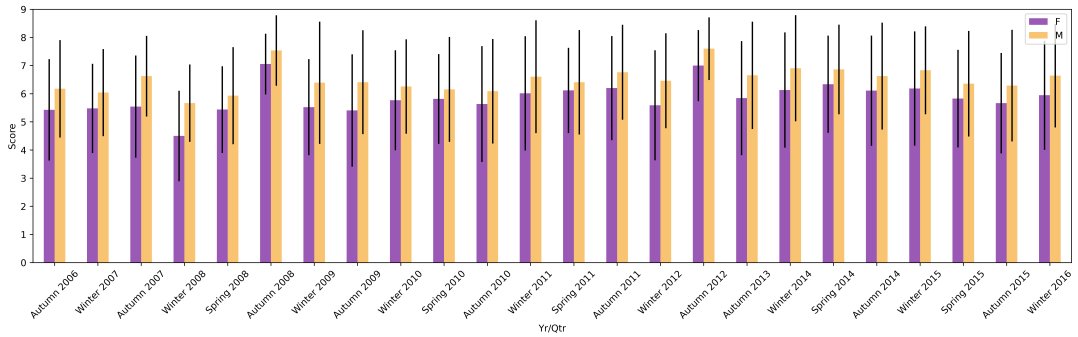
Table 5.4: Average student scores on the three pretests by intersection of ethnicity and gender. The first four ethnicity groups fall under URM status. (Note: the gender group *Not Indicated* was omitted from this table due to too few samples in each group.)

set includes all demographic features in our data set, the average score on each question and the linguistic complexity quantified as grade level for each question. The features do not include variables that indicate either the proficiency level of a student, or the skills needed to answer a question correctly. Thus any improvement over baselines will be indicative of demographic and/or linguistic features being predictive of performance. We first present results for each pretest individually, which allows us to investigate whether the importance of demographic features changes over the course of instruction. We also present analysis for all three pretests combined to identify features that remain important across the course of instruction.

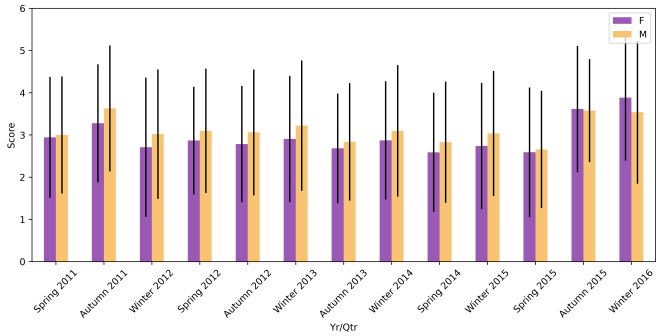
To examine individual questions, we also train models to predict performance for each question



(a) Pretest 1



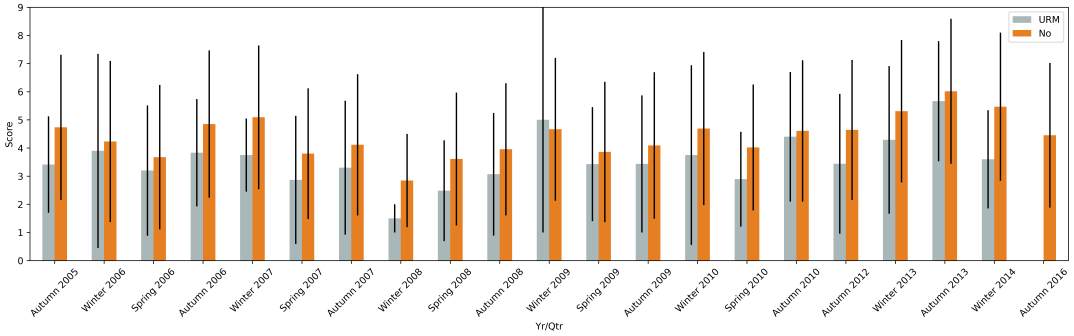
(b) Pretest 2



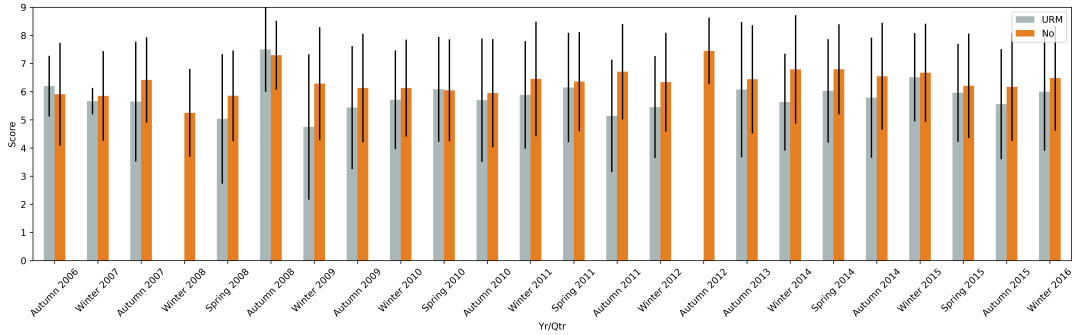
(c) Pretest 3

Figure 5.2: Average pretest scores by quarter taught for gender groups female and male.

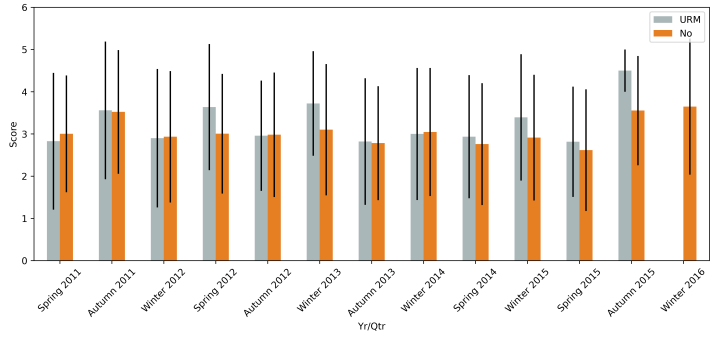
individually. We compare against baseline metrics to identify questions where demographics are predictive of performance, and identify the highest importance features for these questions.



(a) Pretest 1



(b) Pretest 2



(c) Pretest 3

Figure 5.3: Average pretest scores by quarter taught for URM status. Missing data indicates too few students in the group.

5.3.1 Predictor variables

Two sets of predictor variables are used in the analysis, question level and student level. We also include whether the test was taken in the autumn, winter or spring quarter. For question-level variables, while we do not include cognitive features for questions (content knowledge required to answer the question), we include the average score on the question (Qn-Avg) and a measure of language difficulty. For language difficulty, we can use either hand-crafted features that indicate difficulty of the text (e.g. average length of sentences, ratio of sub-clauses to clauses), or we can use automated methods that directly predict linguistic complexity without hand crafted features. We take the later approach, using a neural network based system, bidirectional context with attention (BCA), for predicting the difficulty level of the questions developed in Chapter 3, which gives the state-of-the-art performance on shorter texts like questions. The system quantifies linguistic complexity as grade level between 0 (Kindergarten) and grade 12, and can output both discrete and continuous numbers for that range. For the analysis, we get the continuous grade level predictions from the pre-trained system[4] for each question (L-NN). For the text of a question, we count everything that has to be read to answer the question, including the preamble for the question (that might set the context for multiple questions), the question itself, and answer choices where present.

Including the average score on the question for all students will allow the model to account for the basic difficulty level of the question. We incorporate the predicted linguistic complexity of the question to assess whether language difficulty shows an impact on student performance. Using models that account for feature interdependence will allow us to see whether it interacts with other predictor variables, such as student demographics. For student-level variables, we use gender, ethnicity, and URM status. The predictor features are shown in table 5.5, and the total number of data points are presented in table 5.6, which is the total number of student respondents times the number of questions for each pretest.

5.3.2 Prediction model

The target variable is binary, whether the student scored correctly (1) or not (0), for which we use a classification model. Most work in the space of student performance prediction, including

[4]https://github.com/Farahn/Liguistic-Complexity

Feature scope	Features	Values
Pretest	Quarter the test was taken	3 (Autumn, winter, spring)
Question	Question avg. score (Qn-Avg)	Continuous number between 0 and 1
	Linguistic complexity (L-NN)	Continuous number between 6.4 and 9.3
Student	URM status (URM)	3 (Yes, no, not indicated)
	Gender (Gen)	3 (Female, male, not indicated)
	Ethnicity (Eth)	8 (Afro-Am, Amer-Ind, Haw/Pac, Hisp, Asian, Caucasn, Intrnl and Not Ind)
Predicted	Score on a question	0 (incorrect), 1 (correct)

Table 5.5: The features used as predictors for predicting whether a student responds correctly or incorrectly on a question.

(Hazari et al., 2007; Salehi et al., 2019) use linear models, thus we present a logistic regression baseline. However, linear models rely on the assumptions that the features are independent and that the relation between the predictors and the target is linear. We demonstrated in Chapter 4 that both of these assumptions are problematic, and further, do not allow us to study the interaction between features. Thus for this analysis we use random forests, which are an ensemble of decision trees, and allow us to capture feature inter-dependence.

5.3.2.1 Experimental setup

We use the random forest implementation from sklearn (Pedregosa et al., 2011). For the experimental set-up, we use five-fold CV. All data points from a quarter are kept in one split, and the quarters in each split are randomly selected so that each of the five splits is approximately the same size. For each iteration, we use four folds for training, and one for testing. We use the four training folds to perform a grid search in a five-fold cross-validation (CV) setting to find the optimal parameters of

Dataset	Total data points
Pretest 1	46197
Pretest 2	66978
Pretest 3	27108

Table 5.6: Number of data points (number of student respondents x number of questions on a test) for the prediction models.

maximum tree depth and number of estimators in the ensemble. The value of Qn-Avg is computed using only the four folds of training data, not the test data. We then use all the training data to train the model with the best parameters, and report results on the held out test set. This process is repeated five times, training five models. We report the average classification accuracy on the five CV test folds. This set up is repeated individually for each pretest.

While we demonstrated that repeated hold-out, an alternate hold-out testing method, gave more reliable results for the analysis of middle school items in Chapter 4, in this case we want to control for the split to have full quarters in either training or test, which limits the additional variability we can get from random sampling. Thus we present results for repeated CV (RCV), in which we create 20 distinct 5-fold CV sets, obtain the 5 test split accuracy for each setting, and present the average score for the 20x5 test sets, 100x RCV. For all cases, CV and RCV, we compute weighted average of the test accuracy, since the test splits are not the same size as we control for quarters not being divided across the splits.

We have two baselines: i) predicting the majority class for the data and, ii) using a one feature logistic regression model in the same set up as five-fold CV and 20x RCV using only the average question difficulty as a predictor (LR (Qn-Avg)).

		RCV Test Accuracy		CV Test Accuracy	
Dataset	**Majority**	**LR (Qn-Avg)**	**RF**	**LR (Qn-Avg)**	**RF**
Pretest 1	51.7%	56.7%	59.2%	56.1%	58.6%*
Pretest 2	70.4%	78.7%	74.3%	78.7%	73.9%
Pretest 3	50.1%	60.4%	60.7%	60.5%	60.9%*

Table 5.7: Random forest classifier and baseline performance for the three pretests. The logistic regression (LR (Qn-Avg)) baseline uses one feature, the average question score computed on the training data. (* indicates significantly better performance than the baseline at $p < 0.005$ for CV.)

5.3.2.2 Results

We report the mean CV test accuracy, the baseline accuracy when always predicting the majority class, and the logistic regression with the average question difficulty baseline (LR (Qn-Avg)) in table 5.7. The results show that the classifier performs better than the majority class baseline for all pretests. However, to study the gains from adding demographic and linguistic features, the random forest baseline with the average question difficulty as the predictor provides a better comparison. For pretest 1, additional features improve performance over this baseline by 2.5%, and for pretest 3 by 0.5% (both significant with $p < 0.005$), while performance for pretest 2 is lower than the baseline.[5] We see similar trends for the repeated CV setting. This indicates that student demographics and language difficulty are predictive of performance for pretest 1 and 3. For pretest 2, these factors are not predictive of performance. The results show that additional features are most predictive at the start of instruction, and this effect decreases later in the course.

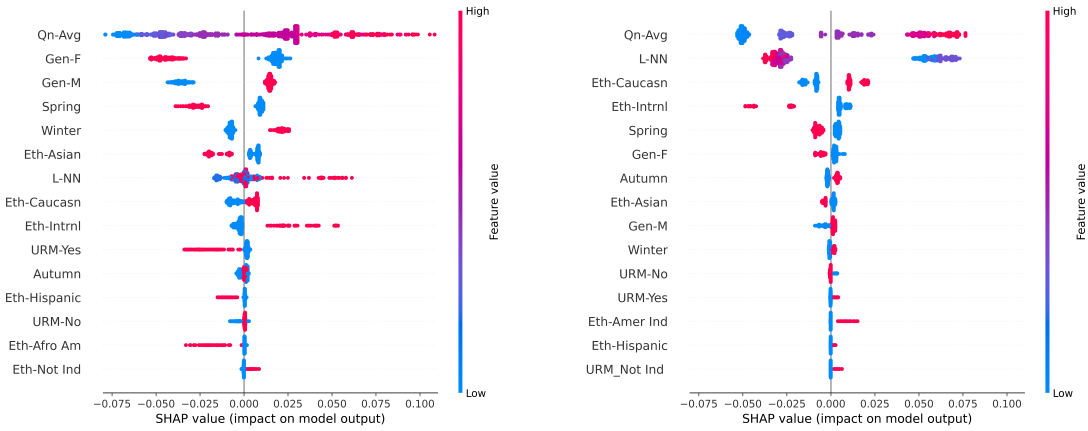
[5]The logistic regression baseline gives similar levels of accuracy as a random forest trained using only Qn-Avg, with the same trends for significance.

5.3.3 Feature importance

To visualize importance of features in contributing to the prediction model, we use the Tree Explainer from (Lundberg et al., 2020), which uses the SHapley Additive exPlanation (SHAP) framework for interpreting complex models (Lundberg and Lee, 2017). For this, we train the model using all the data with the best parameters learned in a CV setting using the same five splits as above, then use the explainer to visualize the contribution of each feature for each of the three pretests for the trained RF model.

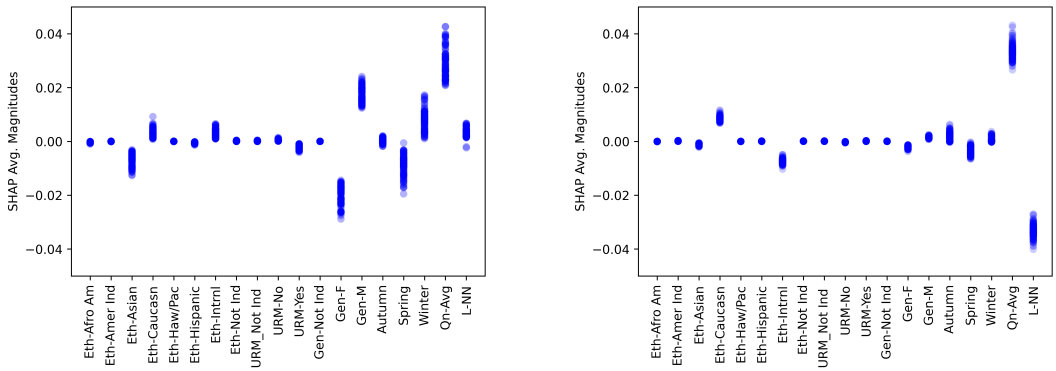
We compute SHAP values, which are visualized in figure 5.4 for pretests 1 and 3, which show a higher performance than the two baselines, displaying the top 15 features. The figure illustrates the distribution of SHAP values for individual items (width indicates density of items with a particular SHAP value). For this model, which gives us the highest gain over the baseline, we see that the average score on a question (Qn-Avg) is the strongest predictive factor, which we expect. The next two predictive features are gender, with a negative impact associated with being female, and a positive impact associated with being male on the probability of responding correctly. This ties in with gender based performance difference for pretest 1 (table 5.3). The third set of features ranked by importance are whether the test was taken in Spring or Winter quarter, with students performing better in Winter. For pretest 3, the top feature is L-NN, showing that lower reading level contributes to a higher probability of scoring correctly. The second feature is Qn-Avg, followed by the ethnicity indicator for international students, with a positive value tied to lower probability of scoring correctly.

For pretests 1 and 3, where our feature set performs better than the baseline, we also use SHAP to study the variation in the contribution of each feature in the RCV setting. We compute the average SHAP magnitude for the features for each of the 100 models trained for the 20x5 training instances, shown in figure 5.5. The figure for pretest 1 (left) shows the highest magnitudes for Qn-Avg and the gender groups male and female. Pretest 3 (right) shows that the highest magnitude features are Qn-Avg and L-NN, both of which show a low variance. Comparing the two figures, we see that when L-NN has a high impact (pretest 3), the demographic identifier for international students shows a negative impact. On the other hand, for pretest 1, where L-NN does not have a high contribution, the identifier for international students has a positive impact on the probability of correct response.



(a) Pretest 1 (b) Pretest 3

Figure 5.4: SHAP analysis of the different features as used in the random forest predictor trained on the full data set with parameter selection for pretest 1 (right) and pretest 3 (left).



(a) Pretest 1 (b) Pretest 3

Figure 5.5: SHAP average magnitude for features as used in the 100 random forest predictors trained in the RCV setting for pretests 1 and 3.

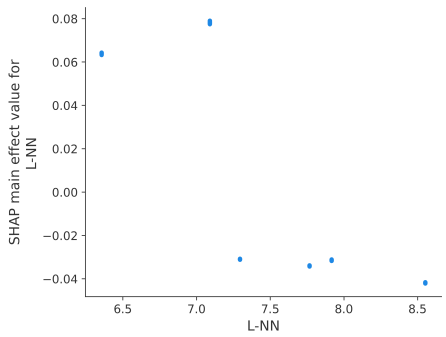


Figure 5.6: SHAP main effect of neural network grade-level prediction for pretest 3 on the probability of correct response for the random forest classifier.

Using the SHAP framework, we can examine in more detail the main effect of each feature on the probability of correct response. The SHAP values for the main effect of the L-NN for pretest 3 is shown in figure 4.6. For this pretest, we see an impact that indicates that higher linguistic difficulty decreases the probability of correct response, with the threshold slightly above 7, which is similar to the effect for NAEP grade 8 science assessments as shown in Chapter 4. The main effect indicates that there is a non-linear relation between L-NN linguistic complexity and probability of correct response, which strengthens the argument for using non-linear models for analysis.

5.3.4 Performance prediction for the combined pretest data

We repeat the above analysis for predicting student score, this time combining the data from all three pretests. To account for the pretests, we include one additional pretest level feature, the pretest number (‘Test-1’,‘Test-2’ and ‘Test-3’). This analysis allows us to look at which features are important across all three pretests. The total data points for this analysis are 140283, sum of all samples in table 5.6. The results are presented in table 5.8. The result for our model is worse than the logistic regression (LR (Qn-Avg)) baseline, showing that when all tests are taken together, our feature set is not predictive of performance.

		RCV Test Accuracy		CV Test Accuracy	
Dataset	**Majority**	**LR (Qn-Avg)**	**RF**	**LR (Qn-Avg)**	**RF**
Pretest 1,2&3	59.2%	67.7%	65.6%	65.5%	64.1%

Table 5.8: Random forest classifier performance for all three pretests together. The logistic regression (LR (Qn-Avg)) baseline uses one feature, the average question score computed on the training data.

5.3.5 Analysis of individual questions

We repeat the analysis of individual pretests,[6] this time taking the data for one question at a time. We do not include the linguistic complexity or the average question difficulty, since, for a given question, the values are constant for the entire training/test data. We compare the results with the majority class baseline, which is the percentage of times students respond correctly (or incorrectly) to the question for all the data. There were 6 questions for which the random forest with the full feature set performed significantly better than the majority vote, 4 questions in pretest 1, and 1 question each in pretests 2 and 3. For these questions, we then ran the SHAP analysis, and identified the top 3 features and the associated average SHAP values, which are presented in table 5.9. We see that for pretests 1 and 2, the demographics of gender and ethnicity are ranked highest by the individual models; for pretest 3, the quarter in which the test was taken is the most important feature.

5.4 Summary

This Chapter presented an exploratory study into student performance in introductory calculus-based physics. Looking at performance by demographics indicates that at the start of instruction, students exhibit performance differences linked to their gender and ethnicity. The performance gaps persist across the data we have looked at from 2005 to 2016. The intersection of gender and ethnicity indicates that female students tend to start with lower proficiency for all ethnicity groups,

[6]The three pretests are shown in appendix A.

with the lowest average scores for underrepresented female students. Later in the academic quarter the differences decrease, and are no longer significant for URM status. This trend is similar across all quarters we look at. One hypothesis for the decrease in the performance gap is that the initial difference is due, in part, to the incoming level of high school preparedness, which becomes less relevant over the course of the quarter. This can be further explored by looking at performance for pretests for topics not taught in high school, so we can see if the differential persists for those topics. Another possibility is that weaker female students dropped the class. A limitation of our data is that we have intersecting data for only three quarters for all three pretests, and we have missing data, which makes it challenging to study attrition and individual performance for a student as they move from pretest 1 to pretest 3.

We further investigate whether student demographics and linguistic complexity are predictive of student performance by training models to predict whether a student will respond correctly or incorrectly to a question. Our results show that for pretest 1, gender is a predictor of performance, with the trained model outperforming the majority class baseline. For both the later pretests, we see that demographics have a smaller impact on the model prediction, indicating again that the impact of demographics decreases over the course of instruction. We saw that for pretest 3, linguistic complexity is predictive of performance. Looking at the main effect of linguistic complexity for pretest 3, we see a pattern similar to that for NAEP grade 8 science questions presented in Chapter 4. This shows a consistent, non-linear, impact of language difficulty across middle school and college level STEM assessments, which has not been explored in prior work.

Our results provide insight into demographic based performance differentials at the start of instruction, which decrease as the quarter progresses. We observe that coming into the course, female and URM students persistently perform worse than male and URM students. The fact that this gap decreases with instruction is promising, and can be further explored looking at data from additional tests, as well as grades on the course. A limitation of our study is that we do not account for incoming academic preparedness of students. The analysis can be further strengthened by accounting for the academic level of incoming student in addition to demographics, to look for interaction effects. Automatically predicted linguistic complexity shows an impact on student performance for two of our three pretests. Here, it may be interesting to look at interaction effects with international student demographics. This work provides a method to study the non-linear impact of linguistic complex-

ity on student performance, and shows that for pretest 3, higher levels of complexity are tied to a lower probability of correct response. Predictive models with additional cognitive features can help explore this impact further, and to see whether language difficulty interacts with other aspects of a question, as it does for middle school assessments.

	Question	CV test accuracy	Majority vote	Top features	
	#	%	%	Feature	SHAP val.
Pretest 1	**1a**	57.0	51.7	Spring	-0.025
				Eth-Asian	-0.019
				Gen-F	-0.017
	3	59.9	50.8	Gen-F	-0.042
				Gen-M	+0.034
				Spring	-0.027
	5	60.1	52.1	Gen-F	-0.039
				Gen-M	+0.032
				Spring	-0.030
	8	58.7	56.0	Gen-F	-0.036
				Gen-M	+0.036
				Eth-Asian	-0.015
Pretest 2	**11**	59.8	56.7	Gen-F	-0.027
				Gen-M	+0.024
				Eth-Caucsn	+0.013
Pretest 3	**6**	54.8	51.8	Spring	-0.024
				Autumn	+0.010
				Winter	+0.009

Table 5.9: Random forest classifier performance for individual questions, with the average SHAP values for the top three features. (*Note:* only questions where the random forest model performed better than the majority class baseline are shown.)

Chapter 6

GENDER REPRESENTATION IN EDUCATIONAL TEXTS

This chapter presents our work on extracting gendered language from educational texts. Educational content, including textbooks and articles, often contain descriptions of people, either describing real people, e.g. historical figures, scientists, public figures etc. or fictional characters. Increasingly, STEM assessment items are situated in real-world scenarios. These contextualized questions, or "story problems," also often feature characters.

An example of an item is presented in figure 6.1, which shows the context for a science assessment question part of the Program for International Student Assessment (PISA) exam for the year 2006 (Pisa, 2015). The context includes four people in laying out the background for the questions that follow. All four of the people in this example are gendered. The context is centered around a woman, Mary Montagu, and features a man, Edward Jenner. The text of the question starts with a description of the physical appearance of Mary Montagu. The verbs used when the character

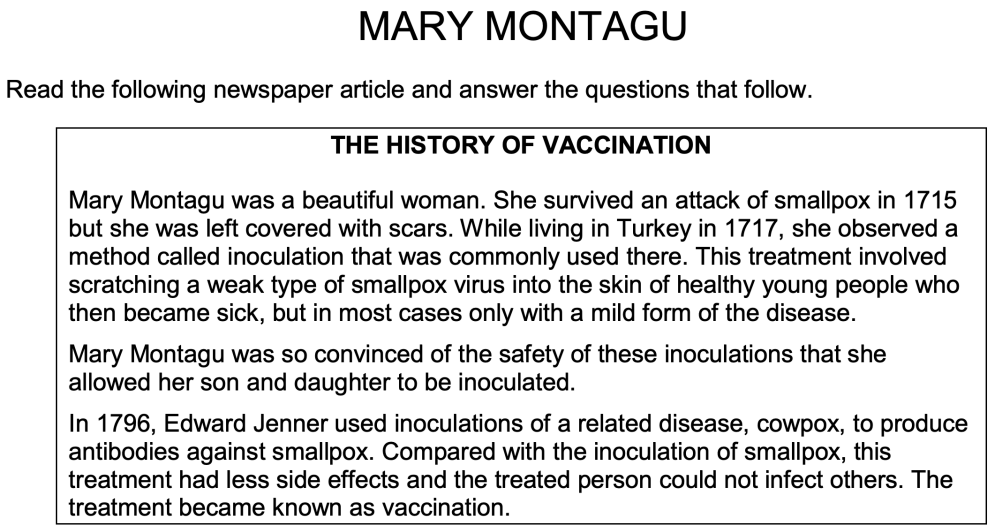


Figure 6.1: Question context for a PISA science assessment.

appears as subject include `survived`, `observed` and `allowed`. These verbs imply a lack of agency for the agent (subject), in this case Mary Montagu. On the other hand, the verb used when Edward Jenner appears as subject is `used`, which denotes agency for the agent. Another aspect that can be seen in the question is that the inoculation from Turkey is implied to be unsafe, while the vaccination developed by Edward Jenner, a name associated with Caucasian racial identity, is described as being safer.

The above example highlights that even standardized assessments, which are carefully controlled, can unevenly portray characters and contain cultural biases. In this context, research in education has explored biases in textbooks and assessments, including representation of gender, culture and ethnicity. This is important since such stereotypes can negatively impact students, both in terms of performance and in terms of retention, specially in STEM.

Existing work in this space looks at small samples, relying on expert annotation. While this allows for nuanced exploration of cultural and other biases, these methods do not scale to large collections. The ability to automate the analysis of language usage to assess for biases can help in examining existing sets of content and assessments, specially as tools that educators can use to profile content and items. Our work makes the following important contributions in this space:

1. Extending prior work on automatic extraction and analysis of agency and authority associated with gendered language to STEM education contexts;

2. Developing automatic activity classification algorithms that lend further insight into gender bias; and

3. Introducing a method for assessing text for bias based on gender prediction confidence, and leveraging SHAP analysis methods (Lundberg et al., 2020) to provide interpretable classifier results

Our proposed profiling tools can be used to evaluate the bias in gender representation and provide interpretable analysis of features that contribute to the bias, which can be applied to a collection of texts or items, as well as to individual items. We examine five datasets of educational content and four corpora of science and mathematics assessment items. Our results show that across all the

content datasets, mentions of masculine characters are twice as frequent as those of feminine characters. We further show that for one of the assessment data sets, feminine characters are associated with lower agency and authority, consistent with findings for movie scripts presented in (Sap et al., 2017). Our results on math assessment questions indicate that, in addition to power and agency, there may be differences in the activity contexts in which gendered characters are represented.

The rest of the chapter is organized as follows: section 6.1 presents the background with related work in education and NLP. Section 6.2 presents the methods used in our work, including tools for extraction of gendered language, automatic classification of activity categories, and evaluation of biased gender representation. Sections 6.3 and 6.4 present results for distributional analysis and gender prediction tasks. Analysis of word associations are presented in section 6.5, followed by the summary in section 6.6.

6.1 Background

This section gives an overview of research on the impact of gender representation, specifically negative stereotypes, on students studying STEM subjects, which motivates our work. We then present a research study that created a framework for examining power and authority of characters presented in movie scripts, which we use in our work, together with an overview of other studies that also apply these methods for automatic extraction of gendered language.

6.1.1 Impact of gender representation

Decades of research in social psychology has confirmed that performance can be undermined when a person is triggered by a negative stereotype about their identity group when that identity is salient. This includes prescriptive gender stereotypes (Prentice and Carranza, 2002), discrimination and sexism (Glick and Fiske, 2001), stereotype and social identity threat (Steele and Aronson, 1995; Murphy et al., 2007; Sekaquaptewa and Thompson, 2003), and even unconscious gender-stereotypical cues in the environment (Cheryan et al., 2009). In science education, alignment with stronger gender–science stereotypes (implicit associations and endorsement of male superiority in science) leads to women identifying less with science and, in turn, weaker science career aspirations (Cundiff et al., 2013). Similarly, stronger implicit math-male stereotypes corresponds with more negative implicit and explicit math attitudes for women than positive attitudes for men (Nosek et al., 2002). In other

words, women were negatively impacted by the stereotypical representations to a greater extent than men were positively impacted, suggesting that the target group, in this case women, is more susceptible to the effect of bias.

Gender stereotypes, norms, and roles in assessment show that testing as a social instrument fails to reflect gender equality values (Gayles, 2011). The impact of these associations is not limited to impressions. For example, students who encounter stereotype threats may hold less motivation to provide a correct response or simply experience cognitive depletion, which can lead to under performance of woman and girls of color in achievement tests (Gayles, 2011). Studies in (Galdi et al., 2014; Davies et al., 2002) have also demonstrated the negative impact of stereotypes on the academic performance of girls and women in mathematics. Thus analysis of gender bias in STEM assessments is important both for assessing impact on student performance, as well as expanding the discussion beyond achievement scores to the social consequences of testing. This can help build assessment tools carefully designed to combat discrimination.

6.1.2 Connotation frames for power and agency

Sap et al. (2017) introduced a method to quantitatively study gender biases in movie scripts by identifying patterns of agency and power implied by the usage of verbs associated with female and male characters. The work presented connotative frames of power and agency, where connotation frames present a formal method to capture the information, including assumed facts and sentiment, implied by a verb predicate (Rashkin et al., 2016). The approach leveraged a lexicon of power and agency which can be used to annotate verbs in a text, and whether that power and agency lie with the subject or object. The study found differences in terms of both power differential and agency of female vs. male characters. Positive agency and power were significantly associated with male characters, while negative agency was significantly associated with female characters.

Leveraging the connotation frames of power and agency, subsequent work has looked at automatic extraction of gendered mentions in movie summaries, news articles and fictional novels (Parthasarthi et al., 2019), as well as large scale analysis of whether men and women are talked about differently in faculty reviews and celebrity news, which found significant differences (Chang and McKeown, 2019). Work in (Lucy et al., 2020) applied the connotation framework to a set of

U.S. history textbooks used in Texas, showing negative agency and no power associated with mentions of women. To our knowledge, there has been no large scale automatic analysis of STEM educational corpora in terms of gender representation.

6.2 Methods

Our work focuses on analysis of feminine and masculine representation in educational texts, with a breakdown of roles and activities. This section presents the data used in our work and the methods for extraction and analysis of gendered language. We also introduce a gender classification model that serves as a tool for profiling gender bias.

6.2.1 Data

Two main types of datasets are used in this work, educational content and assessments.

- **Content**
  - Open source textbooks and articles for K-12 (Michigan; Siyavula; CK12) (compiled for linguistic complexity classification presented in Chapter 3).
  - Reading samples provided in the Common Core Standards (CCS) Appendix B,[1] which are used as examples for grade appropriate texts, both literary and informational, for K-12.
  - Three sources of educational articles for K-12 students, WeeBit (Vajjala and Meurers, 2012) (articles from BBC Bite-size and Scholastic Weekly Reader), OneStopEnglish (Vajjala and Lučić, 2018) and News ELA (Staff).
- **Items**
  - PISA science assessments (Pisa, 2015) typically taken by 15 year-old students nationally and internationally.[2]

[1] http://www.corestandards.org/assets/Appendix_B.pdf

[2] https://nces.ed.gov/surveys/pisa/countries.asp

- National Assessment of Educational Progress (NAEP) science[3] and math assessments, administered nationally.[4]
- Algebra problems from the AQuA dataset, which was compiled for automatic question answering (Ling et al., 2017).

The AQuA data set contains dedicated training and test sets. For our work we use the training set, which consists of approximately 100,000 samples. For the analysis, we filter the questions to include only those which include people in the text, including neutral, feminine and masculine mentions, resulting in 33k items (a third of the training dataset). This subset is then used in all the subsequent analysis.

6.2.2 Extraction of gendered language and associated roles

For extraction of gendered language, we first use three NLP tools: i) automatic part-of-speech (POS) tagging, associating words with dependency parsing labels, ii) automatic named entity recognition (NER), which involves identifying entities like people, places, object etc. in a text, and iii) dependency parsing. All three tasks are established NLP tasks, and both POS tagging and NER are able to achieve high accuracy (Manning et al., 2014; Yadav and Bethard, 2018). Specifically, for our automatic identification of characters and associated gender information in educational corpora, we use spaCy (Honnibal and Montani, 2017),[5] an open source python library for NLP. It was used to annotate the texts with POS tags to identify pronouns, nouns and verbs, as well as to identify persons in the text using NER. The dependency parsing allowed extracting associated verbs and syntactic role of each person mention.

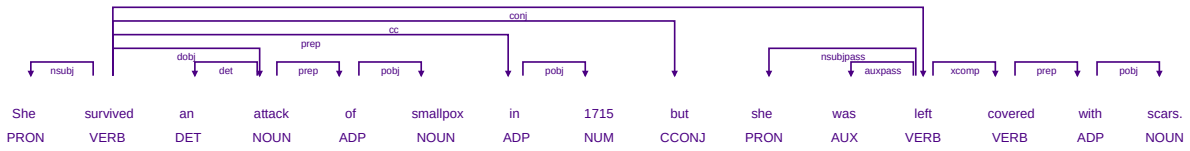
Part-of-speech tagging and named entity recognition allow for extraction of pronouns, names and nouns that indicate people in a text. To assign gender to these mentions, we use a database of names with associated gender (Michael, 2007), and lists of gendered and neutral pronouns and nouns.[6] For compound names such as `Edward Jenner`, we use the first name to get the gender,

[3]The same set of items used for analysis in Chapter 4

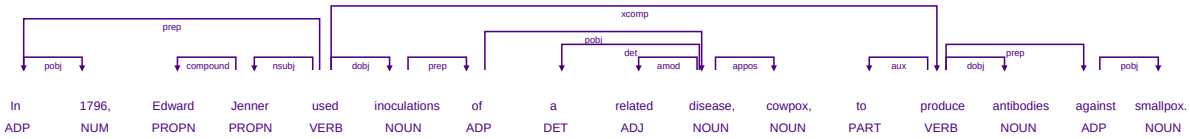
[4]https://nces.ed.gov/nationsreportcard/

[5]https://spacy.io/

[6]Available in Appendix C.



(a) "She survived an attack of smallpox in 1715 but she was left covered with scars."



(b) "In 1796, Edward Jenner used inoculations of a related disease, cowpox, to produce antibodies against smallpox."

Figure 6.2: Dependency parse examples, showing POS tags and synctactic relations.

since the second name is almost always a masculine name. For nouns, the identified words were matched to lists denoting people (e.g. `child`, `women`, `men`, etc.), categorized into three groups, masculine, feminine and neutral. For categorizing the names, the python package Gender Guesser was used, which uses lists of 50 thousand names from (Michael, 2007). Each name is matched to one of five groups: feminine, mostly feminine, androgynous, mostly masculine and masculine. Neutral mentions include nouns like `children`, `scientists`, `teacher` etc., as well as pronouns like `I`, `you`, `they`, `them`, `us`, `we`, and androgynous names. To try and reduce the ambiguity in pronouns, we tried using coreference resolution using spaCy, however the results were not sufficiently accurate for downstream tasks. While the neutral category contains mentions that do not convey gender information, for the datasets used in our study, it does not specifically contain mentions of gender neutral characters. Thus our subsequent work uses only two gender categories, masculine and feminine.

For person mentions we automatically extract features including the POS tag, the syntactic role (subject, object, other), the POS tag of the root, associated agency and authority if the root is a verb, and whether the word appears in conjunction with others for each gendered mention. If the identified POS tag of the root was `Verb`, we used the power and agency lexicons from (Sap et al., 2017) to extract the associated annotations if the lemmatized verb was present in the lexicons. Figure 6.2

shows two examples of dependency parsing for the question on vaccines. For example, in 6.2 (a) the root of the first `she` is the verb `survived`, and it has the role subject. For figure 6.2 (b), the root for `Edward Jenner` is `used`, again with the role subject.

Automatic extraction of gendered mentions and associated linguistic features allows us to conduct distributional analysis of the number of times feminine, masculine and gender-neutral characters are mentioned, the frequency they appear as subjects or objects in the text and the power/agency statistics. In addition, the automatically extracted information is used in a gender classifier, which serves as a bias profiling tool. Finally, the gendered mentions are used to partition sentences for word cloud analysis.

6.2.3 Text classification for activity categorization

The math assessments involve contexts that reflect a variety of activities, which we hypothesized might have gender representation bias. Therefore we develop a classifier to automatically label the activities presented in the question context, if any. For this task we trained text classification systems using the pretrained transformer model, BERT (Devlin et al., 2019). Since there is no annotated dataset available for this task, we created a set of 2135 annotated samples randomly selected from the AQuA training dataset. The activity categories were defined by graduate student collaborators in educational measurement. The initial set consisted of 11 categories, later merged into 10 due to a small number of samples. The categories are: academics, arts and creativity, business and finances, vehicles and travel, exercise and sports, food and drink, home maintenance, industry and farming, science and research, and social relationships. Detail about the categories and examples are presented in Appendix B.

Since the activity classifier output will be used as input to the gender classifier, we removed gender information in the training, development and test data by replacing all names with a generic <name> token and all pronouns with a generic <pronoun> token for the activity classifiers, so that the classifiers do not learn explicit mentions of gender.

Since an item can have one or multiple labels, ten independent binary classifiers were trained for each of the activity categories. The first round of annotation was done by hand, using crowdsourcing, followed by reexamining the items that were assigned no labels. The 2135 annotated

samples were then used to train ten classifiers, with 80% used for training, 10% for dev and 10% for testing. For each category, the 10% held out test set was used to find the probability threshold that maximized the F1 score (p_{F1}).

To increase the labelled data, an iterative thresholding approach was used. The trained classifiers were used to obtain predictions for the entire AQuA training corpus. The labels were then used to find additional positive and negative samples for each category by manually identifying predicted posterior probability thresholds, starting with the threshold identified for maximizing F1 score. The samples with probabilities higher than twice the threshold ($2p_{F1}$) were ranked in order of increasing probability, and randomly inspected to see whether the samples belonged to the category. The threshold was doubled until all the inspected samples belonged to the assigned category. These samples were then assigned a positive label for the category. For negative samples, the items with a probability below $0.5p_{F1}$ were examined, ranked in descending order. The threshold was halved until all inspected samples did not belong to the assigned category, and were assigned to the negative labels for the category. This created larger annotated datasets for each of the ten categories, which were then used to retrain the classifiers, again using an 80/10/10 split. The trained classifiers were used to get probabilities for the ten categories for all the AQuA training items, as well as the NAEP math assessment items. The accuracies for the ten categories range from 55% to 84% on the held out 10% test data annotated via this process.

6.2.4 Automatic prediction of gender and bias profiling

To further study the interaction between the features and as a means of providing a text profiling tool, we use the extracted features to predict whether the mention is feminine or masculine using a random forest classifier and compare the majority vote as a baseline. Since the majority class for 6 of our 9 datasets are much higher than 50%, which itself is indicative of bias, we measure the classifier performance using the classifier cross-entropy normalized by the entropy of the uniform (binary) distribution. Specifically, for each classifier (C), we first calculate the cross-entropy using the predicted probabilities as:

$$CE = -\frac{1}{N_f + N_m}\left(\sum_{e_i \in E_f} \log p_C(f|x_i) + \sum_{e_j \in E_m} \log p_C(m|x_j)\right)$$

where N_f is the number of feminine mentions, N_m is the number of masculine mentions, E_f are the feminine mentions, E_m are the masculine mentions, and $p_C(g|x_i)$ for mention e_i is the probability of the classifier C predicting gender g given the input features x_i associated with mention e_i. The cross-entropy gives us a measure of how different the classifier predictions are from the actual distribution. A lower value of cross-entropy indicates that the classifier predictions are close to the actual distribution of labels. NCE is defined as

$$NCE = \frac{E_M - CE}{E_M}$$

where E_M for a uniform binary distribution is given as

$$E_M = -\frac{1}{2}\left(\log\frac{1}{2} + \log\frac{1}{2}\right) = \log 2$$

A value of NCE that is close to zero indicates that the CE is high, and the classifier is not predictive of the gender. A higher value of NCE indicates that the classifier is more predictive of gender, with the maximum possible value of NCE being 1, indicating that the classifier perfectly predicts gender. The NCE can also be negative, indicating that the classifier is performing worse than random chance. For texts with person mentions that are not gender balanced, the majority vote classifier has NCE $>$ 0. For texts where the NCE is higher than the majority vote baselines, gendered language use plays a role. For these cases, we present the top features used by the prediction model and whether the impact is negative or positive on the probability of predicting a feminine mention.

For most datasets, we consider only the authority/agency and word category features, i.e. whether the word is subject or object, appears in conjunction with other mentions and the syntactic relation with the root. Additionally, for the two math assessment sets, AQuA and NAEP, we have the predicted probabilities for each of the ten activity categories from the trained text classifiers. Note that the set of predicted probabilities is the same for all mentions that appear in one item, since the predictions are on an item level. For these two datasets we present three sets of prediction experiments: i) using all features (RF-All), ii) using just the word category and agency/authority features (RF-POS) and, iii) using just the predicted activity probabilities (RF-Act).

For all the prediction experiments, we use a five-fold CV setting, training and tuning the maximum tree depth on 4 folds and testing on the fifth fold, repeating five times, similar to the method used in Chapter 5. The results are reported for the average test fold accuracy. As in previous chapters, the sklearn (Pedregosa et al., 2011) random forest implementation is used for the experiments,

and analysis of feature importance for the trained models is done using the SHAP tree explainer framework (Lundberg et al., 2020).

6.3 Distributional Analysis

6.3.1 Feminine, masculine and gender-neutral representation

We first present the distributional analysis of person mentions for all 9 datasets (content and items), followed by detailed analysis of content and items. We extract the counts for neutral, feminine and masculine characters in each of our corpora. The results are shown in table 6.1. The sample size for each dataset varies. For the textbooks, the sample size is a paragraph, which is the smallest sample for the content data. For CCS, the sample is an excerpt from a book. For the WeeBit, NewsELA and OneSop corpora, the sample is an entire article. For the items, the sample is the entire question, including the question context and answer choices where present. We see that across all content datasets, there are fewer feminine mentions than masculine mentions. The trend also holds for item datasets, with the exception of NAEP Science, which has slightly more feminine mentions as compared to masculine mentions. We test whether the difference between the feminine and masculine mentions is significantly different from 50% using a binomial test. We find that for all content datasets and for AQuA, the differences are significant. For PISA and the two NAEP datasets, the differences are not significant.

6.3.2 Analysis of content

For a more detailed analysis, for all the mentions identified as masculine or feminine, we extract the syntactic relation of the word to the head word using the dependency tree obtained using spaCy. The most frequent syntactic roles are presented in table 6.2. The values are the percentage calculated individually for feminine and masculine characters for each dataset. We see consistent trends across all five datasets in terms of a higher fraction of masculine characters appearing as subject. We show only subject and object, and the additional annotation for conjunction since other POS tags are less frequent and more noisy. The row "Delta" shows the difference between the percentage of times mentions appears as subject vs object. Delta is greater for masculine characters vs. feminine characters for all datasets. In other words, for all datasets, masculine characters occur more frequently

Dataset	Sample size	Num. Samples	% Samples w/ no mentions	Avg. mentions/ sample	All mentions (%)		
					N	F	M
Content							
Textbooks	Para.	33575	38.8	3.6	85.9	3.9*	10.3
CCS App. B	Excerpt	265	1.1	32.9	60.1	12.7*	27.2
WeeBit	Article	10488	1.8	18.6	71.7	10.8*	17.5
NewsELA		9565	0.0	60.6	61.7	12.7*	25.6
OneStop		570	0.0	49.6	66.1	9.5*	24.4
Items							
PISA Science	Item	48	10.6	9.9	58.7	19.8	21.5
NAEP Math		444	64.9	3.4	40.8	28.7	30.6
NAEP Science		133	32.6	3.4	79.8	11.4	8.8
AQUA Algebra		33031	0.3	3.3	39.2	16.4*	44.4

Table 6.1: Percentage of neutral (N), feminine (F) and masculine (M) mentions in educational content and assessment corpora including all nouns, pronouns and names. * indicates the frequency of F is significantly lower than M using a binomial test with $p < 0.001$. (Average mentions per sample are calculated using only the samples with mentions.)

as subjects than objects when compared to feminine characters. There is also a trend for conjunction, indicating that feminine characters are more frequently presented in groups than masculine characters.

For the subset of the gendered mentions where the authority annotation is available for the root verb, table 6.3 presents the distribution of the authority being with the agent (subject), theme (object)

	Texbooks		CCS App. B		WeeBit		NewsELA		OneStop	
Gender	F	M	F	M	F	M	F	M	F	M
Total Count	2822	7510	1092	2335	20637	33458	73279	148424	2669	6851
Subject	39.2	39.8	45.4	48.6	49.0	51.2	60.2	64.4	56.4	61.1
Object	20.4	18.6	19.6	16.7	20.8	18.7	14.4	11.8	19.6	11.6
Delta	18.8	21.2	25.8	31.9	28.2	32.5	45.8	52.6	36.8	49.5
Conjunction	7.5	5.5	2.3	1.5	4.4	3.0	2.8	1.7	3.8	2.2

Table 6.2: Percentage of POS tags from dependency parsing for the gendered nouns, pronouns and names. Delta shows the difference between percentages for subject and object. (Percentage is calculated using the total count for each column).

or being equal is presented for when the mention appears as a subject. For four of the five datasets, the percentage of times the authority lies with the agent is greater for masculine mentions compared to feminine mentions. The agency of the subject can be positive, negative or equal. Since the trends are similar to that of authority, they are not shown here.

6.3.3 Analysis of items

We repeat a similar analysis for the item datasets. Since the NAEP and PISA items come from standardized tests (more thoroughly vetted), we expect less evidence of bias in these sets. Table 6.4 shows the distribution for POS tags for the four item datasets. We see that, unlike the content datasets, the difference between subject and object mentions does not show a consistent trend, and is smaller for the masculine mentions for two of the four corpora.

The authority distribution is shown in table 6.5. Here the outlier is the PISA dataset, where we see that the percentage of cases where the agency lies with the masculine mentions is more than double that of feminine mentions. The values for agency show a similar trend, and are not shown here.

	Texbooks		CCS App. B		WeeBit		NewsELA		OneStop	
Gender	F	M	F	M	F	M	F	M	F	M
Total Count	956	2943	522	1079	9457	15951	40871	87339	1398	3903
Agent	42.7	41.0	27.4	33.7	31.4	34.5	24.9	25.1	23.3	25.7
Equal	19.4	20.3	23.6	23.3	25.4	25.0	44.0	47.5	45.1	48.3
Theme	14.3	12.7	7.7	9.7	13.4	13.2	10.9	10.4	8.7	10.7

Table 6.3: Percentage of cases where authority lies with agent, theme or is equal when the character is the subject in the text (percentage is calculated using the total count for each column which includes cases where the mention is neither subject nor object).

	PISA Science		NAEP Math		NAEP Science		AQuA Algebra	
Gender	F	M	F	M	F	M	F	M
Total Count	82	91	149	159	39	30	17637	47712
Subject	61.0	69.2	69.8	64.8	76.9	56.7	56.5	55.8
Object	22.0	5.5	2.7	2.5	-	6.7	20.1	19.5
Delta	39	63.7	67.1	62.3	76.9	50.0	36.4	36.3
Conjunction	-	14.3	3.4	5.0	2.6	-	13.7	5.0

Table 6.4: Percentage of POS tags from dependency parsing for the gendered nouns, pronouns and names. Delta shows the difference between percentages for subject and object. (percentage is calculated using the total count for each column).

	PISA Science		NAEP Math		NAEP Science		AQuA Algebra	
Gender	F	M	F	M	F	M	F	M
Total Count	43	58	89	83	26	16	8494	24456
Agent	23.3	56.9	57.3	62.7	61.5	68.8	56.9	54.2
Equal	34.9	6.9	19.1	21.7	23.1	6.2	15.3	14.8
Theme	18.6	20.7	13.5	10.8	15.4	6.2	9.2	9.6

Table 6.5: Percentage of cases where authority lies with agent, theme or is equal when the character is the subject in the text (percentage is calculated using the total count for each column).

6.4 Gender Prediction

To see whether syntactic features and power/agency frames vary across the feminine and masculine character representation in our datasets, we train models using these features to predict the gender (F/M) of the mention. We use four features, the syntactic relation (including the indicator for `Conj`), the POS tag for the head word, and binary indicators of agency and authority when present as input to the random forest (RF-POS).

We use the majority vote as the baseline, for which $p(g|x_i)$ is the relative frequency of mentions of gender (masculine or feminine) in the CV training set. We also present the NCE for each model. A higher NCE indicates that the model is predictive of gender, and the dataset contains biased gender representation. The results are presented in table 6.6. We see that for two of our datasets, we get a slight but significant ($p < 0.01$) improvement over the majority vote. For all datasets, we have an NCE greater than zero, indicating that the classifier is predictive of gender when normalized against a uniform distribution for feminine and masculine mentions in the text.

The important features used by the classifier are a reflection of differences we see in tables 6.2 and 6.3. For the CCS and NewsELA, the top feature is the POS being subject, with a negative impact on probability of feminine mention. The POS being object is the top feature for OneStop, with a

positive impact on probability of feminine mention. For the WeeBit text, the top feature is authority lying with agent, again with a negative contribution for the probability of feminine mention. For the textbook dataset, the top feature is the lack of authority label, which has a positive effect on probability of feminine mention.

	% Acc. (NCE)	
Dataset	**Majority vote**	**RF-POS**
Textbooks	72.7 (0.15)	72.8 (0.16)
CCS App. B	68.1 (0.10)	67.8 (0.10)
WeeBit	61.9 (0.04)	62.1 (0.05)
NewsELA	67.0 (0.09)	67.0 (0.09)
OneStop	72.0 (0.14)	72.0 (0.15)

Table 6.6: Average CV test results for the random forest classifier predicting gender (F/M) for different datasets.

For the prediction task for items, we again use the POS feature set, and report results for five-fold CV compared to the majority vote (RF-POS). For the two math item sets, since we have the additional feature set of ten predicted activity category probabilities, we conduct experiments with (RF-All) and without this set (RF-POS). We also include results for just using the activity features (RF-Act). The results, shown in table 6.7, indicate that, unlike the content datasets, we get sizeable improvements over the majority baseline for three of the four cases. This is also reflected in the values of NCE for PISA, NAEP math and AQuA datasets, which are higher than that for the content datasets. For the NAEP math RF-POS experiment, we see that the NCE is negative, which indicates that the classifier is doing worse than chance.

We look at the top features identified as important and the associated impact on the probability of feminine mention. For PISA, consistent with the distribution for authority in table 6.5, the top 3 features are tied to agency and authority, while for NAEP math and AQuA, the best results are for

	% Acc. (NCE)			
Dataset	**Majority vote**	**RF-POS**	**RF-All**	**RF-Act**
PISA Science	52.6 (-0.003)	78.0 (0.28)	-	-
NAEP Science	56.5 (-0.005)	55.1 (0.02)	-	-
NAEP Math	51.6 (-0.001)	46.8 (-0.12)	76.3 (0.28)	82.1 (0.50)
AQuA Algebra	73.0 (0.16)	73.8 (0.19)	78.1 (0.31)	80.1 (0.40)

Table 6.7: Average CV test results for the random forest classifier predicting gender (F/M) for different datasets.

RF-Act, and the top features are predicted activity class probabilities. The top three features and associated SHAP effects for these three datasets are shown in table 6.8. We see that the category Exercise & Sports has a negative impact for AQuA, but positive impact for NAEP. This indicates that questions about exercise & sports in the AQuA dataset more often involve masculine characters, while for NAEP this category more often involves feminine characters. While the NAEP RF-POS performance being worse than the baseline suggests that the dataset is controlled for gender representation, the 30 point improvement in accuracy with the activity categories indicates that there may be unintended biases in the dataset associated with the context the question is framed in, irrespective of syntactic roles.

6.5 Word Association

Motivated by the increase in prediction accuracy of gender with the addition of activity probabilities, we examine the words associated with the feminine and masculine mentions. Unlike random forests where we can use SHAP for feature analysis, for the activity classifiers we use BERT, which is not interpretable. Looking at associated words allows us to examine patterns in word usage that may contribute to the differences in activity contexts. For the content dataset we get the highest NCE for the gender prediction for the textbook data, for which we also explore word associations.

Dataset	**Best Model**	Top Features	
		Feature	SHAP Val.
PISA Science	RF-POS	Positive Agency	-0.063
		Authority w/ Agent	-0.047
		Equal Authority	0.043
NAEP Math	RF-Act	Home Maintenance prob.	-0.056
		Social Relationships prob.	-0.049
		Exercise & Sports prob.	+0.047
AQuA Algebra	RF-Act	Arts & Creativity prob.	+0.016
		Academics prob.	+0.016
		Exercise & Sports prob.	-0.016

Table 6.8: Random forest classifier top features for PISA, NAEP Math and AQuA datasets in predicting gender, with average SHAP values associated with the probability of feminine mention.

For the cases in the NAEP math dataset, where the mention appears as a subject, the most frequent verbs when associated with a feminine or masculine mention are shown in figure 6.3. The word size is indicative of frequency, with larger words appearing more frequently; the color variation is for improving readability. The usage of different verbs may reflect the differences in the activity contexts, which leads to improved classifier performance for RF-Act. We also see that for feminine mentions, the words seem to be either very frequent (larger) or infrequent (smaller), while for masculine mentions, the verbs have a more uniform distribution.

A similar analysis for the verbs in AQuA math items is shown in figure 6.4, and figure 6.5 shows the most frequent verbs that appear in the textbook data when the character appears as subject. Again, we can see that there is an observable difference in the verbs across the two groups. Across all 3 cases presented above, we see that feminine mentions are strongly associated with `work`.

(a) Feminine mentions

(b) Masculine mentions

Figure 6.3: Verbs for feminine and masculine mentions in the NAEP math data.

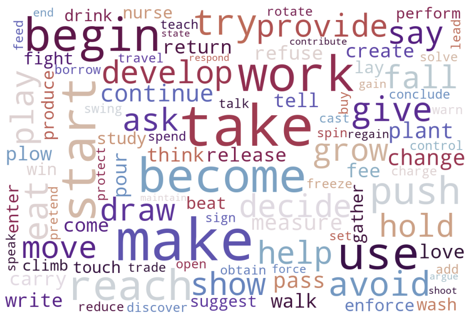
(a) Feminine mentions

(b) Masculine mentions

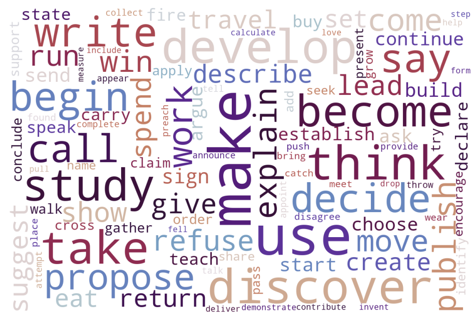
Figure 6.4: Verbs for feminine and masculine mentions in the AQuA math data.

For the math problems, where there is an emphasis on finances, females are also more strongly associated with `earn`. In examining the verbs associated with communication, masculine mentions are more frequently associated with `write`. This is most notable in the textbook data, but it also shows up in NAEP math. In the textbook data, masculine characters also `publish`. The textbook data also shows a much richer set of verbs for communication for men (`declare`, `describe`, `explain`, `speak`, `state`, `present`, etc.) as compared to those for women (`say`, `suggest`, `tell` etc.).

For the verbs associated with knowledge in the math items, masculine mentions are associated



(a) Feminine mentions



(b) Masculine mentions

Figure 6.5: Verbs associated with feminine and masculine mentions in the textbook data.

with `know`, `think`, `believe` and `remember`, while for feminine mentions we see `claim` and `wonder` `Think` also appears more frequently for masculine mentions in the textbook data. Looking at verbs associated with the scientific process in the textbook data, we can see many words notably more frequent for masculine mentions than for feminine mentions, including `discover`, `develop`, `study`, `think`, `write`, `lead`, `describe`, `create`, `suggest` and `argue`, as well as words that are not present for feminine mentions `propose`, `publish`, `explain`, `establish` and `build`. Conversely, words that show up more strongly for feminine mentions are `grow` and `measure`. For feminine mentions, we also see words like `start`, `avoid` and `need`.

6.6 Summary

Gender stereotypes in assessments and educational content can negatively impact both student performance and retention in STEM. Our work provides automated tools to profile texts and identify these biases, building on prior work on analysis of agency and authority. The automatically identified mentions of characters, gender information, syntactic roles, agency and authority labels and activities reveal patterns in the educational corpora studied in this work. For all content datasets, a consistent trend is that masculine characters appear, on average, twice as often as feminine characters. We also see that, when normalized for the number of total mentions, masculine characters appear more frequently as subjects in text compared to objects. This is also flagged in the models trained for predicting the gender of mentions. This is a concern, since the datasets, including

articles, are used by students, and recommended by educators. For example, NewsELA is a popular resource for English language learners, which include students typically underrepresented in STEM. The verbs in the textbook data across the two gender groups highlights the differences in the portrayal of characters; masculine characters are shown as actively contributing to academic activities, e.g. discovering, proposing and publishing, while feminine characters are far more likely to be associated with tasks like learning, finding and growing. This is indicative of the implicit societal biases in gender roles, and shows that as early as Kindergarten, students are exposed to these norms through educational content. For standardized PISA and NAEP assessment datasets, we observe that the items are better controlled for in terms of frequency of mentions. However, for PISA items we flag problematic trends in agency and authority, with feminine characters being associated with low agency and power relative to masculine mentions.

We further provide insight into unintended biases that may be present in math assessments through automatic activity classifiers. The analysis for AQuA dataset, which is used for training question-answering systems, shows two trends: i) a much smaller representation of feminine characters, and ii) a high accuracy in predicting gender when using information about activities presented in the text. While this is not a standardized assessment set, this dataset was collected from existing educational resources, and it is used to train machine learning systems. This indicates that existing educational resources used for training NLP systems may contain gender biases, which will be reflected in the trained systems. Our work also finds results that indicate that the NAEP math assessments, which are well controlled for in the frequency of syntactic roles of gendered mentions, have differences in terms of the activity genres the characters are presented in. Analysis of words associated with feminine and masculine mentions in the math assessment and textbook datasets provides additional insight into how mentions differ across these two groups.

For a set of assessments of a collection of texts, the NCE gives us a tool for evaluating the gender bias in the data, with a high NCE indicating biased text. The associated SHAP analysis of which features contribute to the gender prediction provides a method to flag aspects of the text that can be problematic. For individual items, the confidence of the prediction can be used as an indicator of bias. If the probability is close to 50%, it indicates a low confidence in predicting gender, and shows that the item does not have an imbalanced representation. Like the aggregate dataset, SHAP feature values for individual data samples can provide a tool to examine which features are relevant for an

individual mention if the prediction confidence is high. Thus our tool can be used to evaluate both aggregate datasets and individual text samples.

Chapter 7

CONCLUSIONS

Our work explores automated methods building on NLP tools to automatically analyze language use in K-16 STEM education, and proposes methodologies to examine how these aspects of language impact student performance. Our focus is on linguistic complexity and gender representation. We conclude with a summary of the research presented in this thesis including results and contribution, and the broader implications of the proposed methods. We then go over some recommendations for future work.

7.1 Summary

For linguistic complexity quantification, we propose a new hierarchical neural network structure, bidirectional context with attention, trained with a dataset of K-12 open source textbooks. The model provides state-of-the art performance on short texts, particularly assessment questions. In addition to a novel text classification algorithm, a key contribution of our work is bridging the research methodologies in the fields of NLP and educational measurement. We provide downstream validation of automatically extracted linguistic complexity via our proposed system with student performance, and show consistent patterns of linguistic complexity effects on student performance in both K-12 and college STEM assessments. The linguistic complexity classifier allows us to examine effects of language difficulty across multiple sets of assessments, showing findings that extend across multiple STEM assessments.

Our work examines item difficulty prediction in depth, introducing machine learning methodologies to this established educational measurement task. We present a comprehensive survey of existing literature on analysis and prediction of item difficulty for STEM assessments, and build a link to standard machine learning terminology and methodology. This allows us to present a methodology for robust analysis of what makes science assessments difficult, including the response types and science practices being assessed. The work utilizes method of analyzing the importance of fea-

tures in contribution to item difficulty for ensemble methods (Lundberg et al., 2020), giving insight into feature contribution for more complex non-linear prediction models. We provide a deeper exploration of the impact of linguistic complexity on question difficulty than done in previous work, showing a non-linear relation of linguistic complexity to difficulty and the interdependence of linguistic and other question features, which contributes to the lack of significant findings in previous literature for the impact of language difficulty.

We use these methodologies to analyzes physics assessments, looking at student performance by gender and ethnicity, providing an alternative to standard differential item function (DIF) analysis approaches used extensively in educational literature (Osterlind and Everson, 2009), which allows us to look at a combination of demographics for items. Our work indicates that for students coming into the course, there are differences in proficiency based on demographics. Using the task of predicting student performance on individual assessment questions, we investigate the impact of demographics and language difficulty. The models indicate that demographics become increasingly less predictive of performance over the course of instruction. Further, when used as a feature in predicting student performance, automatically extracted linguistic features showed an impact similar to that for middle school assessments when significant.

We provide analysis of gender representation within K-12 STEM texts using NLP techniques, providing profiling tools that can be used by educators to identify biased context and assessments. We examine 9 datasets, 5 for educational content and 4 for science and math assessment items. Across the content corpora, including textbooks, articles and sample excerpts provided by the Common Core Standards, masculine characters are mentioned twice as often as feminine characters. There are consistent differences in the fraction of characters appearing as subject vs. objects in the text, with masculine characters mentioned more frequently as subjects when compared to feminine characters. For item datasets, the frequency of mentions are better controlled, however we observe differences in authority and agency, which are significantly predictive of gender. For math assessment corpora, we further demonstrate that the category of activity presented in the question context (e.g. social relationships, business and finance, home maintenance etc.) is predictive of gender, indicating that there may be biases in the representation of gendered characters across various activity categories. Using the different features as input to a classifier for predicting gender of a person mention, we provide a mechanism to assess gender bias in text that considers a variety of factors in

combination.

In summary, this work has developed state-of-the-art tools for automatically analyzing STEM texts in terms of linguistic complexity and gender bias. Further, it bridges elements of educational measurement and machine learning research. Situated in the context of making STEM educational content and assessments less biased, we present tools that support educators by removing the need for extensive time consuming expert annotation of educational corpora, which can have an impact in increasing equity in K-16 classrooms.

7.2 Broader Implications of Proposed Methods

We present a framework using random forests and SHAP for analysis of item difficulty and individual student performance. This can be applied to DIF analysis to flag questions for which students belonging to a focal group perform significantly better or worse than the reference group. As opposed to conventional DIF analysis, using RF and SHAP analysis allows us to investigate a combination of features and student demographics that impact student performance.

Recent work in (Lucy et al., 2020) presents analysis on representation of ethnic groups in history textbooks via extracting mentions of people, and manually categorizing them for ethnicity by matching with predefined lists. This can be applied to our work, extending the analysis to account for roles, authority and agency associated with different ethnic identities, and building tools that can be used by educators.

In exploring the utility of the models presented in this work, we found that the model for predicting linguistic complexity (BCA) also lends well to the task of automated essay scoring (AES), as shown in (Nadeem et al., 2019), which is summarized below. The BCA model allows us to train discourse aware models without feature engineering, which is useful for essay scoring. We looked at two essay datasets, the ETS Corpus of Non-Native Written English from the Linguistic Data Consortium (LDC) (Blanchard et al., 2013) consisting of 12,100 TOEFL essays[1], and sets 1 and 2 of the Automated Student Assessment Prize (ASAP) Competition.[2]

For smaller training data sets which is typical for AES, we introduce two pretraining tasks to

[1] https://catalog.ldc.upenn.edu/LDC2014T06

[2] http://www.kaggle.com/c/asap-aes

improve performance: natural language inference and discourse marker prediction. The results show that discourse centered pretraining improves performance for AES. The model performs better than feature based models for tasks with larger training datasets. As an alternative to pre-training tasks, we incorporate pretrained text representation to work around the challenge of smaller datasets. BCA, in combination with contextualized word representation from BERT (Devlin et al., 2019), achieved state-of-the-art performance on automated scoring on the test of English as a foreign language (TOEFL) essays.

7.3 Future Work

Our research has indicated several directions for future work. The proposed model for predicting linguistic complexity has shown promising results when used as a factor for predicting student performance and overall question difficulty. However, like other neural models, the model suffers from a lack of interpretability, which is a considerable limitation in the context of educational applications. While we examined attention weights, our analysis does not indicate patterns that can identify elements of text that can help educators, consistent with other studies (Jain and Wallace, 2019). A possible direction for interpretability is causal inference. Recent work in (Alvarez-Melis and Jaakkola, 2017) presents a model agnostic framework for interpreting predictions, applied to sequence to sequence tasks. The model creates explanations that consist of input and output tokens that are causally related based on the neural model. The ability to provide causal explanations about what makes a text difficult would help educators in modifying the text as needed.

For analysis of physics assessments, a direction for future work is analysis of pretests that focus on concepts new to students, where high school academic preparedness plays a smaller role. This will allow us to further explore whether demographic based performance differences continue to decrease over the course of instruction. In addition, it would be useful to explicitly account for academic level, e.g. high school grades and/or scores on SAT/ACT, to determine whether student demographics are less significant indicators of performance when these factors are taken into consideration. This is a potential direction of future work for this analysis, and will help put our work in context with studies that control for academic preparedness.

For item difficulty analysis, a direction for future work is the analysis of PISA items, which are longer than the questions analyzed in our work. Our pilot studies have shown that the item difficulty

prediction of PISA items benefits from the addition of features related to linguistic complexity. For the analysis of gendered language, we automatically extract features that allow us to demonstrate gender biases in educational corpora. Our analysis indicates patterns of gender representation in PISA questions that show male characters with more authority and agency. Being able to use these patterns in conjunction with student demographics in predicting performance will allow us to directly examine the impact of gendered language on performance. This can provide a framework to examine impact that can generalize across different assessments.

As mentioned in Chapter 6, language can be indicative of biases based on culture and ethnicity. Research in (Deckman et al., 2018) has also shown that racial biases are present in educational content, and it is important to address these issues. However, current techniques for automatically extracting information about cultural and ethnic biases and identity are not well developed, and are not accurate enough for downstream analysis. In this context, building tools that can be used to accurately extract these nuanced features would be useful for educators. Another direction of future work is building systems to automatically correct for biases found in the data. For fictional person mentions in a text, this can be done by changing the gender of the mention to balance the representation. A challenge with this method would be to account for historical figures and real persons, and to make sure that the changes do not apply to them. For non-fictional person mentions, an alternative solution to correct the biased representation would be to use the features flagged as predictive of gender by the profiling tool to examine the way gendered mentions are presented and then rewrite the text.

BIBLIOGRAPHY

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Jamal Abedi. Psychometric issues in the ELL assessment and special education eligibility. *Teachers College Record*, 108(11):2282, 2006.

Jamal Abedi and Carol Lord. The language factor in mathematics tests. *Applied Measurement in Education*, 14(3):219–234, 2001. doi: 10.1207/S15324818AME1403_2. URL http://dx.doi.org/10.1207/S15324818AME1403_2.

Ayana Allen, Lakia M Scott, and Chance W Lewis. Racial microaggressions and african american and hispanic students in urban schools: A call for culturally affirming education. *Interdisciplinary Journal of Teaching and Learning*, 3(2):117–129, 2013.

David Alvarez-Melis and Tommi Jaakkola. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, 2017.

Yigal Attali and Jill Burstein. Automated essay scoring with E-Rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3), 2006.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. TOEFL 11: A corpus of non-native English. *ETS Research Report Series*, 2013(2):i–15, 2013.

Leo Breiman, Jerome Friedman, Richard A Olshen, and Charles J Stone. Classification and regression trees. *New York: Chapman & Hall*, 1984.

Eric Brewe, Vashti Sawtelle, Laird H Kramer, George E O'Brien, Idaykis Rodriguez, and Priscilla Pamelá. Toward equity through participation in modeling instruction in introductory university physics. *Physical Review Special Topics-Physics Education Research*, 6(1):010106, 2010.

Brent Bridgeman. A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement*, 29(3):253–271, 1992.

Susan M Brookhart and James H McMillan. *Classroom assessment and educational measurement*. Routledge, 2019.

Aoife Cahill, James H Fife, Brian Riordan, Avijit Vajpayee, and Dmytro Galochkin. Context-based automated scoring of complex mathematical responses. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 186–192, 2020.

Jeanne Sternlicht Chall and Edgar Dale. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, 1995.

Serina Chang and Kathleen McKeown. Automatically inferring gender associations from language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5750–5756, 2019.

Danqi Chen and Christopher Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, 2014.

Sapna Cheryan, Victoria C Plaut, Paul G Davies, and Claude M Steele. Ambient belonging: how stereotypical cues impact gender participation in computer science. *Journal of personality and social psychology*, 97(6):1045, 2009.

Michelene TH Chi, Paul J Feltovich, and Robert Glaser. Categorization and representation of physics problems by experts and novices. *Cognitive science*, 5(2):121–152, 1981.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.

CK12. CK-12 Free Online Textbooks, Flashcards, Adaptive Practice, Real World Examples, Simulations. https://www.ck12.org, 2007.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Meri Coleman and Ta Lin Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283, 1975.

Kevyn Collins-Thompson and James P. Callan. A language modeling approach to predicting reading difficulty. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2004, Boston, Massachusetts, USA, May 2-7, 2004*, pages 193–200, 2004. URL http://aclweb.org/anthology/N/N04/N04-1025.pdf.

Michael J Cortese and Maya M Khanna. Age of acquisition ratings for 3,000 monosyllabic words. *Behavior Research Methods*, 40(3):791–794, 2008.

National Research Council et al. *In the mind's eye: Enhancing human performance*. National Academies Press, 1992.

Victoria Crisp and Rebecca Grayson. Modelling question difficulty in an A level physics examination. *Research Papers in Education*, 28(3):346–372, 2013.

Victoria Crisp, Ezekiel Sweiry, Ayesha Ahmed, and Alastair Pollitt. Tales of the expected: the influence of students' expectations on question validity and implications for writing exam questions. *Educational Research*, 50(1):95–115, 2008.

Matthew A Cronin, Cleotilde Gonzalez, and John D Sterman. Why don't well-educated adults

understand accumulation? A challenge to researchers, educators, and citizens. *Organizational Behavior and Human Decision Processes*, 108(1):116–130, 2009.

Ronan Cummins and Marek Rei. Neural multi-task learning in automated assessment. *arXiv preprint arXiv:1801.06830*, 2018.

Jessica L Cundiff, Theresa K Vescio, Eric Loken, and Lawrence Lo. Do gender–science stereotypes predict science identification and science career aspirations among undergraduate science majors? *Social Psychology of Education*, 16(4):541–554, 2013.

Paul G Davies, Steven J Spencer, Diane M Quinn, and Rebecca Gerhardstein. Consuming images: How television commercials that elicit stereotype threat can restrain women academically and professionally. *Personality and Social Psychology Bulletin*, 28(12):1615–1628, 2002.

Beatriz de los Arcos, Robert Farrow, Rebecca Pitt, Martin Weller, and Patrick McAndrew. Adapting the curriculum: How K-12 teachers perceive the role of open educational resources. *Journal of Online Learning Research*, 2(1):23–40, 2016.

Sherry L Deckman, Ellie Fitts Fulmer, Keely Kirby, Katharine Hoover, and Abena Subira Mackall. Numbers are just not enough: a critical analysis of race, gender, and sexuality in elementary and middle school health textbooks. *Educational Studies*, 54(3):285–302, 2018.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Ezekiel J Dixon-Román, Howard T Everson, and John J McArdle. Race, poverty and SAT scores: Modeling the influences of family income on black and white high school students' SAT performance. *Teachers College Record*, 115(4):1–33, 2013.

Jason Drinkwater. Newsela: Teaching current events in the ESL classroom. *Helen Solórzano*, 39 (2):66, 2016.

Myroslava Dzikovska, Rodney Nielsen, and Claudia Leacock. The joint student response analysis and recognizing textual entailment challenge: making sense of student responses in educational applications. *Language Resources and Evaluation*, 50(1), 2016.

Yasmine El Masri, Steve Ferrara, Peter Foltz, and Jo-Anne Baird. Predicting item difficulty of science national curriculum tests: The case of Key Stage 2 assessments. *Curriculum Journal*, 28 (01):59–82, 2017.

Mary K. Enright and Kathleen M. Sheehan. Modeling the difficulty of quantitative reasoning items: Implications for item generation. In Sidney H. Irvine and Patrick C. Kyllonen, editors, *Item Generation for Test Development*, chapter 5, pages 129–157. Routledge, 2002.

Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. Neural automated essay scoring and coherence modeling for adversarially crafted input. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 263–271, 2018.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics, 2010.

Steve Ferrara and Teresa Duncan. Comparing science achievement constructs: Targeted and achieved. *Educational Forum*, 75(2):143–156, 2011. ISSN 0013-1725.

Steve Ferrara and Jeffrey T. Steedle. Item response demands, predicting item difficulty, and validity of inferences from test. In *Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY*, 2018.

National Center for Science and Engineering Statistics. Women, minorities, and persons with disabilities in science and engineering: Special report NSF 19-340. 2019.

Roy Freedle. Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review*, 73(1):1–43, 2003.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Cary Funk and Kim Parker. Women and men in STEM often at odds over workplace equity. https://vtechworks.lib.vt.edu/bitstream/handle/10919/92671/WomenSTEEMWorkplace.pdf, 2018.

Silvia Galdi, Mara Cadinu, and Carlo Tomasetto. The roots of stereotype threat: When automatic associations disrupt girls' math performance. *Child development*, 85(1):250–263, 2014.

Joy Gaston Gayles. *Attracting and Retaining Women in STEM: New Directions for Institutional Research, Number 152*, volume 124. John Wiley & Sons, 2011.

Peter Glick and Susan T Fiske. An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American psychologist*, 56(2):109, 2001.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36(2):193–202, 2004.

Arthur C. Graesser, Danielle S. McNamara, and Jonna M. Kulikowich. Coh-metrix. *Educational Researcher*, 40(5):223–234, 2011. doi: 10.3102/0013189X11413260. URL http://dx.doi.org/10.3102/0013189X11413260.

Zahra Hazari, Robert H Tai, and Philip M Sadler. Gender differences in introductory university physics performance: The influence of high school physics preparation and affective factors. *Science Education*, 91(6):847–876, 2007.

Rachel Henderson and John Stewart. Racial and ethnic bias in the force concept inventory. *2017 PERC Proceedings*, 2018.

Paula R. L. Heron. Effect of lecture instruction on student performance on qualitative questions. *Phys. Rev. ST Phys. Educ. Res.*, 11:010102, Jan 2015. doi: 10.1103/PhysRevSTPER.11.010102. URL https://link.aps.org/doi/10.1103/PhysRevSTPER.11.010102.

Marian Hickendorff. The language factor in elementary mathematics assessments: Computational skills and applied problem solving in a multidimensional irt framework. *Applied Measurement in Education*, 26(4):253–278, 2013.

Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on empirical methods in natural language processing*, pages 1373–1378, 2015.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

Dietmar Höttecke, Markus Sebastian Feser, Lena Heine, and Timo Ehmke. Do linguistic features influence item difficulty in physics assessments? *Science Education Review Letters*, 2018:1–6, 2018.

Bill Hussar, Jijun Zhang, Sarah Hein, Ke Wang, Ashley Roberts, Jiashan Cui, Mary Smith, Farrah Bullock Mann, Amy Barmer, and Rita Dilig. The condition of education 2020. nces 2020-144. *National Center for Education Statistics*, 2020.

M Ikonomakis, S Kotsiantis, and V Tampakas. Text classification using machine learning techniques. *WSEAS transactions on computers*, 4(8):966–974, 2005.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, 2019.

Ridong Jiang, Rafael E Banchs, and Haizhou Li. Evaluating and combining name entity recognition systems. In *Proceedings of the Sixth Named Entity Workshop*, pages 21–27, 2016.

Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

Daniel Jurafsky and James H Martin. Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing. *Upper Saddle River, NJ: Prentice Hall*, 2008.

Rachel Kachchaf, Tracy Noble, Ann Rosebery, Catherine O'Connor, Beth Warren, and Yang Wang. A closer look at linguistic complexity: Pinpointing individual linguistic features of science multiple-choice items associated with English language learner performance. *Bilingual Research Journal*, 39(2):152–166, 2016. doi: 10.1080/15235882.2016.1169455. URL http://dx.doi.org/10.1080/15235882.2016.1169455.

Michael Kane. Validity. In R.L. Linn, editor, *Educational Measurement*, pages 17–64. American Council on Education, Macmillan Publishing, New York, 2006.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Beethika Khan, Carol Robbins, and Abigail Okrent. The state of U.S. science and engineering 2020. nsb-2020-1. *National Science Foundation*, 2020.

Tushar Khot, Niranjan Balasubramanian, Eric Gribkoff, Ashish Sabharwal, Peter Clark, and Oren Etzioni. Markov logic networks for natural language question answering. *arXiv preprint arXiv:1507.03045*, 2015.

Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53(11):3735–3745, 2009.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, 2014.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Beata Beigman Klebanov, Christian Stab, Jill Burstein, Yi Song, Binod Gyawali, and Iryna Gurevych. Argumentation: Content, structure, and relationship with essay quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 70–75, 2016.

Diane Larsen-Freeman and Michael H Long. *An introduction to second language acquisition research*. Routledge, 2014.

Florence Le Hebel, Pascale Montpied, Andrée Tiberghien, and Valérie Fontanieu. Sources of difficulty in assessment: example of PISA science items. *International Journal of Science Education*, 39(4):468–487, 2017. ISSN 0950-0693.

Michael Leachman, Kathleen Masterson, and Eric Figueroa. A punishing decade for school funding. *Center on Budget and Policy Priorities*, 29, 2017.

Claudia Leacock and Martin Chodorow. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405, 2003.

Hee-Sun Lee, Amy Pallant, Sarah Pryputniewicz, Trudi Lord, Matthew Mulholland, and Ou Lydia Liu. Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Science Education*, 103(3):590–622, 2019.

Jackie FK Lee and Andy CO Chin. Are females and males equitably represented? A study of early readers. *Linguistics and Education*, 49:52–61, 2019.

Soo Lee and Youngsuk Suh. Lord's wald test for detecting DIF in multidimensional irt models: A comparison of two estimation approaches. *Journal of Educational Measurement*, 55(2):328–353, 2018.

Min Li, Maria Ruiz-Primo, Dongsheng Dong, Jim Minstrell, Xiaoming Zhai, and Phonraphee Thummaphan. Issues in developing science contextualized items. In *NCME annual conference*, 2017.

Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3): 18–22, 2002.

Zhouhan Lin, Minwei Feng, Cicero Nogueira do Santos, Mo Yu, Bing Ziang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *Proc. ICLR*, 2017.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, 2017.

Chris Little and Keith Jones. The effect of using real world contexts in post-16 mathematics questions. In *Proceedings of the British Congress for Mathematics Education April 10*, pages 137–144, 2010.

Jiawei Liu, Yang Xu, and Lingzhe Zhao. Automated essay scoring based on two-stage learning. *arXiv preprint arXiv:1901.07744*, 2019.

Edward Loper and Steven Bird. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70. Association for Computational Linguistics, 2002.

Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in texas us history textbooks. *AERA Open*, 6(3):2332858420940312, 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.

Saoussan A Maarouf. Supporting academic growth of English language learners: Integrating reading into stem curriculum. *World Journal of Education*, 9(4):83–96, 2019.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

Sandra P Marshall. *Schemas in problem solving*. Cambridge University Press, 1995.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. Supervised and unsupervised neural approaches to text readability. *arXiv preprint arXiv:1907.11779*, 2019.

Elijah Mayfield and Alan W Black. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162, 2020.

Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 109–142, 1980.

Laura McCullough. Gender, context, and physics assessment. *Journal of International Women's Studies*, 5(4):20–30, 2004.

Lillian C McDermott, Mark L Rosenquist, and Emily H Van Zee. Student difficulties in connecting graphs and physics: Examples from kinematics. *American Journal of Physics*, 55(6):503–513, 1987.

Vanes Mesic. Modeling the discrimination power of physics items. *European Journal of Physics Education*, 2(3):5–19, 2011. ISSN 1309-7202.

Vanes Mesic and Hasnija Muratovic. Identifying predictors of physics item difficulty: A linear regression approach. *Physical Review Special Topics - Physics Education Research*, 7(1):010110–Physics Education Research, 2011, Vol.7(1), p.010110–1–010110–15, 2011. ISSN 1554-9178.

Jörg Michael. 40000 namen, anredebestimmung anhand des vornamens. *c't*, pages 182–183, 2007.

Michigan. Michigan Open Book Project. http://textbooks.wmisd.org/, 2014.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

Rosamond Mitchell, Florence Myles, and Emma Marsden. *Second language learning theories*. Routledge, 2013.

Kristin M. Morrison and Susan E. Embretson. Abstract: Using cognitive complexity to measure the psychometric properties of mathematics assessment items. *Multivariate Behavioral Research*, 49 (3):292–293, 2014. ISSN 0027-3171.

Mary C Murphy, Claude M Steele, and James J Gross. Signaling threat: How situational cues affect women in math, science, and engineering settings. *Psychological science*, 18(10):879–885, 2007.

Farah Nadeem and Mari Ostendorf. Estimating linguistic complexity for science texts. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55, 2018.

Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. Automated essay scoring with discourse-aware neural models. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 484–493, 2019.

National Research Council. *Knowing what students know: The science and design of educational assessment*. National Academies Press, 2001.

Jessica Nelson, Charles Perfetti, David Liben, and Meredith Liben. Measures of text difficulty: Testing their predictive value for grade levels and student performance. *Council of Chief State School Officers, Washington, DC*, 2012.

Huy V Nguyen and Diane J Litman. Argument mining for improving the automated scoring of persuasive essays. In *Proc. AAAI*, pages 5892–5899, 2018.

Tracy Noble, Catherine Suarez, Ann Rosebery, Mary Catherine O'Connor, Beth Warren, and Josiane Hudicourt-Barnes. "i never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science. *Journal of Research in Science Teaching*, 49(6):778–803, 2012.

Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. Math= male, me= female, therefore math$\neq$ me. *Journal of personality and social psychology*, 83(1):44, 2002.

Steven J Osterlind and Howard T Everson. *Differential item functioning*, volume 161. Sage Publications, 2009.

Nitya Parthasarthi, Sameer Singh, et al. Genderquant: Quantifying mention-level genderedness. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2959–2969, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Sarah E Petersen and Mari Ostendorf. A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89–106, 2009.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, 2012.

Douglas Pidgeon and Alfred Yates. *An introduction to educational measurement*. Routledge, 2018.

OECD Pisa. Draft science framework, 2015.

Deborah A Prentice and Erica Carranza. What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of women quarterly*, 26(4):269–281, 2002.

Daniel T Pyburn, Samuel Pazicni, Victor A Benassi, and Elizabeth E Tappin. Assessing the relation between language comprehension and performance in general chemistry. *Chemistry Education Research and Practice*, 14(4):524–541, 2013.

Jill L Quilici and Richard E Mayer. Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, 88(1):144, 1996.

Hannah Rashkin, Sameer Singh, and Yejin Choi. Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1030. URL https://www.aclweb.org/anthology/P16-1030.

Martin Reisslein, Roxana Moreno, and Gamze Ozogul. Pre-college electrical engineering instruction: The impact of abstract vs. contextualized representation and practice on learning. *Journal of Engineering Education*, 99(3):225–235, 2010.

Brian Riordan, Sarah Bichler, Allison Bradford, Jennifer King Chen, Korah Wiley, Libby Gerard, and Marcia C Linn. An empirical investigation of neural methods for content scoring of science explanations. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 135–144, 2020.

Camelia V. Rosca. *What makes a science item difficult? A study of TIMSS -R items using regression and the Linear Logistic Test Model*. PhD thesis, Boston College, 2004.

Amjed Abu Saa, Mostafa Al-Emran, and Khaled Shaalan. Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques. *Technology, Knowledge and Learning*, 24(4):567–598, 2019.

Shima Salehi, Eric Burkholder, G Peter LePage, Steven Pollock, and Carl Wieman. The impact of

incoming preparation and demographics on performance in physics i: a multi-institution comparison. *arXiv preprint arXiv:1905.00389*, 2019.

Maria Veronica Santelices and Mark Wilson. Unfair treatment? The case of Freedle, the SAT, and the standardization approach to differential item functioning. *Harvard Educational Review*, 80 (1):106–134, 2010.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, 2017.

Richard Scheines, Matt Easterday, and David Danks. Teaching the normative theory of causal reasoning. *Causal learning: Psychology, philosophy, and computation*, pages 119–38, 2007.

Sarah E Schwarm and Mari Ostendorf. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics, 2005.

Denise Sekaquaptewa and Mischa Thompson. Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. *Journal of experimental social psychology*, 39(1):68–74, 2003.

Julia Shaftel, Evelyn Belton-Kocher, Douglas Glasnapp, and John Poggio. The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment*, 11(2):105–126, 2006.

Kathleen M Sheehan, Irene Kostin, and Hilary Persky. Predicting item difficulty as a function of inferential processing requirements: An examination of the reading skills underlying performance on the NAEP Grade 8 reading assessment. In *Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA*, 2006.

Kathleen M Sheehan, Michael Flor, and Diane Napolitano. A two-stage approach for generating unbiased estimates of text complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pages 49–58, 2013.

Sandip Sinharay. An ncme instructional module on data mining methods for classification and regression. *Educational Measurement: Issues and Practice*, 35(3):38–54, 2016.

Siyavula. Open Textbooks — Siyavula. https://www.siyavula.com/read, 2014.

Guillermo Solano-Flores, Chao Wang, Rachel Kachchaf, Lucinda Soltero-Gonzalez, and Khanh Nguyen-Le. Developing testing accommodations for English language learners: Illustrations as visual supports for item accessibility. *Educational Assessment*, 19(4):267–283, 2014.

Minjung Song and Roger Bruning. Exploring effects of background context familiarity and signaling on comprehension, recall, and cognitive load. *Educational Psychology*, 36(4):691–718, 2016.

NewsELA Staff. NewsELA: Bring context and relevance to any ELA curriculum. https://newsela.com.

Claude M Steele and Joshua Aronson. Stereotype threat and the intellectual test performance of African Americans. *Journal of personality and social psychology*, 69(5):797, 1995.

AJ Stenner, Ivan Horabin, Dean R Smith, and Malbert Smith. The lexile framework. *Durham, NC: MetaMetrics*, 1988.

Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, 2016.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, 2015.

Rex Taibu and Franca Ferrari-Bridgers. Physics language anxiety among students in introductory physics course. *EURASIA Journal of Mathematics, Science and Technology Education*, 16(4): em1835, 2020.

Rex Taibu, David Schuster, and David Rudge. Teaching weight to explicitly address language ambiguities and conceptual difficulties. *Physical Review Physics Education Research*, 13(1): 010130, 2017.

Catherine S Taylor and Susan Bobbitt Nolen. *Classroom assessment: Supporting teaching and learning in real classrooms*. Prentice Hall, 2005.

Ross Turner and Ray J Adams. Some drivers of test item difficulty in mathematics: an analysis of the competency rubric. In *Paper presented at the Annual Meeting of the American Educational Research Association (AERA)*, volume 13, page 17. Citeseer, 2012.

Ross Turner, John Dossey, Werner Blum, and Mogens Niss. Using mathematical competencies to predict item difficulty in PISA: A MEG study. In *Research on PISA*, pages 23–37. Springer, 2013.

Sowmya Vajjala and Ivana Lučić. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304, 2018.

Sowmya Vajjala and Detmar Meurers. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173. Association for Computational Linguistics, 2012.

Sowmya Vajjala and Detmar Meurers. Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL-International Journal of Applied Linguistics*, 165(2):194–222, 2014.

Sheila W. Valencia, Karen K. Wixson, Terry Ackerman, and Elizabeth Sanders. Identifying text-task-reader interactions related to item and block difficulty in the NAEP reading assessment. In *A publication of the NAEP Validity Studies Panel, San Mateo, CA: American Institutes for Research*, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,

Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017.

Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xuanjing Huang. Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 791–797, 2018.

Karen White and Susan Rotermund. Elementary and secondary mathematics and science education. science & engineering indicators 2020. nsb-2019-6. *National Science Foundation*, 2019.

Mikyung Kim Wolf and Seth Leon. An investigation of the language demands in content assessments for English language learners. *Educational Assessment*, 14(3-4):139–159, 2009.

Vikas Yadav and Steven Bethard. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, 2018.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.

D Yogatama, C Dyer, W Ling, and P Blunsom. Generative and discriminative text classification with recurrent neural networks. In *Thirty-fourth International Conference on Machine Learning (ICML 2017)*. International Machine Learning Society, 2017.

Haoran Zhang and Diane Litman. Co-attention based neural network for source-dependent essay scoring. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 399–409, 2018.

Muhua Zhu, Yue Zhang, Wenliang Chen, Min Zhang, and Jingbo Zhu. Fast and accurate shift-reduce constituent parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 434–443, 2013.

Appendix A

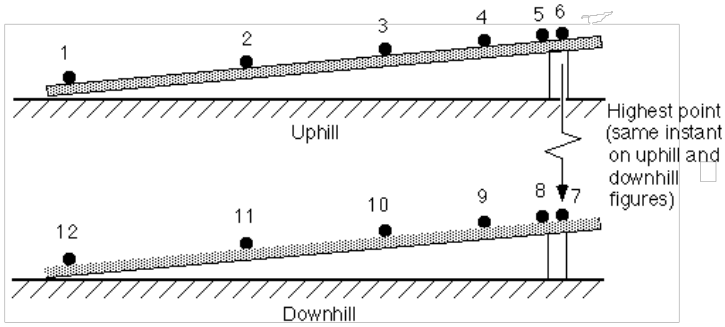
PHYSICS PRETESTS

A.1 Pretest 1

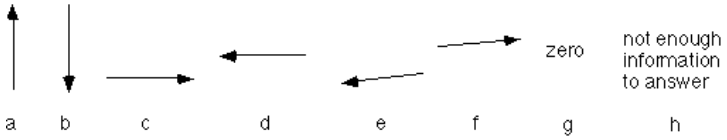
(A1D)U4a

A ball is given a quick push such that it rolls up and then down a ramp. The motion of the ball **when the hand is no longer pushing the ball** is demonstrated at right.

Part A: The strobe diagram at right represents the motion of the ball as it rolls up and then down the track. (In a strobe diagram, the position of an object is shown at instants separated by equal time intervals.)



For the following questions, choose the arrow from the list below that best represents the direction of the acceleration of the ball.



Question 1.
For each of the following locations of the ball, which of the arrows above best represents the instantaneous acceleration of the ball?

	a	b	c	d	e	f	g	h
Location 3					**Y**			
Location 6					**Y**			
Location 10					**Y**			

Question 2.
Explain the reasoning you used to obtain the direction of the instantaneous acceleration of the ball at each location.

Question 3.
How does the magnitude of the instantaneous acceleration change as the ball moves uphill (locations 1 to 6)?

Increases	Decreases	**Remains the same**	Others

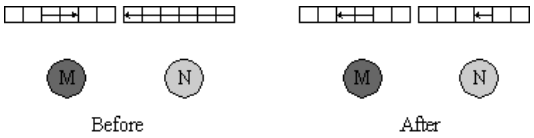
Question 4.
Briefly explain your reasoning for your previous response.

Question 5.
How does the magnitude of the instantaneous acceleration change as the ball moves downhill (locations 7 to 12)?

Increases	Decreases	**Remains the same**	Others

Question 6.
Briefly explain your reasoning for your previous response.

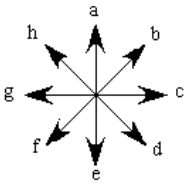
Part B: Two pucks collide on frictionless table. The diagram shows the velocity vectors of the pucks just before and just after the collision. The velocity vectors are drawn to scale.



Question 7.

Which of the arrows at right best represents the average acceleration of puck M?

Ans: g

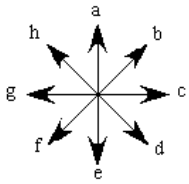


i. acceleration is zero
j. other or not enough information

Question 8.

Which of the arrows at right best represents the average acceleration of puck N?

Ans: c



i. acceleration is zero
j. other or not enough information

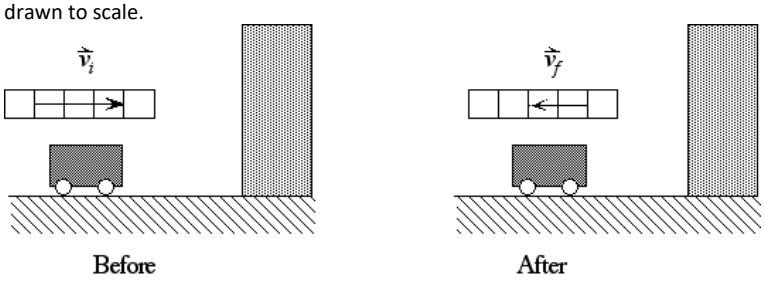
Question 9.
Explain your reasoning for your previous two responses.

Question 10.
Is the *magnitude* of the average acceleration of puck M during the collision greater than, less than, or equal to the magnitude of the average acceleration of puck N?

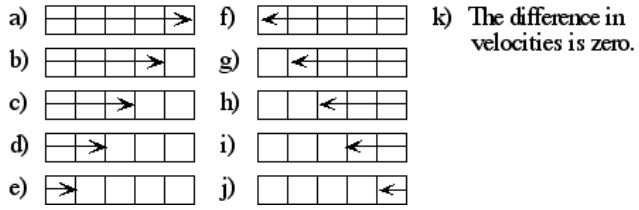
Greater than	**Less than**	Equal to	Not enough information

Question 11.
Briefly explain your reasoning for your previous response.

Part C: A cart rolls towards a wall on a level, frictionless table. The diagram shows the velocity vector of the cart just before and just after it collides with the wall. The velocity vectors are drawn to scale.



Question 12.
Which of the arrows below best represents the quantity $\vec{v}_f - \vec{v}_i$ the difference between the final and initial velocity vectors?



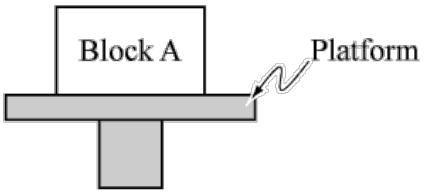
k) The difference in velocities is zero.

Ans: f

Question 13.
Briefly explain your reasoning for your previous response.

A.2 Pretest 2

(N23)U13a



Part I: Block A is on a platform as shown. Consider the three cases described below.

Case 1: Both the block and the platform remain at rest.

Question 1.
In Case 1, is the magnitude of the force exerted on block A by the platform *greater than, less than,* or *equal to* the magnitude of the force exerted on the platform by block A?

Greater than	Less than	**Equal to**	Not enough information

Question 2.

Consider the magnitudes of the forces you ranked above for Case 1. Are any of the magnitudes equal to zero? (Select all that apply.) **(None were correct)**

mag. of force on A by platform = 0	mag. of force on platform by A = 0

Question 3.

Explain the reasoning you used to answer the two previous questions about Case 1.

Case 2: The platform is now moving upward with *constant speed.*

Question 4.
In Case 2, is the magnitude of the force exerted on block A by the platform *greater than, less than,* or *equal to* the magnitude of the force exerted on the platform by block A?

Greater than	Less than	**Equal to**	Not enough information

Question 5.
Consider the magnitudes of the forces you ranked above for Case 2. Are any of the magnitudes equal to zero? (Select all that apply.) **(None were correct)**

mag. of force on A by platform = 0	mag. of force on platform by A = 0

Question 6.

Explain the reasoning you used to answer the two previous questions about Case 2.

Case 3: The platform is now moving upward and *speeding up.*

Question 7.
In Case 3, is the magnitude of the force exerted on block A by the platform *greater than, less than,* or *equal to* the magnitude of the force exerted on the platform by block A?

Greater than	Less than	**Equal to**	Not enough information

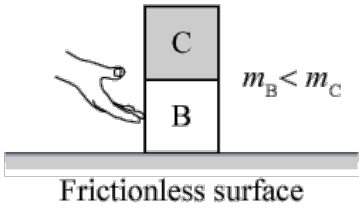
Question 8.
Consider the magnitudes of the forces you ranked above for Case 3. Are any of the magnitudes equal to zero? (Select all that apply.) **(None were correct)**

mag. of force on A by platform = 0	mag. of force on platform by A = 0

Question 9.

Explain the reasoning you used to answer the two previous questions about Case 3.

Part II: Two blocks, B and C, of unequal mass ($m_B < m_C$) are placed on a frictionless surface. A hand pushes block B with a constant force to the right, as shown. Block C **does not** slip on block B. At the instant shown, the velocity of block B is non-zero and to the right.



Question 10.
At the instant shown, which of the following statements best characterizes the magnitudes of the velocities of blocks B and C (v_B and v_C, respectively)?

$v_B > v_C > 0$	$v_B > v_C = 0$	$v_C > v_B > 0$
$v_C > v_B = 0$	$\boldsymbol{v_B = v_C > 0}$	$v_B = v_C = 0$

Question 11.
At the instant shown, which of the following statements best characterizes the magnitudes of the accelerations of blocks B and C (a_B and a_C, respectively)?

$a_B > a_C > 0$	$a_B > a_C = 0$	$a_C > a_B > 0$
$a_C > a_B = 0$	$\boldsymbol{a_B = a_C > 0}$	$a_B = a_C = 0$

Question 12.
Explain the reasoning you used to answer the two previous questions.

Question 13.
At the instant shown, which of the following statements best characterizes the magnitudes of the net forces on blocks B and C ($F_{net\ B}$ and $F_{net\ C}$, respectively)?

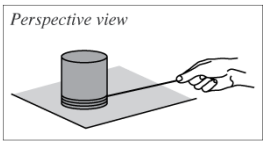
$F_{net\ B} > F_{net\ C} > 0$	$F_{net\ B} > F_{net\ C} = 0$	**$F_{net\ C} > F_{net\ B} > 0$**
$F_{net\ C} > F_{net\ B} = 0$	$F_{net\ B} = F_{net\ C} > 0$	$F_{net\ B} = F_{net\ C} = 0$

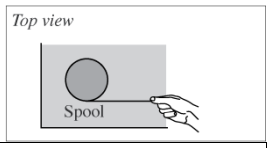
Question 14.
Explain the reasoning you used to answer the previous question.

A.3 Pretest 3

(DRB)U8c
Part I

A spool is pulled across a frictionless table as shown. The hand pulls horizontally on the thread. The thread has been wound around the bottom of the spool many times.



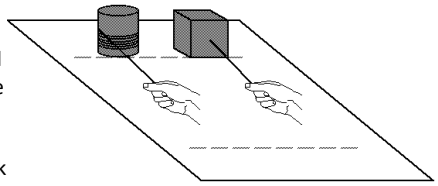


Question 1.

Which of the following best describes the motion of the spool?

The spool will rotate, but its center will remain in place.	**The spool will rotate and move across the table.**
The spool will move across the table, but not rotate.	Other.

Part II



Two objects, a block and a spool, are each pulled across a level, frictionless surface by a string. The block and the spool have the same mass.

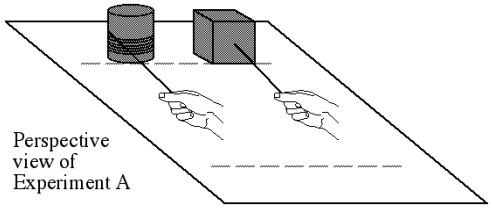
The string pulling the block is tied to a small hook at the center of the front face of the block. The string pulling the spool is wrapped many times around the spool and may unwind as it is pulled.

Two experiments are performed with the block and spool, Experiment A and Experiment B.

In Experiment A, the two hands start pulling at the same time so that the hands *move together*. That is, the experimenters make sure that, at every instant, their hands have each moved the same distance. Each hand pulls with a constant tension, but these tensions may or may not be the same as each other.

In Experiment B, the strings are pulled with the *same constant tension*. In this case, the hands may or may not move together as they pull the strings.

(Make the approximation that the strings and the hook are massless.)



Question 2.
For Experiment A (where the *hands move together*), which of the following options best describes when the spool crosses the finish line?

The spool crosses the finish line before the block.	The spool crosses the finish line at the same time as the block.
The spool crosses the finish line after the block.	The center of the spool stays in the same place and does not cross the finish line at all.
The center of the spool moves backwards, away from the finish line, and does not cross the finish line at all.	

Question 3.
Explain your reasoning for your previous response.

Question 4.
For Experiment B (where the strings are pulled with the *same tension*), which of the following options best describes when the spool crosses the finish line?

The spool crosses the finish line before the block.	**The spool crosses the finish line at the same time as the block.**
The spool crosses the finish line after the block.	The center of the spool stays in the same place and does not cross the finish line at all.
The center of the spool moves backwards, away from the finish line, and does not cross the finish line at all.	

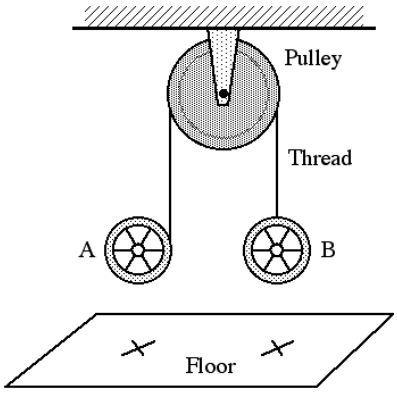
Question 5.
Explain your reasoning for your previous response.

Part II

Two identical spools are held the same height above the floor. A thread is wrapped many times around spool A. The same thread passes over a pulley, and is attached to a fixed point on spool B, so that spool B will not rotate. An X is marked on the floor directly below each spool.

Both spools are released from rest at the same instant.

(Assume that the pulley and thread are massless and that the axle of the pulley is frictionless.)



Question 6.
Is the tension in the part of the thread just above **spool A** *greater than, less than,* or *equal to* the tension in the part of the thread just above **spool B** (after the spools are released but before either spool hits the floor)?

Greater than	Less than	**Equal to**

Question 7.
Explain your reasoning for your previous response.

Question 8.
Is the magnitude of the acceleration of the center of mass of spool A *greater than, less than,* or *equal to* the magnitude of the acceleration of the center of mass of spool B?

Greater than	Less than	**Equal to**

Question 9.
Explain your reasoning for your previous response.

Question 10.
Will spool A hit the floor *before, after,* or *at the same instant* as spool B?

Before	After	**At the same instant**

Question 11.
Explain your reasoning for your previous response.

Appendix B

MATH ITEM CONTEXT CATEGORIES

Category	Definition	Example
Academics	Students, teachers, classrooms, courses, majors, grades, tests	The average age of students of a class is 15.8 years. The average age of boys in the class is 16.4 years and that of the girls is 15.4 years. The ratio of the number of boys to the number of girls in the class is?
Arts and creativity	Expression, music, jewelry, painting, crafts, singing, CDs, clothes	To fill an art exhibit, the boys in an art course are assigned to create one piece of artwork each in the following distribution: 1/3 are sculptures, 1/8 are oil paintings, 1/2 are watercolors, and the remaining 10 pieces are mosaics. How many boys are in the art class?
Business and finances	Investment, banking, finances, retail sales, income/salary	Anthony and Michael sit on the six member board of directors for company X. If the board is to be split up into 2 three-person subcommittees, what percent of all the possible subcommittees that include Michael also include Anthony?
Vehicles and travel	Cars, driving, transportation, trains, ships, travel	How many seconds will a 220 meter long train take to cross a man running with a speed of 8 km/hr in the direction of the moving train if the speed of the train is 80 km/hr?

Exercise and sports	Walking/running, team sports, cycling, coaching, fitness, games	A and B go around a circular track of length 600 m on a cycle at speeds of 18 kmph and 48 kmph. After how much time will they meet for the first time at the starting point?
Food and drink	Food items, groceries, dining, cooking, snacks, beverages	John's Ice Cream Shop sells ice cream at m cents a scoop. For an additional n cents, a customer can add 2 toppings to his or her sundae. How much would a sundae with 2 scoops and 2 toppings cost, in terms of m and n?
Home maintenance	Home repair, non-commercial build-ing/construction, appliances, garden, fencing, yard work	Thomas bought a table for Rs. 2000 and a chair for Rs. 1000. After one month of usage he sold table to David for Rs. 1800 and chair to Michael for Rs. 900. What is the total percentage of loss incurred to Thomas?
Industry and farming	Manual labor or industrial work setting, machine work, work, farming	If 12 welders work at a constant rate, they complete an order in 8 days. If after the first day, 9 welders start to work on the other project, how many more days the remaining welders will need to complete the rest of the order?
Science and research	Lab science, space, experimentation, survey research, ratios, speed/acceleration calculations	From a group of 16 astronauts that includes 7 people with previous experience in space flight, a 3-person crew is to be selected so that exactly 1 person in the crew has previous experience in space flight. How many different crews of this type are possible?

Social relationships	Social behavior or interactions, family dynamics, friendship, elections, villages, cities, teams, pets, groups, family	My grandson is about as many days as my son in weeks, and my grandson is as many months as I am in years. My grandson, my son and I together are 140 years. Can you tell me my age in years?

Table B.1: Item Context Categories and Definitions with AQuA Examples.

Appendix C

GENDERED AND NEUTRAL NOUNS AND PRONOUNS

Lists of nouns and pronouns used in extraction of gendered language.

C.1 Pronouns

Feminine:

She, her, hers, herself

Masculine:

He, him, his, himself

Personal (neutral):

I, me, we, us, myself, ourself, ourselves

Neutral:

They, them, their, you, themself, themselves

C.2 Nouns

Neutral:

Adult, adults, person, people, child, children

Feminine:

Girl, girls, woman, women, mrs, ms, mother, mothers, sister, sisters, grandmother, wife

Masculine:

Boy, boys, man, men, mr, father, fathers, brother, brothers, grandfather, husband

Occupations (neutral):

Accountant, accounts assistant, accounts clerk, accounts manager, accounts staff, acoustic engineer, actor, actress, actuary, acupuncturist, adjustor, administration assistant, administration clerk, administration manager, administration staff, administrator, advertising agent, advertising assistant, advertising clerk, advertising contractor, advertising executive, advertising manager, advertising staff, aerial erector, aerobic instructor, aeronautical engineer, agent, air traffic controller, aircraft designer, aircraft engineer, aircraft maintenance engineer, aircraft surface finisher, airman, airport controller, airport manager, almoner, ambulance controller, ambulance crew, ambulance driver, amusement arcade worker, anaesthetist, analyst, analytical chemist, animal breeder, anthropologist, antique dealer, applications engineer, applications programmer, arbitrator, arborist, archaeologist, architect, archivist, area manager, armourer, aromatherapist, art critic, art dealer, art historian, art restorer, artexer, artist, arts, assembly worker, assessor, assistant, assistant caretaker, assistant cook, assistant manager, assistant nurse, assistant teacher, astrologer, astronomer, attendant, au pair, auction worker, auctioneer, audiologist, audit clerk, audit manager, auditor, auto electrician, auxiliary nurse, bacon curer, baggage handler, bailiff, baker, bakery assistant, bakery manager, bakery operator, balloonist, bank clerk, bank manager, bank messenger, baptist minister, bar manager, bar steward, barber, barmaid,

barman, barrister, beautician, beauty therapist, betting shop, bill poster, bingo caller, biochemist, biologist, blacksmith, blind assembler, blind fitter, blinds installer, boat builder, body fitter, bodyguard, bodyshop, book binder, book seller, book-keeper, booking agent, booking clerk, bookmaker, botanist, branch manager, breeder, brewer, brewery manager, brewery worker, bricklayer, broadcaster, builder, builders labourer, building advisor, building control, building engineer, building estimator, building foreman, building inspector, building manager, building surveyor, bursar, bus company, bus conductor, bus driver, bus mechanic, bus valeter, business consultant, business proprietor, butcher, butchery manager, butler, buyer, cab driver, cabinet maker, cable contractor, cable jointer, cable tv installer, cafe owner, cafe staff, cafe worker, calibration manager, camera repairer, cameraman, car dealer, car delivery driver, car park attendant, car salesman, car valet, car wash attendant, care assistant, care manager, careers advisor, careers officer, caretaker, cargo operator, carpenter, carpet cleaner, carpet fitter, carpet retailer, carphone fitter, cartographer, cartoonist, cashier, casual worker, caterer, catering consultant, catering manager, catering staff, caulker, ceiling contractor, ceiling fixer, cellarman, chambermaid, chandler, chaplain, charge hand, charity worker, chartered, chartered accountant, chauffeur, chef, chemist, chicken chaser, child minder, childminder, chimney sweep, china restorer, chiropodist, chiropractor, choreographer, church officer, church warden, cinema manager, circus proprietor, circus worker, civil engineer, civil servant, claims adjustor, claims assessor, claims manager, clairvoyant, classroom aide, cleaner, clergyman, cleric, clerk, commissioned, consultant, coroner, councillor, counsellor, dealer, decorator, delivery

driver, doctor, driver, economist, editor, employee, employment, engineer, english teacher, entertainer, envoy, executive, farmer, fireman, floor layer, floor manager, florist, flour miller, flower arranger, flying instructor, foam convertor, food processor, footballer, foreman, forensic scientist, forest ranger, forester, fork lift truck driver, forwarding agent, foster parent, foundry worker, fraud investigator, french polisher, fruiterer, fuel merchant, fund raiser, funeral director, funeral furnisher, furnace man, furniture dealer, furniture remover, furniture restorer, furrier, gallery owner, gambler, gamekeeper, gaming board inspector, gaming club manager, gaming club proprietor, garage attendant, garage foreman, garage manager, garda, garden designer, gardener, gas fitter, gas mechanic, gas technician, gate keeper, genealogist, general practitioner, geologist, geophysicist, gilder, glass worker, glazier, goldsmith, golf caddy, golf club professional, golfer, goods handler, governor, granite technician, graphic designer, graphologist, grave digger, gravel merchant, green keeper, greengrocer, grocer, groom, ground worker, groundsman, guest house owner, guest house proprietor, gun smith, gynaecologist, hgv driver, hgv mechanic, hairdresser, handyman, hardware dealer, haulage contractor, hawker, health advisor, health and safety, health care assistant, health consultant, health nurse, health planner, health service, health therapist, health visitor, hearing therapist, heating engineer, herbalist, highway inspector, hire car driver, historian, history teacher, hod carrier, home economist, home help, homecare manager, homeopath, homeworker, hop merchant, horse breeder, horse dealer, horse riding instructor, horse trader, horse trainer, horticultural consultant, horticulturalist, hosiery mechanic, hosiery worker, hospital consultant, hospital doctor, hospital

manager, hospital orderly, hospital technician, hospital warden, hospital worker, hostess, hot foil printer, hotel consultant, hotel worker, hotelier, househusband, housekeeper, housewife, housing assistant, housing officer, housing supervisor, hygienist, hypnotherapist, hypnotist, it consultant, it manager, it trainer, ice cream vendor, illustrator, immigration officer, import consultant, importer, independent means, induction moulder, industrial chemist, industrial consultant, injection moulder, inspector, instructor, instrument engineer, instrument maker, instrument supervisor, instrument technician, insurance agent, insurance assessor, insurance broker, insurance consultant, insurance inspector, insurance staff, interior decorator, interior designer, interpreter, interviewer, inventor, investigator, investment advisor, investment banker, investment manager, investment strategist, ironmonger, janitor, jazz composer, jeweller, jewellery, jockey, joiner, joinery consultant, journalist, judge, keep fit instructor, kennel hand, kitchen worker, knitter, labelling operator, laboratory analyst, labourer, laminator, lampshade maker, land agent, land surveyor, landlady, landlord, landowner, landworker, lathe operator, laundry staff, laundry worker, lavatory attendant, law clerk, lawn mower, lawyer, leaflet distributor, leather worker, lecturer, ledger clerk, legal advisor, legal assistant, legal executive, legal secretary, letting agent, liaison officer, librarian, library manager, licensed premises, licensee, licensing, lifeguard, lift attendant, lift engineer, lighterman, lighthouse keeper, lighting designer, lighting technician, lime kiln attendant, line manager, line worker, lineman, linguist, literary agent, literary editor, lithographer, litigation manager, loans manager, local government, lock keeper, locksmith, locum pharmacist, log merchant, lorry

driver, loss adjustor, loss assessor, lumberjack, machine fitters, machine minder, machine operator, machine setter, machine tool, machine tool fitter, machinist, magician, magistrate, magistrates clerk, maid, maintenance fitter, make up artist, manicurist, manufacturing, map mounter, marble finisher, marble mason, marine broker, marine consultant, marine electrician, marine engineer, marine geologist, marine pilot, marine surveyor, market gardener, market research, market researcher, market trader, marketing agent, marketing assistant, marketing coordinator, marketing director, marketing manager, marquee erector, massage therapist, masseur, masseuse, master mariner, materials controller, materials manager, mathematician, maths teacher, matron, mattress maker, meat inspector, meat wholesaler, mechanic, medal dealer, medical advisor, medical assistant, medical consultant, medical officer, medical physicist, medical practitioner, medical researcher, medical secretary, medical student, medical supplier, medical technician, merchandiser, merchant, merchant banker, merchant seaman, messenger, metal dealer, metal engineer, metal polisher, metal worker, metallurgist, meteorologist, meter reader, microbiologist, midwife, military leader, milklady, milkman, mill operator, mill worker, miller, milliner, millwright, miner, mineralogist, minibus driver, minicab driver, mining consultant, mining engineer, money broker, moneylender, mooring contractor, mortgage broker, mortician, motor dealer, motor engineer, motor fitter, motor mechanic, motor racing, motor trader, museum assistant, museum attendant, music teacher, musician, nanny, navigator, negotiator, neurologist, newsagent, night porter, night watchman, nuclear scientist, nun, nurse, nursery assistant, nursery nurse, nursery worker, nurseryman, nursing assistant, nursing auxiliary, nursing manager, nursing

sister, nutritionist, off shore, office manager, office worker, oil broker, oil rig crew, opera singer, operations, operative, operator, optical, optical advisor, optical assistant, optician, optometrist, orchestral, organiser, organist, ornamental, ornithologist, orthopaedic, orthoptist, osteopath, outdoor pursuits, outreach worker, packaging, packer, paediatrician, paint consultant, painter, palaeobotanist, palaeontologist, pallet maker, panel beater, paramedic, park attendant, park keeper, park ranger, partition erector, parts man, parts manager, parts supervisor, party planner, pasteuriser, pastry chef, patent agent, patent attorney, pathologist, patrolman, pattern cutter, pattern maker, pattern weaver, pawnbroker, payroll assistant, payroll clerk, payroll manager, payroll supervisor, personnel officer, pest controller, pet minder, pharmacist, philatelist, photographer, physician, physicist, physiologist, physiotherapist, piano teacher, piano tuner, picture editor, picture framer, picture reseacher, pig man, pig manager, pilot, pipe fitter, pipe inspector, pipe insulator, pipe layer, planning engineer, planning manager, planning officer, planning technician, plant attendant, plant driver, plant engineer, plant fitter, plant manager, plant operator, plasterer, plastics consultant, plastics engineer, plate layer, plater, playgroup assistant, playgroup leader, plumber, podiatrist, police officer, polisher, pool attendant, pools collector, porter, portfolio manager, post sorter, postman, postmaster, postwoman, potter, practice manager, preacher, precision engineer, premises, premises security, press officer, press operator, press setter, presser, priest, print finisher, printer, prison chaplain, prison officer, private investigator, probation officer, probation worker, procurator fiscal, produce supervisor, producer, product installer, product

manager, production engineer, production hand, production manager, production planner, professional boxer, professional racing, professional wrestler, progress chaser, progress clerk, project co-ordinator, project engineer, project leader, project manager, project worker, projectionist, promoter, proof reader, property buyer, property dealer, property developer, property manager, property valuer, proprietor, psychiatrist, psychoanalyst, psychologist, psychotherapist, public house manager, public relations of?cer, publican, publicity manager, publisher, publishing manager, purchase clerk, purchase ledger clerk, purchasing assistant, purchasing manager, purser, quality controller, quality engineer, quality inspector, quality manager, quality technician, quantity surveyor, quarry worker, racehorse groom, racing organiser, radio controller, radio director, radio engineer, radio operator, radio presenter, radio producer, radiographer, radiologist, rally driver, receptionist, recorder, records supervisor, recovery vehicle coordinator, recreational, recruitment consultant, rector, reflexologist, refractory engineer, refrigeration engineer, refuse collector, registrar, regulator, relocation agent, remedial therapist, rent collector, rent offcer, repair man, repairer, reporter, representative, reprographic assistant, research analyst, research consultant, research director, research scientist, research technician, researcher, resin caster, restaurant manager, restaurateur, restorer, retired, revenue clerk, revenue officer, riding instructor, rig worker, rigger, riveter, road safety officer, road sweeper, road worker, roadworker, roof tiler, roofer, rose grower, royal marine, rug maker, saddler, safety officer, sail maker, sales administrator, sales assistant, sales director, sales engineer, sales executive, sales manager, sales

representative, sales support, salesman, saleswoman, sand blaster, saw miller, scaffolder, school crossing, school inspector, scientific officer, scientist, scrap dealer, screen printer, screen writer, script writer, sculptor, seaman, seamstress, secretary, security consultant, security controller, security guard, security officer, servant, service engineer, service manager, share dealer, sheet metal worker, shelf filler, shelter warden, shepherd, sheriff, sheriff clerk, sheriff principal, shift controller, ship broker, ship builder, shipping clerk, shipping officer, shipwright, shipyard worker, shoe maker, shoe repairer, shooting instructor, shop assistant, shop fitter, shop keeper, shop manager, shop proprietor, shot blaster, show jumper, showman, shunter, sign maker, signalman, signwriter, site agent, site engineer, skipper, slater, slaughterman, smallholder, social worker, software consultant, software engineer, soldier, solicitor, song writer, sound artist, sound engineer, sound technician, special constable, special needs, speech therapist, sports administrator, sports coach, sports commentator, sportsman, sportsperson, sportswoman, spring maker, stable hand, staff nurse, stage director, stage hand, stage manager, stage mover, station manager, stationer, statistician, steel erector, steel worker, steeplejack, stenographer, steward, stewardess, stock controller, stock manager, stockbroker, stockman, stocktaker, stone cutter, stone sawyer, stonemason, store detective, storeman, storewoman, street entertainer, street trader, stud hand, student, student nurse, student teacher, studio manager, sub-postmaster, sub-postmistress, supervisor, supply teacher, surgeon, surveyor, systems analyst, systems engineer, systems manager, tv editor, tachograph analyst, tacker, tailor, tank farm operative, tanker driver, tanner, tattooist, tax advisor,

tax analyst, tax assistant, tax consultant, tax inspector, tax manager, tax officer, taxi controller, taxi driver, taxidermist, tea blender, tea taster, teacher, teachers assistant, technical advisor, technical analyst, technical assistant, technical author, technical clerk, technical co-ordinator, technical director, technical editor, technical engineer, technical illustrator, technical instructor, technical liaison, technical manager, technician, telecommunication, telecommunications, telegraphist, telemarketeer, telephone engineer, telephonist, telesales person, television director, television engineer, television presenter, television producer, telex operator, temperature time, tennis coach, textile consultant, textile engineer, textile technician, textile worker, thatcher, theatre manager, theatre technician, theatrical agent, therapist, thermal engineer, thermal insulator, ticket agent, ticket inspector, tiler, timber inspector, timber worker, tobacconist, toll collector, tool maker, tour agent, tour guide, town clerk, town planner, toy maker, toy trader, track worker, tractor driver, tractor mechanic, trade mark agent, trade union official, trading standards, traffic warden, train driver, trainee manager, training advisor, training assistant, training co-ordinator, training consultant, training instructor, training manager, training officer, transcriber, translator, transport clerk, transport consultant, transport controller, transport engineer, transport manager, transport officer, transport planner, travel agent, travel clerk, travel consultant, travel courier, travel guide, travel guide writer, travel representative, travelling showman, treasurer, tree feller, tree surgeon, trichologist, trinity house pilot, trout farmer, tug skipper, tunneller, turf accountant, turkey farmer, turner, tutor, typesetter, typewriter engineer, typist, tyre

builder, tyre fitter, tyre inspector, tyre technician, undertaker, underwriter, upholsterer, valuer, valve technician, van driver, vehicle assessor, vehicle body worker, vehicle engineer, vehicle technician, ventriloquist, verger, veterinary surgeon, vicar, video artist, violin maker, violinist, voluntary worker, wages clerk, waiter, waitress, warden, warehouse manager, warehouseman, warehousewoman, watchmaker, weaver, weighbridge clerk, weighbridge operator, welder, welfare assistant, welfare officer, welfare rights officer, wheel clamper, wholesale newspaper, window cleaner, window dresser, windscreen fitter, wine merchant, wood carver, wood cutter, wood worker, word processing operator, works manager, writer, yacht master, yard manager, youth hostel warden, youth worker, zoo keeper, zoo manager, zoologist