

THE USE OF LARGE LANGUAGE MODELS TO PREDICT ITEM PROPERTIES

By

Francis Smart

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Measurement and Quantitative Methods – Doctor of Philosophy

2024

ABSTRACT

Calibrating items is a crucial yet costly requirement for both new tests and existing ones as items become outdated due to changing relevance or overexposure. Traditionally, this calibration involves giving items to a large number of participants, a process that requires substantial time and resources. To reduce these costs, researchers have sought alternative calibration methods. Before the emergence of Large Language Models (LLMs), these methods mainly relied on expert opinions or computational analysis of item features. Yet, the accuracy of experts in predicting item performance has varied, and computational approaches often struggle to capture the intricate semantic details of test items.

The emergence of LLMs might offer a new avenue of addressing the need for item calibration. These models, popularized by OpenAI (like the GPT series), have shown remarkable abilities in mimicking complex human thought processes, and performing advanced reasoning tasks. Their achievements in passing sophisticated exams and executing cross-language translations underline their potential. However, their capacity for predicting item properties in test calibration has not been thoroughly investigated. Traditional calibration relies heavily on direct human interaction, such as pretesting and expert assessment, or on statistical modeling of item features through resource intensive machine learning algorithms.

This dissertation explores the potential of LLMs to predict item characteristics, tasks that have traditionally required human insight or complex statistical models. With the increasing accessibility of high-performance LLMs from organizations like OpenAI, Meta, and Google, and through open-source platforms such as HuggingFace.com, there is promising

ground for investigation. This study examines whether LLMs could replace human efforts in item calibration tasks.

To evaluate the effectiveness of LLMs in predicting item properties, this dissertation implements a training and testing framework, focusing on assessing both the relative and absolute difficulties of items. It undertakes three theoretical investigations: firstly, examining the ability of LLMs to predict the relative difficulty of items; secondly, assessing the feasibility of using multiple LLMs as substitutes for test-takers and attempts to use their responses as predictors of item difficulty; and thirdly, applying a search algorithm, guided by LLM predictions of relative difficulty, to ascertain absolute difficulties.

The findings indicate that the models have statistical significance in predicting relative item difficulty, limited by modest explanatory power — with adjusted R-squared values around 5-10%. However, the application of LLMs in predicting relative item difficulties through pairwise comparisons proves to be more promising, achieving a pairwise accuracy of about 62% and demonstrating predicted correlations with item difficulty ranging between 0.36 and 0.42.

This suggests that whereas LLMs show potential in certain aspects of item calibration, their effectiveness varies depending on the specific task. This demonstrates a potential promising result that warrants further exploration into the capabilities of LLMs for item calibration, potentially leading to more efficient and cost-effective methods in the field of test development and maintenance.

This dissertation is dedicated to my beloved family. To my wife, for her unwavering support through countless adventures; to my parents, for their unconditional love; to my siblings, for always having my back; and to the children we have been privileged to care for in our home. I also extend my gratitude to my in-laws, who have welcomed me warmly into their family.

Finally, this work is dedicated to the disenfranchised children we have cared for, as well as those who remain victims of a dysfunctional system. Their resilience in the face of adversity is a constant source of inspiration.

ACKNOWLEDGEMENTS

I extend my deepest gratitude to Dr. Kimberly Kelly, my major professor, for her unwavering support, expert guidance, and insightful feedback. Her dedication was pivotal in the completion of this dissertation.

I am also thankful to my committee members, Dr. Kenneth Frank, Dr. Alicia Alonzo, and Dr. Christopher Nye, for their invaluable suggestions and critical insights, which greatly enhanced this work. The financial support from the Institute of Education Sciences, U.S. Department of Education (Award # R305B090011) funded this work and my studies and was crucial and is deeply appreciated.

Special thanks to Dr. Nathan Bos for his camaraderie and shared wisdom, which made this endeavor both productive and enjoyable. I would also like to thank Jeremy, a polygrapher at the CIA, as well as Mary Peyton with Montgomery County whose loose relationship with integrity has inspired me to do better, to keep pushing for positive change in the world.

Finally, I would like to thank my family and friends for their endless encouragement and understanding. A special thanks to my parents, whose belief in me kept me motivated through challenging times, and to my wife whose patience and support were invaluable.

Thank you all for your contributions and support. This dissertation would not have been possible without you.

TABLE OF CONTENTS

| | |
|---|-----|
| CHAPTER I: INTRODUCTION | 1 |
| CHAPTER II: LITERATURE REVIEW | 17 |
| CHAPTER III: RESEARCH METHODS..... | 47 |
| CHAPTER IV: RESULTS..... | 87 |
| CHAPTER V: DISCUSSION AND CONCLUSION | 104 |
| BIBLIOGRAPHY..... | 122 |
| APPENDIX A: PROMPT SCORING | 131 |
| APPENDIX B: PROMPT DESIGN - RELATIVE DIFFICULTY EVALUATION..... | 135 |
| APPENDIX C: PROMPT DESIGN - COLLATERAL ITEM INFORMATION..... | 144 |
| APPENDIX D: PROMPT SELECTION ALGORITHMS..... | 146 |
| APPENDIX E: SENTITIVITY ANALYSIS | 150 |
| APPENDIX F: ALGORITHMS ILLUSTRATED..... | 154 |

CHAPTER I: INTRODUCTION

1.1 Introduction

Calibrating test items is a crucial yet costly part of developing and maintaining testing programs. As test items age or become overexposed, they must be replaced, demanding the creation and calibration of novel items—a resource-intensive process. Pretesting novel items traditionally involves a large number of test takers, which incurs substantial expense.

Consequently, more cost-effective calibration methods are in high demand. Researchers have been investigating alternative calibration methods since the 1950s. These methods have typically hinged on evaluations by subject matter experts (SMEs), or computational tools focused on observable item features like word count, diction, and statistical correlations between infrequent word usage and item parameters. While SMEs can assess the nuanced content of items, their predictions on item difficulty and other parameters have been variable (Hambleton et al., 1998). On the other hand, computational tools, typically traditional machine learning models, rely upon having many items to train on, often hundreds to thousands of similar examples. Additionally, computation models are typically purely complex statistical models that do not have predictive power outside of the items they have been trained on. However, recent advancements in natural language processing using Large Language Models (LLMs) may offer a shift in how we approach this challenge.

Current developments using LLMs have been significant. These models have the capability to perform complex reasoning and produce clear, human-like responses. Traditionally, item calibration leaned heavily on human intervention—in both pretest-based approaches and expert judgment methods. The goal of this dissertation is to explore to what

extent these advanced models might be used to reduce the need for expensive human input in the item calibration process.

Large Language Models (LLMs) have shown remarkable capabilities, achieving impressive outcomes in areas beyond their initial training. Notably, without any specific preparation, they have scored in the top percentiles of professional exams like the Bar and LSAT (OpenAI, 2023). Researchers have deployed LLMs in activities somewhat akin to the scope of this dissertation, by being assessed on their ability to tackle a wide array of test items, demonstrating their proficiency in solving problems once believed to necessitate human intelligence and knowledge. This process of solving various problem sets with known outcomes serves as a benchmark for evaluating and comparing LLMs, illustrating their aptitude in emulating human cognition (OpenAI, 2023; Anil et al., 2023; Lewis et al., 2019; Devlin et al., 2018).

Despite their proficiency in generating human-like responses, thanks to extensive training on vast datasets of human language, the potential of LLMs to predict test item characteristics has not been thoroughly investigated. The task of assessing item features has traditionally depended on human inputs, either through expert judgment or through empirical data gathered from pretesting, a practice rooted in the seminal works of Thurston (1925), Zimowski et al. (1996), Lorge & Kruglov (1952), Thorndike (1982), and Bejar (1983). Alternatively, computational methods have sought to forecast item characteristics by analyzing observable attributes, an approach surveyed by Benedetto et al. (2023). These computational methods typically require large datasets to function effectively. However, the variabilities in

expert reliability alongside the substantial data requirements of computational approaches, hint at the benefits an automated method—less reliant on massive datasets—could provide.

This dissertation investigates the use of Large Language Models (LLMs) as potential surrogates for certain human tasks in the procedure of item calibration. Given the proven capabilities of LLMs to mimic human responses and their widespread availability through platforms such as OpenAI's GPT-3 and 4, Meta's Llama 2, Google's Gemini-Pro, and the resources at HuggingFace.com, there is a promising potential for these models to substitute some of the human efforts traditionally necessary in assessing item attributes. These attributes, previously quantifiable only through extensive manual labor or, to a lesser extent, through intricate statistical models, present a new frontier for LLM application.

For the empirical aspect of this investigation, item content and statistical data have been sourced from the National Assessment of Educational Progress (NAEP), focusing on a selected set of 462 released items from Mathematics and Science domains. The selection process favored items amenable to LLM analysis, thus excluding those heavily reliant on visual inputs, referencing current facts, or for which the alternative text, for those visually impaired, is poorly encoded might hinder LLM evaluation.

Beyond examining theoretical models, this dissertation delves into practical applications, specifically, the use of a genetic algorithm to enhance the process of selecting prompts when determining the relative difficulty of items. This exploration begins without a predetermined notion of the most effective prompt but proposes to identify it through experimentation with various prompt recommendations. These techniques range from incorporating additional contextual information and adjusting input and output instructions, to

modifying the prompt's 'temperature', all recommended by experts in the field as potentially influencing an optimal prompt design.

1.2 Understanding the Challenges in Item Calibration

To build an effective model for analyzing item parameters, it is crucial to define *items* and their function. Items are tools designed by experts to gauge the underlying qualities of a test-taker, which may include knowledge, abilities, skills, or a mix thereof.

The field of test theory lays out the features of evaluations, specifically concentrating on fairness, reliability, and validity. Fairness is rooted in the principle of measurement invariance, which insists that tests should yield comparable results across diverse groups and over time (Mellenbergh, 1989; Wicherts & Dolan, 2010; Van de Schoot et al., 2015). Reliability concerns the consistency with which an item measures the intended construct, defined by Lord & Novick (1968) as the ratio of signal (valid results) to noise (errors in measurement) (Borsboom & Molenaar, 2015). Mohajan (2017) further elaborates on the importance of discerning authentic results from distortions. Validity, however, is more complex; its fundamental definition pertains to whether a tool assesses what it's supposed to (Borsboom & Molenaar, 2015), yet it can be interpreted in numerous ways.

An item must closely correspond with the construct domain it is devised to measure for it to be effective. A well-designed test comprises a range of items that jointly span the entirety of the defined construct domain. Importantly, for a test's validity, the results should be influenced more by the construct domain than by any related, yet distinct domains. Take the assessment of personal agility, for instance: a varied set of physical tasks would be more

appropriate than including a driving test portion, since the latter assesses not only similar motor skills but also knowledge of traffic laws, which are extraneous to agility.

The reliability of items is essential when creating both items and tests. A reliable test distinguishes clear measurements from any interference or ‘noise’ that may arise from irrelevant characteristics of the construct. For instance, the reliability of a personal agility test featuring an obstacle course could be affected by uncontrollable elements like weather, showing how reliability can be influenced by factors not intended in the test design.

While there is a wide recognition that items vary in difficulty, the notion of discrimination—the likelihood that an item will be answered correctly more often by individuals at different ability levels—tends to receive less focus, yet it is vitally important in the calibration of test items.

1.2.1 The Need for Item Parameter Identification

In various fields like education, psychology, medicine, and survey design, items are utilized as measuring tools for latent traits or intangible concepts. The effectiveness and reliability of these items as measures are crucial aspects in their development. Comprehending the key features of these items plays a substantial role in several applications.

In the context of education, items are typically employed to evaluate performance at student, class, institutional, and national levels. They are instrumental in identifying gaps in students’ knowledge and serve as a vital metric in determining personal and technical readiness. However, overly difficult, or extremely easy items could potentially discourage or disinterest students. Items with low discriminatory power can be less informative about the students’ abilities compared to those with higher discrimination loadings.

Another motivational aspect for item parameter identification is its usage in high stakes testing, where fairness and functionality are paramount. In such scenarios, different versions of the same test should offer roughly equivalent measures. This means that scores of two test-takers should be comparable for decision-making purposes, such as allocation of university admission slots or scholarship funds.

1.3 Statement of the Problem

Existing methods of item calibration are expensive. These methods most often rely upon the use of pretesting involving hundreds to thousands of test takers from the examinee population. Some existing testing regimes have been able to incorporate a pretesting procedure into their testing paradigms by forcing test takers to take additional items or test sections. Yet this is not a perfect or optimal solution as it forces students to invest effort in solving items that do not contribute to their overall ability estimates. Apart from shifting this cost of test takers, this kind of pretesting introduces a risk factor for which items might be compromised.

Apart from pretesting, various approaches have been proposed to aid in the calibration of items. These methods can largely be divided into 1. subjective inputs by subject matter experts (SMEs) and non-experts to a lesser extent and 2. computational methods. Inputs by SMEs have been shown to have varying degrees of success with some features such as content domain being strongly predicted while other features such as the item parameter discrimination being difficult to predict by subject matter experts. In general SME based approaches are expensive as they involve time and effort by highly skilled individuals. Computational approaches on the other hand often are costly on the front end, as they involve

expert input into designing a predictive computational model but tend to be much less expensive once established. Like SME based approaches, computational models have varying degrees of success depending upon the item parameters or features being estimated as well as the content the domain of the items.

1.4 Purpose of the Study

In this dissertation I propose the use of an alternative method that takes elements of both the subject matter expert (SME) approach and the computational approach. Large Language models (LLMs) have been shown to have some success at predicting human-like response patterns in a wide array of scenarios (OpenAI 2023). They have been shown to be able to generate satisfactory responses to a range of questions previously only answerable by humans even in the context of not having been explicitly trained for that task. As these models are flexible problem solvers capable of handling diverse requests having been trained on source material spanning vast numbers of documents there is reason to believe that LLM models might be effective at predicting item parameters.

1.5 A Theoretical Framework

1.5.1 Large Language Models General Problem-Solving Capabilities

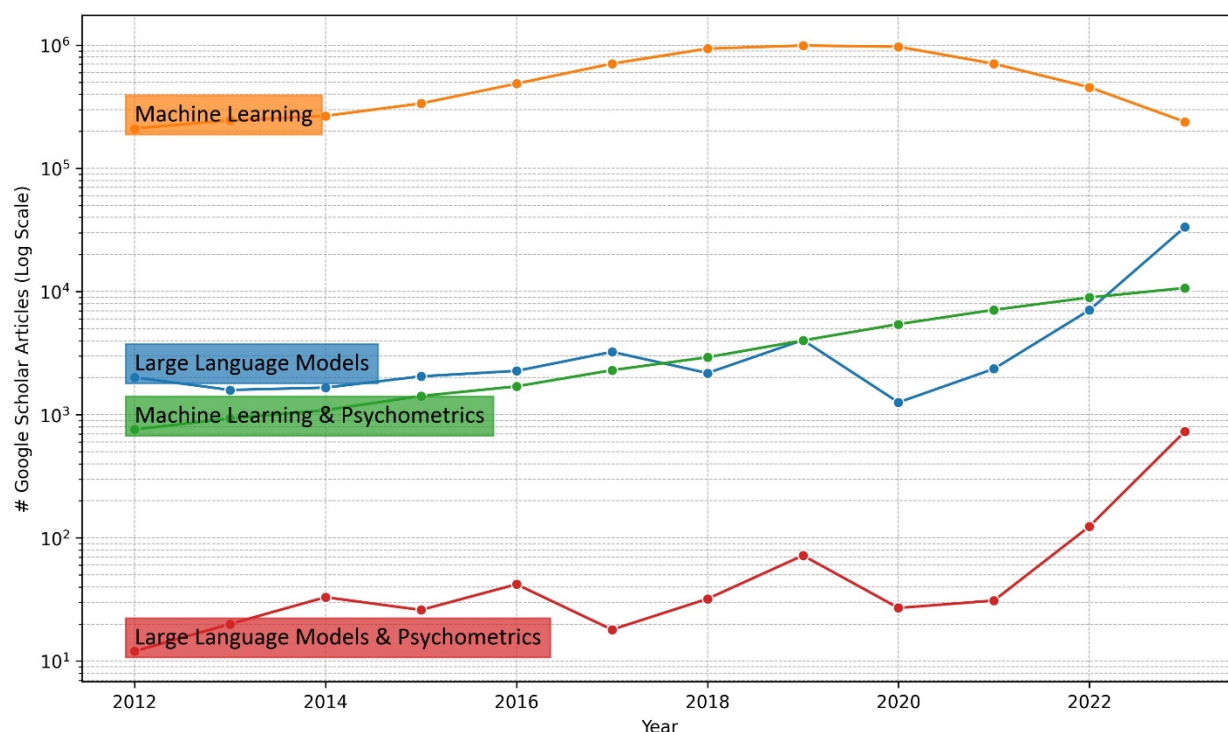
Large Language Models (LLMs) have astonished the world with their ability to flexibly solve a wide array of complex problems for which they were never explicitly trained. Many of these abilities seem to be emergent features of the training procedure. These models are built using transformer-style models (Vaswani et al., 2017) pre-trained to predict the next token in a document. Built on numerous deep layers of billions of simulated neurons and weights, these models are often astonishingly good at generating human like predictions. The very well-known

LLM OpenAI's GPT-3 (Brown et al., 2020) demonstrates "extraordinary language comprehension, fluency, and contextual understanding, enabling it to excel across a wide range of NLP tasks" (D'Souza, 2023: page 1). Its successor OpenAI's GPT-4 (Open AI, 2023) is even more proficient and is currently one of the most sophisticated models available to the public. Many of the large technology companies are competing in this space with the aim of developing similarly complex LLMs such as Google's Gemini-Pro (Anil et al. 2023) as well as Facebook's Llama 2 (Touvron et al., 2023).

The capabilities of LLMs as general problem solvers are relatively new development in the machine learning toolkit with the concept of "Large Language Model" transforming as the models in which spelling, and grammar checking was the primary goal in the early 2010s to models that astonishingly can solve a variety of complex problems in recent years. As such we can see that these models have been growing exponentially in popularity and are being used as a source of research (Figure 1).

FIGURE 1: TRENDS IN SCHOLARLY ARTICLES CITING LARGE LANGUAGE MODELS

These are the returns using Google Scholar searching for the terms. When there is a term that has multiple words the search term is in quotes (for example: “Machine Learning” and “Large Language Models”).



These models have distinguished themselves by being able to solve a wide range of challenges previously reserved for humans. These include GPT-4 (OpenAI, 2023) scoring in the top 90th percentile on a Uniform Bar Exam challenge, 88th percentile on the LSAT, 93rd percentile on the SAT Math, 80th percentile on the GRE Quantitative, 99th percentile on the GRE Verbal, 54th percentile in GRE Writing, 99th-100th percentile in the USABO Semifinal Exam, in the top 20 percentile in AP tests of Environmental Sciences, Macroeconomics, Microeconomics, Physics 2, Psychology, Statistics, Government, US History, as well numerous other exams. While performing astonishingly well in many areas, in some areas GPT-4 is far from mastery. When taking Codeforce's evaluation, a service which evaluated coding ability, it scored only 392 which

is below the bottom 5th percentile. Similarly, it struggled with Leetcode’s examination scoring only 3/45 on Leetcode’s most difficult challenges. GPT-4 scored only in the top 6-12th place on the AMC 10 exam, which is an invitation only math exam that focuses on innovative and challenging math problems. It is noteworthy that GPT-4 is surprisingly good at solving many tasks for which it is not explicitly trained and that it does not seem to benefit significantly from reinforcement learning on similar problems (OpenAI, 2023 Table 8).

While many applications of large transformer models are currently being evaluated and deployed both in research and business, this dissertation seeks to test to what capability these models might provide a feasible replacement for human effort in evaluating and estimating item parameters.

1.5.2 Item Difficulty Prediction

The primary focus of this dissertation is item difficulty, both relative and absolute. Relative item difficulty has a long tradition in item property estimation with early researchers such as Lorge & Kruglov (1952) estimating it along with absolute item difficulty. Absolute item difficulty (D_i) is defined as how likely an item is to be answered incorrectly for a given population. Item i is relatively more difficult than item j if $D_i > D_j$. In more advanced item modelling methods such as item response theory (IRT) the parameter b_i can often be used interchangeably with D when calculating relative item difficulties. Though it is worth noting that more complex models with IRT such as the 3PL model allow for more complex response patterns such that relative item difficulties might change if using D_i rather than b_i to calculate relative difficulties. Relative difficulties do not consider difficulty scaling items as some items might be much easier

or more difficult than other items. By ordering items from lowest difficulty to highest difficulty a spectrum of rank ordered relative difficulties can be established.

When thinking about relative difficulties they appear in some ways more intuitive to that of absolute difficulties. Imagine item A having a difficulty of 60% for a 4th grade population while item B might have a difficulty of 80% for the same population. These numbers are readily interpretable. The relative difficulty of item B is greater than that of item A for 4th graders. Yet there is a hidden level of abstraction that is based on the absolute item difficulties being conditional upon the population in question (that of 4th graders). Let's imagine the same two items being given to 8th graders who now have a new absolute difficulty of 20% for item A and 30% for item B. In this hypothetical example the relative difficulties do not change but that absolute difficulties have changed dramatically with a 40-point change for item A and 50-point change for item B.

Now let us imagine we are asking SMEs to estimate the item absolute difficulty for items A and B for 4th graders. To accomplish this, they would need to first imagine the steps involved in solving the problem, have a mental model of a population of 4th graders and finally have to imagine what percentage of that population would get the item correct. On the other hand, asking a SME to rank the relative difficulties of two items involves the SME mapping out mentally the steps involved in solving item A, then item B, and then evaluating if the steps involved in item A are more or less difficult than those required to solve B. Notice that this latter process does not require the SME to have a mental model of the population so long as we can assume that the relative difficulty of items pairs between populations remains constant (Appendix E.2 explores

the consistency of the rank correlations of item performance between populations groups for the items evaluated in this dissertation).

It is important to note that item difficulty models, both classical and IRT, require absolute item parameters. Fortunately, existing literature provides guidance in methods for mapping item difficulty from relative estimates of item difficulty to absolute (Lorge & Kruglov, 1953) though this dissertation proposes to follow the binary search algorithm similar to that proposed by Attali et al. (2014).

1.5.3 Large Language Model Bias

A key consideration when developing or evaluating any method that produces potentially actionable information is any underlying bias in the LLM being used. This area of focus is of high interest in ongoing development of LLMs (Bai et al., 2024; Liu et al., 2024; Rakshit et al., 2024; Bender et al., 2021). These models are trained using often minimally curated data gathered from the web. Concerns of limiting bias in generated responses is seen as a safety consideration as these models might inadvertently reinforce or introduce harmful prejudices or stereotypes against individuals or groups. This kind of overt prejudice could be very harmful.

However, the use of LLMs to estimate item difficulty is unlikely to be vulnerable to this kind of prejudice yet there also exists a more subtle form of bias that these models might experience. If for instance most training materials for the models is based on the content generated by a particular group of individuals, then inference made on the basis of that content might be more likely to reflect the perspective of that group rather than that of other groups. Some aggregate statistics suggest that girls tend to have stronger reading, writing, and

communication skills than boys (Rieley et al., 2019) whereas boys have some slight advantage over girls in math (McGraw et al., 2006). Free and Reduced-Price Lunch (FRP) ineligible students (not poor) having an advantage over FRP eligible (poor) students (Marchant, 2015). To what extent any of these differences is an effect of or simply correlated with gender, race, English as a second language status, disability, socio-economic status, or any of the other many demographic groups students are divided into is a sensitive matter open to ongoing debate.

However, being cognizant that though an LLM is a conglomeration of numerous voices it generally only represents a single voice at a given time. As such it might be vulnerable to bias in how it returns responses to a task. One of the methods evaluated in this dissertation requires the LLM to make an evaluation of the relative difficulty between item pairs. In many cases we could imagine that there might be no difference in expected relative difficulties between item pairs based on subpopulation group. This might be the case when the form and content of two different items is similar. However, if one item was to contain very few words while another item has a lengthy reading passage that needed to be interpreted, we might expect that the perspective underlying the behavior of the LLM to be predictive of how relatively difficult or easy it finds these different items. Fortunately, this bias is testable as the items evaluated in this dissertation have varying indicators of their different difficulties based on demographic statistics.

LLMs, however, present at least one additional source of bias based on how they find solutions and what kind of challenges they face that are distinct from those faced by students. For example, LLMs, by definition, have a vast trove of linguistic information to draw on. As such they are likely to perform much better on items involving recall or word recognition than

students. Conversely items involving some visual content such as graphs, figures, photo, charts, or maps are not content generally comprehensible to purely language based LLMs. As such, items that rely upon visual material are likely to be much more difficult for LLMs to solve than items that are easily coded in pure linguistic terms.

It is unclear to what extent the advantages or disadvantages of LLM cognitive strengths would affect the relative difficulty ranking of pairwise items. The effect of this bias will likely be to what extent LLMs hold an internal representation of these items and the solution steps required to solve them and then can compare that internal representation between the two sets of items. Li et al. (2023) trained a GPT model using Othello game transcripts and argued that the model sustains a continuous representation of the state of the game board. Likewise, Gurnee & Tegmark (2023) explored temporal and spatial representations with LLMs and found that the LLM seems to generate geographic encodings mirroring latitude and longitude coordinates. That said both these studies use very tangible and easily bounded representations of an internal space while holding the representation of an item's complexity seems an order of complexity greater of a challenge.

Overall, LLMs are black boxes with numerous parameters creating complex spaces difficult to understand and in the case of propriety models restricted from being directly observed. As the second study in this dissertation involves inferring item difficulty based on how much difficult LLMs have at solving the items, these potential sources of structural bias might be important limiting factors.

1.6 Research Questions

This dissertation explores the capacity of Large Language Models (LLMs) to generate responses that mimic those of humans. Specifically, it investigates whether LLMs can accurately predict outcomes for tasks traditionally undertaken by subject matter experts and test-takers. The research focuses on two principal areas: the estimation of relative difficulty of items by the LLMs and the inference of item parameters based on the performance of a variety of LLMs in attempting to solve these items.

1.6.1 Can LLM Models Predict Relative Item Difficulty

In this question I seek to understand to what extent relative item difficulty can be predicted using the current state of LLMs. If LLMs can successfully predict relative item difficulties, then depending upon the accuracy of such predictions this might provide a useful input to other approaches to item calibration or provide sufficient accuracy to greatly reduce the need for additional calibration.

1.6.2 Do LLMS Simulate Student/Test Taker Responses

Model complexity, typically measured by the number of parameters in the model (ranging from a low end of 10s of millions to a high end of trillions), is generally perceived as corresponding with a higher ability of the LLMs to solve more difficult problems. In this study I will use model complexity as a proxy for test taker ability to study to what extent item difficulty can be predicted by the success of the LLMs at solving problems.

1.6.3 Binary Search Algorithm Estimation of Absolute Item Difficult

This dissertation, building on the proposed algorithm of Attali et al. (2014) will test if a series of LLM guided relative difficulty prompts will lead to viable estimates of absolute item difficulty.

1.7 Significance of the Study

Should LLMs prove effective, they could serve as a cost-efficient tool for item calibration, potentially reducing the expenses involved in maintaining current testing programs and developing new testing measures. If the parameter estimates generated by LLMs turn out to be less reliable for direct application in high-stakes testing scenarios, their cost-effective nature means that even moderate success could still offer valuable support in lower-stakes contexts, such as online learning environments. Moreover, possessing even approximate estimates of item difficulties could help lower pretesting costs. This potential to streamline the pretesting process aligns with the research area known as "optimal test design," which aims at reducing the financial and logistical burdens of test development.

CHAPTER II: LITERATURE REVIEW

2.1 Introduction

In this chapter I review the various methods used to estimate item parameters with a particular focus on methods which employ subject matter experts (SMEs). With regards to these studies, I introduce a general conceptual framework intended to aid in explaining why some studies might have been successful while others struggled. In this chapter estimating item difficulty is divided into three buckets: pretesting, SME estimation, and computational methods. Building on these methods, Large Language models (LLMs) offer a potential mechanism for substituting in the input of SMEs or test-takers in these processes. To help set the stage for this I review some of the recent uses of LLMs in education and psychology.

2.2 Search Description

Item parameter estimation has a long history in psychometrics, education, and psychology. Formalized first in the defining and estimation of classical test theory, concepts of item difficulty independent of test taker ability and later refined in item response theory (IRT) item calibration has been explored in many forms.

In this study I explore three research fields in how they approach item calibration: education, psychology, and computer science. In general education and psychology have relied upon pretesting and SME inputs whereas computer science approaches have focused on computational methods.

2.3 Conceptual Framework for Relative item Difficulty Estimation

Many researchers have attempted to use subject matter experts and non-experts (SMEs) to predict relative item difficulty. It is also common for researchers to use expert raters

to estimate item difficulty (Attali et al., 2014; Bejar, 1983) and item domain knowledge (personal experience in writing items in that domain). In this paper I introduce the non-parametric prediction equation $\Psi(\eta_i, \gamma_i, Z_1, Z_2)$ which estimates the likelihood of successfully classifying an item as more or less difficult than an accompanying item. Ψ is a function of item rater skill ability (η_i) in the construct domain as well as item prediction ability (γ_i). Ψ is also a function of the features of item 1 and 2 (Z_1 and Z_2).

Theoretical predictions of the model are that as SME overall ability increases that generally this will lead to better predictions of item difficulty. This need not be true globally, but we should expect SMEs to perform worse at relative item difficulty estimation if their ability levels are much lower than the item difficulty levels. In practice this means that we might be able to expect a high schooler to be a reasonable estimator of the relative item difficulties of basic arithmetic items but not expect a grade schooler to be a good estimator of the relative difficulties of calculus items.

$$\partial\Psi(\eta_i, \gamma_i, Z_1, Z_2)/\partial\eta_i > 0$$

Additionally, we can make the axiomatic assertion that as SME item difficulty ranking skill increases that the probability of correctly identifying the items' relative difficulties also increases.

$$\partial\Psi(\eta_i, \gamma_i, Z_1, Z_2)/\partial\gamma_i > 0$$

To make additional predictions we need to specify some additional parameters. At this point it is helpful to specify what is relative difficulty. I will define relative difficulty

$(R(Z_1, Z_2, \theta))$ of two items as the difference in the expected probability of getting item 1 correct relative to that of item 2 for a given population θ .

$$\begin{aligned} R(Z_1, Z_2, \theta) &= E[P(X_1 = 1) - P(X_2 = 1)|\theta] \\ &= E[P(X_1 = 1)|\theta] - E[P(X_2 = 1)|\theta] \\ &= E[P(X_1|\theta)] - E[P(X_2|\theta)] \end{aligned}$$

In general, it seems intuitively correct that many achievement items, which have large differences in relative difficulties, would retain the same rank ordering of difficulties regardless of the population studied. This assumption I argue is the implicit underlying theoretical basis of much relative difficulty estimation studies.

For example, an SME might be an effective estimator of the difficulty of two math items $(12+13=?)$ and $(256+124=?)$ not because most SME simulate the relative difficulties of the two different items for the target population (say 4th graders) but because they can evaluate how difficult, that is how many steps would be required, for themselves to solve the items and then assume that the population of interest would face similar challenges. This assumption might not hold up if the average test taker deploys a different strategy than that of the SME.

Using this model, I make the first testable prediction. As the absolute size of R increases the likelihood of correctly predicting item difficulties increases.

$$(1) \quad \partial \Psi(\eta_i, \gamma_i, Z_1, Z_2) / \partial |R| > 0$$

Notice that this prediction is irrespective of the underlying item parameter models. To further enrich the model, I include a similarity function $S(Z_1, Z_2)$ that represents the content overlap between the two items. I theorize that the difficulty of items which evaluate knowledge in similar content domains is easier to rank than items of dissimilar content domains. The

underlying justification for this assertion is integrated in the theory of cognitive diagnostic models that explicitly attempt to diagnose and measure the cognitive steps required to solve items. As such, two items which are similar in content domain will be more likely to have similar steps involved in solving the items. The difference in these steps, as one item might have more complexity than another item, provides a basis for asserting that the more complex item is therefore more likely to be more difficult. However, when items are of diverse content domains then it is harder to infer that one item is more difficult than another as the steps involved in solving each are unique to each item.

$$(2) \quad \partial \Psi(\eta_i, \gamma_i, Z_1, Z_2) / \partial S_{1,2} > 0$$

I will test an additional hypothesis regarding relative item difficulty estimation. I argue that items which evaluate constructed knowledge that is explicitly taught and builds on well-known steps are easier to estimate than item difficulties in items which evaluate either implicit knowledge or general non-hierarchical knowledge. The underlying argument for why this would be the case is also built on the cognitive diagnostic approach to item difficulty estimation.

When the steps necessary to solve a problem are well known then it is less complex to evaluate relative item difficulty than for items in which knowledge is non-hierarchically acquired. I

propose function C_i which is a composite measure of the combined constructed complexity of the items being evaluated. As items get more complex, building on accumulated knowledge in an educational setting, I expect items to get easier to rank.

$$(3) \quad \partial \Psi(\eta_i, \gamma_i, Z_1, Z_2) / \partial C_i > 0$$

I evaluate one additional hypothesis. This hypothesis is that items which have higher discrimination a will also be easier to predict relative difficulties. The driving feature behind this hypothesis is that items with poor discrimination introduce a level of randomness which obscures item properties. That is, items which have poor discrimination are items for which low ability students have a non-trivial chance of getting them correct and high ability students have a non-trivial chance of getting them wrong. In the presence of this noise, I hypothesize that the ability to predict difficulty ranking may be compromised.

$$(4) \quad \partial\Psi(\eta_i, \gamma_i, Z_1, Z_2)/\partial a_2 > 0 \text{ and } \partial\Psi(\eta_i, \gamma_i, Z_1, Z_2)/\partial a_1 > 0$$

2.4 Review of Research

2.4.1 The Need for Item Calibration

Psychometric items either achievement or psychological are designed to measure one or more mental construct. Measuring constructs is often considered of importance in education where the acquisition of student knowledge and skills is considered a primary goal. To build effective instruments for measuring student performance item are typically written and designed by item writing experts, evaluated for validity by expert panels, and calibrated on test takers sampled from the population for which the instrument will be used. High stakes test design and calibration is often quite costly. The SAT for example, takes between 18 and 30 months to develop a new form costing approximately \$1 million US (Dudley, 2016).

As high stakes exams are often used to determine eligibility for schools, scholarships, and professional certification there is always an incentive for bad actors to steal and release items. As a result, many testing regimes are constantly developing and calibrating new

replacement items. There is a great potential benefit to finding methods of reducing the effort associated with calibrating new items. Stocking (1990) demonstrates that it is possible to select examinees if the skill level of examinees can be known in such a way as to reduce the number of examinees required to calibrate an item. The study of optimal test design has been applied to item calibration to select items appropriate for each examinees estimated ability level (Berger, 1992; Jones and Jin, 1994; Buyske, 2005; Lu, 2014; Zheng, 2014; Van Der Linden and Ren, 2015; Ren et al., 2017; Berger, 2017; Hassan and Miller 2019; He and Chen, 2020; Hassan and Miller, 2020; among others).

This dissertation explores new methods made possible by LLMs to generate item parameter estimates.

2.4.2 Item Calibration Methods/Models

Pretesting

The gold standard method for estimating item parameters is by administering those items to populations comparable to those whom the instrument is meant to evaluate. This can often be accomplished by adding additional items or sections to an existing instrument when a testing program has already been established. New testing programs typically would not have access to this low-cost pool of ready test takers. Regardless of the method of initial calibration, much care is needed in monitoring how items perform with ongoing administration to identify potential “item drift,” either through overexposure or changes in the testing population.

Under pretesting, examinee response is used to calibrate items relative to other already calibrated items. How many examinees are required to calibrate an item is a function of several factors including the required precision of the item being calibrated as well as the complexity of

the model being estimated. The simplest IRT model, the Rasche model might be calibrated with as little as 30 examinees if the precision of estimation of ± 1 logit is sufficient (Linacre, 1994). However, more complex models or applications requiring higher precision might require thousands of examinees to sufficiently estimate item parameters.

Item pretesting is the gold standard for item calibration. It allows for the most direct measure of item performance. It also allows testing professionals to identify variations in item performance based on different population groups. Test validity requires that items perform similarly across testing groups dependent only on the latent trait being examined rather than other factors which might predict item performance. Methods to identify differences in item performance based on population groups fall under the literature identified as “differential item analysis” or simply “diff.”

This ability to identify differences in performance by population group performance is a major advantage of pretesting which neither expert judge item rankings, the computational methods reviewed in this dissertation, nor the LLM approaches presented in this dissertation provide substitutes for.

Subject Matter Experts Review

Numerous studies starting with Tinkelman (1947) followed by subsequent researchers (Lorge and Kruglov 1952, 1953; Ryan, 1968; Thorndike, 1982; Bejar, 1983; Cross et al., 1984; Melican, 1989; Yao, 1991; Fernandez et al., 2003; Hambleton et al., 2003; Lu et al., 2007; Attali et al., 2014) have attempted to estimate item properties through use of expert human judges. These approaches have had varying levels of success. Typically, individual judges perform

poorly when estimating item properties, yet the average performance across judges does better. There is significant variation between studies, however.

While finding viable alternatives to pre-testing is desirable, the use of judges tend to be expensive, and it is unclear if providing additional training to them results in better estimates of item properties (Bejar 1983). Currently expert judges are frequently used in scale development and validation (Boateng et al., 2018; Hardesty and Bearden, 2004). It is unclear to what extent they are used to estimate item parameters, such as difficulty and discrimination, in professional testing programs. That said, not all uses of items involve high stakes testing. Many studies use imperfect measures of item difficulty based on item difficulty estimates generated by subject matter experts, fully accepting their lack of precision. Yao (1991) for example examines computer adaptive testing when item parameters are imprecisely estimated while others use item difficulty estimates generated by subject matter experts for the purpose of directing automated tutoring content (Fernandez et al., 2003; Lu et al., 2007).

The rest of this section will examine some of the notable results from the use of judges to estimate item properties. Thorndike (1982) has human judges assign item difficulty on a scale between 1 (would be passed by 75% or more of examinees) and 9 (would be passed by 30% or less of examinees). This study used the largest panel of expert judges to review and found reasonable success. Overall, across twenty judges he estimates correlations of 0.83, 0.74, and 0.72 among the average of 20 human raters and the empirical difficulty estimate. This is much higher than the single judge rating of between 0.23 to 0.32. Similar work by other researchers across different item domain fields found item difficulty estimates by subject matter experts correlated with empirical difficulties of between 0 and 0.49 with most estimates

having single rater correlation of less than 0.3 (Bejar, 1983; Melican et al., 1989; Cross et al., 1984).

The following section presents five different research efforts into the use of item experts (and in one case non-experts) at predicting item features. Overall, these methods show promise but ultimately seem to lack the precision to be adopted into a professional high stake testing environment.

TABLE 1: SUMMARY TABLE OF EXPERT JUDGE ESTIMATES

The following table shows the results as found in the following papers. These results unfortunately are spotty as each paper presents different estimators of how well their method performed. In this table Interrater Corr and Rank Corr / Corr represents the Interrater Correlation as well as the Rank Correlation and Item Parameter Correlation while MAE represents Mean Absolute Error of the estimators.

| Paper | Year | Method | Item Type | Interrater Corr | Rank Corr / Corr | MAE |
|-----------------|------------|--|--------------------------------|-----------------|------------------|------------------------|
| Lorge & Kruglov | 1952 | Judges 8 (Test Writing Class) | 8th Grade Arithmetic | .73 / .46 | .83 / .84 | 12.51 / 14.15 |
| Lorge & Kruglov | 1953 | Judges 14 (Advanced Degree in Teaching Math) | 8th Grade Arithmetic | | .65 - .74 | |
| Lorge & Kruglov | 1954 | Judges 14 (Advanced Degree in Teaching Math) | 8th Grade Arithmetic | | | 23 - 24.1 / 2.1 - 12.2 |
| Bejar | 1981, 1983 | Judges 4 | Test for Standard English | .95-.91 | .16-.30 | |
| Mislevy et al. | 1993 | Mixed | Pre-Professional Skills Test | | 0.49 | |
| Attali et al. | 2014 | 24 ETS Judges | 8 Subject Groups of Math Items | | .50-.80 | |

Lorge, Kruglov, & Diamond (Item Similarity: High, Complexity: Medium to High)

Early research into estimating item difficulties through expert judges seemed to demonstrate promising results. Lorge & Kruglov (1952) propose and test a method of estimating classical test theory (CTT) item difficulty and relative difficulties. They split their rater pool of eight PhD candidates taking a test writing class into two studies: 1. In which the raters receive 30 items out of the 150 in which CTT difficulty was provided and 2. The raters received no specific reference information and rated all 150 items. The items were 8th grade arithmetic items.

They found that the raters in study one was much more highly correlated with each other when it came to predicting the two tasks: item ranked difficulty and CTT Difficulty (pass rate) than those who received no reference information. On average, intercorrelations were 0.73 for study one and 0.46 for study two, indicating that having the additional information led to significant improvements in agreement of the judges. Unfortunately, although the judges agree with additional item framing information, they appear to be no better at predicting true difficulty than judges who did not receive that information. Overall, both groups were good at predicting item relative difficulties with correlation scores of 0.84 and 0.83. However, they systematically underestimated absolute item difficulties even in the case of study one in which 30 items had absolute item difficulty estimates provided for reference.

Lorge & Kruglov (1953) follow up their study with another study attempting to address the underestimation of item difficulty in a multistage manner. They split item judges into two groups A and B and for two 45 item tests Test I and Test II. The judges were asked to rank items

in terms of difficulties. At the first stage no information was given and judges were asked to estimate the difficulty of items. In the second stage, group A had 10 items revealed (22% of the items) and was asked to give a new assessment of the non-revealed items' difficulties. This is followed by a third stage of evaluations, when judges were asked to rank the difficulties of items in Test II. The same procedure for the second stage was repeated with judges in group B. Overall, this procedure was meant to assess to what extent judges can learn from mistakes and improve on their difficulty predictions. Revealing the subset of items did improve the estimates for both relative and absolute difficulties.

While this early research was promising, there were and still are concerns as to how well these results map from 8th grade arithmetic items to other types of items. NAEP for example has five general content areas for eighth grade items: "Number properties and operations"; "Measurement"; "Geometry"; "Data analysis, Statistics, and Probability"; and "Algebra." Of these, only a small subset of one of these five content domains, "Number properties and operations" even have some items characterized as "arithmetic items," – though many items would have the necessity of solving arithmetic as part of a solution – finding evaluation.

Lorge & Diamond (1954) continued exploring the possibility of using judges to estimate item difficulty building on the Lorge & Kruglov (1953) paper by further examining how relative difficulty rankings of items can be leveraged as a linear projection into absolute difficulty space. Under three simplifying assumptions they find that using the mean linear projection produces better difficulty estimates than taking the average pass rate estimated directly from judges. Using the linear projection, they were able to estimate the mean absolute error of absolute

item difficulties as 23-24.1 for the item estimates without revealed item difficulties and 2.1 – 12.2 for the remainder of items after revealing 10 item difficulties.

Arbuckle & Cuddy (Item Similarity: High, Item Constructed Complexity: Medium)

Arbuckle & Cuddy (1969) in an unrelated study of item difficulty estimation by judges deploy a two-part study attempting to see: 1. if four experienced judges could predict item difficulties for recall items and 2. if naïve judges could predict those same item difficulties. Four experienced student judges with recall items were asked to first estimate how difficult they would find a series of sets of items (160 sets – 100 5 item sets and 60 6 item sets) to recall. Items for which there existed agreement among judges were kept (105 items) and then given to those same judges. The judges' predictions from the first session were 62% to 72% accurate (50% being random) across the eight judge and item-set pairs giving evidence that experienced examinees could predict item difficulty. Interestingly the four judges were allowed to guess how likely their answers were to be correct and that corresponded with an 84 to 92% accuracy.

To test if naïve student judges without experience with recall items could also predict item difficulties, 150 students with no experience with recall items rank how difficult they expected the items to be. In a second study in the same paper, they split the 150 students into two groups who each reviewed or one of two alternative sets of items (15 items each per set).

The investigation revealed interesting findings about the subjects' capability to predict recall. In Experiment I, practiced subjects were able to predict their recall with an accuracy that was significantly greater than chance. Experiment II further substantiated that even naïve subjects demonstrated a reliable decrease in recall probability that aligned with their predictions along the "very likely" to "very unlikely" scale. This consistency in prediction and

recall, regardless of subjects' prior experience with PA learning, suggested a robust ability of subjects to judge associative strength immediately after presentation. The methodology of immediate predictions and evaluations instructed test takers to minimize the use of rehearsal strategies or other memory aids, focusing instead on the subjects' intrinsic assessment capabilities.

Interestingly, the study showed that the frequency of non-expert judges' predictions of correct recall was consistent with the judges' assessments of item difficulty. Higher difficulty ratings corresponded to lower predictions of correct recall. The subjects' predictions seemed influenced by the apparent difficulty of the PA pairs. Despite some individual variability among subjects, a correlation was found between the two independent assessments (judges' difficulty ratings and subjects' predictions).

This study is of note in that it demonstrates a common task in which both experts and non-experts can estimate item difficulty with some degree of success. However, the study is also problematic in that it does not estimate normed item difficulties nor does not provide any other statistics for comparison with other methods with regards to the correlation of relative item difficulties nor does it provide mean error estimates for item predictions.

Bejar (Item Similarity: Medium, Item Constructed Complexity: Low)

Bejar (1981 and 1983) conducted a study intended to encourage item experts to pool their knowledge to predict more precise item estimates. The study was broken into two parts with the first part dedicated to training expert judges while the second part involved rating items in terms of difficulty, discrimination, and factors that contributed to difficulty. Four professional item writers with between three and 20 years of experience were recruited as

raters. The raters were assembled as a group and asked to work individually writing down their difficulty and discrimination estimates. After the judges revealed their estimates and discussed among themselves the rationale behind their ratings, the raters rated the items a second time. Then difficulty, discrimination, and other statistical information about items was revealed including the distribution of response patterns across students for each distractor, as well as mean criterion score of those choosing each distractor.

The raters were instructed to rate items using a delta difficulty index ($\Delta = \phi^{-1}(1 - p)$) with ϕ^{-1} being the inverse normal CDF and p being the proportion answering correctly. They were also instructed to rate the item using the biserial correlation ($r = \frac{M_R - M_W}{S_T} \frac{p(1-p)}{y}$) with M_R and M_W being mean score for students getting the item right and those getting it wrong, while S_T is the standard deviation of the criterion scores, y is the ordinate of the normal density function corresponding to the $\min(p, 1 - p)$.

The raters were trained on three sets of 20 usage items and three sets of 10 sentence correlation items taken at random and assembled into booklets. Empirical item statistics were calculated as the estimated equated delta and the estimated biserial correlation based on a random sample of 2000 students. A total of three rating sessions were held. Interrater reliability was also calculated on the ratings before and after each discussion as well as after each session. After discussion, the interrater reliability of difficulty estimates increased but in while between sessions the interrater reliability decreased.

The items evaluated dealt with 24 different major error categories in English. These error categories had been previously identified at the time of composition. The error categories in the sample items tested by Bejar only exhibited 19 of the major errors the items are designed

to identify. Bejar used these error categories to create error bands in terms of estimation of the mean difficulty and the discrimination. The mean difficulty for most items were within a narrow window. Some error categories were more difficult while a few appear to be easier than other items.

Information on error-type average difficulty and discrimination was provided to the judges for the final item estimation phase in which all four judges rated two sets of 50 items each. Each set contained 35 usage items followed by 15 sentence correction items. Using the item feature information associated with the type of error, Bejar estimated and projected difficulty and discrimination values. These values were then correlated with the true difficulty and discrimination. Overall, the results were quite mixed with the judges doing better than the error-projected values in difficulty for one set of items but worse for the other.

The highest correlation in any of the phases observed, that was among five different phases with each item if phase 1-3 being rated twice, with any of the time types or methods, was a 0.63 while the correlation between the average rating and the empirical rank was 0.16 and 0.30. Bejar concluded that this is too low for usage in a functional evaluation. He suggested that including more judges might help reduce noise and improve parameter estimates but this approach would be cost prohibitive.

An interesting feature of Bejar's 1981 paper was four reflections by the item judges on how difficult and frustrating it was to predict item properties. It is likely that the items evaluated in this experiment posed a more difficult challenge than those related to estimating the relative difficulty of arithmetic items. Yet, not having to take into account the discomfort of item-raters is a distinct advantage of the LLM methods proposed in this paper.

Mislevy et al. (Item Similarity: Medium, Item Constructed Complexity: Medium to High)

Mislevy et al. (1993) present research into estimating item difficulties that combined both expert review and annotation, indexes, and machine learning models. They presented a study that explored a statistical methodology for equating tests when traditional methods are constrained by the unavailability of examinee response data. The study postulated that while standard equating practices are reliant on large pools of examinee responses, various alternative sources of data, like content specifications, expert opinions, or theories related to psychological processes involved in solving test items, could provide valuable insights into item characteristics.

Mislevy et al. explored item data from the Pre-Professional Skills Test (now known as the Praxis) from 1985 and 1990 measuring reading, writing, and math skills for prospective teachers during college years. One of the ways Mislevy et al. integrate expert judgement into their equating procedure was by leveraging the insights of subject matter experts who were adept at predicting item properties. Mislevy et al. utilized this expert judgement by coding items based on content and cognitive processing features that are then incorporated into their item parameter prediction model. They emphasized that although expert judgement does not account fully for item difficulty variance, it served as a substantial component in the absence of traditional examinee data.

The study methodically categorized test items based on their content and cognitive process demands, as judged by experienced item developers. Items assessed include selection and placement tests that tap into the reading, mathematics, and writing skills of prospective teachers, as evidenced by the analysis of the Pre-Professional Skills Test. The researchers

assigned ratings to various item features, reflecting their potential contribution to the difficulty and effectiveness of the items. These features included aspects such as the number of syllables per word, sentence length, the presence of concealed information, and more intricate elements such as the number of rules present in a problem-solving task versus the number needed for its resolution.

The success of incorporating subject matter expert inputs into equating was quantified by the variance accounted for in item parameters through multiple regression models. For example, the predictive model, which Mislevy et al. developed, was designed to predict IRT parameters such as item discrimination (slope), difficulty (intercept), and guessing (lower asymptote) by correlating them with collateral information variables provided by the experts. While the authors recognize that the predictions did not match the precision of traditionally obtained item parameter estimates, they were beneficial in forming a tentative equating function in which they presented a representation of uncertainty.

Overall, the study demonstrated a pragmatic approach to test equating in conditions where examinee data are sparse or non-existent. The utilization of expert input was hoped to establish a new paradigm that allowed for fruitful intersection between psychometric theory and practical constraints. Although the resulting psychometric properties are less precise than those derived from large scale pretesting data, the contribution of subject matter experts' insights offered an alternative method for equating test items..

This study was interesting in that it deployed judges in a more complex way than the previous studies by both having judges rank items as well as generate additional collateral information on items which can then be used jointly to predict item difficulty. The studies

presented in this paper build on this work by likewise leveraging LLMs to generate additional features of items and item pairs (collateral information) which is then used to aid in the estimation both item difficulties and predictions of likelihood of correctly estimating relative item difficulties.

Attali et al. (2014) (Item Similarity: Medium to High, Item Constructed Complexity: High)

Attali et al. (2014) hypothesized that by asking judges to rank items relative to one another in short item sets, a more precise assessment could be achieved. In their study, a total of 26 subject matter experts from Educational Testing Service (ETS) were employed, including SAT and GRE test developers, experienced item writers, and relatively new item writers. These judges were required to rank sets of SAT mathematics items, which covered eight major content areas such as algebraic problem-solving and geometry. Item ranking was done within each of these content areas. Their procedure had these experts arrange seven items per set in a rank order from easiest to hardest, balancing cognitive load against efficiency. The sorting provided indirect information on the relative difficulty through 21 paired comparisons per set. To enhance the variety of comparison types, the easiest and hardest items were oversampled among a pool of 28 released multiple-choice items per content area. The items were then randomly ordered in booklets for the examination. The study's findings suggested that the judges could successfully rank order the items across various content areas, with a median Spearman rank-order correlation of 0.79 between their judgments and actual item difficulties.

In the study, the team conducted analyses in two stages. Initially, descriptive analyses of the rankings were performed to assess the correlations between the judges' assessments and the equated delta values of the items. However, these were not deemed unbiased estimates

due to the non-random selection of items. Thus, a more robust analysis was conducted focusing on individual paired comparisons. Each complete ranking by a judge led to 21 comparisons, where the primary outcome was whether the judge could accurately identify the harder item. A hierarchical general linear model was applied to the binary outcome of comparison correctness as a function of the empirical difficulty difference between compared items. The study found little influence of judges' background on their ability to discriminate item difficulties, and the probability of success in these paired comparisons increased on average with the empirical difficulty difference between items.

To propose a potential implementation strategy for their comparative judgment approach, Attali et al. outlined a binary search algorithm. This simulated procedure involved judging the difficulty of a novel item which is of the same family of to a series of anchor items that had known difficulties. Starting with an anchor item at the median difficulty level, subsequent anchors were chosen incrementally based on the outcome of each prior comparison—akin to a binary search strategy. This enabled a translation of pair-wise comparisons into a numerical estimate of difficulty, with each comparison refining the estimated difficulty level on a novel item. For example, after three comparisons, raters could categorize an item among eight difficulty percentiles.

To assess this method's practicality, Attali et al. performed a simulation with 10,000 items and examined correlations between true item difficulties and averaged difficulty judgments from various numbers of comparisons and raters. The findings revealed that the correlation increased with the number of comparisons but plateaued beyond three, whereas the number of raters contributed more significantly to the correlation even when more than

four or five were rating. Impressively, the study detailed that five raters utilizing three comparisons could replicate the accuracy of empirical difficulty estimates traditionally derived from a sample of 100 test-takers. This result underscores the potential efficacy and efficiency of using comparative judgments in estimating item difficulties as opposed to relying on large-scale field trials.

Overall, this paper presents some successful results with correlations between item difficulty and true difficulty at a similar level as that of Lorge and Kruglov (1952). While promising and potentially attributable to advances in item-ranking algorithms by judges, I suspect the differences in types of items being evaluated has more explanatory power than the underlying ranking algorithm. One reason is that “easiest and hardest items were oversampled to increase the likelihood of all types of comparisons across the difficulty spectrum.” While this meant that they could get more precise estimates for the relative difficulty ranks on the lower and upper tails of the difficulty spectrum, it also likely led to a greater number of large pairwise item difficulties than would be expected in a random sample of items.

Also, like Lorge and Kruglov, Attali et al. were able to confine the relative item ranking by judges to items of a narrow subject window for eight different sets of items from the SAT. These items are also those items in which skills for learning them are potentially taught in a well-defined and potentially linear manner. This is quite different than the language misuse items evaluated by Bejar (1981, 1983) or the Pre-Professional Skills Test items which covered a range of subjects evaluated by Mislevy et al. (1993).

Computational Models

Indexes

As human based methods of estimating item parameters are inherently expensive, alternative means of estimating item parameters have been extensively explored. These include estimates of reading complexity through linguistic complexity scales to more recent developments in large language models specifically trained for the estimation of item parameters.

Linguistic Complexity Scales such as the Flesch (1948), Farr-Jenkins-Paterson the first computer implemented readability index (Danielson & Bryan, 1963), and the subsequent Flesch-Kincaid (Kincaid et al., 1975) have long been used as methods of estimating text difficulty and readily map to estimating the difficulty of reading comprehension items with various levels of success (Rafatbakhsh and Ahmadi, 2023). Brown (1998) finds that readability indexes are only weak predictors of the difficulty of cloze items for students who have English as a foreign language. Though he does find various grouping of lexicological features such as average number of syllables per sentence, frequency of words longer than 7 characters, and the percentage of function words combined were strong predictors of difficulty. Freedle and Kostin (1993) test numerous measures of text quality including vocabulary level, paragraph length, number of paragraphs, abstractness of the text, as well as traditional readability indexes and find that they are able to predict 46% to 59% of the variance in item difficulty in a smaller set of items which mapped to 21% to 29% using a larger set of items.

While linguistic complexity are intuitive and natural predictors of difficulty with reading comprehension items and other linguistic items it is not clear to what extent they should be

significant predictors of item difficulty for items outside of this domain except in that they measure incidental difficulty caused by linguistic complexity tangential to the primary scale being measured.

Machine Learning Models

In recent years pretrained transformer models have risen in popularity as a potential method for estimating item difficulty. These models often start with a generalized pretrained transformer model such as BERT (Devlin et al., 2018) or BART (Lewis et al., 2019) which are then adapted through “fine-tuning” for a specific use case such as predicting item parameters when item parameters are known for the training data. This method and similar training than validating specialized predictive models has been shown to be generally successful in some use cases such as predicting construct loading (Hernandez and Nie, 2022).

Before the recent advances in LLMs, many machine learning and natural language processing explorations involved developing supervised models trained on source data with a specific output purpose in mind.. These models might have built on a generalized model such as BERT or BART but their applications are usually extremely specific such as predicting a specific item feature, typically item difficulty. In the next two subsections I will summarize two recent papers with the same lead author Bendetto though one paper looks at a recent development of a computation model to predict item response parameters while the other summarizes recent computational methods in the field primarily driven by computer scientists.

Bendetto et al. (2020)

Bendetto et al. (2020) build and train an NLP model to predict item difficulty and discrimination for multiple choice items by extracting meaningful features from the items and

using them as a predictive model. They introduce a framework for estimating item newly created items in three steps: 1. Estimating latent traits of items, 2. Extracting meaningful features from items, and 3. Estimating item properties from those features. Their framework allows for these steps to be done separately with different items, though presumably the underlying construct must be unidimensional. They also present an ablation study to support their choice of features. They also do a validation study predicting student responses using estimated item properties for an observable ground truth.

They use the two-parameter logistic model (2PL) in their item parameter estimates. They note that most studies like theirs use the item “wrongness” or CTT difficulty as the primary predictive variable of interest rather than IRT item difficulty. Item features extracted from the items are stored in a Q matrix which is then used in a linear model to predict item parameters. They use two random forest regressions to predict item difficulty and discrimination. They divide the features to extract into three components i) Readability Features, ii) Linguistic Features, and iii) Information retrieval features.

The readability features they use are: Flech Reading Ease (1948), Flesch-Kincaid Grade Level (Kincaid et al, 1975), Automated Readability Index (Senter and Smith, 1967), Gunning FOG Index (Gunning, 1968), Coleman-Liau Index (Coleman, 1965), and SMOG Index (Mc Laughlin, 1969).

Linguistic features are similar to readability features (motivated by DuBay, 2004) and use Word Count Question, Word Count Correct Choice, Word Count Wrong Choice, Sentence Count Question, Sentence Count Correct Choice, Sentence Count Wrong Choice, Average Word

Length Question, Question Length divided by Correct Choice Length, Question Length Divided by Wrong Choice Length.

For Information Retrieval features making the assertion that the words used in the text must imply a relationship with the latent trait being measured. They preprocess the text using standard NLP steps then consider the text of the question and the possible choices by grouping the text together they then use Term Frequency-Inverse Document Frequency (TF-IDF) selecting a two-part threshold tuned with cross-validation- to remove both too frequently-used words and too uncommon words.

In addition to using Random Forest (RF) they also tested, Decision Trees (DT), Support Vector Regression (SVR), and Linear Regression (LR). Hyperparameter tuning was preformed via a 10-fold randomized cross-validation. The results of their experiments were reported in both Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

TABLE 2: DIFFICULTY AND DISCRIMINATION FROM LITERATURE

Bendetto et al. (2020)

| | DIFFICULTY | | | | DISCRIMINATION | | | |
|-----|------------|------|----------|------|----------------|------|----------|------|
| | VALIDATION | | TEST SET | | VALIDATION | | TEST SET | |
| | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE |
| RF | .739 | .575 | .753 | .587 | .393 | .296 | .369 | .287 |
| DT | .748 | .586 | .826 | .636 | .393 | .295 | .375 | .290 |
| SVR | .797 | .632 | .804 | .629 | .394 | .298 | .379 | .296 |
| LR | .752 | .599 | .779 | .607 | .397 | .298 | .378 | .293 |

Overall, their model outperforms other recent models such as Qiu et al. (2019), Huang et al. (2017), and Yaneva et al. (2019). Though comparison is difficult to perform as the models are not publicly available and they report using item CTT RMSE on difficulty rather than IRT.

Benedetto et al. use a scaling formula “relative RMSE” defined as $\frac{RMSE}{difficulty_{max} - difficulty_{min}}$.

Using relative RMSE they find a substantial improvement over methods predicting item difficulty.

Benedetto et al. (2023)

Benedetto et al.'s (2023) survey paper serves as current reference for understanding the development of computational methods in estimating item difficulty. It reviews 18 computational models which have all been published since 2015 predicting item difficulty using Natural Language Processing (NLP) techniques. These NLP approaches offer the potential to some limitations of earlier methods by automating and refining the estimation process, thus enhancing the scalability, objectivity, and consistency of question calibrations.

The review presents a taxonomy based on question characteristics, which is pivotal in organizing and comparing different difficulty estimation approaches. Specifically, the categorization distinguishes between Language Assessment (LA) and Content Knowledge Assessment (CKA), while considering the formats and contexts of questions such as reading comprehension, listening comprehension, vocabulary knowledge, and sentence knowledge. By choosing this taxonomy, Benedetto et al. organize recent research, providing a structured means for discussing and comparing varied methodologies in item difficulty estimation.

Computational methods for estimating item parameters, which are central in item difficulty estimation, often involve feature extraction from the question texts. Various machine learning models, such as support vector machines (SVMs), random forests, neural networks, and state-of-the-art approaches like transformer-based models (e.g., BERT), have been explored. These methods have shown some success in capturing semantic representations and

syntactic structures of question texts. The survey acknowledges the transition from traditional feature engineering to leveraging pre-trained language models, which allows for higher levels of generalization and has shown promising results in gauging question difficulty for low stakes applications.

The survey by Benedetto et al. also sheds light on the challenges associated with the evaluation and reproducibility of difficulty estimation models. Due to the scarcity of publicly available educational datasets and the privacy concerns tied to such data, the direct comparison of various algorithms remains difficult. As a result, consistent evaluation metrics and standardized protocols for model validation are lacking. The authors stress the need for more communal data sharing to corroborate the reliability of item difficulty estimation systems in diverse educational contexts.

Concluding their survey, Benedetto et al. draw attention to the implications and future directions in the field of question difficulty estimation. While significant strides have been made, there are areas ripe for improvement, such as exploring the effect of multimodal data (e.g., visual content associated with questions), enhancing the interpretability of the model predictions, and developing methodologies that can generalize across various domains and languages. Their work relates to models deployed in live educational environments, where having an idea of item difficulty can enhance student learning outcomes.

The current state of computation models seems to suggest that they are sufficient in their use case, low-stakes online learning environments, but still need to demonstrate much higher accuracy before being ready to be used in high stakes testing environments.

2.4.3 The Generative Large Language Model Revolution

The Use of Generative Large Language Models in Education

Advancements in large language models (LLMs), like GPT-3, have paved the way for new research possibilities in education. These models have been leveraged for several educational purposes, including the creation of automated questions (Bezirhan & von Davier, 2023; Raina & Gales, 2022; Wang et al., 2022; Settle et al., 2020; von Davier, 2019), the production of educational materials (Hocky & White, 2022; Moore et al., 2022; Walsh, 2022), the scoring of responses (Mizumoto & Eguchi, 2023; Wu et al., 2023), and providing feedback to students (Matelsky et al., 2023; Peng et al., 2023). The capabilities for language fluency, adaptability, and user-friendliness exhibited by LLM have enhanced their role in educational innovations.

A key example of this innovation comes from Settles et al., (2020) who effectively employed a LLM to generate a vast number of linguistic items. Subsequently, a second model was used to predict the difficulty of these items by examining features like their length, word log-likelihood, and Fischer score.

LLMs excel in various knowledge-based and problem-solving tasks, outperforming expectations in areas where they haven't received explicit training. Rae et al., (2021) put a LLM to the test across 152 varied tasks, with 57 tasks pertaining directly to educational interests, covering subjects like High School Chemistry and Astronomy. Moreover, White et al., (2023) explored the potential of LLMs in interactive tutoring, assessing their proficiency in solving specialized chemistry coding challenges.

Despite the notable advantages of LLMs in enhancing educational experiences, concerns about their misuse remain significant. Critics, such as Rudolph et al., (2023), emphasize the

relative ease with which these models could be misappropriated to generate inauthentic student work or for teachers to craft insincere responses, highlighting the need for cautious and considerate implementation in educational settings.

The Use of Generative Large Language Models in Content Evaluation

The ability of large language models to correctly predict the semantic content of language combined with their ability adapt to a variety of challenges has made them a promising potential tool for estimating content properties. However, the degree to which these models can reliably discern content quality and attributes is still an active area of research.

Moore et al. (2022) attempted to use of fine-tuned GPT-3 model variants to identify low-quality student generated chemistry questions as well as predict Bloom's revised Taxonomy of item complexity (Yahya et al., 2012; Bloom, 1956) in the context of online learning for which results are compared with expert judgement. In terms of high and low-quality items the GPT-3 fine-tuned model agreed only 40% of the time with an overestimation of item quality in 85/86 instances of mismatch between the expert reviewers and the model. They also use a fine-tuned LLM to predict Bloom's revised Taxonomy levels for 120 questions. The model only agreed with the expert reviews 38/120 times (32%). Overall, the model demonstrated excessive optimism with regards to item quality while predicting item complexity labels at a less than desirable level.

Yang and Menczer (2023) leverage LLMs to predict the credibility of over 7,000 news sources. ChatGPT produced a Spearman correlation of 0.54 in ratings with those of experts. It is unclear to what extent it demonstrates the ability of the model to estimate content quality as it seems to rely on the training knowledge about the new sources rather than directly reading

articles and predicting their quality from their content. Furthermore, while these results are statistically significant a correlation of 0.54 still leaves a sizable number of sources misclassified.

Bewersdorff et al. (2023) have more success deploying GPT-3 and GPT-4 as a rater attempting to identify student errors in student experimental protocols. They find that their system successfully identified many types of student's errors such as when students are focusing only on the expected outcome and not on the dependent variable (accuracy = .9), when students change trials during an ongoing study (accuracy = 1), but struggled more with identifying when a student is conducting a valid control study (accuracy = .6).

Concerns Over Generative Large Language Models in Content Classification

Concern over the reliability of LLMs to classify content is an emerging research field. The recent study by Reiss (2023) provides an analysis of ChatGPT's reliability in the domain of text annotation and classification, both methods of which are used in this dissertation. The paper, entitled "Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark," addresses the non-deterministic nature of ChatGPT and its implications on the consistency of annotation outputs. Reiss emphasizes the variations that can occur in ChatGPT's performance even with identical inputs, attributing such fluctuations to the model's inherent randomness and the sensitivity of its response to different prompt instructions and model parameters, such as temperature. The paper demonstrates how minor modifications in the prompt or changes in the model's temperature settings can result in varied outputs, thereby raising questions about the deterministic reliability that is often assumed of computational methods in text analysis.

In conducting the study, Reiss tests and quantifies the influence of these variations using 234 German-language website texts, classified as 'news' or 'not news' with ten different sets of instructions and temperature settings at 0.25 and 1. The author notes that while pooling outputs from repetitions could improve consistency, the consistency based on a single output generation did not meet the Krippendorff's Alpha (2011) threshold of 0.8 for acceptable reliability. This finding highlights the need for concern over the use of tools such as ChatGPT for content classification. Reiss' investigation offers caveats on the limitations and potential biases that may emerge when using LLMs. The study stresses the importance of majority decision protocols based on multiple repetitions of input configurations to improve consistency resulting in increased costs.

CHAPTER III: RESEARCH METHODS

3.1 Introduction

This chapter outlines the methodologies used to investigate the use of large language models (LLMs) for estimating item properties. It details empirical tests to assess the effectiveness of LLMs in predicting item difficulties. The first theoretical exploration involves a model evaluating the probability that an LLM accurately predicts relative item difficulty. This is complemented by a study leveraging multiple LLMs working together to estimate absolute item difficulty, substituting for human test-takers. Finally, the chapter describes an approach that combines the model's relative item difficulty prediction with a binary search algorithm to predict absolute difficulties. These methods offer novel insights into how LLMs may assume roles traditionally occupied by human judgment and effort in item calibration.

3.2 Experimental Design

Training and Testing Framework

The training, testing, and validation framework is a common method used in machine learning for model construction. In these approaches, models are flexibly fit and selected based on their predictive performance for the outcome variable. A prime concern with this approach is overfitting, which can result in a loss of inference due to the testing of numerous parameters. The typical solution is to divide the data into two sets: one for developing the model (training) and another for testing the selected model's performance.

Although the approach used in this dissertation does not follow the traditional machine learning model training, it does experiment with and test different model inputs. The goal is to select inputs that yield the highest performance in terms of the pairwise ranking of items.

These inputs, referred to as “prompts,” are known to be highly sensitive to seemingly small changes in their content.

3.3 Research Questions

3.3.1 Study One: Can Generative LLMs Predict Relative Item Difficulty?

In this dissertation, I am primarily concerned with whether LLMs can predict relative item difficulties. To test this, I conduct a series of pairwise comparisons between items, prompting the model to infer which item is more difficult. The item data is split into two pools: training and testing. The training data is used to evaluate different prompts, optimizing them to determine which produces the best output performance (see Appendix B). The top-performing prompt is then applied to the testing data, and the output is analyzed in this study and in Study Three.

I estimate the ability of LLMs to gauge pairwise relative item difficulty using a linear regression model. This model uses the correct ranking ($RankingCorrect_{i,j}$) of an item relative to another as a binary outcome, coded as either 1 (correct) or 0 (incorrect). The following equation is applied at the item pair level, with subscripts representing the two different items being compared.

$$\begin{aligned}
 (3.3.1.1) \quad & RankingCorrect_{i,j} = \beta_0 + \\
 (Row\ 2) \quad & \beta_1 SecondItemHarder_{i,j} + \beta_2 |D_i - D_j| + \beta_3 \hat{a}_i + \beta_4 \hat{a}_j + \\
 (Row\ 3) \quad & \beta_5 Grade_{12} + \beta_6 Grade_8 + \beta_7 Math + \\
 (Row\ 4) \quad & \gamma_1 Complexity_i + \gamma_2 Complexity_j + \\
 (Row\ 5) \quad & \gamma_3 |Complexity_i - Complexity_j| +
 \end{aligned}$$

$$(Row\ 6) \quad \gamma_4 NumberOfSteps_i + \gamma_5 NumberOfSteps_j +$$

$$(Row\ 7) \quad \gamma_6 |NumOfSteps_i - NumOfSteps_j| +$$

$$(Row\ 8) \quad \gamma_7 Similarity_{i,j} + \epsilon_{i,j}$$

In equation (3.3.1.1) β_0 is the constant while the next row is composed of values derived directly from the NAEP data. The binary indicator variable (*SecondItemHarder_{i,j}*) is 1 if the second item presented to the LLM is more difficult than the first item. This term captures any bias that might exist if the model is more likely to rank the second item as either easier or more difficult. The second exogenous predictor ($|D_i - D_j|$) is the absolute difference in empirical difficulties. The assumption is that the greater the difference in difficulty between two items, the more accurately the LLM will rank them.

The next two predictors (\hat{a}_i and \hat{a}_j) are proxies for item discrimination parameter (a) which is not provided by NAEP. I hypothesize that items with stronger discrimination will be easier to classify. The hypothesis is that items with stronger discrimination are easier to classify. Items with higher discrimination indicates that there is less construct irrelevant noise than such that those who pass the item are more likely to have higher abilities than those who fail the item relative to an item with the same construct difficulty but lower discrimination parameters. The proxy variable for discrimination is calculated as the population average ability level for item i who answered the item correctly ($\bar{\theta}_{Yi=1}$) minus the population average ability level of those who answered incorrectly: ($\bar{\theta}_{Yi=0}$). Thus:

$$(3.3.1.2) \quad \hat{a}_i = (\bar{\theta}_{Yi=1} - \bar{\theta}_{Yi=0})$$

The proxy variable for item discrimination \hat{a}_i are normalized with a mean of zero and a standard deviation of 1 across all items in each subject, grade, year combination. The hypothesis is that the coefficients β_3 and β_4 will be positive, indicating that items with higher discrimination are easier to rank.

With NAEP data, the population average ability level for each option selected by test-takers is listed, along with the proportion that selected each option. As a result, calculating $\theta_{Yi=0}$ takes the additional step of calculating the weighted average between ability levels for test takers who chose a distractor d . With P_{id} defined as the proportion who chose that distractor, we can calculate the weighted average in the following way:

$$(3.3.1.3) \quad \bar{\theta}_{Yi1=0} = \frac{\sum_{d=1} P_{id} \bar{\theta}_{Yi1=d}}{\sum_{d=1} P_{id}}$$

Here is an example of the calculation of \hat{a} for an item in which population average performance metrics are provided for each option taken. For an item with four distractors the average performance on the test for those who got the item correct is 500 while for those who chose the distractors 1, 2, and 3 respectively are 200, 300, and 400. Also, the proportion of those who chose each distractor was 5%, 20%, and 15%. The difficulty of the item is 40% meaning 60% of those taking the item got it correct. The calculation of the weighted average of those who chose an incorrect option would therefore be $\bar{\theta}_{Yi1=0} = (.05 * 200 + .2 * 300 + .15 * 400) / (.05 + .2 + .15) = 325$. The proxy variable for a_i in turn would be calculated as $\hat{a}_i = (\bar{\theta}_{Yi=1} - \bar{\theta}_{Yi=0}) = 500 - 325 = 175$.

Let's compare this item with another hypothetical item j which has the same difficulty (40%) and the same proportion choosing each distractor (5, 20, and 15) but has a different

population average performance. Let's say those who chose the correct answer got a 525 on average and those who chose each of the distractors got a 100, 200, and 300. Now $\bar{\theta}_{Yj1=0} = (.05 * 200 + .2 * 200 + .15 * 300) / (.05 + .2 + .15) = 237.5$. The proxy variable for a_j in turn would be calculated as $\hat{a}_j = (\bar{\theta}_{Yj=1} - \bar{\theta}_{Yj=0}) = 525 - 237.5 = 287.5$. From the above example of item i and j we can see that \hat{a} seems to be a reasonable proxy for the discrimination parameter a in that as the difference in overall performance between those who got an item correct and those who got the item incorrect gets larger so does \hat{a} .

The variables in (3.3.1.1) row 2 are all values for which there might be missing or incomplete information if deploying these methods in a test development setting and therefore should be framed as context setting. However, row 3 does not suffer from these issues as it would be known to the test developer prior to calibration what grade and subject individual items are intended for. The remaining explanatory variables would also be known to the test developer, but they are generated via an initial feature generation step by the LLM.

For each item (i) the LLM assigns during separate prompts a complexity ranking from 0 to 10 as well as a list of steps required to solve the item. The complexity ranking is directly extracted from the LLM's predicted complexity level when presented with an item (see Appendix C). I hypothesize that items which are more complex will also be items which are easier to rank. Likewise, when item pairs have a large disparity in item complexity

$(|Complexity_i - Complexity_j|)$, I hypothesize that these items will be also easier to rank. This is due to items which have greater complexity disparity being a proxy of sorts for perceived item difficulty and therefore in turn they also have larger predicted item difficulties.

In an equivalent manner to item complexity, the number of steps required to solve an item (*NumberOfSteps_i*) is generated by the LLM and used as a predictor. Unlike complexity, the number of steps is indirectly calculated by first prompting the LLM to list the steps required to solve the item then counting the numbers provided (see Appendix C). Unlike complexity, the number of steps has no upper limit. Like higher complexity, a higher number of steps indicates items which are more cognitive and therefore hypothesized to be easier to rank.

3.3.2 Study Two: Do LLMs Response Pattern Simulate Student Responses?

In this study I seek to understand if multiple LLMs working together can be used to predict absolute item difficulty. This model will attempt to leverage the problem solving “skill” of different LLMs to predict how difficult items are. In this approach I will use model performance from an array of LLMs as a linear predictor of empirical item difficulty. In addition to the average performance will of various LLMs on each item I will include additional computational scales such as word count (*WordCount*), Flesch-Kincaid Index (*FleschKincaidIndex*), and average syllable count (*SyllablesPerWord*) as predictors in the linear model. Empirical item difficulty (*EmpiricalItemDifficulty_i*) in this equation is the percentage of students who selected a distractor or no response.

In this study, I aim to understand if multiple large language models (LLMs) working together can predict absolute item difficulty. This hypothesis seeks to leverage the problem-solving abilities of different LLMs to foresee how challenging items will be. The model utilizes performance data from various LLMs as a linear predictor of empirical item difficulty. Alongside the average performance of various LLMs on each item, I include additional computational scales such as word count (*WordCount*), Flesch-Kincaid Index (*FleschKincaidIndex*), and

average syllable count (*SyllablesPerWord*). These predictors, when incorporated into a linear model, provide an equation that can be estimated through logistic regression. Empirical item difficulty (*EmpiricalItemDifficulty_i*) in this context is the percentage of students who selected a distractor or provided no response.

$$\begin{aligned}
 (3.3.2) \quad & EmpiricalItemDifficulty_i = \beta_0 + \beta_1 Grade_{12} + \beta_2 Grade_8 + \beta_3 Math + \\
 (Row2) \quad & \beta_4 WordCount + \beta_5 FleschKincaidIndex + \beta_6 SyllablesPerWord + \\
 (Row 3) \quad & \gamma_1 \overline{Prompt_1 Score_{GPT3.5_i}} + \gamma_2 \overline{Prompt_1 Score_{GPT4_i}} + \\
 (Row 4) \quad & \gamma_3 \overline{Prompt_1 Score_{Llama7B_i}} + \gamma_3 \overline{Prompt_1 Score_{Llama13B_i}} + \\
 (Row 5) \quad & \gamma_4 \overline{Prompt_1 Score_{Llama70B_i}} + \gamma_5 \overline{Prompt_1 Score_{GeminiPro_i}} + \\
 (Row 6) \quad & \kappa_1 \overline{Prompt_2 Score_{GPT3.5_i}} + \kappa_2 \overline{Prompt_2 Score_{GPT4_i}} + \\
 (Row 7) \quad & \kappa_3 \overline{Prompt_2 Score_{Llama7B_i}} + \kappa_4 \overline{Prompt_2 Score_{Llama13B_i}} + \\
 (Row 8) \quad & \kappa_5 \overline{Prompt_2 Score_{Llama70B_i}} + \kappa_6 \overline{Prompt_2 Score_{GeminiPro_i}} + \epsilon_i
 \end{aligned}$$

In the above equation, the bar over a variable indicates the mean score ($\sum_{i=1}^N \frac{var_i}{N}$) of that variable for that item. For example, if GPT-3.5 got an item correct on two out of three attempts under prompt 1, then the mean score for that item under that prompt would be $\overline{Prompt_1 Score_{GPT3.5_i}} = .66$. Since LLMs do not have memory of previous attempts, we can consider their mean performance score—derived from repeated attempts on the same item—as estimators of the “true” performance score for that LLM on that item. The use of LLM response APIs in this dissertation is distinct from the use most are familiar with via ChatGPT or other consumer facing LLM applications. Those applications have processes in which they capture prior prompts and response and feed those values into subsequent queries. Through this mechanism they have a working memory. This dissertation does not use this kind of continuous prompt feed to generate responses.

I predict the coefficients on β_0 through β_6 to be positive except for the coefficient on Math which I have no hypothesis for. In general, higher graded items tend to be more difficult than lower graded items even for students at that grade level.

This model tests whether the difficulty a student faces when attempting an item corresponds to the difficulty the LLM faces with the same item. If LLMs encounter similar challenges attempting items as students, we would expect the coefficients on the LLMs (γ_{1-5} and κ_{1-5}) to be, on average, positive. Due to the similarity in their training data and design, different LLMs may exhibit some collinearity, with their responses correlated. This introduces a concern about near-multicollinearity, which could potentially obscure the effects of individual LLM responses in the model.

However, the overall goal of (3.2.2) is less about interpreting the coefficients on LLM performance scores so much as it is about building a model using those performance scores to predict empirical item difficulty ($\widehat{EmpiricalItemDifficulty}_i$) for a given item (i). As such the estimation of empirical difficulty as a sum of performance scores of LLMs can be thought of as the weighted average of scores with the weights selected via Ordinary Least Squares (OLS) to maximize their ability to predict the empirical item difficulty. The test to measure how well the LLMs performance is at jointly predicting item difficulty is an F-test with $F =$

$$\frac{\text{explained variance}}{\text{unexplained variance}}.$$

"The fact that some or all predictor variables are correlated among themselves does not, in general, inhibit our ability to obtain a good fit nor does it tend to affect inferences about mean responses or predictions of new

observations, provided these inferences are made within the region of observations.” – Kutner et al. 2004

It is worth noting that OLS makes no requirement that the coefficients need to be positive on the explanatory variables to get a good prediction of the dependent variable. However, it would be very strange if there was a negative correlation between LLM performance and item difficulty.

3.3.3 Study Three: Binary Search Estimation of Absolute Item Difficulty

In this dissertation, I aim to explore the use of a series of binary relative item difficulty classifiers as an alternative method for estimating "absolute" item difficulty. I employ two different methods to estimate absolute item difficulty. The first method relies solely on the pairwise responses from Study One.

Method One: Joint estimation of Item Parameters

Coding the pairwise responses as either 1, -1, or 0 we can input them as explanatory variables in a logistic regression with the outcome variable defined as 0 or 1 depending upon if item i is ranked as more difficult than item j . Using the following procedure, we can generate a matrix of explanatory variables whose coefficients are estimates of item difficulty. The following is the value of item i in column i and row associated with the item pair.

$$(3.3.3) X_i = \begin{cases} \text{if } i \text{ is first item then} \rightarrow & 1 \\ \text{if } i \text{ is the second item then} \rightarrow & -1 \\ \text{if } i \text{ is not in the pairwise comparison} \rightarrow & 0 \end{cases}$$

Here is a clarifying example: imagine there are four item pairs evaluated, 1 with 2, 2 with 1, 1 with 3, and 3 with 1. Using the above equation, we can convert those values to the following matrix:

$$X = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & -1 \\ -1 & 0 & 1 \end{bmatrix}$$

Now we just need to convert our pairwise comparisons returned from the LLM into outcome variables. I do this by coding the dependent variable as 1 if the first item is ranked as more difficult than the second item and zero otherwise. Let's imagine the above pairings that in terms of difficulty ranking the LLM generated the following results $1 > 2$, $2 < 1$, $1 > 4$, $4 > 1$. This would be coded as the following Y dependent variable column:

$$Y = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Now that we have our explanatory variables defined in this way and our dependent variables, this model is straightforward to estimate using a logistic regression model.

$$Y = \text{logistic}(\hat{\beta}X)$$

The values of the estimated coefficient $\hat{\beta}_i$ is the estimate of the item parameter for item i . Overall, this approach is parsimonious and aligns well with a 1 parameter logistic (1LP) item response theory model estimation procedures. To compare the estimates $\hat{\beta}_i$ with the empirical difficulties I convert the empirical difficulties to be on the same scale as the 1PL model using the sigmoid function. Since both distributions are centered at zero, directly comparing estimates does not present a challenge, though in practice a new batch of items would need be grouped with anchor items which would be used to appropriately scale and position the novel items. While this method makes sense in pretesting operations the next method demonstrates

how pre-calibrated items can be more directly leveraged to aid in the empirical estimation of novel item difficulties.

Method Two: Individual Estimation of Item Parameters

This method draws inspiration from Attali et al. (2014), who suggest using a series of pre-calibrated items to help estimate item difficulties. Their approach approximates a computer adaptive testing procedure, by matching either incrementally more difficult or less difficult items to rapidly converge on the unknown item difficulty. The primary reason for this expediency is the prohibitive cost of subject matter experts' time. However, since LLMs (Large Language Models) tend to generate responses affordably, rather than relying on the smaller set of item pairs needed for Attali et al.'s procedure, I suggest using all known item pairs (within subject and grade) as effective item anchors for each unknown item i .

This analysis is simulation of how this procedure might be used to calibrate a novel item or set of items in the context of an already existing calibrating item pool. Rather than using this method to calibrate a set of unknown items, I am using it to sequentially estimate an item's difficulty, assuming all other items have known properties. While this procedure might feel contrived, it represents the information available to many testing programs. It would not be unusual for a testing program to have dozens or hundreds of pre-calibrated items to which any newly generated item can be compared.

To estimate the difficulty of an item with a set of items with known properties, it is a bit trickier than Method One. To estimate item i 's parameters we need to input the pairwise comparisons for all pairwise item paired with i into the estimation. The following example will demonstrate. Using the same formulation for the \mathbf{X} matrix above we will split it into column

containing the i indicators (X_i) and the matrix not including the i indicators ($X_{\sim i}$). Taking item 1 as the item with unknown parameters and using the same example from method one, we get:

$$X_i = \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} \text{ and } X_{\sim i} = \begin{bmatrix} -1 & 0 \\ 1 & 0 \\ 0 & -1 \\ 0 & 1 \end{bmatrix}$$

Using the same subscript notation for item i parameter estimates ($\hat{\beta}_i$) as well as the known parameters for all other items ($\beta_{\sim i}$) we can specify the logistic regression in the following way:

$$Y = \text{logistic}(\hat{\beta}_i X_i + \beta_{\sim i} X_{\sim i})$$

In this model the only parameters being estimated are $\hat{\beta}_i$ not $\beta_{\sim i}$. This estimation model is more difficult to implement than the model from Method One as most statistical packages has a logistic function built in but do not allow for the direct specification of known coefficients. However, by coding the logistic function, directly inputting known parameters, and specifying the unknown parameter, this equation can be executed.

To test the overall performance of this estimation procedure, I sequentially estimate the item parameter for each item i individually then compare the joint estimation with the sigmoid transformation of the empirical difficulties in the same manner as Method One.

3.4 Item Data

The data used in this study are taken from the National Assessment of Educational Progress (NAEP). NAEP is the largest standardized assessment of student knowledge within primary education in the United States and is administered every two years to grades 4, 8, and 12. NAEP was first administered in 1969. It is given to nationally representative samples of

students typically between the sizes of 15,000 and 26,000 students during each administration (Campbell et al., 2000; Rampey et al., 2009).

From the NAEP exam this dissertation will focus on the released item subjects of Mathematics and Science of which there are a total of 1,633 released NAEP items: 1,012 of which there are multiple-choice items. The National Assessment of Educational Items releases items in nine subjects, seven on which are reported below (Table 3). The two subjects, “Art” and “Technology and Engineering Literacy,” were excluded from consideration due to having very few items. In this dissertation I focus on evaluating the items in Mathematics and Science. Math items are those with greatest populace, while the science items cover a diverse range of subjects in natural science. Reading items were excluded from this review as they have already been thoroughly examined in the literature with observed findings that passage features are highly predictive of item difficulty. Both US History and Geography items were excluded as a sizable portion of the items required the use of visual components such as maps or diagrams. While multimodal LLMs (models which take and evaluate visual components) might be suitable to evaluate these items in the future they are not the focus of this study. Civics and Economics items were excluded due to the small number of items released.

The following table lists the full number of items available to the public. Originally, I intended to use a much larger sampling of items. However, when reviewing the items, I found that many of them had either high dependence on visual components or presented transcription error requiring a need for manual review as with the items collected from the nationsreportcard.gov (National Center for Education Statistics, 2024).

TABLE 3: NATIONAL ASSESSMENT OF EDUCATION PROGRESS ITEMS

| | Civics | Economics | Mathematics | Reading | US History | Geography | Science |
|-------------------------------|--------|-----------|-------------|---------|------------|-----------|---------|
| Total Count | 261 | 112 | 1238 | 720 | 424 | 322 | 395 |
| Easy | 80 | 31 | 472 | 345 | 113 | 121 | 94 |
| Medium | 111 | 56 | 376 | 246 | 163 | 112 | 120 |
| Hard | 70 | 25 | 389 | 128 | 148 | 88 | 177 |
| Grade 4 | 90 | 0 | 448 | 245 | 142 | 104 | 120 |
| Grade 8 | 95 | 0 | 499 | 268 | 146 | 136 | 142 |
| Grade 12 | 76 | 112 | 223 | 154 | 136 | 81 | 132 |
| Content Classifications | 12 | 6 | 17 | 17 | 16 | 6 | 10 |
| Multiple Choice | 197 | 93 | 800* | 394 | 280 | 213 | 212* |
| Short Constructed Response | 51 | 7 | 383 | 246 | 107 | 75 | 141 |
| Extended Constructed Response | 13 | 12 | 54 | 79 | 37 | 33 | 38 |

* Items considered for deployment in this study.

In this dissertation, I examine multiple choice items only, as they are the most common format of items. They also allow for the exploration of item difficulty under the specifications of study two, which the other formats do not allow for given a separate procedure would be required to evaluate whether a response is correct or sufficient (as is the case in constructed response item). All items are initially filtered by using an automated system (involving word match) to remove items which are anticipated to be a poor match for an LLM due to certain features such as being excessively visually dependent (e.g. reading a map, interpreting a graph, explaining a figure, etc.) or involves contemporary facts at the time of item administration (such as the name of the current president), these items were removed. Between the 212 Science and 800 Math items only 462 items survived the initial filter. Most removed items involved

some visual component in the form of an image in either the question body, selection choice, or both. All items have visual components described using Section-508 compliant alternative text.

All 462 remaining items surviving the automatic filtering process were then manually evaluated for transcription error in the formulation of the alternative text or for excessive reliance of visual components. Surprisingly, 10.2% of the items had some kind of substantive error in transcribing the 508-alternative text which would have made the items either more difficult or impossible to answer correctly. Another 4.5% of items were flagged as having an excessively high dependence on visual components. Nathan Bos, a professor at John Hopkins and employee at Mitre, contributed to this analysis by reviewing NAEP items for errors and visual dependency. The interrater reliability for error flagging was 88% while the interrater reliability for visual dependency exclusion was 85%. I chose to exclude items in which either rater was flagged for excessive visual dependency, or which contained an alternative text transcription error. After manual review 388 items remained for evaluation with 300 of them in mathematics and 88 of them in science.

These 388 remaining items were taken from a total of NAEP 36 subject/grade tests administered between 1990 and 2019 (National Center for Education Statistics, 2024). Each of the tests was assigned to either training or testing data. Training data was used to evaluate different prompt options and select the prompt which performed the best, (see Appendix B, C, and D). Within each test (year, subject, and grade combination) items were paired together to create binary relative difficulty item combinations. Each item pair was evaluated twice with each of two items taking turns in the first position and second position (e.g. “is A more difficult

than B”, and “is B more difficult than A”). Table 4 shows the summary counts and combination counts across all tests and items.

TABLE 4: TEST ASSIGNMENT INTO TESTING AND TRAINING SUMMARY TABLE

| Type | Subject | Count | Binary Combinations |
|----------|-------------|-------|---------------------|
| Testing | Mathematics | 198 | 1609 |
| Testing | Science | 60 | 245 |
| Training | Mathematics | 102 | 594 |
| Training | Science | 28 | 104 |

Tests were assigned either to testing or training such that, when possible, most items were placed in the testing pool. Also, tests were distributed evenly over the time periods, grades, and subjects administered. Table 5 shows the individual test assignment to type groups. The number of binary combinations is calculated by using the standard formula $F(n, k) = \frac{n!}{k! \times (n - k)!}$ with n being number of items in the test and k being 2 as items are matched in pairs.

TABLE 5: TEST ASSIGNMENT INTO TESTING AND TRAINING

| Type | Subject | Grade | Year | Count | Binary Combinations |
|----------|-------------|-------|------|-------|---------------------|
| Training | Mathematics | 4 | 1990 | 14 | 91 |
| Testing | Mathematics | 4 | 1992 | 22 | 231 |
| Testing | Mathematics | 4 | 1996 | 7 | 21 |
| Testing | Mathematics | 4 | 2003 | 28 | 378 |
| Training | Mathematics | 4 | 2005 | 12 | 66 |
| Training | Mathematics | 4 | 2007 | 17 | 136 |
| Training | Mathematics | 4 | 2009 | 9 | 36 |
| Testing | Mathematics | 4 | 2011 | 16 | 120 |
| Testing | Mathematics | 4 | 2013 | 12 | 66 |
| Testing | Mathematics | 8 | 1990 | 13 | 78 |
| Training | Mathematics | 8 | 1992 | 5 | 10 |
| Training | Mathematics | 8 | 1996 | 4 | 6 |
| Testing | Mathematics | 8 | 2003 | 16 | 120 |
| Training | Mathematics | 8 | 2005 | 16 | 120 |
| Testing | Mathematics | 8 | 2007 | 19 | 171 |
| Training | Mathematics | 8 | 2009 | 3 | 3 |
| Testing | Mathematics | 8 | 2011 | 15 | 105 |
| Testing | Mathematics | 8 | 2013 | 11 | 55 |
| Training | Mathematics | 12 | 1990 | 15 | 105 |
| Testing | Mathematics | 12 | 1992 | 19 | 171 |
| Testing | Mathematics | 12 | 1996 | 6 | 15 |
| Testing | Mathematics | 12 | 2005 | 13 | 78 |
| Training | Mathematics | 12 | 2009 | 7 | 21 |
| Testing | Mathematics | 12 | 2013 | 1 | 0 |
| Training | Science | 4 | 2000 | 9 | 36 |
| Testing | Science | 4 | 2005 | 10 | 45 |
| Testing | Science | 4 | 2009 | 6 | 15 |
| Testing | Science | 8 | 2000 | 9 | 36 |
| Training | Science | 8 | 2005 | 5 | 10 |
| Training | Science | 8 | 2009 | 3 | 3 |
| Testing | Science | 8 | 2011 | 3 | 3 |
| Testing | Science | 8 | 2019 | 5 | 10 |
| Testing | Science | 12 | 2000 | 7 | 21 |
| Testing | Science | 12 | 2005 | 15 | 105 |
| Training | Science | 12 | 2009 | 11 | 55 |
| Testing | Science | 12 | 2019 | 5 | 10 |

3.5 General Item Properties

Item Properties Directly Taken / Calculated from NAEP Parameters

This section explores some of the item properties used as either the explanatory or dependent variables in this dissertation. On average the item difficulty of for all items is 46.2 with items in lower grades being slightly less difficult and items in higher grades being slightly more difficult (6). The most difficult set of items by grade and subject are 12th grade science items.

Table 6 shows some summary statistics on item difficulty, with lower difficulty indicating that more students got this item correct and higher difficulty indicating that fewer students got the item incorrect. The data is segmented by either: grade and subject, or item assignment grouping type (testing or training data), or ungrouped level. The columns SD, P10, and P90 reference standard deviation, percentile 10, and percentile 90 respectively.

TABLE 6: ITEM DIFFICULTY PARAMETERS

| Grade | Subject | Type | Min | Mean | Max | SD | P 10 | Median | P 90 | Count |
|-------|-------------|-----------|-----|-------|-----|-------|------|--------|------|-------|
| 4 | Mathematics | | 5 | 44.76 | 92 | 19.16 | 17.6 | 46 | 69.0 | 137 |
| 4 | Science | | 13 | 39.28 | 81 | 18.64 | 20.6 | 33 | 68.8 | 25 |
| 8 | Mathematics | | 6 | 46.49 | 84 | 17.98 | 21.1 | 50 | 66.9 | 102 |
| 8 | Science | | 13 | 45.60 | 73 | 13.69 | 28.0 | 47 | 64.2 | 25 |
| 12 | Mathematics | | 7 | 46.79 | 88 | 20.45 | 23.0 | 45 | 74.0 | 61 |
| 12 | Science | | 30 | 54.87 | 85 | 15.40 | 32.7 | 56 | 78.5 | 38 |
| | | Testing | 5 | 46.94 | 92 | 18.86 | 20.7 | 48 | 72.0 | 258 |
| | | Training | 9 | 44.81 | 88 | 18.25 | 21.0 | 43 | 67.3 | 130 |
| | | Ungrouped | 5 | 46.22 | 92 | 18.69 | 21 | 47 | 71 | 388 |

The \hat{a} parameter is the proxy estimate of the item response theory parameter a and is calculated using equation (3.3.1.2) by taking the overall population test score by item choice

(NAEP, 2024) between the difference between the population mean test score for those who answered the item and chose the correct response minus the weighted mean test score for those who answered the item and chose the incorrect response.

Table 7 shows average difference in population scores between those who got the item correct and those who got the item incorrect (\hat{a}).

TABLE 7: \hat{a} PARAMETER ESTIMATES

| Grade | Subject | Type | \hat{a} | | |
|-------|---------|----------|-----------|-------|-------|
| | | | min | mean | max |
| 4 | Math | | 0.6 | 24.99 | 80.1 |
| 4 | Science | | 11.6 | 25.65 | 42.4 |
| 8 | Math | | 4.9 | 39.18 | 292 |
| 8 | Science | | 12.3 | 21.87 | 44.2 |
| 12 | Math | | 11.5 | 49.91 | 304 |
| 12 | Science | | 5.4 | 22.26 | 37.1 |
| | | Testing | 4.6 | 32.70 | 304 |
| | | Training | 0.6 | 31.24 | 159.5 |
| | | | 0.6 | 32.21 | 304.0 |

Item Properties Estimating Using LLMs

Complexity and number of steps were generated from prompting the LLM Gemini-Pro to provide an estimate of the complexity of the item for a student four grades below the item level. The purpose of instructing the LLM to assess the item from the perspective of a student four grades lower than the administered level was due to LLMs generally overestimating the ability of students.

LLMs overestimating the ability of students at a particular grade level is a commonly observed bias in interacting with these models. It is unknown why this bias exists, but it is almost certainly based on the training corpus. Interestingly it is a bias that is shared by subject

matter experts who tend to overestimate the ability of students at particular grade levels (Lorge and Kruglov, 1952 and 1953).

Having the LLM assess the item from the perspective of students four grades lower (with first being the lowest grade), the average complexity was 6.77 with items assigned to higher grades being seen as slightly more complex than those assigned to lower grades (Table 8).

The number of steps were generated by prompting Gemini-Pro to list the number of steps required to solve the item then counting those steps. More work could have been done in cleaning the number of steps listed as often the LLM was often prone to generate non-meaningful steps for low complexity items. When items required very few steps to find a solution the LLM often generated standard steps that were generally required for all items (for example: “look at the question,” “read the potential answers”, “think about how to solve the question”, etc.).

This is the reason there is not more variation between the number of steps required to solve the lower grade items and those required to solve the higher-grade items. This filling in of steps with generic content might also be a contributing factor as to why science items had more steps listed than those of mathematics items for which procedural steps are common (for example: “think about the water table”, “conceptualize the relationship between predators and prey”, etc.).

Assignment to different item types either training or testing was based on Table 5 and produced distributions of relative item difficulties which were similar (Figure 2.). The goal of assignment of items to different types was to favor the testing group with more items to aid in

final inference and results when possible while still having sufficient items as well as subject, year, and grade coverage in the training data in order to make inference on which item prompts functioned optimally.

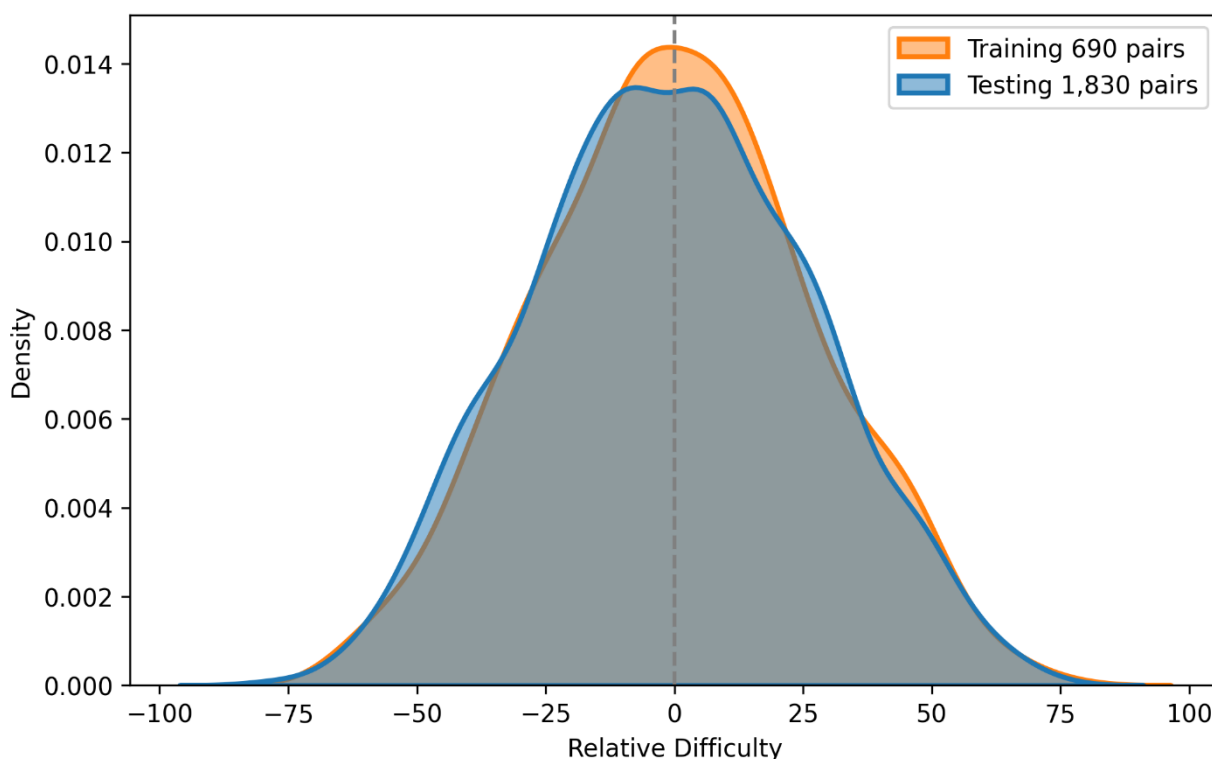
Table 8 shows two item properties complexity and number of steps generated independently for each item from Gemini-Pro.

TABLE 8: ITEM COMPLEXITY, NUMBER OF STEPS, AND ALPHA PARAMETERS

| Grade | Subject | Type | Complexity | | | Number of Steps | | |
|-------|---------|----------|------------|------|-----|-----------------|------|-----|
| | | | min | mean | max | min | mean | max |
| 4 | Math | | 1 | 5.66 | 10 | 2 | 3.78 | 9 |
| 4 | Science | | 1 | 7.32 | 8 | 1 | 4.04 | 6 |
| 8 | Math | | 1 | 7.13 | 10 | 2 | 4.01 | 9 |
| 8 | Science | | 5 | 7.64 | 8 | 2 | 4.20 | 8 |
| 12 | Math | | 3 | 7.41 | 9 | 1 | 3.82 | 9 |
| 12 | Science | | 6 | 7.89 | 9 | 1 | 4.13 | 8 |
| | | Testing | 1 | 6.83 | 10 | 1 | 3.99 | 9 |
| | | Training | 1 | 6.66 | 10 | 2 | 3.79 | 9 |
| | | | 1 | 6.77 | 10 | 1.00 | 3.93 | 9 |

FIGURE 2: ASSIGNMENT PAIRWISE RELATIVE DIFFICULTIES BY ASSIGNMENT TYPE

This figure shows that the general distribution of relative item difficulties by assignment type has similar distributions.



3.6 Item Properties by Demographic Groups

A key consideration when developing any large-scale assessment is to consider differences in response patterns by distinct groups. Generally, speaking, some differences in response patterns can be attributable to “true” variation in underlying population differences while some differences in response patterns might be attributable to unintended features of the item which makes it easier or more difficult for certain groups. Using the high-level statistics provided by NAEP it is unlikely that we can identify which of these factors is driving differences in response patterns that we observe in the data.

Figure 3 demonstrates that there is a slight difference in response patterns with females finding items more difficult than males. While Figure 4 shows that in general American students who identify as demographically Asian/Pacific Islander find the items the least difficult followed by White, Hispanic, and Black.

FIGURE 3: DISTRIBUTION OF ITEM DIFFICULTIES BY GENDER

This figure shows the density of item difficulties stacked by the item difficulty faced by different genders. Curves that load densities further to the left indicate that this population group finds the items easier while those to the right find the items more difficult.

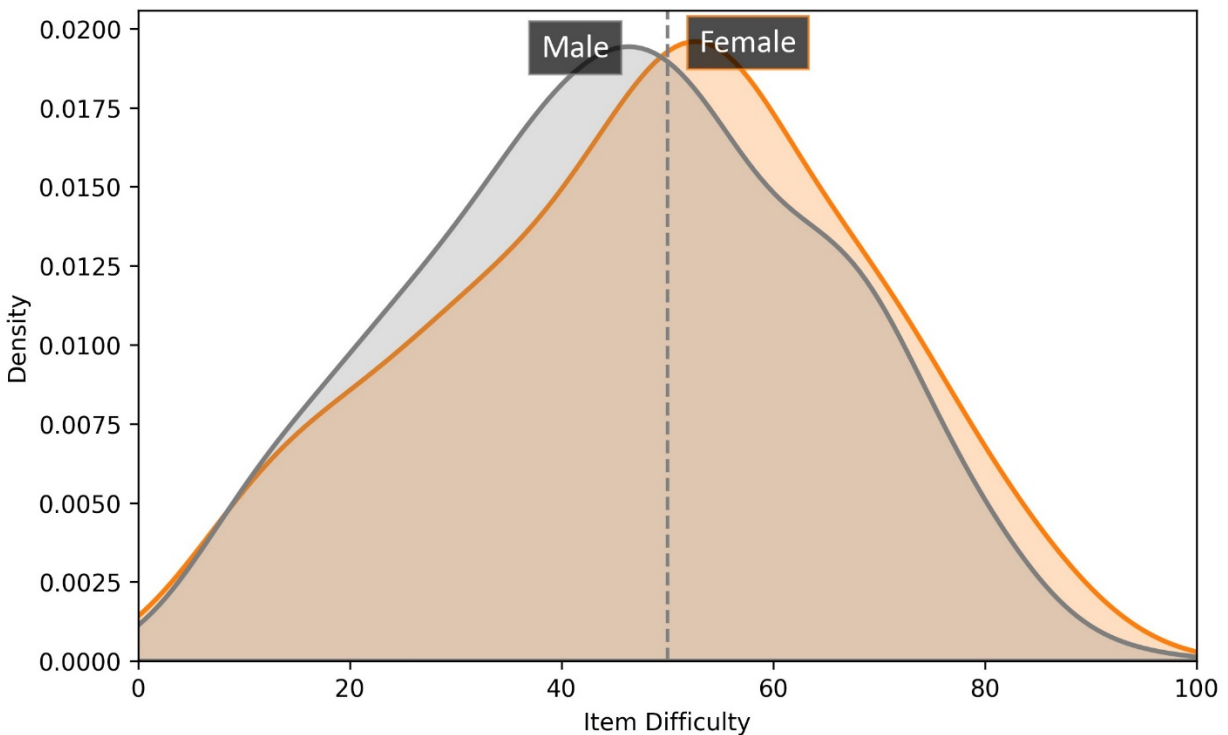
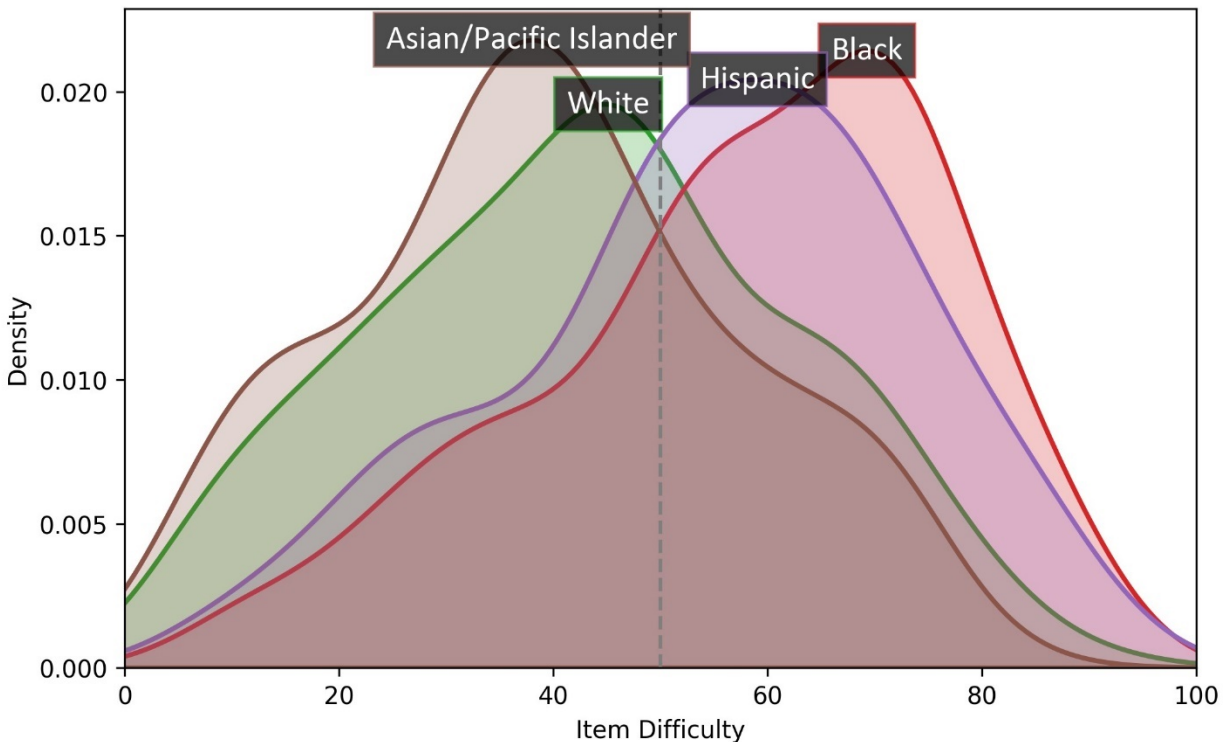


FIGURE 4: DISTRIBUTION OF ITEM DIFFICULTIES BY RACE

This figure shows the density of item difficulties stacked by the item difficulty faced by that population group. Curves that load densities further to the left indicate that this population group finds the items easier while those to the right find the items more difficult.



3.7 Prompt Template Selection

Following the process outlined in Appendix B, C, and D binary choice prompt templates were randomly generated across the ten different options. In total 40 prompts were randomly generated. A variety of LLMs were explored with these models including GPT3.5, GPT4, Google's Gemini Pro, as well as several smaller open-source models (Llama 7B, Llama 13B, and Llama 70B). Overall, the difference in performance of the ranking task between GPT 3.5, GPT4, and Gemini-Pro was slight. However, Gemini-Pro offered a cost advantage (at this time Gemini-

Pro is zero cost for development). As a result, Gemini-Pro was used as the primary model to optimize template design and evaluate LLM pairwise classification performance.

To aid in reducing training noise, for the first generation the training data pairwise groups were split into two equal sized buckets: high contrast pairs and low contrast pairs¹. High contrast pairs included item pairs in which the difference in difficulty was greater than or equal to 19 percentage points while those in low contrast pairs were item pairs in which the difficulty in which the difference in difficulty was less than 19 percentage points. Items which had the same difficulty (NAEP rounds difficulty to whole percentage points) were excluded from the training data since the accuracy of their responses could not be determined. The LLM responses to pairwise comparisons are divided into three categories: one correctly ranked, 0 incorrectly ranked, and null for uninterpretable responses in which the LLM returned a response that could not be evaluated. These responses varied but often amounted to a statement like, “the two items have the same difficulty” or “the difference in difficulty of the items could not be determined”.

The overall performance of the first forty prompt templates using 60 random item pairs was 59% accurate averaged across all models when ignoring uninterpretable responses and only 46% accurate when treating uninterpretable responses as misses. However, the top ten performing responses did much better with an average correct response of 68% and 64% even when considering uninterpretable responses.

¹ Note that this is similar in some ways to the procedure used by Attali et al. (2014) in which they oversampled the highest and lowest performing items which would likely lead to larger disparities in the pairwise difference in difficulties of items.

The top ten performing templates were then randomly “bred” with each other creating a total of roughly 30 new children, some of which were dropped if duplicated. These children in addition to their parents were tested against another 30 randomly selected high contrast item pairs in addition to the 30 underwent by the parents. The average performance on these items pairs was 63% excluding nulls and 47% when nulls are included. Nulls occur when the LLM does not produce an interpretable binary response, often in the form of statements like "I cannot determine" or "both items are equally difficult." The top ten from this generation had a 72% correct prediction result excluding nulls and 71% including nulls.

New generations were added, in a similar manner, and additional item pairs continued to be added while maintaining the evaluation of existing prompt templates for a total of four generations and 137 prompt templates. Prompt template performance converged quickly with the top two performing templates on all the training data being from the second generation (template 61 and 60). Overall, the genetic algorithm led to selection of higher performing templates. All but one template in the top twenty were generated from cross breeding through the genetic algorithm. If the genetic algorithm had acted as simple random prompt generator, then we would expect that roughly 29% of the prompts in the top 20 would be from the first generation rather than the 5% we observe.

For the top twenty performing models all binary item pairs were assigned for evaluation creating a total of 698 item pairs which were evaluated. These item pairs included items in which the absolute difference in difficulty was less than 19 percentage points. The average correct ranking of item difficulty for the top performing model (template 61) across all the training data was 63%. Taking the top model template and evaluating it on the 1854 item pairs

in the testing data produces a similar performance of just over 62%. The slight drop in model performance might be due to random variation or slight “overfitting” due to evaluating so many models in the training data. For studies 1 and 3 template sixty-one is used exclusively.

Table 9 shows the parameters used in template 61, while Table 10 shows an example prompt, followed by Table 11 which shows an example response.

TABLE 9: TEMPLATE 61 PARAMETERS

The following table shows the parameters used in top performing template (Template 61). See Appendix B for a list of parameter variants.

| Variable | # | Value |
|----------------------------|---|--|
| Question Naming | 4 | Question a, Question b |
| Task Title | 0 | <i>None</i> |
| Persona | 0 | <i>None</i> |
| Task Introduction | 2 | You will find two different questions ahead. Evaluate their difficulty based on depth and breadth of knowledge required, and cognitive demand. |
| Context | 4 | Item Content Tags: {Content Tags} |
| N-shot | 0 | <i>None</i> |
| Perspective | 3 | Evaluate the content from the perspective of a {max(grade – 4, 1)}} th/st grade student. |
| Task Approach Instructions | 4 | Detail the sequence of actions needed to solve this relative difficulty ranking challenge. |
| Item Shared Context | 0 | <i>None</i> |
| Task Instructions | 4 | Here are two questions for your consideration. Please rate them in order of difficulty. |
| Task Output | 0 | The more difficult question is: {{Question 1}}/{{Question 2}}. |

To get a sense of how this template works we can see an example in Table 9 in which two items from Grade 8 in 1990 are compared.

TABLE 10: TEMPLATE 61 MODEL PROMPT EXAMPLE

The following box shows an example of a pairwise relative difficulty task framed in the perspective of template design sixty-one followed by some item statistics.

| |
|--|
| <p>You will find two different questions ahead. Evaluate their difficulty based on depth and breadth of knowledge required, and cognitive demand.</p> <p>Evaluate the content from the perspective of a fourth grade student.</p> <p>Detail the sequence of actions needed to solve this relative difficulty ranking challenge.</p> <p>Question a:</p> <p>Item Content Tags: Probability, Ratio, Fraction, Random Selection, Mathematics</p> <p>There are 15 girls and 11 boys in a mathematics class. If a student is selected at random to run an errand, what is the probability that a boy will be selected?</p> <p>A. $\frac{4}{26}$</p> <p>B. $\frac{11}{26}$</p> <p>C. $\frac{15}{26}$</p> <p>D. $\frac{11}{15}$</p> <p>Question b:</p> <p>Item Content Tags: Area, Perimeter, Rectangle, Length, Width</p> <p>Four rectangles; first is 1 by 1, second is 1 by 3, third is 1 by 2, and the fourth is 1 by 3.</p> <p>A certain rectangle has an area equal to the sum of the areas of the four rectangles shown above. If its length is 4, what is its width?</p> <p>A. 2</p> <p>B. $2\frac{1}{4}$</p> <p>C. $2\frac{1}{2}$</p> <p>D. 3</p> <p>Here are two questions for your consideration. Please rate them in order of difficulty</p> <p>Please provide your evaluation in the following format:</p> <p>The more difficult question is: Question a/Question b</p> |
| <p>Item Year: 1990</p> <p>Item Grade: 8th</p> <p>Item 1 ID: 1990-8M7 #18 M017001</p> <p>Difficulty: 62.0</p> <p>Item 2 ID: 1990-8M9 #19 M028831</p> <p>Difficulty: 83.0</p> |

TABLE 11: TEMPLATE 61 EXAMPLE RESPONSE

| |
|---|
| <p>**Question a:**</p> <p>* Depth and breadth of knowledge required:** Requires basic understanding of probability and ratios, including concepts of random selection and proportions.</p> <p>* Cognitive demand:** Level 2 (Basic Application of Skills and Concepts)</p> <p>**Question b:**</p> <p>* Depth and breadth of knowledge required:** Requires knowledge of geometry and area concepts, including formulas for rectangles (length x width) and understanding of perimeter and area.</p> <p>* Cognitive demand:** Level 3 (Strategic Thinking and Complex Reasoning)</p> <p>**From the perspective of a 4th/5th grade student:**</p> <p>Both questions would be considered challenging, but Question b would be more difficult due to the need for more complex understanding of geometry concepts.</p> <p>**Sequence of actions for difficulty ranking challenge:**</p> <ol style="list-style-type: none"> 1. Read and analyze each question carefully. 2. Identify the specific concepts and skills required for each question. 3. Determine the depth and breadth of knowledge required. 4. Evaluate the cognitive demand of each question. 5. Rank the questions in order of difficulty based on the above criteria. <p>**The more difficult question is: Question b**</p> |
|---|

We might wonder how unique the parameters are chosen using template 61 compared with the other top performing templates. Looking at Table 12 we can see that model sixty-one happens to share the same parameters as the mode of the ten top performing models of all models evaluated. All the parameters varied between the top ten performing models except for the N-shot (See Appendix B, Table 34) examples which was universally excluded. Variants of persona, task title, item shared context, and task output were also generally excluded in the top performing models (See Appendix B: Table 30, 31, and 40).

Exclusion of example tasks from the item prompt (N-shot) and exclusion of item shared context is particularly interesting as it diverges from the standard recommendations in the

literature which often emphasize that the more context for a task provided in the prompt either through examples or additional context the better.

TABLE 12: THE PARAMETERS OF THE TOP TEN PERFORMING MODELS

Models are listed from best performing (left) to least (right) for pairwise comparison of items using the LLM Gemini-Pro.

| Model | 61 | 60 | 118 | 59 | 106 | 105 | 51 | 126 | 49 | 10 | Mode |
|----------------------------|-----------|-----------|------------|-----------|------------|------------|-----------|------------|-----------|-----------|--------------|
| Question Numbering | 4 | 0 | 0 | 1 | 3 | 4 | 3 | 2 | 1 | 4 | [4] |
| Task Title | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | [0] * |
| Persona | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 3 | 4 | 3 | [0] * |
| Task Introduction | 2 | 2 | 3 | 3 | 1 | 1 | 1 | 2 | 4 | 2 | [2] |
| Context | 4 | 4 | 4 | 4 | 4 | 0 | 4 | 0 | 2 | 0 | [4] |
| N-shot | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | [0] * |
| Perspective | 3 | 3 | 3 | 0 | 3 | 3 | 3 | 3 | 2 | 3 | [3] |
| Task Approach Instructions | 4 | 4 | 3 | 3 | 0 | 0 | 0 | 4 | 0 | 4 | [0,4] |
| Item Shared Context | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | [0] * |
| Task Instructions | 4 | 4 | 3 | 3 | 4 | 4 | 0 | 2 | 3 | 4 | [4] |
| Task Output | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | [0] * |
| Temperature | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 3 | 3 | 4 | [0] * |
| Generation | 2 | 2 | 4 | 2 | 4 | 4 | 2 | 4 | 2 | 1 | [1] |

* Indicates this prompt feature was rejected from final prompt selection.

It should be noted that the top ten prompts shown in Table 12 seem to be the best performing prompts using the LLM Gemini-Pro. Other LLMs or even future releases of Gemini-Pro will have different parameter weights in which a slightly or completely different top prompt template will turn out to be more effective. As such, the contribution of this research is less about the specific prompts which turned out to be more effective and more about the approach of using optimization tools (such as a genetic algorithm) to improve the performance of prompts. There is no reason, based on reviewing the literature on prompt design, that this set of prompts (rejecting context and persona features) would have been expected to be

successful. However, by implementing an optimizing search algorithm over different template features the performance of the LLM classifier was able to be significantly improved.

3.8 LLM Pairwise Performance Under Independence

There is no reason to think that errors associated with the responses from Large Language Models (LLMs) are independent and identically distributed (IID). However, if they were then there is an appropriate sample size of pairwise comparisons which would lead to a correct estimation of pairwise item difficulty at any given confidence value. Table 13 shows under IID how many pairwise comparisons would be needed in the case of 90% confidence. Thus, under IID even a weak signal such as that found when the relative difficulty in difference is between 1-9 could create a strong signal.

TABLE 13: PAIRWISE COMPARISON AND IDEAL SAMPLING

This table shows how many random samples for each item would be needed to be 90% confident that the average difficulty is correctly classified.

| Relative Difference | Correct* | Pairwise Counts in Sample | Sample needed for 90% Conf. |
|---------------------|----------|---------------------------|-----------------------------|
| 1-9 | 53.6% | 526 | 311 |
| 10-19 | 54.2% | 904 | 223 |
| 20-29 | 61.6% | 880 | 31 |
| 30-39 | 63.0% | 541 | 23 |
| 40-49 | 72.5% | 454 | 7 |
| 50-59 | 81.8% | 214 | 3 |
| 60-69 | 83.9% | 112 | 3 |
| 70-79 | 88.9% | 18 | 3 |
| 80+ | 100.0% | 4 | 1 |

* The correct % is based on the average percent correct for each set of pairwise comparisons.

However, the predictions of LLMs are not independently sampled but instead suffer from correlation in error due to shared predictive bias. Looking at just the item pairs when the

difficulty difference in items is less than 1%, in which theoretically the prediction should be close to 50% accurate (random chance). I find that the two-way t-test with a 90% threshold ends up rejecting the null 88% of the time instead of the expected outcome under IID of 10%.

This is a limitation of using LLMs as expert classifiers. A benefit of LLMs is that they have the advantage of not remembering their previous response to the same task. Unfortunately, the informational weights which led to their response does not change, though there is stochasticity in the model output. Thus, responses and the error in those responses are serially correlated. LLMs have an advantage over human experts who may remember their previous response and for repeated requests for classification likely will be biased by both having consistent internal weights (or personal experiences) as well as recalling their previous response.

Looking at the pairwise comparisons by grade and subject (Table 14) we see that the average percent correct of guesses hovers between 58-62% correct with pairwise comparisons across grades 4, 8, and 12 in both Math and Science with the single substantive deviation for Math grade 12 in which the pairwise comparison is instead averaging at 73% correct.

TABLE 14: PAIRWISE COMPARISON AND IDEAL SAMPLING

| Grade | Subject | Percent Correct | Pairwise Comparisons |
|-------|---------|-----------------|----------------------|
| 4 | Math | 61.6% | 1621 |
| 4 | Science | 59.5% | 116 |
| 8 | Math | 59.4% | 1040 |
| 8 | Science | 60.9% | 92 |
| 12 | Math | 73.0% | 514 |
| 12 | Science | 58.1% | 270 |

3.9 Study One: Item Parameters

Parameters used to estimate the models in Study One are taken either from NAEP, binary predictions using the LLM prompt model, or independently generated using a different prompt (Table 15). The “Correct” variable is the likelihood that the LLM binary prediction model successfully predicted the relative item difficulty of two items. On average for the validation data the model is 62% correct. The minimum for this variable is 50% (under random guessing) while the maximum is 100% under perfect knowledge.

TABLE 15: PAIRWISE COMPARISON ITEM PARAMETERS

This table shows statistics on the various item parameters evaluated in Study One. The $|\text{var}|$ refers to the absolute value function while the Δ refers to the difference between the two pairwise values.

| Variable | Source | Mean | SD | Min | Max |
|--|-----------------------------------|-------|-------|-------|-------|
| Correct | Binary Prediction | 0.62 | 0.48 | 0 | 1 |
| \hat{a}_1 | NAEP | 0.05 | 1.15 | -2.2 | 6.49 |
| \hat{a}_2 | NAEP | -0.08 | 0.94 | -2.2 | 6.49 |
| Second Item Harder | NAEP | 0.51 | 0.5 | 0 | 1 |
| Diff ₁ | NAEP | 46.28 | 19.86 | 5 | 92 |
| Diff ₂ | NAEP | 47.02 | 18.72 | 5 | 92 |
| Complexity 1 | Independent LLM Prediction | 6.41 | 2.2 | 1 | 10 |
| Complexity 2 | Independent LLM Prediction | 6.71 | 1.92 | 1 | 10 |
| $ \text{Complexity } \Delta $ | Independent LLM Prediction | 1.83 | 2 | 0 | 9 |
| # of steps 1 | Independent LLM Prediction | 3.86 | 1.35 | 1 | 9 |
| # of steps 2 | Independent LLM Prediction | 3.98 | 1.44 | 1 | 9 |
| $ \text{\# of steps } \Delta $ | Independent LLM Prediction | 1.5 | 1.25 | 0 | 8 |
| $ \text{Diff } \Delta $ | NAEP | 22.12 | 15.67 | 1 | 78 |
| $ \text{Diff } \Delta : \hat{a}_1$ | NAEP | 0.35 | 4.66 | -14.2 | 47.37 |
| $ \text{Diff } \Delta : \hat{a}_2$ | NAEP | 0 | 3.79 | -14.3 | 50.61 |
| $ \text{Diff } \Delta : \text{Complexity 1}$ | NAEP & Independent LLM Prediction | 13.95 | 11.43 | 0.1 | 64.8 |
| $ \text{Diff } \Delta : \text{Complexity 2}$ | NAEP & Independent LLM Prediction | 14.65 | 11.52 | 0.2 | 69.3 |
| $ \text{Diff } \Delta : \text{Complexity } \Delta $ | NAEP & Independent LLM Prediction | 4.49 | 6.97 | 0 | 49.5 |
| $ \text{Diff } \Delta : \text{\# of Steps 1}$ | NAEP & Independent LLM Prediction | 8.49 | 6.95 | 0.2 | 44.1 |
| $ \text{Diff } \Delta : \text{\# of Steps 2}$ | NAEP & Independent LLM Prediction | 8.77 | 7.34 | 0.2 | 54.9 |
| $ \text{Diff } \Delta : \text{\# of Steps } \Delta $ | NAEP & Independent LLM Prediction | 3.5 | 4.46 | 0 | 39.2 |
| $ \text{Diff } \Delta : \text{similarity}$ | Independent LLM Prediction | 5.93 | 6.81 | 0.1 | 52 |

Many variables are referred to with either a 1 or 2 which indicates that these variables apply to either variable 1 or variable 2 in the binary variable comparison. The parameters directly derived from NAEP are: difficulty of the first item (Diff₁) and the difficulty 2 of the second item (Diff₂), the absolute value in the difference in difficulties between the items ($|\text{Diff}$

$\Delta|)$ as well as the proxies for the discrimination parameters (\hat{a}_1 and \hat{a}_2). Difficulties are oriented in the reverse of classical test theory item difficulties, more in line with item response theory difficulties (b parameter) such that greater difficulty indicates more test taker chose an incorrect response. Difficulties with zero indicating 0 percent of the population taking the chose wrong response while a difficulty of 100 indicates 100 percent of the population getting the item wrong.

These difficulties are not used directly in the estimation of the classification performance though they enter the model in two ways: 1. They define what the correct answer to the pairwise comparison is and 2. Their absolute difference enters as an explanatory variable ($|Diff \Delta|$). Though it appears that a single variable is acting as both the explanatory variable and the dependent variable there is no issue as the $sign(Diff \Delta)$ is independent of the $|Diff \Delta|$ for all $|Diff \Delta| > 0$. The other NAEP derived parameters are \hat{a}_1 and \hat{a}_2 which refer to the proxy of item discrimination (See equations 3.3.1.2 and 3.3.1.3).

The variable parameters that are independently generated by the LLMs are item complexity, number of steps, and item pair similarity. Item “Complexity” is generated by asking the LLM (Gemini-Pro) in this case to generate and estimate of item complexity between 1 and 10 with ten being most complex. The “Number of Steps” is generated by asking the LLM (Gemini-Pro) to list the steps required to solve the item, then counting those steps. While the previous two items are generated on the individual item level “Similarity of Item Pair” is generated on the item pair level by prompting the LLM (GPT 3.5²), “On a scale of 1 to 10 with 1

² Relative similarities of items ran afoul of Gemini-Pro’s over ambitious “safety rules”. Much more frequently than expected this prompt came back with no-response which led to using GPT 3.5 to generate these values.

being very dissimilar and 10 being very similar how similar is the content covered by these different questions?”. Like the absolute value of the difference of item difficulty ($|\Delta \text{Diff}|$) the absolute value of the difference of item complexity ($|\Delta \text{Complexity}|$) and number of steps ($|\Delta \text{Steps}|$) is calculated.

3.10 Study Two: Item Parameters and Prompt Design

Unlike the comparison of item pairs (Study One) which used only one LLM model (Gemin-Pro) this study explores how a variety of LLMs perform when attempting to solve NAEP items. I then ask the question, “are the items that the LLM finds more difficult also more difficult for students?” Unlike the previous study in which hundreds of prompt formulations were imagined and explored, this study only made use of only two prompts:

- Prompt 1: {{Item Content and Answers}}. What is the correct response?
- Prompt 2: {{Item Content and Answers}}. Please provide a step-by-step explanation of how to get to the solution.

Each model is given all 388 items each at least three times. The correct answers are scored as one and wrong as 0. For each model, the mean score of the various trials is calculated. Looking at Table 16 we can see the outcomes for the two different prompt attempts. The least complex model Llama 7b gets the correct answer only 43% of the time for prompt 1 and 48% for prompt two. The most complex model, GPT4 however gets the correct answer 76% of the time under prompt 1 and 96% of the time under prompt two. Most of these models show a statistically significant but small negative correlation between item difficulty and model performance on that item. This provides some modest support that LLMs respond similarly to that of students and find more difficult items also more difficult to solve. The model

that has the lowest correlation between difficulty and performance was GPT-4 which performed well (above 75% correct) under prompt 1 and very well under prompt 2 (95% correct).

TABLE 16: LLM ATTEMPTING EACH ITEM

This table displays the outcomes from querying a range of Large Language Models (LLMs) using two distinct prompt models. Here, the score represents the model's average accuracy percentage, while Corr(Score #, Diff) indicates the correlation between the empirical difficulty of the items and the model's score. It is important to note that with random selection, the expected score is 25%. The mean empirical scores for 4th, 8th, and 12th grades are 56.1, 53.7, and 50.1% respectively and 53.7% across all items.

| | Prompt one | | Prompt two | |
|------------|------------|------------------------|------------|-----------------------|
| LLM | Score one | Corr (Score one, Diff) | Score two | Corr(Score two, Diff) |
| GPT 3.5 | 0.649+ | 0.125** | 0.851+ | 0.166*** |
| GPT 4 | 0.764+ | 0.067 | 0.954+ | 0.064 |
| Llama 7b | 0.430- | 0.130** | 0.438- | 0.181*** |
| Llama 13b | 0.532- | 0.137*** | 0.513- | 0.186*** |
| Llama 70b | 0.568+ | 0.094* | 0.666+ | 0.123** |
| Gemini-Pro | 0.604+ | 0.114** | 0.813+ | 0.105** |

*10% significance, **5% significance, ***1% significance.

+ Indicates the LLM outperforms average test taker.

- Indicates the LLM underperforms the average test taker.

Despite 8 out of 12 of the LLM prompt combinations outscoring the average student the correlations between the difficulty and the model scores (Corr(Score #, Diff)) are relatively small. Their largest value is 0.186 for Llama 13b under prompt 2. To get an idea of whether these performance scores and correlations with item difficulties were in line with what is

possible to observe under item response theory I simulated 388 3PL items with random difficulties drawn from a normal distribution. I also simulated 500 random test takers with theta drawn also from a standard normal to estimate the classical test theory difficulty. I sampled a range of theta's treating each model as having its own ability parameter with each theta attempting each item 3 times then taking the average score for that item. I correlated that score with the empirical item difficulties to find the correlation. Then I repeated the entire simulation 200 times to find 95% confidence intervals and medians for both the score and $\text{Corr}(\text{Score \#, Diff})$ under the scenario (Table 17).

To my surprise, I found that under $a = 0.35$ and $c = 0.35$ (admittedly cherry-picked parameters) it is possible to find matches within the 95 percentiles of both score and ρ for all six LLM models and both prompts except in the case of GPT 4 with prompt 1. It is very difficult to get the observed $\text{Corr}(\text{Score 1, Diff})$ of 0.067 while having a score 53% (in the case of $\theta = -3.5$). This of course does not prove that LLMs performance simulates student performance. However, it does show that the performance of LLMs can approximate those simulated under item response theory at least by these measures.

TABLE 17: SIMULATING LLM PERFORMANCE UNDER 3PL

This table presents the results of simulating a 3PL model with 388 random items across various ability distributions. It includes 200 random performance samples with different Θ abilities. The columns Q5, Q50, and Q95 represent the 5th, 50th, and 95th percentiles of the scores. A model and prompt pair is considered a match if both the observed score and the empirical correlation with model performance (ρ) fall within the Q5 to Q95 range.

| | Score | | | ρ | | | LLM Match | Prompt | Match |
|-----------------|-------|-------|-------|--------------|-------|-------|------------|--------|-------|
| | Q 5 | Q 50 | Q 95 | Q 5 | Q 50 | Q 95 | | | |
| $\Theta = -5.0$ | 0.428 | 0.453 | 0.479 | 0.008 | 0.100 | 0.173 | Llama 7B | 1 | 1 |
| $\Theta = -4.5$ | 0.443 | 0.465 | 0.490 | 0.036 | 0.114 | 0.181 | Llama 7B | 2 | 1 |
| $\Theta = -4.0$ | 0.463 | 0.485 | 0.506 | 0.046 | 0.119 | 0.199 | | | |
| $\Theta = -3.5$ | 0.482 | 0.505 | 0.529 | 0.055 | 0.133 | 0.203 | Llama 13B | 2 | 1 |
| $\Theta = -3.0$ | 0.501 | 0.527 | 0.547 | 0.070 | 0.143 | 0.220 | Llama 13B | 1 | 1 |
| $\Theta = -2.5$ | 0.523 | 0.547 | 0.571 | 0.080 | 0.151 | 0.233 | | | |
| $\Theta = -2.0$ | 0.550 | 0.572 | 0.595 | 0.086 | 0.164 | 0.241 | Llama 70B | 1 | 1 |
| $\Theta = -1.5$ | 0.570 | 0.597 | 0.620 | 0.085 | 0.169 | 0.247 | Gemini-Pro | 1 | 1 |
| $\Theta = -1.0$ | 0.601 | 0.623 | 0.646 | 0.115 | 0.181 | 0.255 | | | |
| $\Theta = -0.5$ | 0.626 | 0.649 | 0.674 | 0.111 | 0.191 | 0.275 | GPT 3.5 | 1 | 1 |
| $\Theta = 0.0$ | 0.655 | 0.678 | 0.702 | 0.118 | 0.194 | 0.254 | Llama 70B | 2 | 1 |
| $\Theta = 0.5$ | 0.686 | 0.706 | 0.725 | 0.113 | 0.194 | 0.271 | | | |
| $\Theta = 1.0$ | 0.711 | 0.734 | 0.753 | 0.124 | 0.196 | 0.274 | | | |
| $\Theta = 1.5$ | 0.741 | 0.762 | 0.778 | 0.105 | 0.205 | 0.273 | | | |
| $\Theta = 2.0$ | 0.765 | 0.786 | 0.803 | 0.112 | 0.194 | 0.275 | GPT 4 | 1 | 0 |
| $\Theta = 2.5$ | 0.789 | 0.809 | 0.828 | 0.112 | 0.189 | 0.267 | | | |
| $\Theta = 3.0$ | 0.814 | 0.832 | 0.845 | 0.104 | 0.182 | 0.262 | Gemini-Pro | 1 | 1 |
| $\Theta = 3.5$ | 0.831 | 0.851 | 0.865 | 0.105 | 0.193 | 0.273 | GPT 3.5 | 2 | 1 |
| $\Theta = 4.0$ | 0.852 | 0.869 | 0.887 | 0.102 | 0.178 | 0.256 | | | |
| $\Theta = 4.5$ | 0.873 | 0.887 | 0.901 | 0.093 | 0.163 | 0.244 | | | |
| $\Theta = 5.0$ | 0.886 | 0.902 | 0.915 | 0.078 | 0.164 | 0.238 | | | |
| $\Theta = 5.5$ | 0.903 | 0.916 | 0.928 | 0.076 | 0.153 | 0.235 | | | |
| $\Theta = 6.0$ | 0.916 | 0.929 | 0.939 | 0.048 | 0.147 | 0.226 | | | |
| $\Theta = 6.5$ | 0.926 | 0.938 | 0.949 | 0.044 | 0.133 | 0.228 | | | |
| $\Theta = 7.0$ | 0.936 | 0.947 | 0.958 | 0.053 | 0.120 | 0.200 | GPT 4 | 2 | 1 |

Note: Item parameters b drawn from random normal, $a = .35$, and $c = .35$.

I will briefly explore to what extent these responses might be leveraged to predict item difficulty. In addition to using LLM performance as an explanatory variable I generate several computational variables including “Word Count,” “Flesch-Kincaid Index,” and (average) “Syllables per Word” (Table 18). These values are calculated for the full item text as well as the option text though item numbering and option numbering are excluded in these calculations (for example: “Item 1: What is 7% of 200. A. 7, B. 14, D. 20, E. 21” becomes “What is 7% of 200. 7, 14, 20, 21”).

TABLE 18: DEPENDENT VARIABLES AND EXPLANATORY VARIABLE STATISTICS

| | Mean | SD | Min | Max |
|----------------------|-------|-------|------|------|
| Easiness | 53.78 | 18.71 | 8 | 95 |
| Word Count | 38.63 | 18.22 | 7 | 105 |
| Flesch Kincaid Index | 4.58 | 3.43 | -2.8 | 25.2 |
| Syllables per Word | 1.29 | 0.2 | 0.78 | 2.2 |

CHAPTER IV: RESULTS

4.1 Study One: Predicting Item Pair Rankings

In this study, I predict the performance of the classification model using available explanatory variables. The theoretical justification for these estimates can be found in section 3.3.1. The Table 19 shows the results of estimating how likely the LLM (Gemini-Pro) will be at correctly ranking which item in an item pair is more difficult. For interpretational purposes the results column, “model 4” will be the column for which these results are discussed. The other columns are primarily presented for the purposes of sensitivity analysis.

The results for all non-binary variables are standardized (mean of zero and standard deviation of 1) and can be directly interpreted as the expected change to likelihood of the LLM of correctly ranking the item pairs. For example, the coefficient on the absolute value of the differences in complexity score ($|\text{Complexity } \Delta|$), is 0.0423 which indicates a one standard deviation difference in estimated complexity of the two items results in a 4.23 percentage point increase in the likelihood of the LLM to correctly rank the two items.

The coefficients of binary variables on the other hand, are best interpreted as holding all else equal how much would moving the binary variable value from 0 to 1 change the expected rank performance. For example, the “Second Item Harder” variable is a binary flag which indicates when the second item is harder than the first item. The "Second Item Harder" binary variable consistently demonstrated a statistically significant positive coefficient of around .10-.11 which means that the LLM is about 10 to 11 percentage points more likely to predict the correct item ranking if the second item in a pair happens to be more difficult. This implies that the LLM exhibits a bias towards predicting the second-listed item as more difficult.

Results from the OLS models, as presented in Table 19, indicate several noteworthy findings. A total of 1,825 item pairs were successfully evaluated by the LLM, named Gemini-Pro, with the model refusing to return results for 19 pairs due to sensitivity concerns. The absolute difference in item difficulties ($|\text{Diff } \Delta|$), showed a statistical significance increase of around 8 percentage point increase in the likelihood of correctly ranking item pairs for a 1 standard deviation increase in the $|\text{Diff } \Delta|$.

Grade-level variables for Grade 12 showed positive and significant coefficients across models, indicating their relevance in ranking, compared with the omitted category Grade 4. Grade 8 did not have a significantly different ranking performance compared with the omitted category Grade 4. Contrary to expectations, Math items on average are estimated to be around 5 percentage points more difficult to rank than science items. The difficulty of Grade 12 items was rank, consistent with my predictions, as these items are less reliant on intuition and more on structured or acquired knowledge, according to the model's estimates.

Table 19 shows the likelihood of correctly ranking pairwise items under different linear explanatory variable choices. There is a total number of pairs successfully evaluated of 1825. Each of these pairs is evaluated with each item being listed first or second. The total number of pairs attempted to be evaluated was 1854. Unfortunately, the LLM used for these comparisons (Gemini-Pro) pairs refused to return results for 19 pairs. These pairs triggered its “safety” algorithm as they touched on sensitive topics such as evolutionary selection and gender.

TABLE 19: PREDICTING THE LIKELIHOOD OF CORRECTLY RANKING AN ITEM PAIR

This table shows a series of ordinary least square fitted model estimates. In most models the majority of coefficients are statistically significant at varying levels.

| Model: | 1 | 2 | 3 | 4 | 5 |
|---------------------------------------|-----------|-----------|-----------|-----------|-----------|
| No. observations: | 3650 | 3650 | 3650 | 3650 | 365 |
| R-squared: | 0.043 | 0.047 | 0.049 | 0.054 | 0.054 |
| Adj. r-squared: | 0.043 | 0.046 | 0.047 | 0.051 | 0.051 |
| F-statistic: | 82.16 | 36.29 | 26.67 | 14.91 | 18.79 |
| Prob (f-statistic): | 1.26E-35 | 2.19E-36 | 6.70E-36 | 1.20E-35 | 4.01E-37 |
| Coefficient Estimates: | | | | | |
| Intercept | 0.5706*** | 0.5624*** | 0.5559*** | 0.5497*** | 0.5673*** |
| Second Item Harder | 0.1014*** | 0.0995*** | 0.1155*** | 0.1177*** | 0.108*** |
| Diff Δ | 0.0858*** | 0.0842*** | 0.0837*** | 0.0789*** | -0.076** |
| Grade : 12 | | 0.0787*** | 0.0766*** | 0.0810*** | |
| Grade : 8 | | 0.0006 | -0.0034 | 0.0046 | |
| Math | | -0.0607** | -0.0602** | -0.0474** | |
| \hat{a}_1 | | | -0.0168** | -0.0174** | |
| \hat{a}_2 | | | 0.0089* | 0.0083* | |
| Complexity 1 | | | | 0.0175** | |
| Complexity 2 | | | | 0.0067* | |
| Complexity Δ | | | | 0.0423*** | |
| # of steps 1 | | | | 0.0108* | |
| # of steps 2 | | | | 0.0002 | |
| # of Steps Δ | | | | 0.0086* | |
| Similarity | | | | -0.0049 | |
| Diff Δ : Complexity 1 | | | | | 0.0522*** |
| Diff Δ : Complexity 2 | | | | | 0.0574*** |
| Diff Δ : Complexity Δ | | | | | 0.0657*** |
| Diff Δ : # of steps 1 | | | | | 0.0258** |
| Diff Δ : # of steps 2 | | | | | 0.0072 |
| Diff Δ : # of steps Δ | | | | | 0.0200** |
| Diff Δ : \hat{a}_1 | | | | | -0.0143** |
| Diff Δ : \hat{a}_1 | | | | | 0.0006 |
| Diff Δ : Similarity | | | | | -0.0147* |

The R-squared values across the six models range from 0.043 to 0.059, and the corresponding adjusted R-squared values range from 0.043 to 0.053, highlighting a modest fit to the data. These values, along with highly significant F-statistics (p-values well below the 0.05 threshold), suggest that the models capture a portion of the variability in the LLM's ability to rank item difficulty correctly though their overall predictive accuracy power is low.

The inclusion of discrimination proxy variables (\hat{a}_1 and \hat{a}_2) demonstrated a complex and difficult to explain effect on the ability to estimate item relative difficulties. I had hypothesized that items with greater discrimination should be easier to rank in terms of difficulty than items with low discrimination. This does not appear to be the case. Item pairs which have higher discrimination in the first item (\hat{a}_1) leads to a lower ranking accuracy (1.7 percentage points) while a higher discrimination in the second item (\hat{a}_2) contributes to a higher ranking accuracy (0.8 percentage points). It is interesting that LLM has differences in the item pair rankings due to differences in item discrimination. Overall, though the effects are quite small and contradictory.

However, the item parameters for “complexity” (Complexity 1, Complexity 2, as well as the contrast in the complexity $|\text{Complexity } \Delta|$), aligned with my theoretical hypotheses. Items which have one standard deviation higher complexity in item 1 or item 2 have between a 1.7 and 0.6 percentage point increase respectively in likelihood of correctly ranking the items. Likewise, when there is a one standard deviation difference in complexity ($|\text{Complexity } \Delta|$) between the two items that also corresponds to a 4.2 percentage point increase in the likelihood of correctly ranking the item pair. The complexity coefficients indicate that they are statistically significant predictors of the LLM’s ability to detect relative item difficulty.

In a similar vein, the variable related to number of steps (# of Steps 1, # of Steps 2, and ($|\# \text{ of Steps } \Delta|$) are also positively associated with better predictors of item difficulties though these coefficients tend to be very small. Between these two sets of variables, the hypothesis that more complex or involved items are easier to predict relative difficulty is supported.

In summary, the empirical results support the theoretical expectations posited in section 3.3.1, with the LLM displaying a measurable ability to predict relative item difficulties. Item discrimination and complexity have been substantiated as significant factors. That said the LLM suffered from a clear bias with items appearing second 10 percentage points more likely to be seen as more difficult than items appearing in the first space.

4.2 Study Two: Do LLM Simulate Student/Test Taker Responses?

LLM performance is demonstrated to be a statistically significant if modest predictor of item difficulty (Table 20). Model 2 includes only indicator variables for grade and subject as well as some computational variables, and has an R-squared of 7% and adjusted R-squared of 5.5% while including LLM performance as a predictor under either prompt (Models 3 and 4) increases the R-squared value to 11.9% to 13.1% and the adjusted R-squared value to 9.1% and 10.3% respectively. The individual coefficients on the LLM models are consistently positive and generally statistically significant though prompt specification seems to play an important role. We can interpret the coefficients on prompts directly such that for example under Prompt 1 if Llama 7b got an item correct on all three attempts then we should expect the item to be 6.15 percentage points easier than if it had gotten the item incorrect all three attempts. By adding up the coefficients in the explanatory variables we can get a sense for what the maximum variability in difficulty can be explained through LLM performance. Under Prompt 1 we get a

total of 17.25 percentage points and 20.47 for prompt 2 which when items range in difficulty (easiness) between 8 and 95 this is only about 20% of the range possible.

TABLE 20: PREDICTING ITEM DIFFICULTY FROM LLM PERFORMANCE

The following table has 1-Difficulty as the dependent variable so that the LLM coefficients are positively valued.

| Model: | 1 | 2 | 3 | 4 | 5 |
|--|------------|----------|----------|----------|----------|
| No. observations: | 388 | 388 | 388 | 386 | 386 |
| R-squared: | 0.017 | 0.070 | 0.119 | 0.131 | 0.141 |
| Adj. r-squared: | 0.009 | 0.055 | 0.091 | 0.103 | 0.099 |
| F-statistic: | 2.155 | 4.754 | 4.237 | 4.683 | 3.35 |
| Prob (f-statistic): | 9.29E-02 | 1.09E-04 | 2.76E-06 | 4.06E-07 | 6.12E-06 |
| Coefficient Estimates: | | | | | |
| Intercept | 56.22*** | 60.85*** | 56.16*** | 51.16*** | 51.46*** |
| Grade 12 | -5.78** | -4.44** | -3.00* | -3.72* | -3.25* |
| Grade 8 | -2.36* | -1.76* | -0.47 | -0.44 | -0.26 |
| Math | -0.86 | 0.92 | -3.40* | -1.90* | -3.21* |
| Word count | | -0.23*** | -0.25*** | -0.23*** | -0.24*** |
| Flesch Kincaid Index | | -0.17 | -0.22 | -0.06 | -0.12 |
| Syllables per word | | 3.30 | 0.25 | 0.05 | -0.84 |
| Prompt 1 | GPT35 | | 1.23 | | 0.37 |
| | GPT4 | | 3.03* | | 2.70* |
| | Llama7b | | 6.15** | | 3.69* |
| | Llama13b | | 2.81* | | -0.52 |
| | Llama70b | | 1.33 | | -0.68 |
| | Gemini pro | | 2.70* | | 1.91 |
| Prompt 2 | GPT35 | | | 4.93* | 4.31* |
| | GPT4 | | | 1.13 | 1.00 |
| | Llama7b | | | 4.45* | 3.38* |
| | Llama13b | | | 7.30*** | 6.68** |
| | Llama70b | | | -0.12 | -0.20 |
| | Gemini pro | | | 2.78* | 1.30 |
| *10% significance, **5% significance, ***1% significance | | | | | |

The increase in performance going from using either Prompt 1 or Prompt 2 performance to using both Prompt 1 and Prompt 2 performances is slight in terms of R-squared and minimal

to negative in terms of adjusted R-squared. In terms of the sum of the prompt coefficients we go from Prompt 1 which has 17.25 and Prompt 2 which has 20.47 to a combined coefficient sum of 23.91. The reason for this unexpected decline in model predictive power is that model performance is much closer correlated with other models (Table 21) than it is with the difficulty of the items evaluated (Table 16, page 83). This finding is not entirely surprising as on at least one study (Bejar 1983) expert judges were found to have higher interrater correlations than the correlations between item estimated rankings and empirical rankings.

TABLE 21: PARTIAL CORRELATIONS TABLE BETWEEN MODELS

The following table shows the correlations of the performance on items under models under Prompt 1 as compared to other models under Prompt 1 or models under Prompt 2.

| | | Prompt 1 | | | | | |
|----------|------------|----------|-------|----------|-----------|-----------|------------|
| | | GPT 3.5 | GPT 4 | Llama 7b | Llama 13b | Llama 70b | Gemini-Pro |
| Prompt 1 | GPT 3.5 | * | 0.250 | 0.319 | 0.380 | 0.401 | 0.485 |
| | GPT 4 | 0.250 | * | 0.116 | 0.120 | 0.156 | 0.451 |
| | Llama 7b | 0.319 | 0.116 | * | 0.426 | 0.291 | 0.267 |
| | Llama 13b | 0.380 | 0.120 | 0.426 | * | 0.443 | 0.241 |
| | Llama 70b | 0.401 | 0.156 | 0.291 | 0.443 | * | 0.363 |
| | Gemini-Pro | 0.485 | 0.451 | 0.267 | 0.241 | 0.363 | * |
| Prompt 2 | GPT 3.5 | 0.421 | 0.046 | 0.247 | 0.303 | 0.281 | 0.219 |
| | GPT 4 | 0.072 | 0.127 | 0.021 | -0.004 | 0.008 | 0.008 |
| | Llama 7b | 0.208 | 0.152 | 0.439 | 0.360 | 0.339 | 0.200 |
| | Llama 13b | 0.335 | 0.155 | 0.421 | 0.592 | 0.459 | 0.264 |
| | Llama 70b | 0.283 | 0.153 | 0.305 | 0.435 | 0.467 | 0.293 |
| | Gemini-Pro | 0.299 | 0.155 | 0.227 | 0.276 | 0.329 | 0.498 |
| Mean | | 0.314 | 0.171 | 0.280 | 0.325 | 0.322 | 0.299 |

*Note: These values are 1 but left out so as to not cause issues with the mean calculations.

While the correlation between model performance in this study does not imply that it will hold for future developments of LLMs it does imply a limitation on how far these current

models can be taken in the current prompts. These results suggest that adding additional LLM models or additional prompt variants (in line with those under Prompt 1 and 2) while likely to increase the total explanatory power of the total model are likely to have diminishing marginal effectiveness at predicting item difficulty.

4.3 Study Three: Binary Search Estimation of Absolute Item Difficulty

In this dissertation I propose the use of a series of binary pairwise comparisons to predict item difficulty. These predictions are done through a logistic regression with item indicators as the explanatory variable with the first item listed coded positively and the second item listed coded negatively. This explanatory indicator matrix had 388 columns with values 0, 1, or -1. The choice of which item was chosen as more difficult, the dependent variables, was coded as 1 or 0 depending upon if it was listed first or second. No coefficient was included in the logistic regression.

Using this method I was able to recover estimates for the coefficients for the 258 items which were included in the testing data which I had pairwise estimates of relative difficulty from Study One. This method produced estimates of item difficulty that appear in 1PL item response theory form (Table 22).

TABLE 22: ESTIMATED ITEM DIFFICULTY

| Grade | Subject | Mean | Standard Dev | Q 05 | Q 25 | Q 50 | Q 75 | Q 95 | Count |
|-------|-------------|------|--------------|-------|-------|-------|------|-------|-------|
| 4 | Mathematics | 0.00 | 1.36 | -2.15 | -1.02 | 0.13 | 0.97 | 2.25 | 85 |
| 4 | Science | 0.00 | 0.77 | -1.04 | -0.60 | 0.00 | 0.83 | 1.03 | 16 |
| 8 | Mathematics | 0.00 | 1.27 | -2.06 | -0.96 | 0.13 | 1.01 | 1.84 | 74 |
| 8 | Science | 0.00 | 0.95 | -1.21 | -0.67 | -0.01 | 0.65 | 1.46 | 17 |
| 12 | Mathematics | 0.00 | 1.48 | -2.20 | -1.21 | -0.05 | 1.19 | 2.49 | 39 |
| 12 | Science | 0.00 | 1.19 | -1.62 | -0.80 | 0.01 | 0.74 | 1.84 | 27 |
| | | | | | | | | Total | 258 |

Superficially the item summary parameters from these items appear acceptable but what we would really like to know is how well these item parameters are correlated with the population performance statistics reported on each item by NAEP. To do this, I correlate performance of the pooled population data as well as demographic groups such as gender, race, and free and reduced-price lunch (FRPL). We can see that the pooled population data is 38.9% correlated with the item parameter estimates (Table 23). We would like to see a high correlation between the estimated item scores and the observed item performances. A correlation that falls for our population between the mid-30s and the low 40s while certainly statistically non-random is lower than could be hoped.

TABLE 23: ITEM DIFFICULTY ESTIMATE CORRELATED WITH POPULATION PERFORMANCE

This table gives population mean performance scores averaged across the 258 items as well as the correlation of those scores with the estimated item difficulties generated by using the item pairs as indicators of which item is more difficult. The table to the right also presented the results of a bootstrap simulation of the correlation statistics after random resampling 500 draws from the source data. Bootstrapped headers are standard deviation (SD) and quantiles (Q) 25, 50, and 75.

| Grouping: Population | Population Mean Score | Corr Between Est. Diff & Pop. Score) | Bootstrapped Corr Statistics | | | |
|---|-----------------------|--------------------------------------|------------------------------|-------|-------|-------|
| | | | SD | Q 25 | Q 50 | Q 75 |
| All | 0.530 | 0.389 | 0.058 | 0.353 | 0.394 | 0.429 |
| Gender: Male | 0.544 | 0.398 | 0.058 | 0.363 | 0.400 | 0.440 |
| Gender: Female | 0.519 | 0.343 | 0.062 | 0.308 | 0.347 | 0.386 |
| Race: White | 0.571 | 0.414 | 0.058 | 0.381 | 0.418 | 0.454 |
| Race: Black | 0.421 | 0.363 | 0.062 | 0.325 | 0.368 | 0.403 |
| Race: Hispanic | 0.454 | 0.377 | 0.062 | 0.342 | 0.384 | 0.417 |
| Race: Asian/Pacific Islander | 0.608 | 0.344 | 0.068 | 0.302 | 0.349 | 0.396 |
| Location: City | 0.540 | 0.348 | 0.113 | 0.283 | 0.355 | 0.433 |
| Location: Suburb | 0.582 | 0.354 | 0.113 | 0.287 | 0.358 | 0.439 |
| Location: Town | 0.550 | 0.364 | 0.115 | 0.295 | 0.369 | 0.452 |
| Location: Rural | 0.567 | 0.373 | 0.112 | 0.307 | 0.379 | 0.460 |
| FRPL Eligibility: Eligible | 0.462 | 0.339 | 0.075 | 0.299 | 0.347 | 0.392 |
| FRPL Eligibility: Not eligible | 0.589 | 0.365 | 0.073 | 0.326 | 0.371 | 0.422 |
| FRPL Eligibility: Information not available | 0.589 | 0.360 | 0.074 | 0.319 | 0.369 | 0.415 |
| Pairwise Comparisons Count | | 3,660 | | | | |

Bold Indicates the maximum for the group. None of the differences in Correlation by grouping are large enough to be statistically significant at any level.

From Table 23 we can see that some populations have higher correlations between the item estimates and the population performance. These population groups are male, white, rural, and FRPL: Not eligible. These populations do not correspond with the highest performing population except in the case of male. From the bootstrapped standard errors, it would be

advisable not to read too much into the different correlations by population group as there is no statistical significance of the difference between these different correlations. Even the largest gap which is between white and Asian/Pacific-islander is only .07 of a correlation which is less than one joint standard deviation away ($0.089 = (0.058^2 + 0.068^2)^{1/2}$).

So far in this analysis I have paired each item with each other in each test (segregated by grade, year, and subject) twice with item 1 in the first position in one case and in the second position in the second case. If we were to make the limited assumption that the population performance is stable over time, we can expand our item pairing pool to see to what extent having more items to compare performance against might improve our item difficulty predictions. Table 24 does just this by exploring how predictions of item difficulty change if we pool items across tests to expand our matching pool.

“Pooled estimation” match compares all testing items which share grade and subject and estimates all parameters simultaneously. While “Item 2 Known Parameters” provides the All-Student difficulty estimate for item 2 and generates an estimate for item 1 for all items paired with item 1.

TABLE 24: POOLING GRADE AND SUBJECT ITEMS OVER YEARS

This table shows correlations between the estimated item difficulties pairing up items across any year but keeping within subject and grade.

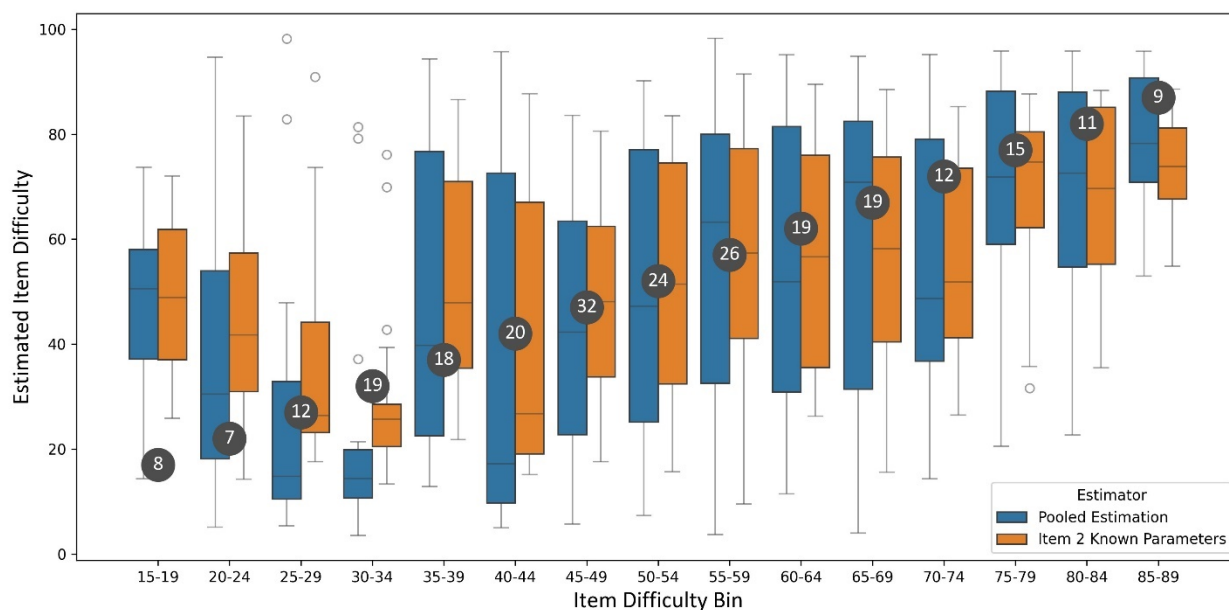
| Grouping | Correlations | | Mean Root Squared Error / Mean Absolute Error | |
|---|-------------------|-------------------------|---|-------------------------|
| | Pooled Estimation | Item 2 Known Parameters | Pooled Estimation | Item 2 Known Parameters |
| All Students | 0.41 | 0.42 | 28.4 / 23.5 | 23.0 / 18.4 |
| Gender: Male | 0.41 | 0.42 | 28.7 / 23.7 | 23.0 / 18.5 |
| Gender: Female | 0.36 | 0.37 | 29.5 / 24.1 | 24.2 / 19.0 |
| Race: White | 0.43 | 0.44 | 28.8 / 23.8 | 23.2 / 18.7 |
| Race: Black | 0.40 | 0.40 | 29.5 / 23.8 | 25.5 / 20.0 |
| Race: Hispanic | 0.40 | 0.41 | 28.8 / 23.4 | 24.2 / 19.2 |
| Race: Asian/Pacific Islander | 0.38 | 0.40 | 31.1 / 25.7 | 24.9 / 20.4 |
| Location: City | 0.42 | 0.41 | 28.1 / 23.6 | 22.2 / 17.7 |
| Location: Suburb | 0.42 | 0.42 | 29.4 / 24.9 | 23.0 / 18.8 |
| Location: Town | 0.43 | 0.43 | 28.2 / 23.6 | 22.5 / 17.8 |
| Location: Rural | 0.44 | 0.44 | 28.7 / 24.0 | 22.7 / 18.1 |
| FRPL Eligibility: Eligible | 0.42 | 0.43 | 28.0 / 23.0 | 23.3 / 18.2 |
| FRPL Eligibility: Not eligible | 0.41 | 0.43 | 30.0 / 24.9 | 24.0 / 19.6 |
| FRPL Eligibility: Information not available | 0.40 | 0.41 | 30.2 / 25.2 | 24.3 / 19.8 |
| Pooled Joint Estimation | 1.00 | 0.98 | 0 / 0 | 8.8 / 7.7 |
| Known Item Parameters (1PL) | 0.98 | 1.00 | 8.8 / 7.7 | 0 / 0 |
| Pairwise Comparisons Count | 30,452 | 30,452 | 30,452 | 30,452 |

The first column of this table shows what happens when we increase our pairwise number of comparisons by pooling across years. This is an 8-fold increase in pairwise comparisons with the number increasing from 3,660 to 30,452. There is no guarantee that increasing the comparisons in this way should work as we are trading increased sample size, and correspondingly decreased sampling error, for the introduction of error due to time related changes in population ability as well as item drift. Fortunately, the gains from pooling items seem to outweigh the introduction of this new error as the pooled estimation correlations

across all tested populations in Table 24 are monotonically equal to or greater than those of the correlation column found in Table 23.

FIGURE 5: ITEM ESTIMATES MAPPED TO ITEM DIFFICULTY BINS

Grouping items into empirical difficulties bins (x axis) and estimated item difficulty (y axis) this figure shows the relationship between the two. The circle represents the mid-point of the empirical bin while the x-axis shows the bin for All Population difficulty. Embedded in the circles are the counts of the number of items in the bin. The center line in the box is the median which the upper and lower box parts are the bottom and upper quartiles while the top and bottom lines are maximum and mins as well as a few points for the outliers.



While evaluating estimator performance it is worth exploring how well item difficulty varies with items in different subjects and grades. Table 25 explores how the statistics between the general population estimate of item difficulty and the estimated item difficulty from the LLM pairwise comparisons perform. It shows that grades and subject correlations are highly

variable with peak performance in grade 12 mathematics with a correlation estimated at around 0.66. While still short of the “target” of 0.8, this point estimate appears promising.

Another way of evaluating item parameter estimates is by looking at the root mean squared error (RMSE) as well as the mean absolute error (MAE). Both these measures give a consistent estimate of the expected absolute “difference” for a random item and its estimate using the two methods. RMSE is a strictly convex estimator while MAE is not. Looking at just the best case, MAE with “Item 2 Known Parameters” we see that the estimated MAE is between 17 and 20 percentage points (Table 23) and between 15 and 22 percentage points (Table 24).

TABLE 25: ESTIMATED ITEM DIFFICULTY STATISTICS BY GRADE AND SUBJECT

This table presents mean correlations, root mean squared error (RMSE), and mean absolute error (MAE). From this table we can see that observed correlation of 12th grade math item difficulty estimates are the most precise with a higher correlation between the all-student difficulty and a lower RMSE and MAE.

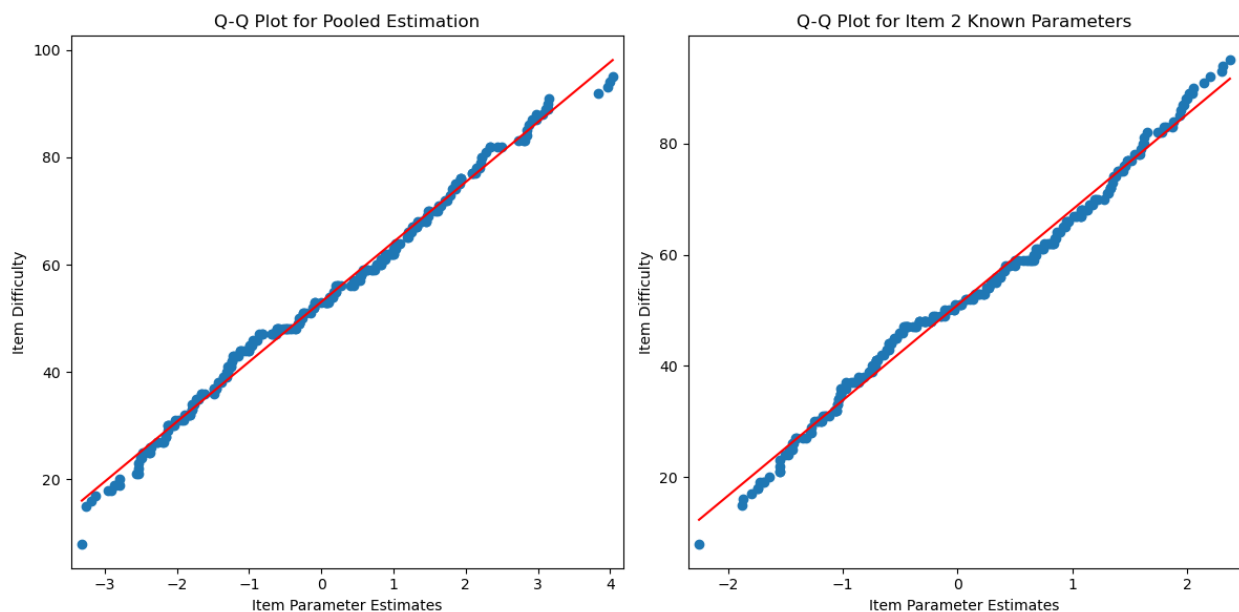
| Grade | Subject | Item Count | Mean Correlation | | Mean Root Squared Error / Mean Absolute Error | |
|-------|-------------|------------|-------------------|-------------------------|---|-------------------------|
| | | | Pooled Estimation | Item 2 Known Parameters | Pooled Estimation | Item 2 Known Parameters |
| 4 | Mathematics | 85 | 0.341 | 0.350 | 29.77 / 24.44 | 24.04 / 18.67 |
| 4 | Science | 16 | 0.459 | 0.378 | 25.26 / 21.26 | 25.46 / 21.73 |
| 8 | Mathematics | 74 | 0.385 | 0.371 | 29.42 / 24.91 | 23.24 / 19.13 |
| 8 | Science | 17 | 0.428 | 0.349 | 26.36 / 22.53 | 21.43 / 18.15 |
| 12 | Mathematics | 39 | 0.659 | 0.650 | 24.83 / 20.54 | 19.40 / 15.09 |
| 12 | Science | 27 | 0.256 | 0.271 | 28.95 / 22.99 | 23.13 / 18.22 |

It is sometimes helpful to look at the QQ plot (quantile-quantile plot) which shows the distributional relationship between two different samples. If both samples are drawn

from the same underlying distribution, allowing for different distribution parameters, then the quantile values plotted against each other should form a linear relationship. Looking at Figure 6 we can see that the quantiles matched against each other for both the pooled and the “Item 2 Known Parameters” estimators have a strong linear relationship suggesting that both observed and estimated item parameters are drawn from the same distribution. That said there is a bit of drift moving away from linear on the upper and lower tails of both plots potentially indicating that these methods are less precise when estimating extreme item parameters.

FIGURE 6: PAIRWISE COMPARISON ITEM PARAMETERS

These figures show a linear relationship between the quantiles of the estimated item parameters and the empirical item parameters.



We might be concerned that item difficulties have drifted over time such that using LLM to estimate items might be less or more accurate depending upon the year the items were administered. Looking at Table 26, while there appears to be some heterogeneity by year of

administration in terms of how effective the LLM estimators are it is unclear to what extent this heterogeneity represents random sampling error. Looking at the overall time trend between average correlation of the estimator and year yields point estimates very close to zero.

TABLE 26: ESTIMATED ITEM DIFFICULTY CORRELATIONS BY YEAR

The following table shows the correlation between All Student's item difficulty and item estimates for the pooled item parameter estimation and the item 2 known parameters grouped by year. The correlation between yearly correlations and year is only -.02 indicating that there does not appear to be a statistically significant trend effects on item performance as measured by correlations with true difficulties.

| Year | Correlations | | Mean Root Squared Error / Mean Absolute Error | | Count | |
|------|-------------------|-------------------------|--|-------------------------|-------|---------|
| | Pooled Estimation | Item 2 Known Parameters | Pooled Estimation | Item 2 Known Parameters | Math | Science |
| 1990 | 0.31 | 0.31 | 31.16 / 25.19 | 26.60 / 21.49 | 13 | |
| 1992 | 0.44 | 0.46 | 28.93 / 22.95 | 24.30 / 18.76 | 41 | |
| 1996 | 0.68 | 0.65 | 26.24 / 23.77 | 19.36 / 16.32 | 13 | |
| 2000 | 0.50 | 0.51 | 25.19 / 20.71 | 21.48 / 17.77 | | 16 |
| 2003 | 0.45 | 0.42 | 28.05 / 22.84 | 22.84 / 18.07 | 44 | |
| 2005 | 0.28 | 0.30 | 27.44 / 22.75 | 22.74 / 17.98 | 13 | 25 |
| 2007 | 0.32 | 0.29 | 26.34 / 22.25 | 21.34 / 16.93 | 19 | |
| 2009 | 0.69 | 0.53 | 16.56 / 14.47 | 21.65 / 19.28 | | 6 |
| 2011 | 0.35 | 0.35 | 32.42 / 27.57 | 24.50 / 19.93 | 31 | 3 |
| 2013 | 0.60 | 0.56 | 28.43 / 24.27 | 21.84 / 17.36 | 24 | |
| 2019 | 0.35 | 0.41 | 29.66 / 25.77 | 21.98 / 18.45 | | 10 |

CHAPTER V: DISCUSSION AND CONCLUSION

5.1 Discussion

This dissertation investigates the potential use of large language models (LLMs) to estimate item difficulty via two distinct methods. The most effective method involves pairwise comparison, where LLMs rank item pairs based on difficulty. The other method uses an ensemble of responses from various LLMs attempting to solve the items, using their success or failure as indicators of item difficulty. Both methods are statistically significant, yet less precise predictors of item fit compared to the requirements need to substitute item pretesting in a professional context. The approach presented here could be compared with three alternative approaches used in practice or explored in the literature. These approaches are pretesting, subject matter expert review, and the deployment of computational models (machine learning approaches).

Each of the existing approaches has its limitations. Pretesting, the gold standard, is expensive; subject matter expert reviews are also costly and imprecise; and computational models require a large number of already calibrated items, typically in the thousands, to build a predictive model. The use of LLMs, however, presents an alternative approach to item calibration. While LLMs have not yet been demonstrated to be sufficiently precise for professional testing contexts, they are inexpensive and require few, if any, already calibrated items.

This initial exploration of using LLMs to calibrate items has shown mixed results, with the most promising among 12th-grade math items. It appears that within 12th-grade mathematics, some item types are well-suited for LLMs, while others are particularly

challenging. My intuition suggests that items relying on visual features may be especially difficult for LLMs.

A logical next step would be to identify which item types among 12th-grade math items LLMs predict difficulty well. Additionally, we should investigate whether this proficiency extends to other higher-level math or reasoning items, such as those used in college-level math, logic, or computer science classes. Furthermore, can LLMs effectively predict the difficulty of items in a professional testing context, such as the Graduate Record Exam (GRE) quantitative reasoning sections, Law School Admission Test (LSAT) analytical reasoning or logical reasoning items, or other professional exams that include a mathematical analysis component?

Numerous scholarly studies on optimal test design for item calibration (Stocking, 1990; Buyske et al., 1998; Jones and Jin, 1994; Buyske, 1998; van der Linden et al., 2015; He and Chen, 2020, among others) emphasize the cost reduction achieved by combining estimates of test taker ability with estimates of item parameters when selecting items. Therefore, if a professional testing program adopts precalibration before pretesting a LLM calibration step, it is likely to reduce the cost of item calibration. The extent of this cost reduction depends on how well these approaches fit with their item bank and the degree to which they have already implemented other item precalibration techniques, such as the computational methods previously discussed.

Outside of professional testing programs, these techniques have ready applications in online interactive tutoring environments. In these settings, the goal is not primarily to optimize the acquisition of information about student abilities (as it is in computer adaptive testing). Instead, the focus is on providing items that are closely targeted to a student's current ability

level, ensuring that the items are neither too easy (which would result in little learning) nor too difficult (which might discourage students).

Even noisy difficulty estimates from LLMs, like those generated in this study, could be sufficiently precise for these applications, particularly for novel items. Once an item has been evaluated by enough test takers, these preliminary estimates can either be discarded or used as priors based on actual student performance.

In both the professional test development context and the online tutoring context, having a method like that provided by LLMs for generating initial estimates of item difficulty offers significant advantages. This approach is low-cost, poses zero exposure risk, and does not depend on training with existing item types.

The following section will discuss several notable features related to using LLMs for item calibration. These include the costs faced in the current market and limitations such as item security, the challenge of correlated LLM responses, and model bias.

5.1.1 Usage Fees of LLMs

An important feature of LLMs to discuss is their cost. In general, the pairwise approach explored in this paper generates large quantities of text. The generation of the pooled data with 30 thousand paired responses for example using 5.5 million words input and generated 4.4 million words out. Using Gemini-Pro as exploration was free but if I had used GPT3.5 this would have cost with current pricing around \$5 for the output and about half that for the input. The next generation LLM by OpenAI (GPT4-Turbo) would have cost about \$40 for the input and \$100 for the output. Using a production plan, Google's Gemini-Pro would have cost a similar amount to that of GPT3.5 at around \$12 for the output and \$4 for the input. As there is a

tremendous amount of ongoing innovation and competitive pressures in this market, I expect new models to continue to come out on an ongoing basis and downward pressure to continue to keep the cost down. This breakdown of costs is a lower estimate of the actual costs of implementing this study as many of the prompts ended up being repeated when the response was uninterpretable or when it was discovered there was some kind of user error. Also, there are other models leveraged in this research such as the various llama models and mistral model, but these models are only lightly used, and their cost is even less than that of GPT3.5.

TABLE 27: COST ESTIMATE OF LLM USAGE

| | | Pooled Items | Training Data |
|---|--------------|--------------|---------------|
| Prompts | Queries | 30,452 | 95,220 |
| | Words | 5,245,852 | 10,482,765 |
| | Characters | 31,614,610 | 60,759,951 |
| | Gemini-Pro* | \$3.95 | \$7.59 |
| | GPT3.5-Turbo | \$1.97 | \$3.93 |
| | GPT4.0-Turbo | \$39.34 | \$78.62 |
| | GPT4.0 | \$118.03 | \$235.86 |
| Responses | Words | 4,408,855 | 5,804,394 |
| | Characters | 31,118,358 | 41,046,116 |
| | Gemini-Pro* | \$11.67 | \$15.39 |
| | GPT3.5-Turbo | \$4.96 | \$6.53 |
| | GPT4.0-Turbo | \$99.20 | \$130.60 |
| | GPT4.0 | \$198.40 | \$261.20 |
| *Gemini-Pro is currently free for development (research). | | | |

5.1.2 Limitations

Item Security

A major drawback not yet discussed in this paper is that the large LLM APIs explored in this paper have no data security as data submitted to the LLM for evaluation may be used for future training. To what extent this security can be exploited is not entirely known. As such it

would be too risky for a test developer to expose sensitive items to the LLMs evaluated in this dissertation. However, this is a well-known limitation of these LLM APIs, and data privacy is likely to be a purchasable feature available in professional applications of LLMs.

However, if secure LLMs become available then the ability for these LLMs to be leveraged to potentially generated estimates of item difficulty which are much more secure than the dominant paradigm of item pretesting.

LLMs' Opinions Being Highly Correlated

Pooling the responses from diverse LLMs initially seems like an ideal way of eliminating sampling and model error. Using LLMs to build an ensemble model to predict item properties is appealing in that new LLMs are constantly under development and being released both through APIs as well as public open-source models. However, a major limitation as revealed in this dissertation is that the responses of LLMs to certain tasks seem to be much more correlated than one would expect. This limitation was a major challenge for using ensemble models to predict item difficulties (Study Two). Unfortunately, while supplementing item performance with additional responses from additional LLMs does appear to improve the performance this improvement has diminishing marginal returns. This is driven by responses across different LLMs being much higher correlated with each other than that of the observed data.

Not presented in this dissertation, I also used pairwise responses from GPT3.5, which was much worse at estimating relative item difficulties, pooled with those of Gemini-Pro. Pooling across these responses did not improve pairwise estimates of item difficulty. Given that errors are positively correlated this is an unsurprising if disappointing result.

The positive correlation of error is an interesting discovery which might limit future applications of this technology. Or it might prove to be a transient feature of the current LLM training paradigm by which various models share temporarily similar training content (for example Wikipedia, Reddit, etc.). It may be a feature of future LLMs that they develop independent additional training content such that their responses become more diverse.

Model Bias

Upon analyzing Tables 4.6.2 and 4.6.3, a discrepancy emerges in the estimated item difficulties across different demographic groups. Notably, estimates for male and white demographic groups exhibit a higher correlation and lower average errors, while there is no statistical evidence to confirm significant disparities between these groups and others. Nonetheless, this trend is concerning, reflecting the perceived societal privilege often associated with these groups. In an ideal scenario, item difficulty estimators should perform uniformly well across all groups. Bucking this problematic result, those students who are identified as rural are fit better in the model than the more affluent, high performing students living in the suburbs. However, the difference in fit between this group is small (2 points) compared with that of gender (5 points) and race (4 points). That said random error will cause any set of point estimators to seemingly favor one group, though large sample sizes should reduce the effect of this kind of random error.

There is not sufficient data in this study to say one way or another if these methods of estimating item difficulties favor or harms any population group. However, that said, this method is flexible and there is the potential prompt optimizing algorithm could generate prompts tailored to each population group being studied. This might be done through explicitly

attempting to align the LLM to simulate a student of a particular type such as, “you are a female grade X student” or by allowing the prompt selection algorithm to find prompt optimums based on the estimated difficulties for that group.

Overall, though this is a topic for future research. It is unknown both how much the LLM model bias is a factor and to what extent it can be mitigated through these methods.

5.2 Future Research

5.2.1 Using a Less Visual and More Cognitive Item Bank

This dissertation investigates the potential of using Large Language Models (LLMs) as tools for estimating item parameters. The study focuses on Math and Science questions at 4th, 8th, and 12th-grade levels, but there are significant constraints to be acknowledged. A primary limitation is that the majority of the National Assessment of Educational Progress (NAEP) items analyzed involve visual elements. Many NAEP items incorporating visuals were not included in this analysis, and even among the selected items, reliance on pictorial features is common. Given the current limitations of visual LLMs compared to their text-based counterparts, I had to resort to using text descriptions (508 alternative text) to inform the LLM about these visual components. This misalignment in the information format may contribute to the less-than-optimal correlation between the LLM-estimated parameters and the actual item parameters. LLMs with “mixed modalities” or vision features is an area of intense ongoing research. Revisiting this study and these items further development in this field has progressed might be fruitful.

My intuition suggests that LLMs might show improved performance when evaluating text-based items. However, this is a hypothesis that needs to be tested. Notably, the LLMs

performed best with 12th-grade mathematics items. This performance bump aligned with my a-priori predictions that the LLMs would perform well on these items due to the complex linearly developed skills needed in high school-level mathematics. In contrast, many 4th-grade items seem to depend more on intuitive understanding acquired through everyday experiences rather than through structured learning. However, science items at the 12th grade level were among the worst items to predict. As a possible future application of this research, it would be useful to apply the approach to additional math or related items at a high school or college level to see if the models continue to perform well in this area.

5.2.2 Optimizing Prompts by Subject and Grade

In the pairwise prompt study explored in this paper a single prompt was selected which optimized classification in the training data across both subject and grades 4, 8, and 12. This prompt was selected through a genetic algorithm that evaluated dozens of prompts and cross bred those prompts to create generations of children prompts. This procedure was highly effective at moving the expected correct classification rate from around 55% to 62%. A more nuanced approach would have been to generate and evaluate prompts specific to each grade and subject combination would have born some additional predictive improvements.

In this dissertation, the same prompt template was used for both science and math items and that prompt performed very differently for the two sets of items with math items being much better difficulty ranked than science items. As math items were also about 4 times more populous it is possible that this divergence is largely driven by the prompt being optimized for the math items over that of the science items. This amalgamation across subjects may have inadvertently led to a less-than-optimal choice of prompts for the science questions.

Study two in this dissertation only explored two different prompts for item solving. However, these prompts resulted in the performance of the underlying model being quite divergent with large differences in how often the items were correctly solved. This was particularly telling with GPT4 which solved 96% of the items correctly under the second prompt. As these the pressure on LLMs is continuously to push them to be better problem solvers it is likely that future LLMs will continue to increase in their performance. As such, the use of even highly effective LLMs to estimate item difficulty by demonstrating incorrect responses might be explored through prompts intended to encourage the model to make mistakes by simulating non-optimal behaviors of students by prompting the models to take on certain personas such as, “as a student rushing through questions thoughtlessly how would you answer...” or through targeted knowledge gaps “you are a student that confuses the formula for a circle’s area with that of the volume how would you answer”?

5.2.3 Applying LLM Estimation Methods to Item Clones

The items released and covered by NAEP in both science and mathematics are quite diverse with topics spanning numerous courses in these subjects such as biology, environmental science, chemistry, and physics in science and arithmetic, geometry, and algebra in mathematics. While it is remarkable that LLMs can have any predictive power estimating the relative difficulty between such diverse items, this kind of pools of diverse items might not be the best application of this approach.

Considering these potential issues, a more effective approach for future research could potentially start with a larger and more granularly categorized pool of items. Within this system, comparisons would preferably be made between items that are more closely related

content-wise. Such a finely-tuned method could also be helpful for evaluating "item clones" variations of original test items. These clones can be created by automated generation tools or crafted by item writers. Although the parameters of the original (parent) item might be known, the corresponding parameters of the clones are typically not. For many applications of these clones, it is not necessary to pinpoint the exact parameters; instead, it is sufficient to ensure they reasonably approximate those of the original item.

Statistical methods can be useful in finding significant deviation in item clone parameters once the item is in circulation, but early detection of irregularities could potentially be enhanced using LLMs. The potential exploration of leveraging LLMs in this context offers the promise of catching aberrant item behavior before clones are administered, providing a proactive potential method of maintaining the integrity of an automated testing framework.

5.2.4 Combining LLM Generated Features with Traditional ML Tools

The use of collateral item information to aid in the estimation of item parameters combined with machine learning flexible models was shown by Mislevy et al. (1993). However, in their paper collateral information was generation though item directly observable item features as well as the input of expert judges. While the knowledge and skills of expert judges cannot be easily replaced, it is also costly, and has been demonstrated LLMs can provide a very affordable if less sophisticated substitute. An ideal future exploration would be to see to what extent the collateral information about items could be generated from LLMs to be used to predict item difficulties. Outside of aiding computational models in predicting item difficulties, Stout et al. (2003) suggest that having collateral information could reduce the examinee pool size needed to calibrate items.

5.2.5 Use of LLMs in Cognitive Diagnostic Assessment

This dissertation has shown that LLMs may show a remarkable ability to pairwise rank items with a correct response rate of 62%, outperforming chance-level performance. This task, which often challenges even human experts, involves synthesizing information from various knowledge sources and shows a significant cognitive feat by the LLMs. However, the precision they achieve falls short of the standards needed as a substitute for item standard item calibration protocols in a professional testing context.

They show limitations, such as decreased accuracy in the presence of information overload—as seen in N-shot examples—and a tendency to show bias towards items presented later in a pair. This bias may be a result of limited attentional focus, or a reflection of common testing practice where easier items precede more difficult ones. That said, beyond mere item ranking, these models generate explanations that could potentially contribute to more than just estimates of item difficulty.

LLMs are capable of not only attempting to solve items but outlined the reasoning involved in those attempts. Mining these reasoning attempts, it might be possible to garnish insight into the underlying reasoning steps taken by students attempting to solve these items. These reasoning steps identified by the LLM need not be obscure or complex but could simply involve outlining some of the procedural features of items required to find the solution. For example, the two items “ $21+43=?$ ” and “ $19+45=?$ ” looks largely the same and have the same answer, yet those solving the items by hand know that the second item involves increasing the 10 digit by 1 (“carrying the 1”) while the first item only involves adding each digit in place. While custom scripts can be written to identify procedural techniques such as “carrying the 1”

for a given item type, writing them and implementing them for the dozens to hundreds of procedural steps typically acquired by students across a variety of item types is a significant burden.

Yet being able to identify the procedural rules required in each item could be extremely helpful for designing tests and test performance reports that not only identify student overall abilities but precisely diagnose and report procedural weaknesses. This is of course not a new ideal in educational measurement as Cognitive Diagnostic Assessments (Bejar, 1984; Huff and Goodman, 2007; Leighton and Gierl, 2007; Sun and Suzuki, 2013; Delgado et al 2019; among many others). And many of these diagnostic procedures rely upon statistical techniques that can infer cognitive diagnostic features, which at times are difficult to interpret and act on. However, LLMs present a new opportunity. Never has such a low-cost option existed for generating a list of steps required to solve items.

By exploiting this opportunity, testing programs might be able to make significant strides in approaching the goal of providing actionable information for which teachers, students, and parents may use to bridge knowledge gaps.

5.2.6 A Reflection on the “Cognitive Capacity” of LLMs

It occurs to me that while the capabilities of LLMs are much praised and criticized it is also poorly understood. This is because LLMs are largely “black boxes” in which their internal complexity is vast, involving billions of interconnected “neurons.” They produce remarkable responses which are praised for both their readability and ability to find reasonable or correct answers. Yet open questions exist such as: “how do these models solve complex problems?”

and “do they possess internal ‘cognitive spaces’ in which to ‘think’ or is their responses entirely limited to syntactic predictions?”

While these questions seem ones for their programmers to address via examination of weights, they might be questions which are similarly difficult to address as the very difficult task of asking neurologists to identify an individual’s abilities based on their neuron patterns. If we would like to understand the capabilities, predispositions, and attributes of individuals asking those individuals various items, achievement or personality, seem to provide a better understanding of those individuals than doing brain scans.

In the same was items might be leveraged to gain insight into the underlying latent traits of LLMs, not just on how often those LLM’s are capable of generating a correct or acceptable response. The performance of LLMs on test items offers an opportunity to probe their underlying cognitive abilities and limitations. Given that LLMs tend to produce responses more like that of other LLMs than that of the general population, an in-depth analysis of their predictions could shed light on their cognitive strengths and weaknesses. Such insights are likely to be increasingly valuable as LLMs become more prevalent in educational settings and society, and as remote learning continues to grow.

Particularly, examining items where LLMs excel or struggle disproportionately, compared to students, could identify their distinctive cognitive patterns. This understanding could be helpful in the evolving landscape of educational evaluation, where tools like essays and exams face new challenges posed by rapidly advancing technologies. Detailed attention to the discrepancies between LLM and student performance could offer novel strategies to detect

and prevent fraudulent behavior in educational assessments acting to preserve the integrity and utility of tools.

Exploring how Large Language Models (LLMs) understand and solve items can provide valuable insights. An intriguing research question investigates whether LLMs use pairwise item rankings involving two distinct techniques to predict item difficulty. The key considerations are:

A. Do LLMs have pretrained embeddings of item features related to difficulty that they can leverage? For instance, some components of items might be inherently known to be more challenging, such as “division” being more difficult than “multiplication.”

B. Do LLMs perform an internal cognitive mapping of challenges involved in solving each item in an item pair, using this representation to predict difficulty? For example, the items “10% of 20 = _” and “10% of 2 = _” superficially involve similar elements (percentages, multiplication, whole numbers). If the LLM relies on just the presence of these components (method A), it might deem the first item harder due to the larger number 20 compared to 2. Typically, items involving 20 and multiplication are more challenging than those involving 2 (e.g., “2*5” vs. “20*5” or “2*345” vs. “20*345”).

If, however, the LLM is using an internal working space to solve both items before estimating difficulty (Method B), we might expect it to consider the second item, “10% of 2 =”, as more difficult. This is because the solution to the second item involves a decimal number, which it might view as a more complex concept than the whole number in the first item.

Nevertheless, this simple example does not provide sufficient evidence that LLMs have a working solution space for reasoning. It is possible that LLMs are using a series of weights tied

to the probability of multiplying single digits by fractions, which could influence their difficulty assessment more than the complexity of multiplying two-digit numbers by probabilities.

To properly investigate whether LLMs use either method, we would need to compose more nuanced items. These items should appear largely the same on the surface but differ significantly in actual solving difficulty, thereby providing a clearer basis for analysis.

Probing deeper into how LLMs navigate pairwise item comparisons could reveal the types of items LLMs excel at or struggle with in ranking relative difficulties. Additionally, leveraging LLMs to both predict item difficulty and attempt to solve items proposes a novel mechanism for evaluating their internal problem-solving processes.

An interesting follow-up study would examine the alignment between LLMs' difficulty predictions and the actual difficulties they encounter. While not definitive, such a study could shed light on the internal workings of LLMs when they predict item difficulties. For example, analyzing items that LLMs predict to be more difficult than empirical estimates from student responses—are these items indeed harder for LLMs to solve? Similarly, do items predicted to be easier than their empirical difficulties prove simpler for LLMs to handle?

This approach involves scrutinizing items that diverge in predicted and empirical difficulties, offering potential insights into how LLMs interpret and tackle problem-solving challenges.

If the LLM's difficulty predictions aligned more closely with its actual performance in attempting the item, this would suggest that the LLM has potentially fully or partially solved the item internally before providing a difficulty ranking. Conversely, if the predictions did not correspond with the LLM's ability to solve the items, this would indicate that the LLM is using embeddings of the item's features, which have some latent difficulty weights associated with

them. If the LLM leverages an internal solution space for ranking items, it would be intriguing for understanding how LLMs approach other tasks where the output is more challenging to evaluate.

5.3 Conclusion

This dissertation examines the feasibility of using Large Language Models (LLMs) for item calibration in test development to potentially reduce the extensive costs and resources currently required. The paper examines the potential for LLM models, like those from OpenAI's GPT series and Google's Gemini-Pro, to simulate human response patterns and reasoning skills, which in turn might enable them to predict test item characteristics.

The data used in this study are 388 math and science items released by NAEP. These items are selected from a much larger set of released items due to their limited dependence on visual components. This study deploys the use of prompt engineering in the form of a genetic algorithm which explores numerous variations of prompts. These variations allowed the pairwise item evaluation to move from an average predictive accuracy around 55% to one closer to 62%. The use of LLMs to calibrate items is a novel application of LLMs. Additionally, this is one of the first studies to deploy genetic algorithms in an educational setting to calibrate the performance of an LLM.

This dissertation proposes and tests two separate predictive models to explore the use of LLMs in predicting item difficulty. The first model estimates how well LLMs can perform pairwise item difficulty rankings. It finds that many of the proposed exogenous variables are statistically significant. However, the model's overall predictive strength is limited, with a maximum adjusted r-squared of less than 5%. This indicates that, while the model is statistically

significant, its ability to use the currently tested variable features to predict whether the LLM can successfully rank two items is of limited predictive power.

In this exploration, I used the ability of LLMs to solve or fail to solve achievement items as a tool to predict item difficulty. I also predicted item difficulty using computational indices such as word count, the Flesch-Kincaid index, and syllables per word. Among these, word count was the only consistently statistically significant predictor. While the individual performance of LLMs on these items was not very predictive of item difficulty, the overall predictive model was statistically significant, with a maximum adjusted R-squared of around 10%.

The final study presented in this dissertation explored the use of pairwise item difficulty comparisons as a predictor of absolute item difficulties. This method was shown to be far from random, with an average correlation between item difficulty and estimated item difficulty in the high .30s to low .40s across methods, subjects, and grades. Despite a particularly strong performance in 12th-grade math (correlation of .66), this method did not achieve the level of correlation (0.8 estimated) required for professional deployment in testing applications.

However, these methods are much lower in cost and easier to deploy compared to the standard approach, which requires hundreds of students pretesting the items. They are also simpler than the two alternative approaches often explored in the literature: using subject matter experts to rank items or employing computational methods that require training on large precalibrated item banks. While the results of this study are limited, the technology is quite new, and more powerful models are actively being developed. As such, the results of this research should be viewed as a preliminary proof of concept with the reasonable expectation

that these methods will improve in predictive ability as the large language models (LLMs) available in the marketplace also improve.

BIBLIOGRAPHY

Anderson, L.W., & Krathwohl, D.R. (2001). A taxonomy for learning, teaching, and assessing, abridged edition. Boston: Allyn & Bacon.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., ... & Wu, Y. (2023). Palm 2 technical report. arXiv preprint arXiv:2305.10403.

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of experimental psychology*, 81(1), 126.

Arvidsson, S., & Axell, J. (2023). Prompt engineering guidelines for LLMs in Requirements Engineering.

Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., & Wanamaker, W. (2014). Estimating item difficulty with comparative judgments. ETS Research Report Series, 2014(2), 1-8.

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2024). Measuring Implicit Bias in Explicitly Unbiased Large Language Models. *arXiv preprint arXiv:2402.04105*.

Ban, J. C., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A Comparative Study of On-line Pretest Item—Calibration/Scaling Methods in Computerized Adaptive Testing. *Journal of Educational Measurement*, 38(3), 191-212.

Bejar, I. I. (1981). Subject Matter Experts' Assessment of Item Statistics. ETS Research Report Series, 1981(2), i-47.

Bejar, I. I. (1983). Subject Matter Experts' Assessment of Item Statistics. *Applied Psychological Measurement*, 7, 303–310.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Benedetto, L., Cappelli, A., Turrin, R., & Cremonesi, P. (2020, June). Introducing a framework to assess newly created questions with natural language processing. In *International Conference on Artificial Intelligence in Education* (pp. 43-54). Cham: Springer International Publishing.

Benedetto, L. (2023). A quantitative study of NLP approaches to question difficulty estimation. *arXiv preprint arXiv:2305.10236*.

Benedetto, L., Cremonesi, P., Caines, A., Buttery, P., Cappelli, A., Giussani, A., & Turrin, R. (2023). A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9), 1-37.

Berger, M. P. (1992). Sequential sampling designs for the two-parameter item response theory model. *Psychometrika*, 57, 521-538.

Berger, M. P. (2017). Item-calibration designs. In *Handbook of item response theory* (pp. 3-20). Chapman and Hall/CRC.

Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E., & Nerdel, C. (2023). Assessing Student Errors in Experimentation Using Artificial Intelligence and Large Language Models: A Comparative Study with Human Raters. *arXiv preprint arXiv:2308.06088*.

Bezirhan, U., & von Davier, M. (2023). Automated Reading Passage Generation with OpenAI's Large Language Model. *arXiv preprint arXiv:2304.04616*.

Bloom, B. S. (1956). *Taxonomy of Educational Objectives*.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quinonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in public health*, 6, 149.

Borsboom, D., & Molenaar, D. (2015). Psychometrics.

Buyske, S. G. (1998). Optimal design for item calibration in computerized adaptive testing: The 2PL case. *Lecture Notes-Monograph Series*, 115-125.

Buyske, S. (2005). Optimal design in educational testing. *Applied optimal designs*, 1-19.

Brown, J. D. (1998). An EFL readability index. *Jalt Journal*, 20(2), 7-36.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.

Campbell, J. R., Hombo, C. M., & Mazzeo, J. (2000). *NAEP 1999 trends in academic progress: Three decades of student performance*. ED Pubs, PO Box 1398, Jessup, MD 20794-1398.

Coleman, E.B.: On understanding prose: some determiners of its complexity. NSFfinal report GB-2604 (1965)

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R.M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. *Journal of Educational Measurement*, 21, 113-129.

- Danielson, W. A., & Bryan, S. D. (1963). Computer automation of two readability formulas. *Journalism Quarterly*, 40(2), 201-206.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- DuBay, W. H. (2004). The principles of readability. *Online Submission*.
- Dudley, R. (2016) 'Massive' breach exposes hundreds of questions for upcoming SAT exams Reuters. August 3, 2016 <https://www.reuters.com/investigates/special-report/college-sat-security/>
- D'Souza, J. A Review of Transformer Models.
- Fernandez, G. (2003, August). Cognitive scaffolding for a web-based adaptive learning environment. In *International Conference on Web-Based Learning* (pp. 12-20). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3), 221.
- Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: implications for construct validity. *Language Testing*, 10(2), 133-170. doi.org/10.1177/026553229301000203
- Gunning, R. (1968). Readability yardsticks. *The Technique of Clear Writing*. New York: McGraw-Hill.
- Gurnee, W., & Tegmark, M. (2023). Language models represent space and time. *arXiv preprint arXiv:2310.02207*.
- Hambleton, R. K., Sireci, S. G., Swaminathan, H., Xing, D., & Rizavi, S. (2003). Anchor-Based Methods for Judgmentally Estimating Item Difficulty Parameters. LSAC Research Report Series.
- Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs. *Journal of business research*, 57(2), 98-107.
- Hassan, M. U., & Miller, F. (2019). Optimal item calibration for computerized achievement tests. *psychometrika*, 84(4), 1101-1128.
- Hassan, M. U., & Miller, F. (2021). An exchange algorithm for optimal calibration of items in computerized achievement tests. *Computational statistics & data analysis*, 157, 107177.
- He, Y., & Chen, P. (2020). Optimal online calibration designs for item replenishment in adaptive testing. *psychometrika*, 85(1), 35-55.

- Hernandez, I., & Nie, W. (2022). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*.
- Hocky, G. M., & White, A. D. (2022). Natural language processing models that automate programming will transform chemistry research and teaching. *Digital discovery*, 1(2), 79-83.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.
- Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., ... & Hu, G. (2017, February). Question Difficulty Prediction for READING Problems in Standard Tests. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1).
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Jones, D. H., & Jin, Z. (1994). Optimal sequential designs for on-line item estimation. *Psychometrika*, 59(1), 59-75.
- Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.
- Kumar, H., Musabirov, I., Reza, M., Shi, J., Kuzminykh, A., Williams, J. J., & Liut, M. (2023). Impact of Guidance and Interaction Strategies for LLM Use on Learner Performance and Perception. *arXiv preprint arXiv:2310.13712*.
- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models*. McGraw-hill.
- Krippendorff, K. (2011). Computing Krippendorff's alpha-reliability.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2022). Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*.
- Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch measurement transactions*, 7, 328.

- Liu, Y., Yang, K., Qi, Z., Liu, X., Yu, Y., & Zhai, C. (2024). Prejudice and Caprice: A Statistical Framework for Measuring Social Discrimination in Large Language Models. *arXiv preprint arXiv:2402.15481*.
- Lord, F.M., Novick, M.R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley
- Lorge, I., & Kruglov, L. (1952). A suggested technique for the improvement of difficulty prediction of test items. *Educational and Psychological Measurement*, 12(4), 554-561.
- Lorge, I., & Kruglov, L. (1953). The improvement of estimates of test difficulty. *Educational and Psychological Measurement*, 13, 34-46.
- Lorge, I., & Diamond, L. K. (1954). The Prediction of Absolute Item Difficulty by Ranking and Estimating Techniques. *Educational and Psychological Measurement*, 14(2), 365-372.
- Lu, F., Li, X., Liu, Q., Yang, Z., Tan, G., & He, T. (2007). Research on personalized e-learning system using fuzzy set based clustering algorithm. In *Computational Science-ICCS 2007: 7th International Conference, Beijing, China, May 27-30, 2007, Proceedings, Part III 7* (pp. 587-590). Springer Berlin Heidelberg.
- Lu, H. Y. (2014). Application of optimal designs to item calibration. *Plos One*, 9(9), e106747.
- Martins, T., Cunha, J. M., Correia, J., & Machado, P. (2023, April). Towards the Evolution of Prompts with MetaPrompter. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)* (pp. 180-195). Cham: Springer Nature Switzerland.
- Matelsky, J. K., Parodi, F., Liu, T., Lange, R. D., & Kording, K. P. (2023). A large language model-assisted education tool to provide feedback on open-ended responses. *arXiv preprint arXiv:2308.02439*.
- Marchant, G. J. (2015). How plausible is using averaged NAEP values to examine student achievement?. *Comprehensive Psychology*, 4, 03-CP.
- McGraw, R., Lubinski, S. T., & Strutchens, M. E. (2006). A closer look at gender in NAEP mathematics achievement and affect data: Intersections with achievement, race/ethnicity, and socioeconomic status. *Journal for Research in Mathematics Education*, 37(2), 129-150.
- Mc Laughlin, G.H.: Smog grading-a new readability formula. *J. Reading* 12(8), 639-646 (1969)
- Melican, G. J., Mills, C. N., & Plake, B. S. (1989). Accuracy of item performance predictions based on the Nedelsky standard setting method. *Educational and Psychological Measurement*, 49, 467-478.

- Mellenbergh, G. J. (1989). Item bias and item response theory. *International journal of educational research*, 13(2), 127-143.
- Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement*, 30(1), 55-78.
- Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), 100050.
- Mohajan, H. K. (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of Spiru Haret University. Economic Series*, 17(4), 59-82.
- Moore, S., Nguyen, H. A., Bier, N., Domadia, T., & Stamper, J. (2022, September). Assessing the quality of student-generated short answer questions using GPT-3. In European conference on technology enhanced learning (pp. 243-257). Cham: Springer International Publishing.
- National Center for Education Statistics (2024) National Assessment of Educational Progress: Search Questions [Data set]. U.S. Department of Education.
<https://www.nationsreportcard.gov/nqt/searchquestions>
- OpenAI (2023) Technical Report <https://doi.org/10.48550/arXiv.2303.08774>
- Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., ... & Gao, J. (2023). Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.
- Qiu, Z., Wu, X., & Fan, W. (2019, November). Question difficulty prediction for multiple choice problems in medical exams. In *Proceedings of the 28th acm international conference on information and knowledge management* (pp. 139-148).
- Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., ... & Irving, G. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.
- Rafatbakhsh, E., Ahmadi, A. Predicting the difficulty of EFL reading comprehension tests based on linguistic indices. *Asian. J. Second. Foreign. Lang. Educ.* 8, 41 (2023).
<https://doi.org/10.1186/s40862-023-00214-4>
- Raina, V., & Gales, M. (2022). Multiple-choice question generation: Towards an automated assessment framework. *arXiv preprint arXiv:2209.11830*.
- Rakshit, A., Singh, S., Keshari, S., Chowdhury, A. G., Jain, V., & Chadha, A. (2024). From Prejudice to Parity: A New Approach to Debiasing Large Language Model Word Embeddings. *arXiv preprint arXiv:2402.11512*.

Rampey, B. D., Dion, G. S., & Donahue, P. L. (2009). NAEP 2008: Trends in Academic Progress. NCEES 2009-479. National Center for Education Statistics.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). sage.

Reilly, D., Neumann, D. L., & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *American Psychologist*, 74(4), 445.

Reiss, M. V. (2023). Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. arXiv preprint arXiv:2304.11085.

Ren, H., van der Linden, W. J., & Diao, Q. (2017). Continuous online item calibration: Parameter recovery and item utilization. *Psychometrika*, 82(2), 498-522.

Rudolph, J., Tan, S., & Tan, S. (2023). War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching*, 6(1).

Ryan, J. J. (1968). Teacher judgments of test item properties. *Journal of Educational Measurement*, 5(4), 301-306.

Settles, B., T. LaFlair, G., & Hagiwara, M. (2020). Machine learning–driven language assessment. *Transactions of the Association for computational Linguistics*, 8, 247-263.

Senter, R., Smith, E.A.: Automated readability index. Technical Report, Cincinnati University, OH (1967)

Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika*, 55(3), 461-475.

Stout, W., Ackerman, T., Bolt, D., Froelich, A. G., & Heck, D. (2003). On the Use of Collateral Item Response Information to Improve Pretest Item Calibration. LSAC Research Report Series.

Thorndike, R. L. (1982). Item and score conversion by pooled judgment. *Test equating*, 309-317.

Thurstone, L.L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16(7), 433–451. <https://doi.org/10.1037/h0073357>

Tinkelman, S. (1947). Difficulty prediction of test items. *Teachers College Contributions to Education*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

- Tran, K. D., Bui, D. V., & Luong, N. H. (2023, October). Evolving Prompts for Synthetic Image Generation with Genetic Algorithm. In 2023 International Conference on Multimedia Analysis and Pattern Recognition (MAPR) (pp. 1-6). IEEE.
- Van de Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Measurement invariance. *Frontiers in psychology*, 6, 1064.
- van der Linden, W. J., & Ren, H. (2015). Optimal Bayesian adaptive design for test-item calibration. *Psychometrika*, 80(2), 263-288.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Von Davier, M. (2019). Training Optimus prime, MD: Generating medical certification items by fine-tuning OpenAI's gpt2 transformer model. *arXiv preprint arXiv:1908.08594*.
- Wang, Z., Valdez, J., Basu Mallick, D., & Baraniuk, R. G. (2022, July). Towards human-like educational question generation with large language models. In International conference on artificial intelligence in education (pp. 153-166). Cham: Springer International Publishing.
- Walsh, J. (2022). Lesson plan generation using natural language processing: Prompting best practices with OpenAI's gpt-3 model.
- White, A. D., Hocky, G. M., Gandhi, H. A., Ansari, M., Cox, S., Wellawatte, G. P., ... & Ccoa, W. J. P. (2023). Assessment of chemistry knowledge in large language models that generate code. *Digital Discovery*, 2(2), 368-376.
- Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice*, 29(3), 39-47.
- Wong, M., Ong, Y. S., Gupta, A., Bali, K. K., & Chen, C. (2023, June). Prompt Evolution for Generative AI: A Classifier-Guided Approach. In 2023 IEEE Conference on Artificial Intelligence (CAI) (pp. 226-229). IEEE.
- Wu, X., He, X., Liu, T., Liu, N., & Zhai, X. (2023, June). Matching exemplar as next sentence prediction (mensp): Zero-shot prompt learning for automatic scoring in science education. In International Conference on Artificial Intelligence in Education (pp. 401-413). Cham: Springer Nature Switzerland.
- Yao, T. (1991). CAT with a poorly calibrated item bank. *Rasch Measurement Transactions*, 5(2), 141.
- Yahya, A. A., Toukal, Z., & Osman, A. (2012). Bloom's taxonomy-based classification for item bank questions using support vector machines. In Modern advances in intelligent systems and tools (pp. 135-140). Springer Berlin Heidelberg.

Yang, K. C., & Menczer, F. (2023). Large language models can rate news outlet credibility. *arXiv preprint arXiv:2304.00228*.

Yaneva, V., Baldwin, P., & Mee, J. (2019, August). Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications* (pp. 11-20).

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023, April). Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-21).

Zheng, Y. (2014). New methods of online calibration for item bank replenishment. University of Illinois at Urbana-Champaign.

Zimowski, M. F., Muraki, E., Mislevy, R., & Bock, R. D. (1996). Bilog-mg. *Multiple group IRT analysis and test maintenance for binary items*.

APPENDIX A: PROMPT SCORING

A.1 Binary Prompts

TABLE 28: AUGMENTED CONFUSION MATRIX

Confusion Matrix with difficulties $D_1 > D_2$ augmented with uninterpretable responses.

| | | Predicted Condition | | |
|----------------|-------|---------------------|---------------------|----------------------------|
| | | Positive | Negative | Uninterpretable |
| True Condition | True | True Positive (TP) | True Negative (TN) | True Uninterpretable (TU) |
| | False | False Positive (FP) | False Negative (FN) | False Uninterpretable (FU) |

Traditional Accuracy (A) of a model is defined as:

$$A = \frac{TP+FN}{Attempts} = \frac{TP+FN}{TP+FN+TN+FP}$$

Prompts will be scored based on the uninterpretable accuracy of its classification.

Uninterpretable accuracy being defined as:

$$Parsable Accuracy = \frac{TP+FN}{Attempts} = \frac{TP+FN}{TP+FN+TN+FP+TU+FU}$$

Responses that cannot be parsed are first repeated either until a response is able to be parsed or a reasonable number of attempts (between 3 and 10) have been made. Responses that do not achieve an interpretable response are considered incorrect in this study, regardless of whether the LLM provides an accurate answer. The reason for this is that a response that cannot be interpreted, cannot be used for analysis.

To avoid the effects of the order in which items are presented, items are tested in both possible sequences (first D1, then D2, and vice versa). In this context, measuring the LLM's performance based solely on accuracy is appropriate. When the LLM is simply making random guesses, we know that its non-predictive accuracy is 50% minus the probability of generating an uninterpretable response. If all responses were interpretable, the chances of a randomly correct answer would be 50%. Because we are presenting items equally in both orders, and the responses are binary, using accuracy as a metric does not cause the issues that it might in traditional machine learning applications, where predicting skewed outcomes can lead to misleadingly high accuracy rates³.

Alternative Measure of Performance: F₁-Score

There are other measures of model performance worth considering. One model is the F₁ score defined as:

$$F_1 = \frac{2TP}{2TP + FN + FP}$$

In the case of ranking two items against each other it is easy to show that the Accuracy (A) converges on the F-score (F₁) when the model is non-predictive (random). Under a random model ignoring uninterpretable responses:

$$E(TP) = E(TN) = E(FP) = E(FN) = \frac{1}{4}$$

³ Accuracy can be misleading when binary outcomes are either very likely or very unlikely to occur. For instance, if an event occurs in the data 90% of the time, a model that predicts an event will always occur ($P(Y=1) = 1$) would seem 90% accurate, even though it actually has no predictive ability.

Therefore, Accuracy would equal:

$$A = \frac{1/2}{1} = 1/2$$

Which, happens to be equal to:

$$F_1 = \frac{2(1/4)}{2(1/4) + 1/4 + 1/4} = 1/2$$

We can see that in the case when responses symmetrically evaluated, and the model performs randomly the expected value of the model is $A = F_1$. As such I instead opt to use the more intuitive measure, Accuracy for evaluating prompt performance.

A.2 Non-Binary Prompts

Prompt formats in which more than two items are ranked for difficulty by the LLM by splitting the items into rank pairs each of which is evaluated for correctness (for example: Rank items: 1,2,3 returns 3,1,2 then the following binary comparisons are checked 3>1, 3>2, and 1>2. Symmetry will also be maintained with the prompt receiving the items in reverse order 3,2,1 to be ranked.

Leveraging Uncertainty

I will reserve for later research potential research applications leveraging uncertainty (See Table 2) with Uncertainty being defined as:

$$|D_1 - D_2| > \varepsilon \text{ where } \varepsilon > 0 \text{ and "small"}$$

In this kind of analysis predicted uncertainty would be a response option available to the LLM as one of three options (1: $D_1 > D_2$, 2: $D_2 > D_1$, 3: $D_1 \sim D_2$) with \sim representing undetermined or undeterminable (ambivalence).

TABLE 29: CONFUSION MATRIX WITH AMBIVALENT WITH DIFFICULTIES $D_1 > D_2$

| | | Predicted Condition | | | |
|----------------|-----------|-------------------------|-------------------------|--------------------------------|---------------------------|
| | | Positive | Negative | Uninterpretable | Ambivalent |
| True Condition | True | True Positive (TP) | True Negative (TN) | True Uninterpretable (TU) | True Ambivalent (TA) |
| | False | False Positive (FP) | False Negative (FN) | False Uninterpretable (FU) | False Ambivalent (FA) |
| | Uncertain | Uncertain Positive (UP) | Uncertain Negative (UN) | Uncertain Uninterpretable (UU) | Uncertain Ambivalent (UA) |

APPENDIX B: PROMPT DESIGN - RELATIVE DIFFICULTY EVALUATION

It has been widely reported that LLMs tend to be highly sensitive to prompt specifications. These specifications have been thoroughly explored with numerous potential factors influencing outcomes. Arvidsson and Axell (2023) provide ten themes for prompt engineers to focus on and recommendations including: context, persona, template, reasoning steps, etc. while Zamfirescu-Pereira et al. (2023) explore some common pitfalls of writing prompts.

This research will build individual prompts templates from a high-level Meta-Template (Figure 7). Some values of this template are filled in for a specific prompt while other values are taken from the content of the items being evaluated. Each of the variables on the Meta-Template level identified with the double curly brackets ({{variable}}) are variables on the template level while single curly brackets ({variable}) are variables on the item level. Variables on the template level may have components within them which are item invariant such as {{Task Introduction}}, which vary on the item level such as {{Item.Context.1}}, or even on the Meta-Template level such as {{Task Output}} which has variable values for the variable names {{Question 1}} and {{Question 2}}.

The Meta-Template is a high-level template that is used to construct individual prompt templates. Variable values are identified with curly brackets {} and are on either of two levels. A single bracket {variable} indicates a variable on the item level while a double curly bracket {{variable}} indicates a variable on the template level. Variables which are italicized are variables which are “optional” or for which their values can take an empty value.

In the Meta-Template there is a total of nine different variables which can be specified on the template level. Each of these variables has a minimum of 3 and up to 8 levels available which I have generated (Tables B.1-B.12). These levels have millions of potential combinations. An example of an individual prompt template might look can be seen in Figure 8. Of which even sampling 1% of them would be both excessively time consuming and costly. Instead, I generate

20 prompt templates. These templates are random combinations of the variable values with each non-default values (white background values) appearing twice and the default values (gray background) appearing for all other cases.

FIGURE 8: AN EXAMPLE OF A PROMPT TEMPLATE READY FOR ITEM CONTENT

| |
|--|
| <p>Assessing Question Difficulty</p> <p>You are a professional item writer.</p> <p>Below are two questions presented for difficulty assessment. Please consider factors such as the complexity of the concepts involved, the amount of information required to answer, and the level of critical thinking or problem-solving skills needed.</p> <p>Outline the process for resolving this relative difficulty ranking issue.</p> <p>Grade: {Grade}, Subject: {Subject}</p> <p>Question A:</p> <p>Bloom's Taxonomy Classification: {Bloom}</p> <p>{Item.Body.1}</p> <p>Question B:</p> <p>Bloom's Taxonomy Classification: {Bloom}</p> <p>{Item.Body.2}</p> <p>After reviewing the content and demands of both questions, determine which question poses the greater challenge.</p> <p>Please provide your evaluation in the following format:</p> <p>"The more difficult question is: Question A/Question B"</p> <p>Model Parameter: Temperature = 0.6</p> |
|--|

The variable {{N-Shot}} refers to items which are given to the LLM as examples of success. The N in N-Shot is the number of examples offered. In this study I test three options: 0-Shot (the default), 1-Shot, and 2-Shot. Each Subject, Grade/Age group has four items randomly selected to be reserved for N-Shot examples. These four items are all from the same year and

compared internally within the group for the example but will generally not share the same year with other items from the same.

The content of Tables 30 through 42 as they enter the template are filled from a combination of invariant content, the item specific content from the source NAEP items, or generated through use of a first stage LLM. Those values calculated in the first stage model are item values that appear in Table 38. The next section of this appendix will go into detail on how those values are calculated.

TABLE 30: VARIANTS ON {{TASK TITLE}}

The first row of this prompt variant is backgrounded in grayed indicating it is the “default value.”

| | |
|---|--|
| 0 | "" |
| 1 | Assessing Question Difficulty |
| 2 | Comparative Difficulty Evaluation |
| 3 | Difficulty Ranking of Questions |
| 4 | Challenge Assessment of Two Questions |
| 5 | Complexity Analysis for Question Pairs |

TABLE 31: VARIANTS ON {{PERSONA}}

| | |
|---|---|
| 0 | "" |
| 1 | You are a professional item writer. |
| 2 | You are an experienced educational psychologist specializing in test difficulty analysis. |
| 3 | You are an academic researcher with expertise in educational assessments. |
| 4 | You are a seasoned teacher familiar with standard testing practices. |
| 5 | You are a curriculum specialist with a focus on standardized test development. |
| 6 | You are a data analyst with a background in comparative educational metrics. |
| 7 | You are a grade school tutor. |

TABLE 32: VARIANTS OF {{TASK INTRODUCTION}}

| | |
|---|--|
| 0 | Below are two questions. Review these questions and rank them in terms of difficulty. |
| 1 | Below are two questions presented for difficulty assessment. Please consider factors such as the complexity of the concepts involved, the amount of information required to answer, and the level of critical thinking or problem-solving skills needed. |
| 2 | You will find two different questions ahead. Evaluate their difficulty based on depth and breadth of knowledge required, and cognitive demand. |
| 3 | Assess the complexity of each question by examining their underlying concepts and the intricacy of the answers they require. |
| 4 | Presented here are two inquiries. Please appraise their level of difficulty taking into account the scope of understanding and analysis needed. |
| 5 | Compare the following two questions and gauge which one necessitates a higher intellectual effort for resolution. |

TABLE 33: VARIANTS ON {{PERSPECTIVE}}

| | |
|---|---|
| 0 | "" |
| 1 | Evaluate the content from the perspective of a {grade}th grade student. |
| 2 | Evaluate the content from the perspective of a {grade - 2}th grade student. |
| 3 | Evaluate the content from the perspective of a { max(grade - 4, 1)}th/st grade student. |
| 4 | Evaluate the content from the perspective of a {max(grade - 8, 1)}th/st grade student. |

TABLE 34: VARIANTS OF {{N-SHOT EXAMPLES}}

| | |
|---|---|
| 0 | "" |
| 1 | <p>Here is an example of two items that have been correctly ranked:</p> <p>Example {{Question 1}}: {Example Item.Body.1 Content}</p> <p>Example {{Question 2}}: {Example Item.Body.2 Content}</p> <p>{{Task Output}} with correct {{Question 1}}/{{Question 2}}</p> |
| 2 | <p>"Here is an example of two items that have been correctly ranked:</p> <p>Example {{Question 1}}: {Example Question 3 Content}</p> <p>Example {{Question 2}}: {Example Question 4 Content}</p> <p>{{Task Output}} with correct {{Question 1}}/{{Question 2}}"</p> <p>"Here is a second example of two items that have been correctly ranked:</p> <p>Example {{Question 1}}: {Example Question 3 Content}</p> <p>Example {{Question 2}}: {Example Question 4 Content}</p> <p>{{Task Output}} with correct {{Question 1}}/{{Question 2}}"</p> |

TABLE 35: VARIANTS OF {{TASK APPROACH INSTRUCTIONS}}

The purpose of these prompts is to prompt a step-by-step, or "Chain of Thought," reasoning process. This method has been demonstrated to enhance the performance of LLMs when tackling tasks that require complex reasoning (Wei et al., 2022).

| | |
|---|--|
| 0 | "" |
| 1 | List the steps involved in solving this relative difficulty ranking problem. |
| 2 | Enumerate the procedures required to tackle this relative difficulty ranking problem. |
| 3 | Outline the process for resolving this relative difficulty ranking issue. |
| 4 | Detail the sequence of actions needed to solve this relative difficulty ranking challenge. |
| 5 | Describe the method to be followed in addressing this relative difficulty ranking task. |
| 6 | Provide the roadmap for navigating through this relative difficulty ranking problem. |

TABLE 36: VARIANTS OF {{ITEM SHARED CONTEXT}}

| | |
|---|--|
| 0 | "" |
| 1 | Test: NAEP |
| 2 | Grade: {Grade} / Age: {Age} |
| 3 | Subject: {Subject} |
| 4 | Grade: {Grade}, Subject: {Subject} |
| 5 | Grade: {Grade}, Subject: {Subject}, Year: {Year} |

TABLE 37: VARIANTS OF QUESTION NAMING ({{QUESTION 1}}/{{QUESTION 2}})

| | |
|---|---------------------------------|
| 1 | Item.Body.1, Item.Body.2 |
| 2 | Question I, Question II |
| 3 | Question i, Question ii |
| 4 | Question A, Question B |
| 5 | Question a, Question b |
| 6 | Question Alpha, Question Beta |
| 7 | First Question, Second Question |

TABLE 38: VARIANTS OF ITEM CONTEXT ({ITEM.CONTEXT.1}/{ITEM.CONTEXT.2})

This content is generated as a first stage of the item calibration from the item contents.

| 0 | "" | (Source) |
|---|---|-----------------|
| 1 | Number of Steps Required to Solve This Problem: {Number of Steps} | LLM (Item.Body) |
| 2 | Steps Required to Solve This Problem: {Steps} | LLM (Item.Body) |
| 3 | Item Complexity Rating: {Complexity} | LLM (Item.Body) |
| 4 | Item Content Tags: {Content Tags} | LLM (Item.Body) |
| 5 | Bloom's Taxonomy Classification: {Bloom.LLM} | LLM (Item.Body) |
| 6 | Bloom's Taxonomy Classification: {Bloom.NAEP} | F(Item.Context) |
| 7 | Item Context: {Item.Context} | Item.Context |

Note: The function here LLM() refers to the LLM's generative function. See Appendix C for more details. While F() refers to the mapping of Item.Context from Table 41.

TABLE 39: VARIANTS OF {{TASK INSTRUCTIONS}}

| 0 | "" |
|---|---|
| 1 | After reviewing the content and demands of both questions, determine which question poses the greater challenge. |
| 2 | In the subsequent section, two questions are listed for analysis. Your task is to deduce which involves greater difficulty for a respondent. |
| 3 | Compare the difficulties of the questions above in terms of their requisites on knowledge and reasoning. |
| 4 | Here are two questions for your consideration. Please rate them in order of difficulty. |
| 5 | Examine each question carefully to establish which one is more difficult requiring more comprehensive depth of understanding and problem-solving ability. |

TABLE 40: VARIANTS OF {{TASK OUTPUT}}

| | |
|---|--|
| 0 | The more difficult question is: {{Question 1}}/{{Question 2}} . |
| 1 | After thorough analysis, it appears that {{Question 1}}/{{Question 2}} is the more demanding question. |
| 2 | Upon evaluation, the more difficult question is determined to be {{Question 1}}/{{Question 2}} . |
| 3 | Considering all factors, {{Question 1}}/{{Question 2}} stands out as the question with greater difficulty. |
| 4 | The assessment concludes that {{Question 1}}/{{Question 2}} poses a higher level of difficulty. |
| 5 | Based on the analysis, I have determined that the question which presents the most complexity is {{Question 1}}/{{Question 2}} . |

TABLE 41: VARIANTS OF {{PROMPT TEMPERATURE}}

| | |
|---|-----|
| 0 | 0.4 |
| 1 | 0 |
| 2 | 0.2 |
| 3 | 0.6 |
| 4 | 0.8 |
| 5 | 1 |

Some of the template design options require additional content related to the items generated at an earlier stage (Table 38). This information might help in the task of predicting item difficulty. How this information is generated will be explored in Appendix C.

TABLE 42: BLOOM'S TAXONOMY MAPPED TO NAEP CONTENT TAGS

This table shows a mapping from the NAEP Content Tags to Bloom's Revised Taxonomy (Anderson and Krathwohl, 2001). The map is imprecise as the original NAEP appears somewhat arbitrarily assigned.

| Bloom's Taxonomy | NAEP Content Tag |
|------------------|--|
| Remember: | locate/recall |
| | knowing |
| | low |
| | historical knowledge and perspective |
| Understand: | comprehends what is read |
| | conceptual understanding |
| | forming a general understanding |
| | interprets what has been read |
| | identifying/describing |
| | moderate |
| | understanding |
| Apply: | applying |
| | practical reasoning |
| | problem solving |
| | using science principles |
| | using scientific inquiry |
| Analyze: | analyzes what has been read |
| | developing interpretation |
| | examine content and structure |
| | examining content and structure |
| | explaining and analyzing |
| | identifying science principles |
| | scientific investigation |
| Evaluate: | critique/evaluate |
| | evaluating and analyzing |
| | evaluate, take, defend |
| | historical analysis and interpretation |
| | reasoning |
| Create: | integrate/interpret |
| | making reader/text connections |

APPENDIX C: PROMPT DESIGN - COLLATERAL ITEM INFORMATION

In this dissertation several of the models make use of collateral item information some of it sourced from large language models (LLMs). One use is in the contents of prompts for Table 37. Following Mislevy et al. (1993), I will refer to this information as “collateral information”. Unlike in the relative item difficulty estimation, with collateral item information I do not have empirically validated values to compare the LLM generated values against. As such, I specify a standard for each piece of collateral information generated.

It is not clear to me what the best scoring criteria for generating the variable {Steps} (steps required to solve a question). Kumar et al. (2023) explores four different prompt strategies with LLMs interacting as a tutor with students. They find that while some slight variation existed between student outcomes based on the prompt strategy used, overall, the differences were slight. My preliminary explorations on this matter seem to suggest that a general criterion for these variables would be: “concise but sufficient.”

This contrasts with the variables {Complexity} and {Bloom.LLM} where we are dealing with a bias of overconfidence of LLMs. This bias is distinct from the commonly known bias in which models will hallucinate reasonable sounding responses that have no real-world bearing (Huang et al., 2023). This confidence is based on “believing” a {grade-N} students should know something that only a fraction of student at that grade actually know. Preliminary experiments with prompting the LLM to evaluate the complexity of items from a {{grade-N} - C} perspective seems promising, where {C} is an integer greater than 1.

TABLE 43: COLLATERAL INFORMATION GENERATED IN A FIRST STAGE

| Variable | Description | Scoring Criteria |
|-------------------|--------------------------------------|--|
| {Steps} | Steps Required to Solve This Problem | Concise but sufficient |
| {Number of Steps} | Number of Steps Required to Solve | Count from {Steps} |
| {Complexity} | Item Complexity Rating | Distribution of Higher-Level Complexity Scores |
| {Content Tags} | Item Content Tags | Concise but sufficient |
| {Bloom.LLM} | Bloom's Taxonomy Classification | Distribution of Higher Bloom Scores |

APPENDIX D: PROMPT SELECTION ALGORITHMS

D.1 Relative Difficulty Estimation – Non-Binary Limited Genetic Algorithm

This research deploys a multistage prompt selection procedure. The most common approach to prompt authoring is a design and debug method in which manually written prompts are passed to a LLM individually and responses are graded by the prompter. Under this framework the top performing prompts are selected for use. However, it has been widely noted that prompt performance can be greatly improved through the use of prompt optimization algorithms. A common algorithm which has been explored to optimize prompt construction is a genetic algorithm (Martins et al., 2023, Tran et al., 2023, Wong et al., 2023 among others).

Prompt selection for the relative difficulty ranking in this paper will follow a similar approach by generating a set of parent prompts, evaluating those prompts, and comparing against known outcomes. The top 50% of performing prompts will then “cross-breed” exchanging attributes and producing offspring. These offspring will then be evaluated and bred. This process will continue for a number of generations, until new offspring do not seem to offer an improvement over previous generation. As interacting with LLMs is costly in time and money the population size is far smaller and the number of generations far less than that typically used in a genetic algorithm⁴.

Final prompt selection will follow the protocol outlined in Table 44.

PROMPT SELECTION PROTOCOL

Using training data I take the following steps.

⁴ The number of generations suggested for genetic algorithms tend to be in the hundreds to thousands for many applications.

1. **Initial Prompt Models:** Prompt models were constructed varying the various features of the prompt model defined in Appendix C with each non-default attribute randomly being selected three times with the default attribute being selected for the remaining pairs. This created an initial pool of twenty-four prompt models across the 12 attributes.
2. **Generation 1:** Using a small number (30) of random item pairs selected from high contrast items (items in the top 50% of differences in difficulty, 19 percentage points different or greater) items were evaluated for relative difficulty estimation. Prompts which did not get a parable response were attempted two additional times to see if the response could be improved.
3. **Generation 2:** Using a genetic algorithm procedure, the top ten prompts from the last generation were selected for further evaluation. Additionally, there was a crossbreeding of approximately thirty children, removing any duplicates, from these ten parents. The surviving parents and roughly thirty children were then evaluated against a new 30 random high contrast items in addition to the 30 random items evaluated in Generation 1.
4. **Generation 3-4:** The procedure outlined in Generation 2 was repeated for Generations 3 and 4 except in addition to the thirty new high contrast items, an equal number of low contrast (items in the bottom 50% of differences in difficulty, less than 19 percentage points).
5. **Final Selection:** Any remaining unevaluated item pairs in the training data are evaluated by the top ten performing templates. The top prompt from these ten is selected to be evaluated against the testing data.

D.2 Collateral Information and Model Answering Ability

For the generation of collateral information and the evaluation of the ability of diverse LLM models to solve items this research will deploy an iterative improvement method. I will start with a base prompt then score it on a subset of items. Performance score for item complexity and bloom's taxonomy will be based on the following score function:

$$Score = - \sum_g^G (\text{count}(g) + \text{count}(\text{unparsable}))^2$$

Where $\text{count}(g)$ is the count of those item values in each of the categories of G . When all of the item responses are uninterpretable the maximum of this function occurs when there is an equal number of items assigned to each complexity score. Uninterpretable output is heavily penalized by adding to the negative score factor for each group.

New features will be added to the base prompt. If the score improves, those features will be kept. If it does not improve those features will be discarded. As part of this research, I will track the iterative changes to the collateral generating prompts over time.

D.3 Prompt Performance Adjustments

Prompt design for evaluating the relative difficulty of items is elaborated in detail in Appendix B. However, this is only a plan. If the prompt generator template (meta-prompt) does not perform as expected in the training data then I plan to adjust the template design with the hope of finding a better performing prompt. Any changes in this way will be documented in the final paper.

FIGURE 9: COMPLEXITY TEMPLATE

| |
|--|
| <p>How complex would the following problem be for a {grade-4}th/st grade student to solve?</p> <p>{Item.Body}</p> <p>Provide a complexity value for a {grade-4}th/st grader attempting to solve this problem as a response (Very Low Complexity, Low Complexity, Medium Complexity, High Complexity, Very High Complexity) by completing the following statement.</p> <p>"The complexity of the problem is:"</p> |
| Model Parameter: Temperature = 0.4 |

APPENDIX E: SENTITIVITY ANALYSIS

E.1 Alternative Model Specifications of Correct Classification

TABLE 44: LOGISTIC REGRESSION MODEL

The dependent variable in this model is the likelihood of correctly predicting the relative difficulty of items using a logistic regression model.

| Model: | 1 | 2 | 3 | 4 | 5 |
|--|-----------|-----------|-----------|-----------|-----------|
| No. observations: | 3650 | 3650 | 3650 | 3650 | 3650 |
| Psuedo-R2 | 0.042 | 0.046 | 0.047 | 0.052 | 0.053 |
| Coefficient Estimates: | | | | | |
| Intercept | -0.256*** | -0.290*** | -0.304*** | -0.967*** | -0.240*** |
| Second Harder | 0.446*** | 0.440*** | 0.499*** | 0.510*** | 0.461*** |
| Diff Δ | 0.025*** | 0.025*** | 0.025*** | 0.023*** | -0.027** |
| Grade : 12 | | 0.371*** | 0.360*** | 0.378*** | |
| Grade : 8 | | 0.007 | -0.009 | 0.028 | |
| Math | | -0.279** | -0.275** | -0.216** | |
| Alpha 1 | | | -0.064** | -0.069** | |
| Alpha 2 | | | 0.040* | 0.036* | |
| Complexity 1 | | | | 0.032* | |
| Complexity 2 | | | | 0.018* | |
| Complexity Δ | | | | 0.096*** | |
| Number of steps 1 | | | | 0.035* | |
| Number of steps 2 | | | | 0.003 | |
| Number of Steps Δ | | | | 0.034* | |
| Similarity | | | | -0.01 | |
| Diff Δ : Complexity 1 | | | | | 0.002** |
| Diff Δ : Complexity 2 | | | | | 0.003*** |
| Diff Δ : Complexity Δ | | | | | 0.005*** |
| Diff Δ : Number of steps 1 | | | | | 0.002** |
| Diff Δ : Number of steps 2 | | | | | 0.001 |
| Diff Δ : Number of steps Δ | | | | | 0.003** |
| Diff Δ : Alpha 1 | | | | | -0.002* |
| Diff Δ : Alpha 2 | | | | | 0.000 |
| Diff Δ : Similarity | | | | | -0.001** |
| *10% significance, **5% significance, ***1% significance | | | | | |

TABLE 45: MULTILEVEL HIERARCHICAL MODEL PREDICTING

The dependent variable in this model is the likelihood of correctly predicting the relative difficulty of items. The table shows the results of a multi-level hierarchical model with a random coefficient on the grouping level at the Diff Δ level this group was defined rounding $|\text{Diff } \Delta|$ by 5 with any values greater than 65 being assigned to the same group. Explanatory variables have been standardized to ease in interpretation.

| Model: | 1 | 2 | 3 | 4 | 5 |
|--|----------|----------|----------|----------|----------|
| No. observations: | 3650 | 3650 | 3650 | 3650 | 3650 |
| Pseudo-R ² | 0.046 | 0.050 | 0.051 | 0.057 | 0.056 |
| Coefficient Estimates: | | | | | |
| Intercept | 0.572*** | 0.564*** | 0.558*** | 0.551*** | 0.568*** |
| Second Harder | | | | | |
| $ \text{Diff } \Delta $ | 0.087*** | 0.085*** | 0.085*** | 0.080*** | -0.074** |
| Grade : 12 | | 0.077*** | 0.075*** | 0.080*** | |
| Grade : 8 | | 0.000 | -0.004 | 0.004 | |
| Math | | -0.061** | -0.060** | -0.048** | |
| Alpha 1 | | | -0.017** | -0.017** | |
| Alpha 2 | | | 0.010* | 0.009* | |
| Complexity 1 | | | | 0.018** | |
| Complexity 2 | | | | 0.006* | |
| $ \text{Complexity } \Delta $ | | | | 0.042*** | |
| Number of steps 1 | | | | 0.011* | |
| Number of steps 2 | | | | 0.000 | |
| $ \text{Number of Steps } \Delta $ | | | | 0.009* | |
| Similarity | | | | | |
| $ \text{Diff } \Delta $: Complexity 1 | | | | | 0.052*** |
| $ \text{Diff } \Delta $: Complexity 2 | | | | | 0.057*** |
| $ \text{Diff } \Delta $: $ \text{Complexity } \Delta $ | | | | | 0.065*** |
| $ \text{Diff } \Delta $: Number of steps 1 | | | | | 0.026** |
| $ \text{Diff } \Delta $: Number of steps 2 | | | | | 0.007 |
| $ \text{Diff } \Delta $: $ \text{Number of steps } \Delta $ | | | | | 0.020** |
| $ \text{Diff } \Delta $: Alpha 1 | | | | | -0.014** |
| $ \text{Diff } \Delta $: Alpha 2 | | | | | 0.001 |
| $ \text{Diff } \Delta $: Similarity | | | | | -0.014* |
| *10% significance, **5% significance, ***1% significance | | | | | |

TABLE 46: ROOT MEAN SQUARE ERROR AND MEAN ABSOLUTE ERRORS

| Grouping | Root Mean Squared Error | | Mean Absolute Error | |
|---------------------------------------|-------------------------|-----------------------------|-------------------------|-----------------------------|
| | Pooled Joint Estimation | Known Item Parameters (1PL) | Pooled Joint Estimation | Known Item Parameters (1PL) |
| Pooled Joint Estimation | 0.000 | 0.363 | 0.000 | 0.311 |
| Known Item Parameters (1PL) | 0.363 | 0.000 | 0.311 | 0.000 |
| All Students | 1.163 | 1.292 | 0.930 | 1.061 |
| Gender: Male | 1.211 | 1.356 | 0.979 | 1.121 |
| Gender: Female | 1.161 | 1.252 | 0.916 | 1.013 |
| Race: White | 1.202 | 1.358 | 0.976 | 1.129 |
| Race: Black | 1.103 | 1.104 | 0.866 | 0.883 |
| Race: Hispanic | 1.109 | 1.190 | 0.880 | 0.959 |
| Race: Asian/Pacific Islander | 1.190 | 1.322 | 0.963 | 1.088 |
| Location: City | 1.079 | 1.193 | 0.835 | 0.948 |
| Location: Suburb | 1.098 | 1.234 | 0.852 | 0.979 |
| Location: Town | 1.044 | 1.136 | 0.804 | 0.889 |
| Location: Rural | 1.082 | 1.216 | 0.842 | 0.967 |
| FRPL Eligibility: Eligible | 1.065 | 1.109 | 0.834 | 0.893 |
| FRPL Eligibility: Not eligible | 1.116 | 1.242 | 0.886 | 1.011 |
| FRPL Eligibility: Info. not available | 1.238 | 1.413 | 1.015 | 1.194 |
| Random normal variables | 1.410 | 1.410 | 1.120 | 1.120 |

Note: All populate parameters and difficulty estimates are standardized (mean=0, var=1).

E.2 Exploring the Consistency of Rank Correlations Across Groups

Looking at the Spearman rank correlation between the 258 items used in the testing data I find that the rank correlation of population performance on individual items has an average rank correlation of 0.96 (Table 48). This high rank correlation is fairly remarkable when considering the 6.48 average absolute performance differences that exist between demographic groups (Table 49). When looking at individual demographic pairs the high rank correlations despite large performance differences appear striking. For example, White and Black students

have a 15-point performance gap on average, yet the rank correlation between those item performances is still 0.95.

TABLE 47: RANK CORRELATION MATRIX OF POPULATION PERFORMANCE

This table shows the correlations between the rank difficulty of the 258 items in the testing data set evaluated in this dissertation.

| | All | White | Black | Hispanic | Asian/ Pacific | Male | Female | Mean |
|---------------------------|------|-------|-------|----------|-------------------|------|--------|------|
| All | | 0.99 | 0.97 | 0.97 | 0.93 | 0.99 | 0.99 | 0.97 |
| White | 0.99 | | 0.95 | 0.96 | 0.93 | 0.99 | 0.98 | 0.97 |
| Black | 0.97 | 0.95 | | 0.97 | 0.90 | 0.95 | 0.97 | 0.95 |
| Hispanic | 0.97 | 0.96 | 0.97 | | 0.93 | 0.96 | 0.97 | 0.96 |
| Asian/Pacific Islander | 0.93 | 0.93 | 0.90 | 0.93 | | 0.91 | 0.94 | 0.92 |
| Male | 0.99 | 0.99 | 0.95 | 0.96 | 0.91 | | 0.96 | 0.96 |
| Female | 0.99 | 0.98 | 0.97 | 0.97 | 0.94 | 0.96 | | 0.97 |
| Mean | 0.97 | 0.97 | 0.95 | 0.96 | 0.92 | 0.96 | | 0.96 |

TABLE 48: ABSOLUTE DIFFERENCE IN POPULATION PERFORMANCE

This table shows the average difference in performance by population group for the 258 items evaluated in this dissertation.

| | All | White | Black | Hispanic | Asian/ Pacific | Male | Female | Mean Abs |
|---------------------------|--------|--------|-------|----------|-------------------|--------|--------|-------------|
| All | | -4.10 | 10.85 | 7.59 | -7.12 | -1.40 | 1.13 | 4.60 |
| White | 4.10 | | 14.95 | 11.70 | -2.82 | 2.54 | 5.12 | 5.89 |
| Black | -10.85 | -14.95 | | -3.26 | -17.59 | -12.31 | -9.73 | 9.81 |
| Hispanic | -7.59 | -11.70 | 3.26 | | -14.74 | -9.11 | -6.54 | 7.56 |
| Asian/Pacific Islander | 7.12 | 2.82 | 17.59 | 14.74 | | 5.47 | 8.10 | 7.98 |
| Male | 1.40 | -2.54 | 12.31 | 9.11 | -5.47 | | 2.53 | 4.77 |
| Female | -1.13 | -5.12 | 9.73 | 6.54 | -8.10 | -2.53 | | 4.74 |
| Mean Absolute | 4.60 | 5.89 | 9.81 | 7.56 | 7.98 | 4.77 | 4.74 | 6.48 |

APPENDIX F: ALGORITHMS ILLUSTRATED

FIGURE 10: ITEM SCRAPING ALGORITHM FLOW CHART

This flow chart shows the steps involved in collecting the NAEP item statistics and primary item content from online.

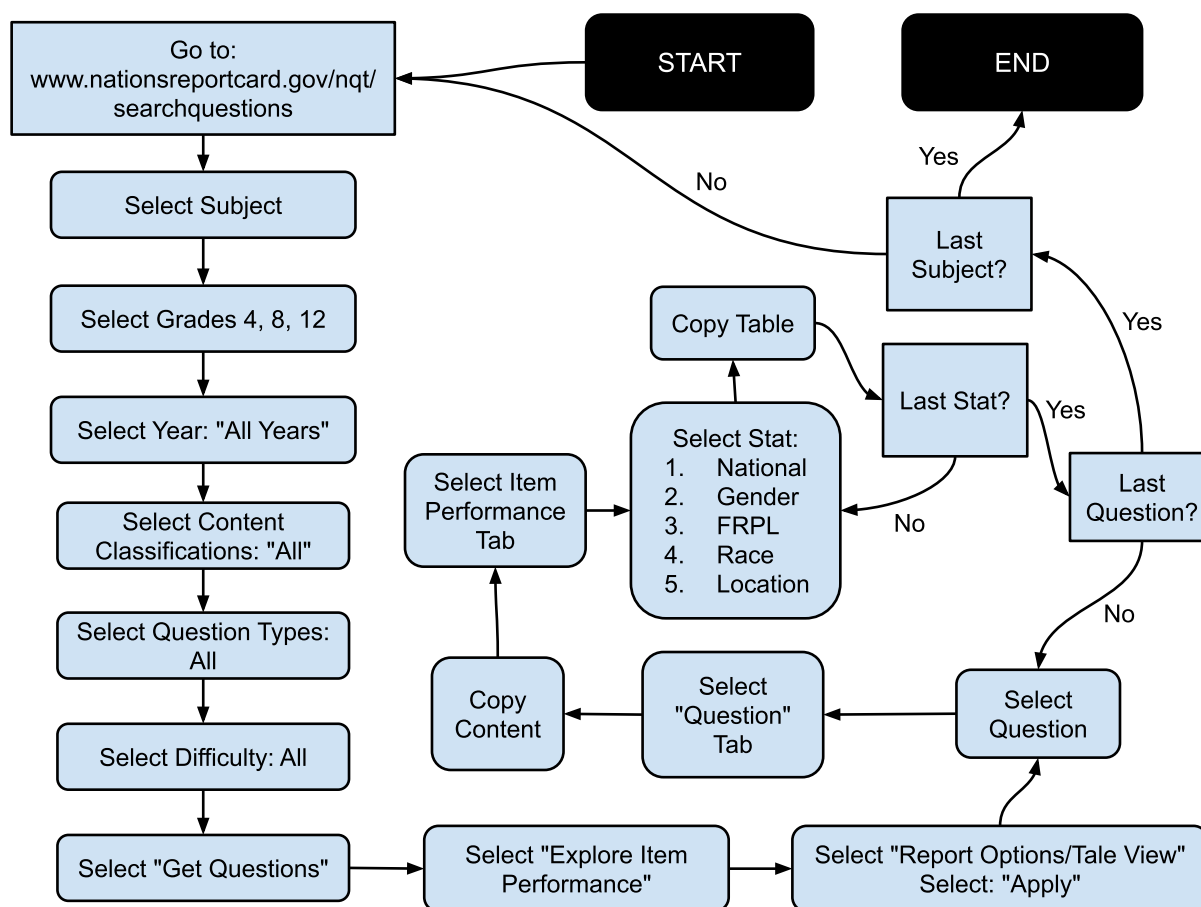


FIGURE 11: PAIRWISE TASK TEMPLATE SELECTION AND EVALUATION FLOW CHART

This flow chart shows a simplified representation of the training of top templates through evaluation of the training data followed by selection of the top template which is then evaluated on the testing data.

