The Anatomy of a High-Return Question: Text, Skills, and the Economics of Achievement Measurement*

Jonathan Moreno-Medina
University of Texas at San Antonio

Eric R. Nielsen
Federal Reserve Board

Viviana Rodriguez
University of Texas at San Antonio

March 2026

PRELIMINARY AND INCOMPLETE

PLEASE DO NOT CIRCULATE

Abstract

Standard test scales aggregate item (question) responses, obscuring the fact that some questions are far more predictive of economically relevant outcomes than others. We digitize item content for approximately 3,500 items and 1 billion student–item records linked to Texas administrative data. We document wide variation in how strongly different items predict later-life earnings, and we show that alternative achievement scales that aggregate items by these estimated "skill prices" (weights) yield estimated racial achievement gaps roughly 50% larger than conventional-score gaps while significantly reordering individual student rankings. To interpret what these prices capture, we bring modern machine learning and AI tools to the item texts. We embed each item using state-of-the-art sentence encoders and estimate a flexible text-to-price mapping. We show that the digitized item texts have substantial incremental explanatory power for variation in these item skill prices, beyond their psychometric characteristics and assessment metadata. Finally, we construct a novel, text-based mapping of items to the Common Core State Standards, highlighting the importance of spatial/geometric, rote/procedural, and automation-exposed mathematics skills in driving outcomes.

Keywords: large language models, machine learning, achievement, measurement, human capital, inequality, test items, psychometrics

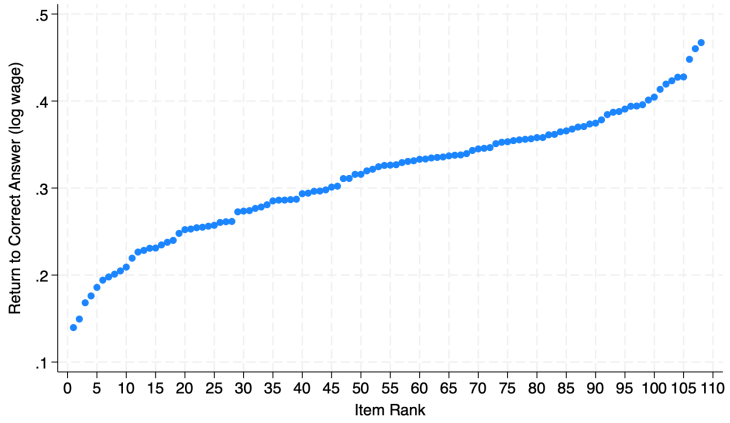
*We would like to thank Samson Alva, Peter Arcidiacono, Elliot Ash, Bocar Ba, Pat Bayer, Stephen Billings, Aimee Chin, Michael Chrzan, Ben Domingue, David Figlio, Josh Gilbert, Havisha Khurana, David Liebowitz, Elaine M. Liu, Nolan Pope, Anthony Rios, Yona Rubinstein, and Ed Rubin for helpful comments and suggestions. Similarly for the participants of the seminars at Politecnico di Milano School of Management, San Diego State University, Politecnico di Milano School of Management, University of Houston, University of Oregon, University of Texas at San Antonio, University of Pennsylvania - GSE, and University of Wisconsin-Madison. All remaining errors are our own. Jaidheer Sirigineedi, Margot Duque, Hannah Landel, and Peter Wilschke provided excellent research assistance. The views and opinions expressed in this paper are solely those of the authors and do not reflect those of the Board of Governors or the Federal Reserve System. This work was made possible through the support of the Student Upward Mobility Initiative, a sponsored project of Rockefeller Philanthropy Advisors that is led by the Urban Institute. Initiative funders include the Walton Family Foundation, Bill & Melinda Gates Foundation, and Joyce Foundation. Contact email: Moreno-Medina: jonmorenomedina@gmail.com; Nielsen: Eric.R.Nielsen@frb.gov; Rodriguez: viviana.rodriguez@utsa.edu.

1 Introduction

Standardized test scores are frequently used in social science and public policy for tracking student progress, evaluating policy, and assessing school and teacher contributions to learning. At their core, these scores collapse rich item-level responses into a single scalar via particular aggregation rules. From the point of view of social science and public policy, the problem is that the aggregation rules underpinning these scores do not leverage the fact that some questions are far more predictive of key outcomes including wages and educational attainment than others.[1]

To illustrate this, Figure 1 displays the raw log-wage difference, roughly one decade after testing, between correct and incorrect responders to each item (question) for the universe of Texas public school 8th graders in 1996. While some questions show an earnings difference of +10%, others have a difference of +50%. This variation suggests that item-level responses contain valuable information that can deepen our understanding of human capital accumulation in K–12 education. To that end, the central question we pursue is: What item characteristics explain this variation, and how can we map such explanatory factors to legible, policy-relevant skills? That is, we seek to explain and understand the variation depicted in Figure 1.

Figure 1: Item-level Raw Wage Returns for 8th Graders in 1996



Notes: This figure plots for each item (test question) administered to 8th graders in 1996, the difference in log wages at age 25 between students who answer each item correctly versus those who do not. Items are ordered in the value of this difference. Panel (a) of Figure 2 presents an analogous graph pooling all items across grade-years in our analytic sample, showing essentially the same range of differential returns.

We propose a two-stage item-anchored framework. In the first stage, adapting and improving the approach taken in Nielsen (2019, forthcoming), we estimate alternative achievement measures by weighting items according to their predictive "skill prices" for a specified outcome (e.g., wages), yielding "item-anchored" achievement measures that (1) are in interpretable outcome units and (2) are responsive to the item-outcome relationships. In the second stage, we interpret these weights (skill prices) through two complementary approaches: one that aggregates the item weights to a *detailed taxonomy of skills* defined by the Common Core State Standard (CCSS) curriculum (quantitative reasoning, reading comprehension, vocabulary, problem solving); and one that uses

[1]To be clear, this is not a criticism of psychometrics *per se*; test designers do not have data on item-outcome relationships nor do they have any remit to consider such information.

artificial intelligence methods applied to the item texts and weights to show show that these texts contain information useful for predicting outcomes above and beyond standard psychometric characteristics. Throughout, we illustrate the framework with wages as the anchor outcome, while noting that the same approach can be applied to other outcomes such as college attendance, high school graduation, and so on.

In order to construct achievement measures that highlight skills linked to economic outcomes, our paper leverages a unique combination of data sources. First, we collect and digitize over 3,500 standardized achievement test questions administered to around 12 million students in grades 3-12 in the state of Texas. We link these text data to approximately 1 billion student–item records that are also linked to longitudinal administrative data (college enrollment, completion, and earnings). Finally, we map individual test questions to detailed skill taxonomies of the CCSS, which codifies over 600 skills in Reading and Mathematics.

We leverage these data in the first stage to estimate item prices, the contribution of each test item to an outcome, expressed directly in outcome units, by projecting the anchor outcome (e.g., log wages) onto the vector of student item response to the math and reading standardized tests, together with covariates.[2] In our baseline linear specification, an item's estimated price (weight) is simply the coefficient estimate on that item's correct-answer indicator in an OLS regression of the anchor outcome on the item indicators and controls. Thus, the item's price is the expected difference in the anchored outcome between answering the item correctly versus incorrectly, holding other items and covariates fixed. We show that our item weight estimates are robust to several alternative specifications (e.g., county vs. commuting-zone fixed effects, restricting to white males, ridge/LASSO), and they provide transparent, interpretable weights that preserve item-level information for the achievement measure we use downstream in the second stage.

We illustrate how the use of our estimated skill prices has significant implications for two standard objects frequently studied in applied work: mean achievement gaps across groups and achievement ranks across individuals. Using achievement scores anchored at the item level to wages, we find white-Hispanic and white-Black gaps that are about 45% larger in relative terms as those calculated from the given, psychometric scores.[3] These differences arise because Black and Hispanic students perform worse on test items that are particularly predictive of earnings yet emphasized relatively less by the given test scale.

The item-anchored scores also significantly reorder student rankings relative to the given test scores. Within a given test score ventile, the 90% range of the item anchored scores often exceeds 2 SD. The item-anchored scores thus disagree fundamentally with the conventional scores as to which students are doing well and which are doing poorly. These empirical results and those on racial/ethnic achievement differences echo those in Nielsen (2019, forthcoming) in a new context with substantially more data and policy relevance.[4]

[2]The covariates include student demographics (race/ethnicity, gender, free or reduced-price lunch (FRPL) status), language status (ESL), year fixed effects, and commuting-zone fixed effects to capture local labor-market access.

[3]We correct for attenuation due to measurement error in the item-anchored scales using a split-half IV reliability adjustment (odd/even items) as in Nielsen (2019, forthcoming). See Section 5 for more details.

[4]Nielsen (2019, forthcoming) uses the National Longitudinal Survey of Youth 1979, a panel survey of roughly

Having constructed economically anchored item weights, we turn to the primary motivating question of the paper: What characteristics of a test item lead to a high item weight – that is, make it highly predictive of long-run outcomes? As stated above, we approach this question using two complementary methods: an item-to-curriculum matching and an "assumption-free" approach which seeks to understand which features of an item, both psychometric and text-based, predict a high weight. Together, these approaches move us from **which questions** matter to **what skills** those questions measure, providing a way to connect outcome-anchored achievement back to instructional content and to potential new skill hypotheses.

To implement the curriculum-based approach, we map each test item to a detailed list of CCSS standards based on the item's content. This step is only possible because we recover the underlying text and structure for the test items and the full list of CCSS standards by digitizing both. To our knowledge, this is the first paper to present a general, algorithmic method for mapping any item to any skill taxonomy based on commonalities in their texts/structures. We then estimate "price" of each CCSS skill as the average item weights of those items mapped to it. We find a striking concentration of high-price CCSS skills in mathematics, and, within math, a clear tilt toward procedural, multi-step computation (e.g., formula application, coordinate representations), rather than more conceptual or interpretive tasks. We further find that spatial reasoning tasks have notably higher weight than non-spatial math tasks. Finally, following Autor et al. (2003), Autor (2013), we construct a measure of the automation exposure of each CCSS skill, and we find that skills more exposed to automation (defined pre-gen-AI) tend to have higher returns. Among reading skills, identification of texts' main ideas has the highest price compared to alternatives such as analyzing tone, citing textual evidence, or determining word meanings. These results provide a first, standards-based account of which scholastic skills are most predictive of long-run outcomes in our setting.

To go beyond predefined skill taxonomies, we convert each digitized question into a machine-readable numerical representation – a location with an LLM embedding space. We then estimate a flexible prediction model that maps this representation (a high-dimensional vector) and vector of psychometric characteristics and metadata (subject, difficulty, discrimination, location in the exam, etc.) to the item's estimated price. This analysis reveals that math items and items occurring later in an exam have consistently higher skill prices (item weights) compared to reading and earlier-test items. Difficult items also have higher skill prices, although the relationship is comparatively modest. Moreover, the item text and structure can explain meaningful variation in item prices beyond the standard item metadata and psychometric characteristics, suggesting that the language of the question contains economically relevant information.

The rest of the paper proceeds as follows. Section 2 introduces the item-anchored framework and defines both anchored achievement and item prices; Section 3 covers the student item-response and earnings data as well as the collection and digitization of the item texts; Section 4 formalizes

12,000 youth who were teenagers in 1980. The test items used in that paper come from the Armed Forces Qualify Test, which was administered to each survey respondent only once.

the estimation of the first stage prices; Section 5 revisits estimated achievement gaps and individual student ranks under item-anchoring versus conventional scales; Section 6 explores the item prices via psychometric and test meta-data features of the items; Section 7 relates the prices to a flexible numerical representation of the item texts; Section 8 interprets prices via the item-common core mapping; Section 9 concludes.

1.1 Prior Literature

This paper touches on many different literatures in economics, education, psychometrics, machine-learning/AI, and policy. We give an overview of these connections in this section.

This paper is most directly connected to Nielsen (2019, forthcoming). That paper is the first to argue that (1) the psychometric aggregation of items throws out economically valuable information and (2) aggregating items based on their observed relationships with economic outcomes can paint a very different picture of both individual and group-level achievement differences. Nielsen (2023b) and Nielsen (2025b) then demonstrate that inferences about the variability of achievement are also sensitive to item aggregation, with item-anchored scores calling into question many widely-cited results.[5]

Bruhn et al. (2025) likewise argue for information loss from item aggregation using more recent cohorts of the Texas data than we do and focusing on shorter-run outcomes such as school discipline, course performance, and school completion. Bruhn et al. (2025) further show that item-level analyses significantly alter teacher value-added estimates and that, absent item content – including text –, item-specific teacher value-added patterns can be used to link substantively similar items across years/cohorts.

Our paper connects also to a burgeoning literature in psychometrics and economics on the use of item-data. On the psychometric side, these papers argue in a variety of contexts that standard scale scores may throw away useful information/nuance in how skills are measured. Gilbert et al. (2025), Ahmed et al. (2025), and Gilbert et al. (2024) show that "item-level heterogeneous treatment effects" (IL-HTE) are pervasive; interventions often differentially impact specific items or subdomains of a test, rather than affecting the latent achievement construct uniformly. Gilbert et al. (2025) proposes to use item-level covariates and sub-scale interactions to explain residual variation in the item effects. Kapoor et al. (2025) shows that LLM embeddings of item text can provide accurate predictions of item difficulty, thus giving a "proof of concept" of the approach we take in this paper to relate such embeddings to our estimates of the economic usefulness of the items. On the economics side, recent research has used item-level data to evaluate the academic returns to cognitive endurance (Reyes, 2025), identify "coachable" items (Reyes et al., 2024), as

[5] One of the results in Nielsen (2019, forthcoming) finds that item-predicted white-black differences in labor market outcomes are equal to observed gaps in these outcomes, but this does not occur for household income gaps. Another finding therein is that individual achievement ranks shift notably under different, economically-motivated aggregation schemes. Nielsen (2023b) argues that males do not consistently show greater variability in achievement using item-anchored scores, while Nielsen (2025b) shows that the variability of item-anchored achievement increases dramatically as children progress through school, with clear implications for analyses that standardize achievement scores to have a unit variance for each grade/age.

well as understand the effect of differential representation on test performance (Lee and Schaelling, 2025). Relative to these literatures, our paper differs in (1) its focus on explaining the economic importance of the items (item-outcome correlations), (2) its greater use of LLMs and machine learning approaches to both digitizing, analyzing, and interpreting the data, and (3) its focus on achievement measurement rather than treatment effect heterogeneity.[6]

Our analysis mapping items and item weights to the Common Core State Standards (CCSS) (Association et al., 2010) speaks to the extensive debate in both academic and policy circles surrounding these standards and curriculum construction more generally.[7] By revealing a completely new aspect of these standards – their widely varying correlations with earnings and other outcomes – our paper both highlights the context-dependent nature of curriculum construction and suggests that some curricular reorientation to high-estimated-return standards might be beneficial. Our exploration and classification of the high- and low-return CCSS skills relates to pre-existing literatures in psychometrics on the "depth of knowledge" (DOK) of a question (Webb, 2002, 2007) and research in economics on measuring the exposure of a skill to pre-gen-AI automation (Autor et al., 2003, Autor, 2013).

Finally our paper relates to the large literatures within economics that seek to understand the role that skills/human capital play in determining earnings, health, and many other outcomes.[8] Relative to this literature, our approach will tend to find an even greater role for skill in explaining outcomes because we generate scoring rules designed to maximize the explanatory power of the test scale for the chosen anchor outcome. Our framework, which assumes that different outcomes *by construction* correspond to different achievement scales and human capital definitions echos Deming (2023) and Deming and Silliman (2025), which both argue that human capital should be understood as multidimensional.

2 Conceptual Framework

Our empirical approach consists of the following steps. First, given student-level data on item responses, demographic and other variables, and an economic outcome, we seek to estimate an "importance vector," $\widehat{\mathbf{\Omega}}$, that measures the economic value of each item as it pertains to predicting the outcome. Second, we assess and justify the particular approach we take to estimating $\widehat{\mathbf{\Omega}}$. The

[6] Our paper contributes also to broader literatures in both economics and psychometrics about the measurement of achievement and the properties of standardized test scores. In particular, past research has argued that psychometric scales do not have a cardinal interpretation (Lord, 1975, Jacob and Rothstein, 2016, Cunha et al., 2021) and that standard results in economics might be sensitive to economically arbitrary scaling choices (Cunha et al., 2010, Bond and Lang, 2018, 2013, 2019, Schröder and Yitzhaki, 2017, Nielsen, 2025a, 2023a).

[7] See, e.g., Costrell (1997), Schmidt and Houang (2012), Porter et al. (2011), Cobb and Jackson (2011), Hiebert and Mesmer (2013), Opfer et al. (2016), Blazar et al. (2020), Hahm (2026).

[8] For earnings see: Heckman and Kautz (2012), Chetty et al. (2011), Hanushek and Woessmann (2008), Heckman et al. (2006), Altonji and Pierret (2001), Cawley et al. (2001), Neal and Johnson (1996), Murnane et al. (1995); health: Cutler and Lleras-Muney (2010), Christopher Auld and Sidhu (2005), Kaestner and Callison (2011), Mocan and Altindag (2014); criminal/antisocial behavior: (Mears and Cochran, 2013, Ttofi et al., 2016, Heckman et al., 2006, Chen et al., 2025); fertility/family outcomes: (Kolk and Barclay, 2019, Mansour and McKinnish, 2014, Shearer et al., 2002), among many others.

bulk of our empirical work then concerns understanding what item-level features lead to a low or high estimated importance.

This section lays out the basic definitions and conceptual framework underlying our approach. Subsequent sections then provide greater details on the methods specific to that section. We start here by describing how we conceptualize the notion of an "importance" or "price" for each test item. Our approach builds on the frameworks in Nielsen (2019, forthcoming) and Bond and Lang (2018).

Introducing some notation, let i index a test-taking student from some population of interest (e.g., from a particular grade/year in the Texas data). The test consists of M dichotomous items with $\mathbf{D}_i$ denoting the vector of i's item responses: $\mathbf{D}_i = [D_{i,1}, \ldots D_{i,M}]$ where $D_{i,m} = 1$ if i gets item m correct and 0 otherwise. For any particular question m, we have $\mathbf{D}_i = [D_{im}, \mathbf{D}_{i,-m}]$ where $\mathbf{D}_{i,-m}$ denotes the $M-1$ length vector of all i's item responses other than to item m.[9]

Implicit to a test scale is the choice of an item *importance* vector, $\mathbf{\Omega} = [\omega_1, \ldots \omega_M]$, containing a weight for each item in the test, which is used to aggregate the full vector of a student's item response ($\mathbf{D}_i$) into a scalar — the test score. One of the simplest scoring rules — percent correct — weights each item equally so that two students with the same number of correct responses will receive the same score regardless of which specific items they answer. Modern psychometric methods such as item response theory (IRT) aggregate in a more theoretically motivated way. For example, the widely-used 3PL IRT model, aggregates items based on the (estimated) difficulty, discrimination, and guess-ability. Two items that are the same on these three dimensions will have equal influence on a student's estimated achievement.[10]

In this paper, we instead seek a definition of $\mathbf{\Omega}$ based on the economic usefulness of each item. To do so, we adapt the item-anchoring approach in Nielsen (2019, forthcoming) which defines achievement A_i as the component of some outcome S_i that is predictable from ideal psychometric data:

$$S_i = A_i + \eta_i. \tag{1}$$

Here, η_i represents determinants of the outcome S_i that are not predictable from psychometric data. Thus, $\mathbb{E}[\eta_i A_i] = 0$ holds by construction. In practice, no test has complete and perfect psychometric data, and so the best we can hope to identify given the observed item data is $\tilde{A}_i \equiv \mathbb{E}[S_i|\mathbf{D}_i, \mathbf{X}_i]$. Thus, $\tilde{A}_i$ is defined relative to the particular items used to assess achievement. $\tilde{A}_i$ will by necessity be a noisy measure of A_i – we suppose that $\tilde{A}_i = A_i + \nu_i$, where ν_i is classical measurement error. This irreducible error ν_i, common in classical testing theory, is important in our context only for the estimation of mean differences in achievement but is not generally important for the definition or estimation of $\mathbf{\Omega}$, our primary object of interest.

[9] In keeping with standard notational conventions, we denote random variables by capital letters, specific values of random variables in lower case, and we denote vectors in bold.

[10] In IRT, the estimated achievement scores will not generally be exactly equivalent to a weighted sum of the item responses. However, for tests with a large number of items, the IRT achievement measure for a student can be well approximated by such a sum. Moreover, in some special cases, such as in the one-parameter Rasch model, the simple sum of the item responses is a sufficient statistic exactly for the MLE achievement estimate.

To actually estimate $\tilde{A}_i$, we suppose that for some known functional form f parametrized by the vector $\mathbf{\Psi}$

$$\tilde{A}_i = f(\mathbf{D}_i, \mathbf{X}_i; \mathbf{\Psi}). \tag{2}$$

Data on outcomes, test items, and controls can then be used to estimate $\mathbf{\Psi}$. The item-anchored scores are then estimated by $\hat{A}_i = f(\mathbf{D}_i, \mathbf{X}_i; \widehat{\mathbf{\Psi}}) = \widehat{\mathbb{E}}[S_i|\mathbf{D}_i, \mathbf{X}_i]$. The test scale defined by $f(\mathbf{D}_i, \mathbf{X}_i; \widehat{\mathbf{\Psi}})$ aggregates individual items based on their observed relationships with the outcome S_i, mediated through the function form assumed for f.[11] Section 4 presents the empirical details of our estimation approach.

The definition of $\mathbf{\Omega}$ that we use in this paper is based on our definition of item-anchored achievement. The weight for item m is defined to be the difference in expected item anchored achievement for those answering the item correctly versus incorrectly, where the expectation is taken over the distribution of $(\mathbf{D}_{-m}, \mathbf{X}_i)$:

$$\omega_m \equiv \mathbb{E}_{\mathbf{D}_{-m},\mathbf{X}}[f(D_m = 1, \mathbf{D}_{-m}, \mathbf{X}; \mathbf{\Psi}) - f(D_m = 0, \mathbf{D}_{-m}, \mathbf{X}; \mathbf{\Psi})]. \tag{3}$$

The definition of ω_m in equation (3) captures well the notion of economic importance: ω_m is the average incremental difference in the expected outcome S for those answering item m correctly versus incorrectly. This definition mirrors the basic structure of how test scales are formed in other contexts. That is, our method is similar to other ways of creating a test score in that it assigns some weight to each item m, but does so in a way that uses the predictive value of those items to long term outcomes; it uses an economic framework to assign weights. Instead of naively assigning equal weight to all items, or using weights driven by psychometric methods that make no use of longer-run economic outcomes, we use the informativeness of the items in predicting the economic outcome of interest.[12]

As defined, the ω_m are population objects, and we must estimate them from our sample. We do this by first estimating $\mathbf{\Psi}$, the parameter vector governing f. This allows us to compute, for each individual i, $\hat{A}_i[D_{im} = 1] = f(D_{i,m} = 1, \mathbf{D}_{i,-m}, X_i; \widehat{\mathbf{\Psi}})$ and $\hat{A}_i[D_{im} = 0] = f(D_{i,m} = 0, \mathbf{D}_{i,-m}, X_i; \widehat{\mathbf{\Psi}})$ where in both cases $\mathbf{D}_{i,-m}$ denotes the actual vector of non-m item responses for student i. We

[11] This scale, unlike traditional test scales, is also cardinal – a given change in scores ΔA corresponds to a fixed change in the predicted value of S, which is itself (by assumption) cardinally interpretable. See Nielsen (2019, forthcoming) for more details.

[12] The $\mathbf{\Omega}$ vector can be alternatively interpreted as capturing an M-dimensional representation of the projection of a latent vector of skills (L). That is, suppose that the vector of responses D_i of each student is generated by some underlying vector of skills of the student, and we want to use D_i as a proxy for that space. Prior literatures have generally employed factor model approaches to decompose skills. In psychometrics, Item Response Theory (IRT) relies on assumed factor structure for skills (e.g., Jöreskog (1969), Reise (2012), Reckase (1985)). While item response data have been seldom used within economics, the dominant conception of skills/achievement is either based directly on IRT or a conceptually similar approach that views observable test scores and behavioral outcomes as noisy measures of latent underlying skills (Heckman et al., 2006, Cunha et al., 2010, Schennach, 2022). These methods generally require the researcher to define *a priori* the correct dimension of the latent space L, and usually require strong parametric assumptions about the data generating process for the observed vector D_i. We see two main advantages of our approach to those mentioned above. First, our method does not require specifying the dimension of the latent space L in advance - which is unlikely to be known by the researcher. Second, our method is also agnostic as to the DGP generating the vector of student responses D_i.

then estimate ω_m using sample averages:

$$\hat{\omega}_m = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{A}_i[D_{im}=1] - \hat{A}_i[D_{im}=0]\right). \tag{4}$$

In the case that f is linear in the item vector, this calculation takes a much simpler form: $\hat{\omega}_m = \hat{\beta}_m$, the estimated coefficient for item m.

Once we have $\widehat{\mathbf{\Omega}}$ in hand, the main part of our analysis seeks to understand how observable characteristics of the items explain (in the statistical sense) these item weights. In particular, we consider two conceptually distinct types of item-level data. First, we denote by R_m all of the non-text characteristics of an item: its subject (e.g., math or reading), IRT parameters (difficulty, discrimination, etc.), learning objective, placement within the test (e.g. question number), etc., and we let $\mathbf{R}$ be the vector of these characteristics across the items m. Second, we denote by E_m the actual text of the item or some mathematical representation of that text, with $\mathbf{E}$ representing the vector of these data across items.[13] Then, we suppose that for some function g and an unknown error vector $\mathbf{\Xi}$

$$\mathbf{\Omega} = g(\mathbf{R}, \mathbf{E}) + \mathbf{\Xi}. \tag{5}$$

Depending on the context, we either assume a particular parametric form for g or that it is well-approximated by a neural network or any other differentiable flexible model. Because we do not observe $\mathbf{\Omega}$, we instead estimate a version of equation (5) substituting $\widehat{\mathbf{\Omega}}$ for $\mathbf{\Omega}$. In doing this, we take account of the first-stage estimation error in the weights. We will discuss the details of this adjustment, as well as other empirical implementation details, in Section 4.

3 Data

Constructing estimates of $\mathbf{\Omega}$ and item-anchored achievement requires student-level data on (1) an interpretable economic outcome (S_i), (2) item responses ($\mathbf{D}_i$), and (3) additional controls ($\mathbf{X}_i$). Furthermore, understanding what characterizes high-return items requires (4) item texts ($\mathbf{E}$), and (5) non-text item characteristics $\mathbf{R}$. To our knowledge, there are no extant data sets with all of these ingredients, for two primary reasons. First, data sources with item responses are quite uncommon. Prior research on item-anchoring (Bond and Lang, 2018, Nielsen, 2019) has relied on survey data such as the NLSY79 and CNLSY that suffer from a number of significant shortcomings. These surveys have comparatively few observations, cover in some cases only a single cohort of youth, and estimate achievement at a single age or at a small number of ages. We transcend the limitations of survey data sources by using instead rich administrative data, containing both item responses and long-term economic outcomes, covering the universe of public school students in Texas from the 1995-1996 to 2018-2019 school years.[14]

[13] We use "E" because our empirical work will use embedding space representations of the text. See Section

[14] Throughout this paper we use the Spring term year to refer to each school year.

Second, even if data sources do contain item responses, they generally do not have high quality information on the items themselves. For example, item-level information included in the Texas ERC data is limited to the item's broad learning objective, it's subject (math or reading), and it's position in the test. Thus, we supplement data in two ways. First, we collect and digitize the test booklets that were administered for statewide testing, which allow us to recover image and text information of each test item. Second, we use the content on the items to link to the taxonomy of skills defined by the Common Core State Standards (CCSS) to obtain a richer picture of skills associated with each question.

Below, we provide a high-level overview of these data sources and key variables. Please refer to Appendix A for additional details.

3.1 Item Response Data

The Texas ERC provides student-level mathematics and reading test data from statewide assessments administered to public school students in grades 3-8 and some high school grades. The purpose of these assessments is to measure the level and evolution of student proficiency and learning in Texas public schools.

The actual assessments used by the state of Texas changed twice during the period covered by our study: the Texas Assessment of Academic Skills (TAAS) (1990-2002), the Texas Assessment of Knowledge and Skills (TAKS) (2003-2011), and the State of Texas Assessments of Academic Readiness (SSTAR) (2012-present). While differing somewhat in design and subjects covered, these assessments are all designed for the same purpose, and all cover mathematics and reading. As mentioned above, in this paper we illustrate our two stage method focusing on drivers of wages – and thus we need to link student tests where enough time has passed to link to early-career adult earnings. Hence, we focus our analysis on the earliest test administration (TAAS) which allows us to observe wages at age 25 while preserving the same assessment design.[15]

A key feature of the state's data collection efforts for our purpose is that individual student item responses are available for all students and all standardized exams starting in 1996. That is, for each student-multiple choice question pair, we observe (1) the answer the student bubbled in, (2) whether their answer was correct, and (3) whether the student skipped the item. In addition to item-level student responses, we also observe the test subject (math or reading) and learning objective associated with each question.[16]

Panel B. of Table 1 presents item-level descriptive statistics for our analysis sample. Overall, the average question is answered correctly by 80% of students. Our sample contains slightly more math questions than reading questions. The raw return to a correct answer (as plotted in Figure 1)

[15] See Section A.2 for an explanation of the anchoring timelines and implications of long-term data availability.

[16] Standardized tests are generally designed to measure broad learning objectives. In Texas, the standardized tests we study were designed to consistently measure these objectives over time. See Table A.3 for a list of testing objectives for both Reading and Mathematics tests under TAAS. Our data allow us to observe these test objectives so that in addition to each student answer, we also observe which learning objective each test item corresponds to.

Table 1: Descriptive Statistics

	All			Digitized Sample		
	(1) Mean	(2) SD	(3) Obs.	(4) Mean	(5) SD	(6) Obs.
Panel A. Student-by-Grade						
Demographics						
Female	0.50	0.50	12,211,377	0.50	0.50	9,349,124
Black	0.14	0.34	12,211,377	0.14	0.35	9,349,124
Hispanic	0.35	0.48	12,211,377	0.35	0.48	9,349,124
White	0.48	0.50	12,211,377	0.47	0.50	9,349,124
Other	0.04	0.19	12,211,377	0.04	0.19	9,349,124
Economically Disadv.	0.44	0.50	12,211,377	0.45	0.50	9,349,124
ESL	0.04	0.20	12,211,377	0.04	0.20	9,349,124
LEP	0.08	0.26	12,198,422	0.08	0.27	9,349,124
Immigrant	0.01	0.10	12,211,377	0.01	0.10	9,349,124
Special Ed.	0.08	0.27	12,211,377	0.08	0.27	9,349,124
Gifted	0.11	0.32	12,211,377	0.11	0.32	9,349,124
Long-Run Outcomes						
Wage at 25	$29,693	$29,010	9,111,275	$29,932	$29,452	6,996,713
Wage at 30	$42,466	$44,479	6,717,698	$42,775	$44,295	4,731,630
Wage at 35	$51,509	$58,171	1,961,384	$51,205	$57,644	965,269
HS Graduate	0.85	0.36	10,776,458	0.85	0.35	8,255,387
Enrolled College	0.67	0.47	11,230,510	0.67	0.47	8,596,239
Panel B. Item-level						
Percent Correct	0.80	0.12	4,739	0.81	0.12	3,555
Math	0.56	0.50	4,739	0.56	0.50	3,555
Digitized	0.75	0.43	4,739	1.00	0.00	3,555
Raw Return	0.33	0.08	3,869	0.32	0.07	2,685
$\hat{\omega}$	0.01	0.02	3,869	0.01	0.02	2,685
Student-Grade-Item Responses		1,235,875,456			937,619,520	

Notes: This table presents descriptive statistics of student-by-grade-level demographics and long-run outcomes in panel A. and item-level characteristics in panel B. All earnings data are converted to 2019 dollars. Columns (1) through (3) present statistics for the universe of students in Texas from 1996-2002 who take standardized tests (grades 3-8 and the exit exam). Columns (4) through (6) present analogous statistics for the grade-year combinations for which test booklets (i.e. text data on questions) were recovered.

for the average item is 33%, while our conditional estimate, $\hat{\omega}$, indicates a 1% return.[17]

Across all items, we are able to recover the text of 75% of the questions administered between 1996-2002 (see Section 3.3.1 for more details). Thus, in columns (4) through (6) we present analogous statistics for the subsample with text data available. Descriptive statistics for this subsample does not differ in any meaningful way from the full sample. Finally, in total, we draw from upwards of 1.2 billion student responses of which 940 million were are able to link text data to.

3.2 Labor Market Outcomes Data

The Texas ERC contains student-level data on employment and earnings through a link to the State of Texas unemployment insurance system. Thus, we can can link item responses to earnings

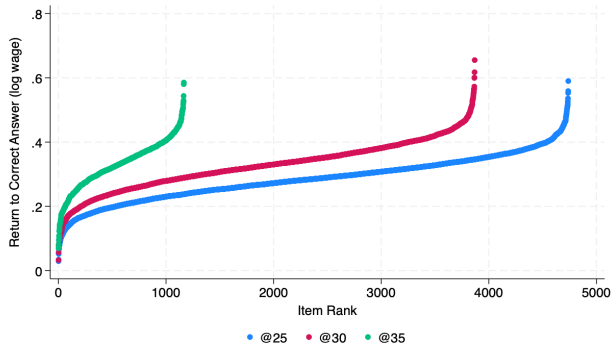
[17] See Section 2 and Section 4 for a discussion on the estimation of these returns.

for any public school student who receives wage/salary income in Texas in adulthood.[18] We cannot observe earnings for individuals who attended public school in Texas but who move out-of-state subsequently. Fortunately, Texas has the lowest outmigration rate of any U.S. state.[19] As such, we are able to recover labor market information for approximately 72% of the universe of relevant test-takers.[20]

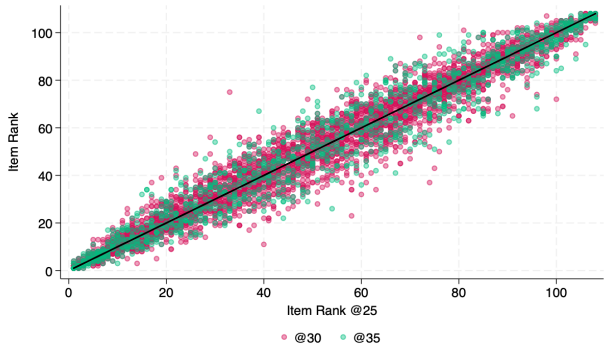
We take as our baseline "anchor" outcome adult earnings as earnings at age 25.[21] Ideally we would be able to measure student earnings at later ages in order to more accurately capture individual variation in lifetime earnings.[22] To avoid pandemic-era wage distortions, we restrict to cohorts whose earnings are observed *pre-COVID*, which prevents us from observing most of the sample at later career ages. Accordingly, we use earnings at age 25, which we we view as a reasonable early-career measure: by that age most individuals have completed schooling and entered the labor market.[23]

Figure 2: Unconditional Wage Item-level Returns for All Grades and Years

(a) Returns at Ages 25, 30, & 35



(b) Rank Correlations Across Ages



Notes: Panel (a) of this figure plots for each item (test question) administered to all students across all years, the difference in log wages (at age 25, 30, and 35) between students who answer each item correctly versus those who do not. Items are ordered in the value of this difference. We are missing earnings data for some grade-year combinations that have yet to become 30 and 35 years of age. This explains differences in sample size between item-level returns at age 25, 30, and 35. Panel (b) shows the correlation between the item rank of these returns at age 25, with returns at age 30 and 35.

To understand the implications of capturing earnings at different labor market ages, Figure 2 plots unconditional item wage returns at ages 25, 30, and 35 for all items administered to Texas public school students during our period of study. Panel (a) of Figure 2 shows that: i) the range or returns is very similar for all wage-specific returns; and ii) capturing earnings at ages 30 and 35, as opposed to 25, implies a loss in the number of items we can link to labor market earnings.[24] Panel (b) presents the correlation of item return rank across ages. This plot indicates a strong rank

[18] TWC data are reported quarterly. We aggregate quarterly earnings for all students to generate yearly earnings.

[19] In 2012, 82% of all Texas-born individuals remained in Texas (Aisch et al., 2014).

[20] Panel (d) of Figure A.4 plots earning match rates for each grade-year combination in our sample. Estimates presented throughout the paper display statistics obtained from grade-year combinations with a match rate to labor market outcomes greater than 20%.

[21] We recover student labor market earnings at age 24, 25, and 26, and average them out for each student observation to approximate earnings at age 25. We then convert earnings to 2019 dollars based on the CPI index.

[22] Studies generally find that earnings around age 40 best proxy for lifetime earnings — see Haider and Solon (2006) and Mazumder (2005).

[23] See Section A.2 for additional discussion of our choice of earnings age.

[24] At age 25, we have a larger coverage of wage data, with a match rate of 71.5%. If instead, we capture students

correlation of returns estimated at different labor market ages. Thus, items that have high (low) returns for wages at 25, continue to have high (low) returns at ages 30 and 35. Given these results, throughout the paper, our preferred specification uses earnings data at age 25.

3.3 Item Characteristics Data

Item-level information included in the Texas ERC data is limited to the item's learning objective, it's subject (math or reading), and it's position in the test. Thus, we supplement these data in three different ways. First, we collect documentation on the test booklets that were administered for every grade and every year in our sample to recover information on the actual content of each question. Second, we use the content of each question to link it to the taxonomy of skills specified by the Common Core State Standards (CCSS). Finally, we recover psychometric parameters — item difficulty and discrimination — for each question via the estimation of a standard Item Response Theory (IRT) model.

3.3.1 Digitizing the Test Booklets

During the period of our study, the Texas Education Agency (TEA) released to the public the actual test booklets both in hard copy and through the internet following each year's test administration cycle - which they subsequently removed from their site. Using the Internet Archive's Wayback Machine, we were able to recover 75% of the test booklets for the years 1996-2002.[25]

With the booklets in hand, we extract question-level information from the PDFs in a four-step process. First, each assessment PDF was systematically split into its respective math and reading sections using a document segmentation approach. Following this, each section was segmented at the page-level and converted to PNG format to ensure compatibility with further processing tasks. In a second step, these images were then used as inputs for OpenAI's GPT-o3 model to perform Optical Character Recognition (OCR) and infer textual content. Through an API call with a predefined prompt, GPT-o3 generated structured JSON outputs containing two distinct data types: items and passages (for reading).[26] We then evaluate the quality of the digitization by groups, between questions without images (86.5%) and those with images (13.5%). We found almost flawless quality for the image-less group (97% accuracy), but lower for those with images. Thus, we decide to have two research assistants manually check and correct all questions with images in them. Thus, overall we obtain a digitization accuracy rate of 97.4%.[27] In the fourth

at age 30, our average match rate lowers to 52.9%. Finally, at age 35, we loose a substantial amount of data. This is mostly driven by the fact that most students in our testing sample have yet to turn 35. See Figure A.4 for match rates for each grade-year combination in our sample.

[25] Table A.4 displays the availability of test booklets.

[26] The output included item-level attributes such as: question number, question stem, question image indicator, and question image description. Passage-level attributes included: passage stem, passage image indicator, and passage image description. A crosswalk linking each question to its corresponding passage was manually coded.

[27] Accuracy rates for questions without images are estimated from a review of a random sample of 100 questions out of 3077. Accuracy rates for questions with images are estimated from the full set of questions with images in our sample. See Figure A.2 for examples of question types and image position.

and final step, we convert the information from each test item into a structured text file with a predefined template suitable for embedding generation. We do so by following a similar structure to Du et al. (2024). The exact template is presented in the Online Appendix.

3.3.2 Common Core State Standards Initiative

We use the digitized content of questions to link them to the taxonomy of skills defined by the Common Core State Standards (CCSS). In 2010, the National Governors Association and the Council of Chief State School Officers sponsored the Common Core State Standards Initiative with the goal of creating clear and consistent grade-level standards in English Language Arts (ELA) and Mathematics. These standards were drafted via work groups composed of policymakers, researchers, and educators in K-12 and higher education.[28]

Crucial for our purposes, the CCSS were written as outcome expectations rather than a prescribed curriculum, which makes them useful as a descriptive "skill language" for characterizing item demands based on item text.[29] This is especially valuable in our setting because the TAAS metadata provide only coarse objective labels that do not vary by grade, whereas the CCSS offer a finer-grained and hierarchically structured set of competencies by subject that can be used to summarize the skills needed to answer the item in a more interpretable way. We digitized the set of CCSS standards from Association et al. (2010).[30] The final list includes 613 standards for both reading and math across grades 3-12.[31]

3.3.3 Psychometric Properties of the Items

In addition to the text of the items, we also consider whether and how "traditional" psychometric item characteristics relate to $\widehat{\mathbf{\Omega}}$. In this section, we thus explain how we construct these psychometric measures, and how these measures vary in our data across grades, subjects, and years.

Using the IRT routines built into STATA, we recover item level estimates of difficulty and discrimination assuming a three-parameter logistic IRT model ("3PL model").[32] The resulting

[28] The work groups also consulted other groups such as community and parent organizations, the business community, civil rights groups, and states. A majority of states adopted the standards after they were released on June 2, 2010. States were given an incentive to adopt the CCSS through Race to the Top grants.

[29] Because our items predate the CCSS, any mapping should be interpreted as an ex post classification of item content into a modern standards framework rather than as an official alignment; nonetheless, it provides a transparent and replicable first pass at linking the language of questions to a well-defined set of academic skills before turning to a fully text-based model.

[30] Math standards can be found here; Reading-ELA standards can be found here.

[31] The Online Appendix presents the full list of standards, along with a short description of the skill used for exposition purposes. Please refer to Association et al. (2010) for the full skill description.

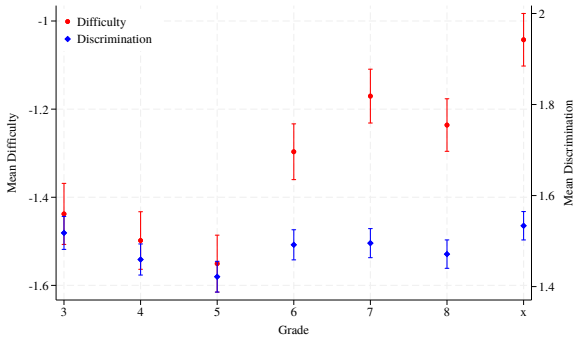
[32] The 3PL IRT model specifies the probability of a correct response as: $P(D = 1 \mid \theta) = c + (1 - c) \cdot \frac{1}{1+\exp[-a(\theta-b)]}$, where a is the discrimination parameter, b is the difficulty parameter, c is the guessing parameter, and student latent ability $\theta \sim \mathcal{N}(0, 1)$. Difficulty and discrimination IRT parameters can be approximated by percent-correct and the correlation of the item response to the total score. Figure A.3 presents the correlation of the IRT parameter estimates with these proxies. We estimate discrimination and difficulty parameters via maximum likelihood for each subject-grade-year combination using STATA's `irt 3pl` command. In our setting, the standard 3PL IRT model did not converge for reading-3rd-2001, reading-6th-2001, reading-8th-1996, reading-8th-2000, math-3rd-1997, math-5th-1996, math-6th-1996, and math-8th-1998. For these subject-grade-year combinations, we estimated a simpler 2PL

item-level parameter estimates yield a number of insights. First, the average item on these TAAS exams is fairly easy, with most items having negative estimated difficulties.[33] This can be seen also in the correct response rates, which average around 80%, albeit with substantial variation. The relatively easy nature of these exams makes sense – they were designed as broad-based assessments of proficiency in academic skills expected of all student and are thus not targeted toward the top of the achievement distribution.[34]

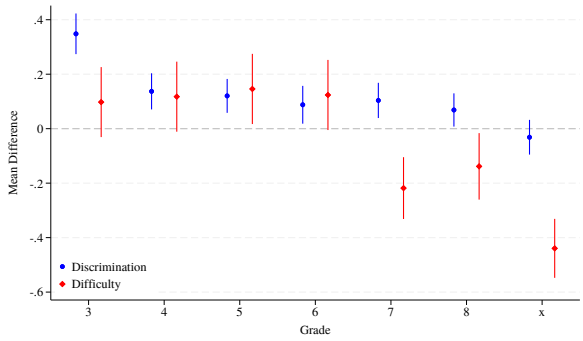
Figure 3 shows mean item difficulty and and discrimination by grade after residualizing with year fixed effects. Panel (a) shows that after remaining roughly flat through grade 5, average item difficulty jumps notably higher in middle school and yet again at the exit exam. By contrast, mean item discrimination is roughly constant across all grades. Turning to differences by test subject, panel (b) shows that reading items are more difficult than math items (on average) in grades 3-6, while math items are more difficult in later grades (and on the exit exam). Finally, reading items are on average more discriminating than the math items across grades 3-8.

Figure 3: Item Difficulty and Discrimination by Grade and Subject

(a) Average Difficulty and Discrimination by Grade



(b) Average Difference: Reading - Math



Notes: Panel (a) plots the average difficulty and discrimination, with 95% confidence intervals, of the items in our analysis sample after removing year fixed effects. Panel (b) plots the corresponding mean differences and 95% confidence intervals in difficulty and discrimination between reading and math items.

4 Estimation of Item Prices ($\widehat{\mathbf{\Omega}}$)

In this section, we describe how we implement our approach empirically. First, we describe how we estimate the item-anchored achievement scales and the corresponding item price or importance vectors $\mathbf{\Omega}$. We then show the robustness of our estimates of $\mathbf{\Omega}$ to alternative model specifications.

IRT model and used its output as a starting point into the estimation of the 3PL model. For all subject-grade-year combinations that had not initially converged, this process solved convergence issues.

[33] Given the normalizations of the 3PL model, a negative difficulty corresponds to a question that differentiates most effectively between test-takers with below-average achievement.

[34] To be clear, even an "easy" test can still differentiate between high performing students, albeit less efficiently than a test targeted towards the upper end of the achievement distribution. As an example, consider an exam where all items have difficulty = -1, discrimination = 1, and guessability = 0.25. Then, students who are one standard deviation above the mean will get about 91% of these items correct, while students two standard deviations above the mean will get 96% correct.

4.1 Estimating $\widehat{\boldsymbol{\Omega}}$ by Ordinary Least Squares

We focus in this paper on log wages of individual i at age 25 as our long-term outcomes of interest (S_i), although the methodology would work just as well for other outcomes. In our baseline approach, we suppose that log wages are linear in item responses and possibly student demographics $\mathbf{X}$. In other words, we assume that f as defined by equation (2) is linear. For each grade-subject-year, we thus estimate via OLS regressions of the form

$$\ln(\text{wage}) = \mathbf{D}'\mathbf{W} + \mathbf{X}'\Gamma + \varepsilon. \tag{6}$$

In this case, equations (3) and (6) imply that an estimate of $\boldsymbol{\Omega}$ is the OLS estimate of $\mathbf{W}$:

$$\widehat{\boldsymbol{\Omega}} = \widehat{\mathbf{W}}. \tag{7}$$

We cluster the standard errors at the school level. Linearity in equation (6) amounts to the assumption that the items do not interact with each other or with demographics. That is, an item is required to have a constant return (skill price) that is independent of the student's demographics and of other tested skills (items).

Return invariance with respect to demographics can be justified both empirically and with reference to the construction of the achievement tests themselves. First, empirically, we find similar estimates of ω within different demographic groups. Specifically, we consider alternative models, including a fully interacted model with race through sample splitting. For these alternative models Table C.1 shows that we can rarely reject the null hypothesis that the individual item weights estimated controlling for demographics in different ways are equal. This is consistent with results presented in Nielsen (2019, forthcoming). Second, because the TAAS exam was designed to provide a reliable measure of basic academic skills pertaining to a fixed set of learning objectives, it is plausible that the constituent items measure broadly useful, basic skills that are economically valuable in a variety of contexts. Indeed, as explained in Section A.1, the TAAS items were carefully vetted for difficulty, content/curriculum alignment, and cultural/racial bias by experienced Texas educators. Thus, a student's item responses likely reflect her skills more-so than skill-irrelevant aspects of her background. Prior research (Neal and Johnson, 1996, Nielsen, 2019) has found in similar settings that psychometric measures correlate with outcomes very similarly for different demographic groups (e.g., white versus Black men), suggesting the labor market discrimination and other non-skill factors play at most a modest role in the translation of skills to earnings. A the same time, recent work has also showed that there might be some detectable effect of item context familiarity with performance (Lee and Schaelling, 2025).

Linearity across the items is perhaps a less obvious assumption. However, as we show below, linear models produce very similar anchored skill and $\boldsymbol{\Omega}$ estimates as models that allow for interactions. Moreover, a series of Monte Carlo simulations, presented in Appendix C.1, suggests that linear models will do quite well in realistic scenarios without "too many" item-item interactions.

In detail, the Monte Carlo experiments in Appendix C.1 generate item response data in a

realistic way by assuming a 3PL IRT model but allowing for outcomes to be determined by items and item interactions. These experiments reveal that the OLS estimates of $\mathbf{\Omega}$ are approximately unbiased across a wide range of data-generating processes (different IRT parameters, skill prices, and interaction specifications). Moreover, the OLS estimates have very similar RMSEs as lasso models which assume the correct order of item interactions (e.g., two-way or three-way), and generally outperform random forest models in terms of bias (though not always in terms of RMSE).

In thinking about the strong performance of linear OLS in these Monte Carlo experiments, it is worth noting that ω_m does not equal the OLS coefficient on item m in the correctly-specified model when there are item-item interactions. Instead, it will reflect both the "direct" linear effect as well as secondary effects that operate through the other items. Analogously, $\widehat{w}_m$, the OLS coefficient for item m in the linear model, will also reflect the same direct and indirect effects. Thus, the asymptotic "bias" in $\widehat{w}_m$ will typically move the miss-specified OLS estimate *closer* to ω_m.

To make this concrete, suppose there are only 2 items and that the item responses are independent of each each other. Suppose further that the true model is given by $S = \psi_0 + \psi_1 D_1 + \psi_2 D_2 + \psi_{12} D_1 D_2 + \varepsilon$ but that the researcher instead estimates a linear model with D_1 and D_2 but no interactions. In this case, plim $\widehat{w}_1 = \psi_1 + \psi_{12}\mathbb{E}[D_2]$. However, in the true DGP, we also have $\omega_1 = \psi_1 + \psi_{12}\mathbb{E}[D_2]$. Thus, in this special case, the linear OLS coefficient on item 1 identifies ω_1, even though the model is miss-specified. With more items and non-independent item responses, plim $\widehat{w}_1$ and ω_m will not generally be exactly equal. However, they will often be close, particularly when the the item responses have broadly similar response rates on average and when the item-item correlations in the student responses are modest.

We thus assume linearity in our baseline estimates, as this assumption comes with a number of substantial benefits. First, linear models are very simple and fast to estimate. Second, the $\widehat{\mathbf{\Omega}}$ can be extracted from the fitted model immediately as the coefficient vectors. Third, linear models allow for the straightforward estimation of $V(\widehat{\mathbf{\Omega}} - \mathbf{\Omega})$ because the sampling covariance matrix for OLS estimates is readily recoverable. Finally, the resulting estimates $\widehat{\mathbf{\Omega}}$ have the straightforward and familiar residual regression interpretation thanks to the Frisch–Waugh–Lovell theorem.

4.2 Controlling for Demographics

Our preferred interpretation of $\widehat{\mathbf{\Omega}}$ is that it is a vector of item prices that themselves include "skill prices" – the labor market returns to the skills assessed by the items. We thus want to control for non-skill factors that affect the level and distribution of earnings. We also want to control, if possible, for confounders – factors that are correlated jointly with item responses and later-life earnings. This is precisely why we included X, which denotes demographic variables and other non-achievement controls, in equation (6).

Because we consider data from a wide variety of years, and because earnings tend to grow over time, we include year fixed effects in X in all our specifications. In addition, we include in X commuting zone fixed effects in our preferred specification to account for differences in the local

labor markets that our test takers have easy access to.[35] Commuting zone fixed effects allow us to compare students with similar labor market opportunities who have different item response patterns. We additionally include indicators for English as a Second Language (ESL) status in order to control for labor market differences by cultural and linguistic background. Finally, we account for the possibility of race and sex discrimination in the labor market, as well as differential labor force participation, by including as a control a full interaction of race and sex. While we simply estimate equation (6) jointly, our procedure is equivalent to estimating $\widehat{\mathbf{\Omega}}$ using just the items and log wages residualized on X.

As an additional robustness check, we follow Nielsen (2019, forthcoming) and estimate equation (6) using only white males – a population that is less likely to experience discrimination in the labor market. The $\widehat{\mathbf{\Omega}}$s that result from this robustness check thus translate item responses into outcomes for everyone as for white males. Overall, although more imprecisely estimated, we find that $\widehat{\mathbf{\Omega}}$s estimated on the sample of white male students, yield qualitatively similar results than our main specification.

4.3 Assessing and Justifying $\hat{\Omega}$

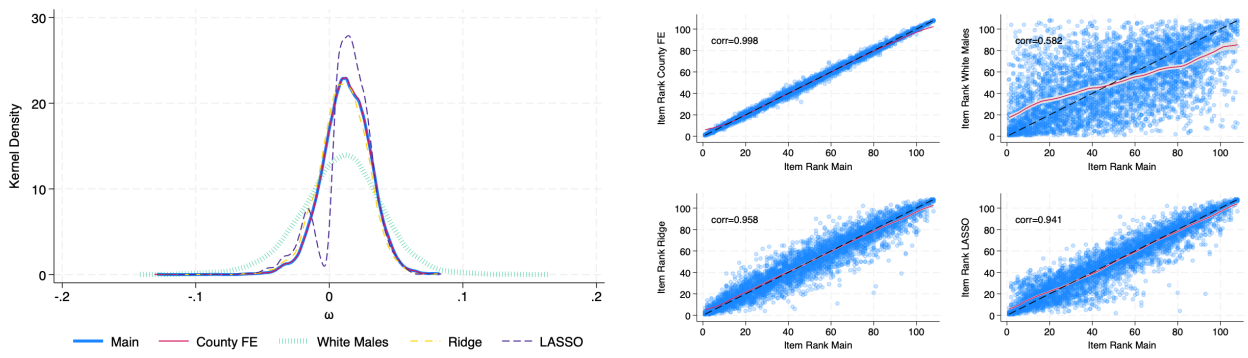
Before delving into our primary analysis understanding what item features explain $\widehat{\mathbf{\Omega}}$, we first explore our estimated item importance vectors across different estimation approaches. Panel (a) of Figure 4 shows the distribution of estimates of ω across different specification. Our baseline regression estimates of the item weights are robust to differences in how we control for geography and race. In particular, "County FE" estimates, ω^{CF}, control for county rather than commuting zone fixed effects in order to consider a larger geographical labor market. "White Male" estimates, ω^{WM}, are derived from the subsample of white males only. Both specifications yield similar distributions of $\widehat{\mathbf{\Omega}}$. Figure 4 also plots estimates obtained from Ridge and LASSO specifications with the relevant λ (penalty parameter) for each model selected via cross-validation. Similarly, the distribution of estimated item-level prices mirrors that of our main specification.

Because the $\widehat{\mathbf{\Omega}}$s depicted in panel (a) of Figure 4 are estimated, some of the apparent dispersion will be due just to estimation error. We verify this in two ways. First, panel (b) of Figure 4 plots the correlation of the rank of item-prices estimated via our main specification, $\hat{\omega}^m$, and the rank of alternative specifications, $\{\hat{\omega}^{m'}\}_{m' \in \{CF, WM\}}$. Overall, these rank-rank correlations are positive and large. However, the comparison with "White Male" seems to be quite noisy. This is likely because we estimate these regressions with only a quarter of the observations used in every other specification. Thus, in our second exercise, we run item-level statistical difference tests across specifications. We find that for approximately 98.5% of items j we cannot reject $H_0 : \omega_j^m = \omega_j^{m'}$. We present these results in Table C.1. Thus, the estimated item weights are quite similar to each other across specifications.

[35] Economic opportunities are quite different in large, growing, dynamic metro areas such as Austin, Houston, and Dallas compared to poorer and more rural parts of the state.

Figure 4: Estimated ω Across Anchor Model Specifications

(a) Distribution of $\hat{\omega}$ Across Specifications $f(\cdot)$ (b) Rank Scatter Plots - $\hat{\omega}^{\text{main}}_{\text{rank}}$ vs. $\hat{\omega}^{\text{alternative}}_{\text{rank}}$



Notes: Panel (a) of this figure plots the distribution of estimated ω across different specifications. Standard errors are clustered by school. Panel (b) shows the correlation between the grade-year item rank of our main specification and the grade-year item rank of the other alternative specifications. A local polynomial fit is added in a red solid line with 95% confidence intervals in gray. The black dashed line represents a 45-degree line. Item rank-rank correlations are added in the top left corner.

5 Item-Anchoring, Achievement Gaps, and Student Ranks

In the previous section, we showed that different, plausible anchor models and methods yield similar estimates $\widehat{\mathbf{\Omega}}$. Moreover, we showed that the individual estimated items prices, $\hat{\omega}_m$, display significant variation – different items predict outcomes very differently. Understanding the drivers of this variation is the primary goal of this paper.

In this section, we further motivate this primary objective by showing that the item-level differences in $\hat{\omega}_m$ have economically and statistically significant implications for a number of relevant empirical questions. In particular, we show that using the item-anchored scores, $\hat{A}_i$, instead of the standard scores dramatically changes (1) estimated white-Black and white-Hispanic achievement gaps and (2) the achievement rankings (percentiles) of individual students. These results extend the findings in Nielsen (2019, forthcoming) to a new context.[36]

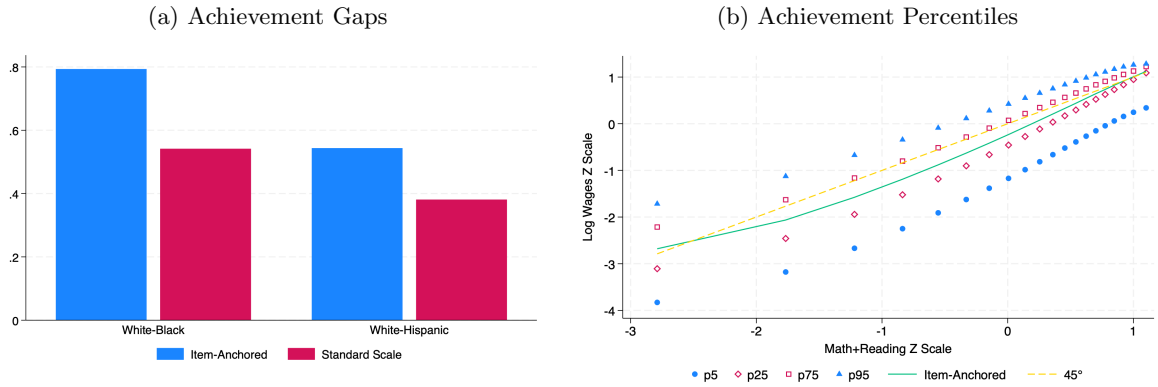
Finding # 1: White-minority item-anchored gaps are larger than standard-scale gaps: Item-anchored achievement gaps in general will differ from gaps estimated using given, psychometric scores. First, trivially, the units will differ – item-anchored gaps are in cardinally interpretable "outcome units," while psychometric scales will be in whatever scale the test-designers construct (which may or may not be cardinal).[37] Second, and more fundamentally, a mean achievement gap could differ across scales because children from one group may do particularly well or poorly on items that are emphasized differently by the item-anchored and given scales.

Panel (a) of Figure 5 shows the average white-Black and white-Hispanic achievement gaps for

[36] Nielsen (2019, forthcoming) focuses on achievement gaps between Black and non-Black, non-Hispanic students as well as between students from high-income versus low-income households. Examining white vs. Hispanic achievement in our setting is motivated by the high share of Hispanic students in Texas – 35% of our sample. Additionally, compared to Nielsen (2019, forthcoming), our analysis covers more recent cohorts over a wider range of grades/ages.

[37] See Cunha et al. (2021) for a discussion of anchoring and cardinality for psychometric scales.

Figure 5: Item-Anchored Achievement Differences



Notes: Panel (a) of this figure shows log wage item-anchored white-Black and white-Hispanic achievement gaps in blue bars and analogous standard math test scale gaps in red bars. Panel (b) plots average given math+reading score ventiles on the x-axis against the ventile-conditional means, 50 percent ranges, and 90 percent ranges of the log wage item-anchored scores on the y-axis.

all grades and years in our sample. Standard scale gaps are presented in red bars, while log wage item-anchored achievement gaps are presented in blue bars.[38] For comparability, both estimates are presented in SD units. In our sample, achievement gaps estimated using traditional test scales yield an average white-Black gap of approximately 0.55 SD and a white-Hispanic gap of 0.38 SD. These findings align with extensive research documenting large racial achievement gaps, with Black students showing particularly pronounced disadvantages.[39] Yet, when achievement gaps are recovered using log wage item-anchored achievement these gaps are estimated to be substantially larger, at around 0.8 SD and 0.55 SD for white-Black and white-Hispanic gaps, respectively. This discrepancy between standard and item-anchored achievement gaps arises because Black and Hispanic students tend to answer correctly at lower rates than their peers high-return questions that are relatively less emphasized by the TAAS scoring rules. These achievement deficits are thus obscured by the aggregation inherent to the given TAAS scale scores. Appendix B discusses this point formally, adding details on how we deal with first stage noise in the estimates of the gaps.

Finding # 2: Anchored achievement scales rank students differently than given scales: In light of the achievement gap results, a natural question is whether student item responses can also affect student ranks that are obtained using traditional test scales. In other words, is our achievement measurement A_i just a rescaled version of the standardized traditional score? To explore this possibility, Panel (b) of Figure 5 plots average given math+reading score ventiles on the x-axis against the ventile-conditional means, 50 percent ranges, and 90 percent ranges of

[38] Item-anchored gaps are adjusted for reliability at every grade-year level using the "split-half IV" method developed in Nielsen (2019, forthcoming). See Appendix B for details on the method and estimation.

[39] In particular, white-Black achievement gaps are typically estimated around 0.5-1.0 standard deviations (SD). See Nielsen (2019), Neal (2006), Bond and Lang (2013), Reardon et al. (2019), Stanford Center for Education Policy Analysis (2012), Quinn (2015), Fryer and Levitt (2004, 2006) among many, many others. While relatively less attention has been paid to white-Hispanic gaps, prior research generally finds gaps about 0.4-0.7 SD. See Hemphill and Vanneman (2011), Reardon and Galindo (2009), Reardon et al. (2019) as well as the National Assessment for Educational Progress (NAEP) achievement gap dashboards available at https://www.nationsreportcard.gov/dashboards/achievement_gaps.aspx.

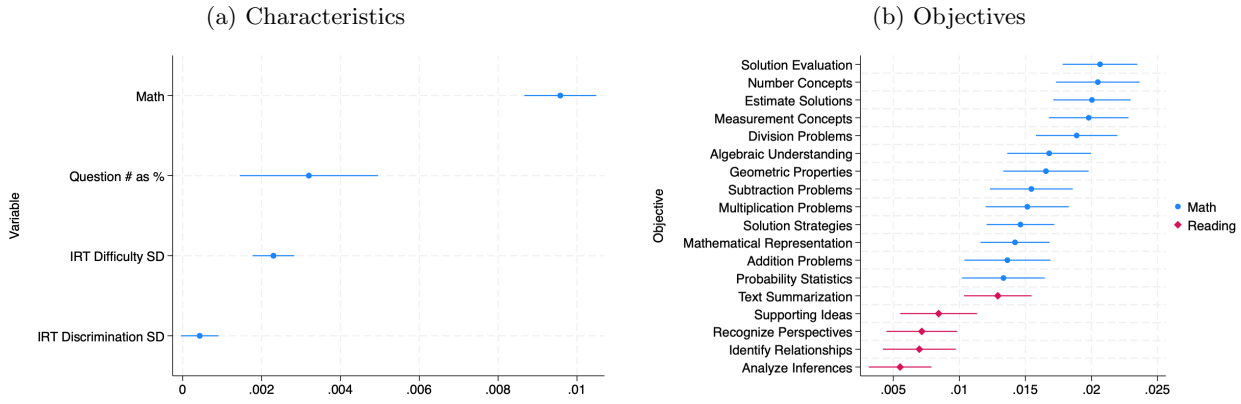
the log wage item-anchored scores on the y-axis. The estimates plotted in this figure suggest a wide variation in the item-anchored scores even among students with very similar given scores – the ventile-conditional 90 percent ranges often cover 2 SD or more. Moreover, this variation reflects more than just measurement error in the item-anchored scales: Wald tests strongly reject equality of the item-anchored scores within each ventile in virtually all grade-years. In fact, we can reject equality in almost all cases even within each percentile or half percentile of the given score distribution.

6 Inside the Item: Observable Features

What differentiates a high-return item from a low-return one? Can we formulate hypotheses about which skills drive long term outcomes based on those differences? In this section, in Section 7, and in Section 8, we explore these questions and develop the paper's central contribution.

In this section, we show how much variation in item weights can be explained by readily-available item-level characteristics such as the subject, difficulty, and broad objectives. To assess the relationship between our estimated item-level returns and standard item-level characteristics we estimate simple item-level OLS regressions of the form $\omega_k = \mathbf{R}'\mathbf{\Theta} + \varepsilon$. We weight observations by the inverse of the variance of ω_k to account for first stage estimatation precision (Hanushek, 1974, Hedges and Olkin, 2014).

Figure 6: Relationship of Item Characteristics to $\hat{\mathbf{\Omega}}$



Notes: Panel (a) of this figure shows estimates of regression coefficients of estimated item-level returns (ω) on observable item-level characteristics. All regressions have grade and year fixed effects and are weighted by the inverse of the square of the SE of ω_K to account for estimate precision (Hedges and Olkin, 2014). Panel (b) presents the exact same regression as (a), but instead of including a subject (math) indicator, that variable is further split into subject objectives as designed by test creators, using the base objective as "Word Meaning - Reading". Figure D.1 presents analogous results for the subset of items for which we were able to collect text data.

Figure 6 presents the estimated coefficients from these regressions. The results presented in panel (a) indicate that math items are associated with higher wage-level returns. On average, math items have a 1 p.p. higher price than reading items. Similarly, items that occur later in the test (have a higher question number) have higher price. Conditional on all other observable characteristics, the last question of a 100-question test has 0.3 p.p higher price than the first. This result goes in line with the research that finds high academic and wage returns to student cognitive

endurance (Brown et al., 2025, Reyes, 2025).

Finally, items that are more difficult, as approximated by a 3PL IRT model, also tend to be associated with higher item-level wage returns. A one SD increase in item difficulty is associated with 0.2 p.p higher item wage returns. However, this is not the case for items that have a higher discrimination parameter. Recall that the discrimination parameter can be approximated by the correlation between responding the item correctly and the overall test score. Given that for a given item j, ω_j estimates are based on conditioning for the full vector of student item-response, $\mathbf{D_{-j}}$, it is not surprising that the discrimination parameter carries no weight in explaining wage returns.

Panel (b) of Figure 6 shows the results of an analogous regression where subjects (math and reading) are allowed to be subdivided into the learning objectives designed by test creators (with *'word meaning'*, a reading objective, as the omitted category). Consistent with earlier evidence, math objectives are generally associated with higher returns on average than reading objectives. Among reading objectives, *'text summarization'* has the highest estimated return, similar in magnitude than the lowest-return math objectives. Moreover, different objectives are often starkly different from each other in terms of their average weight. However, there is no single objective that fully drives high returns.

7 Inside the Item: Item Text Embeddings

This section tests whether the language of the item contains information about item prices beyond psychometric features and test metadata presented in Section 6. As mentioned, one of the key distinguishing features of our setting relative to prior research is that we have digitized item texts. Here we demonstrate that these item texts contain information useful for explaining $\widehat{\mathbf{\Omega}}$ above and beyond the psychometric and test metadata characteristics already considered, and thus provide substantial scope for the item texts to improve our understanding of returns to questions.

Formally, we estimate the sample analogue of equation of equation (5):

$$\widehat{\mathbf{\Omega}} = g(\mathbf{R}, \mathbf{E}) + \mathbf{\Xi}. \tag{8}$$

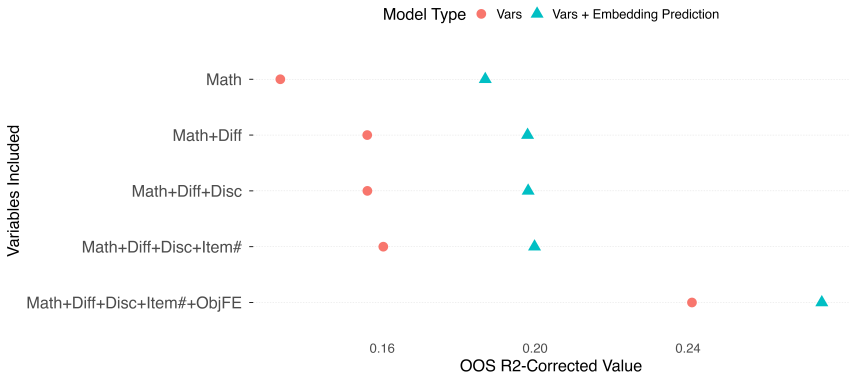
Our basic approach is to compare the fit of this model with the fit of a restricted model which only uses the item psychometric and test metadata – that is, a model in which we use only $\mathbf{R}$ (or subsets of $\mathbf{R}$) and do not use $\mathbf{E}$.

Text representation: In order to implement this procedure, we must first convert the item texts, which are high-dimensional, into lower-dimensional numeric data amenable to quantitative analysis. We accomplish this through a *text embedding*: a function, implemented by a pretrained neural network, that maps an arbitrary piece of text to a fixed-length numerical vector. These vectors are trained so that texts with similar meaning are mapped to nearby points in the embedding space, providing a continuous measure of semantic similarity between any two pieces of text. We provide

further details abuot the embedding model in Section 8.2.[40] We treat the embedding as $\mathbf{E}_m$ for item m.

Prediction model and evaluation: We estimate $g(\mathbf{R}, \mathbf{E})$ flexibly using a five-fold cross-fitting model and a Neural Network (NN).[41] Figure 7 presents the out-of-sample (OOS) R^2s for our NN model for different sets of controls in $\mathbf{R}$ both with and without the text embeddings included as additional controls. Because $\widehat{\mathbf{\Omega}}$ is estimated in a first stage, we adjust the reported R^2 to subtract variation due solely to first-stage estimation error; see Appendix E for the derivation and Figure E.1 for implementation details.[42]

Figure 7: Embeddings Recover New Elements that Explain Returns



Notes: This figure presents the out-of-sample R^2 for each specification that predicts item-level returns with observables. The red dots represent model out-of-sample fit estimates without embeddings, the blue triangles represent model out-of-sample fit estimates that include the embeddings. Regarless of specification, adding the embeddings to the prediction model increased the explanatory power of the model in predicting item-level returns.

Results: Figure 7 reports out-of-sample R^2 for several control sets drawn from $\mathbf{R}$ (subject, objective, position, IRT , and discrimination) with and without embeddings. It reveals two key findings.

- **Incremental explanatory power of text.** Including the text embeddings $\mathbf{E}_m$ always significantly increases the OOS R^2 – roughly 20-60% of the baseline that uses only $\mathbf{R}$. The boost is similar for different non-text controls sets, suggesting further that the information captured by the text embeddings are not well explained by standard psychometric item variables such as difficulty, discrimination, and learning objective.

[40] See Reimers and Gurevych (2019) for a general introduction to text embeddings and Muennighoff et al. (2023) for the Massive Text Embedding Benchmark (MTEB) used to evaluate embedding models.

[41] We assessed multiple other ML models, including XGBoost, random forests, nearest neighbor, etc. but in every case, (i.e., with every control set we considered) NN performed similarly (in the case of XGBoost) or better than the other models. We estimate the NN models via cross-fitting with five folds – each with a random sample of all questions Chernozhukov et al. (2018). We assess the OOS performance for each left-out fold, and the final prediction model is the ensemble of these. See Appendix F for details of the implementation.

[42] In summary, because each item weight is estimated via cross-fitting, we treat the ML predictor for item m as fixed with respect to the first-stage estimation error, conditional on $\mathbf{\Omega}$ (Chernozhukov et al., 2018, Bach et al., 2021). This allows for a straightforward correction to R^2 which removes the variation in $\widehat{\mathbf{\Omega}}$ coming from estimation error which should not itself be explainable by the item texts or other factors.

- **Most variation remains unexplained.** Only a relatively small share of the variation in $\widehat{\boldsymbol{\Omega}}$ is explained by any of these factors, with the corrected R^2s never exceeding 0.25. Most of the variation in the item weights is not explained by any features we observe.

These results establish that item text carries economically relevant information about item prices, and they motivate the standards-based analysis in Section 8, which seeks an interpretable decomposition of that information into recognizable skills.

8 Inside the Item: Standards-Based Skill Mapping

The embedding results in Section 7 show that item language helps explain item prices, but the mapping from text vectors to prices is difficult to interpret. Add to that that the variables analyzed in Section 6 are either very hard to interpret as skills (difficult, discrimination), or too coarse to be helpful for thinking about specific skills (broad learning objectives like 'Number Concepts').

To go deeper, we develop and implement a *text-based mapping* from items to the Common Core State Standards (CCSS) and estimate "skill prices" that summarize the return to each CCSS standard. As mentioned in Section 3.3.2, the CCSS list provides a fine-grained skill taxonomy that are actionable and granular enough to parse the specific skills that are related to future higher wages. Our mapping proceeds in three steps, which we detail below.

8.1 Step 1: Skill Extraction

For each digitized test item, we prompt two large language models, one open source chain-of-thought (Qwen 3-8B) and one closed source (OpenAI's o3), to return a structured description of the item's cognitive demands. In particular, we ask for the skills required to answer the question, the reasoning steps a student would follow, and potential challenges. We ensure that the model receives only the digitized question text in the prompt — no grade label, no curriculum reference, and no CCSS skills — so the extracted skills are fully agnostic to any standards framework.[43] This yields a short text summary per item — typically three required skills, four reasoning steps, and three potential challenges — that characterizes *what the question tests* rather than *what the question says*.[44]

8.2 Step 2: Text Embeddings

We next embed the skill descriptions from Step 1 and the full text of each CCSS standard using **Qwen3-Embedding**, a dedicated open source embedding model based on a chain-of-thought

[43] Specifically, `o3-2025-04-16` with high reasoning effort. We also implement a fully open-source alternative using Qwen3-8B with chain-of-thought reasoning, which produces comparable skill profiles.

[44] This skill-extraction step is distinct from the digitization described in Section 3.3.1. Digitization recovers the question text from scanned booklet images; skill extraction takes that text as input and produces a structured cognitive profile of each question.

equivalent generation model.[45,46] By embedding the LLM-extracted skill descriptions rather than the raw question text, we focus the comparison on cognitive demands rather than surface-level question language. This is important because the CCSS entries are skill-based – we thus translate the items into the kinds of skills that can be plausibly matched to the CCSS taxonomy.

We follow Su et al. (2023) and Muennighoff et al. (2023) and add the instruction "*represent the cognitive skills and knowledge a student needs to answer it correctly*," so that the resulting vectors emphasize skill demands rather than surface-level question language. CCSS standard descriptions are embedded without a prefix, following the asymmetric query–document convention standard in information retrieval.[47]

Figure 8 illustrates the overall flow from a scanned test booklet, through digitization of the question text, to the final numerical embedding. Figure 9 projects the item embeddings into two dimensions via the Uniform Manifold Approximation and Projection (UMAP) method (McInnes et al., 2018). Math (blue) and reading (orange) items occupy clearly distinct regions of the embedding space, confirming that the embeddings capture subject-level semantic structure.

8.3 Step 3: Cosine Similarity

With embeddings for both the item-level skills and the CCSS standards in hand, we measure the semantic proximity of each item to each CCSS standard via cosine similarity. Let E_m be the embedding of the skill description for item m, and let V_j be the embedding of CCSS standard j. The cosine similarity between these embeddings is

$$C_{m,j} = \frac{E_m \cdot V_j}{\|E_m\| \, \|V_j\|} \, . \tag{9}$$

That is, cosine similarity is defined as the normalized dot product of the embedding vectors. Geometrically, $C_{m,j} = \cos(\theta_{m,j})$, where $\theta_{m,j}$ is the angle between the two vectors in the embedding space. Two embeddings will have high similarity whenever they "point" in the same direction. Cosine similarity has been shown to provide a good summary of semantic similarity between the texts associated to the compared embeddings.[48] Let $\mathbf{C}$ denote the $M \times J$ matrix of cosine similarities –

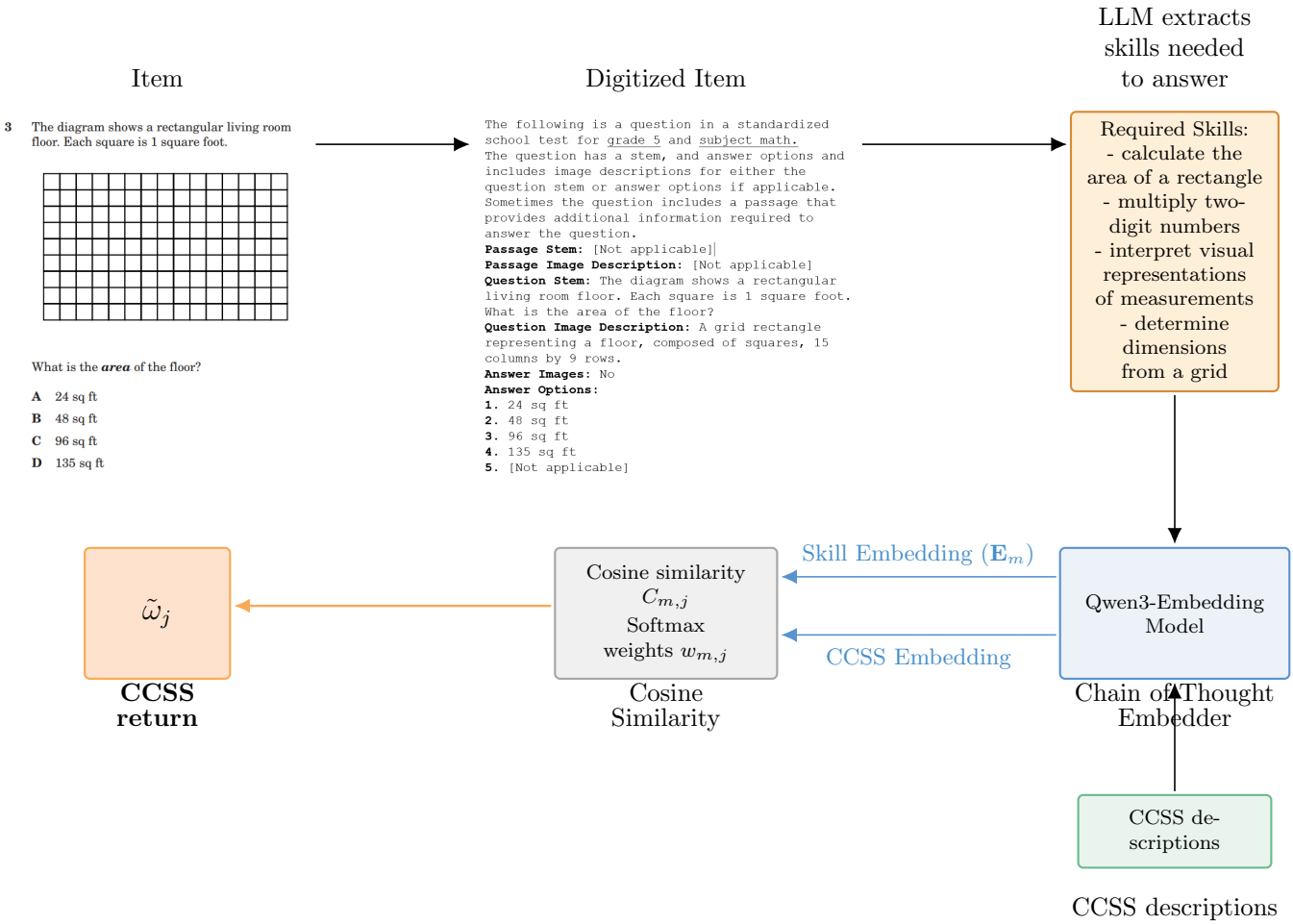
[45] The Qwen3-Embedding (Zhang et al., 2025) is an 8-billion-parameter model purpose-built for text embedding, which maps each input text to a vector of 1,024 real numbers. The model supports Matryoshka Representation Learning (Kusupati et al., 2022), which trains the model so that the leading components of the embedding vector capture the most important semantic information.

[46] Although it is possible to use embeddings from generative LLMs, like the Llama family, general-purpose language models such as these are designed to generate text rather than embed already-existing text. Instead, we use a dedicated embedding model, which is trained specifically to produce vectors that preserve semantic similarity, and consequently achieves substantially higher scores on standard retrieval and matching benchmarks. Qwen3-Embedding-8B scores 70.58 on the MTEB multilingual benchmark, compared to roughly 65 for LLM2Vec (BehnamGhader et al., 2024), an approach that repurposes a generative language model as an encoder.

[47] Instruction-prefixed embedding prepends a short natural-language directive to the input before encoding, directing the model to foreground task-relevant information. Models fine-tuned with instruction prefixes significantly outperform their non-instruction counterparts across nearly all MTEB task categories, with retrieval being a particularly strong beneficiary examples. See Su et al. (2023), Muennighoff et al. (2023).

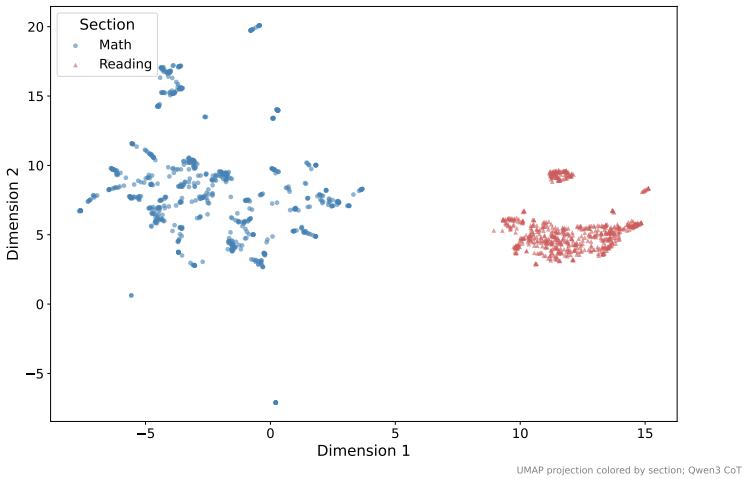
[48] See Ramanujam et al. (2025), BehnamGhader et al. (2024), Thirukovalluru and Dhingra (2025), Schakel and Wilson (2015).

Figure 8: From Test Item to Skill Return



Notes: The figure illustrates the pipeline from a test item to a CCSS skill return. An LLM identifies the skills needed to solve each digitized item. Both item skills and Common Core State Standards (CCSS) descriptions are embedded, cosine similarities $C_{m,j}$ produce softmax weights $w_{m,j}$, and these yield the return $\tilde{\omega}_j$.

Figure 9: Embeddings Differentiate Math and Reading



Notes: The figure shows the 2 dimensional representation of the embedding coordinates by the UMAP method.

C thus connects every item to every CCSS skill, with the magnitude and sign of $C_{m,j}$ determining the "strength" of each item-to-CCSS match.[49]

8.4 Step 4: Defining Skill Returns

For a given CCSS skill we define

$$\tilde{\omega}_j = \sum_m w_{m,j} \times \hat{\omega}_m, \quad \text{where } w_{m,j} = \frac{\exp(\beta \cdot C_{m,j})}{\sum_{m'} \exp(\beta \cdot C_{m',j})}. \tag{10}$$

In words, $\tilde{\omega}_j$ attributes to CCSS skill j a weighted average of all item-level $\hat{\omega}$'s, where the weights follow a multinomial logit (or 'softmax') kernel (McFadden, 1974). The scale parameter β controls how sharply the weighting discriminates between good and poor matches: a higher β means that items with low similarity to skill j are discounted more aggressively relative to better-matched items. In the language of discrete choice, β is the inverse of the error variance in a random utility model (Train, 2009). This formulation has two useful properties. First, it reduces the weighting to a single free parameter β. Second, the log of the denominator—the *inclusive value* from discrete choice theory—provides a natural diagnostic for the quality of the item-to-skill match: standards with low inclusive values have no well-matched items in the data (see Appendix G.4).

We select β by 10-fold cross-validation, stratified by grade and subject, which yields an optimal value $\beta^* = 45$ which we use throughout the remainder of the paper.[50] With this item-to-skill matching procedure, and considering only "substantial" matches with weight greater than 0.01, each item is matched to an average of 5.3 (4.3) CCSS math (reading) skills. The within-item average standard deviation of weight is 0.03 (0.01) for math (reading) items. Figure 10 presents estimates of the CCSS-to-item weights, $w_{m,j}$, from Equation 10 for math and reading items in panel a) and b), respectively. While both reading and math items, match to the other-subject CCSS, own-subject CCSS skills carry the highest weight — meaning that they are the most similar in terms of language. Therefore, our estimation of CCSS returns, $\tilde{\omega}_j$, will be mostly based on the returns estimated for own-subject item returns.

8.5 Results

Figure 11 presents the full distribution of estimated CCSS returns following Equation 10 with the associated 95% confidence intervals. The standard error of each skill return is derived from the item-level standard errors (Section 4) via the variance propagation formula, treating the softmax weights as fixed. We weight all subsequent regressions by $1/\text{SE}(\tilde{\omega}_j)^2$ to account for the fact that some skill returns are estimated more precisely than others (Murphy and Topel, 2002). While Figure 11 labels the five highest-return skills for both math and reading, Table 2 expands this list

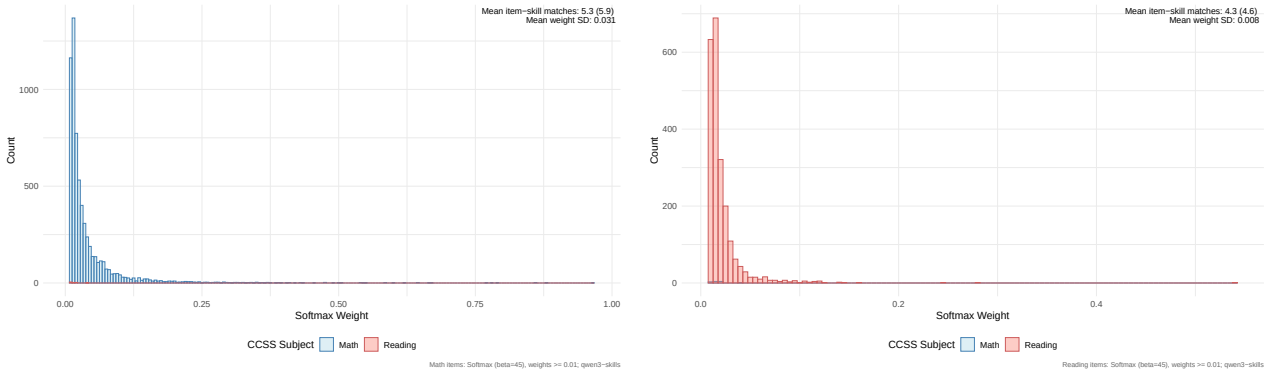
[49] Note that $C_{m,j} \in [-1, 1]$ always.

[50] Figure G.4 reports the CV curve. Our results are robust to using a power kernel $(\max(0, C_{m,j})^p, \quad p \geq 1)$ with the p selected by cross-validation instead.

Figure 10: Item-level Cosine Similarity Weights

(a) Math Items

(b) Reading Items



Notes: Panel (a) and (b) of this figure show the distribution of non-zero weights ($w_{m,j} > 0.01$) specified in Equation 10 ($w_{m,j} = \exp(\beta \cdot C_{m,j}) / \sum_{m'} \exp(\beta \cdot C_{m',j})$) for math and reading items, respectively. Top right shows the mean item-skill matches, and the standard deviation of matched items in parenthesis. Below, it shows the mean *within-item* standard deviation of the weights. For each item, it computes the SD of its weights across all matched CCSS standards (how spread out vs uniform the weights are), then averages that SD across all items. A higher value means the softmax weights are more concentrated on a few CCSS standards, while a lower value means they are more uniformly spread.

to present the top and bottom 20 CCSS skills for each subject. Below, we summarize the main take-aways from this analysis.

Finding #1: Math skills dominate the top rankings, but reading comprehension is also important: Consistent with the results presented in Section 6, Figure 11 shows that math skills cluster at higher returns than reading skills. However, the granularity of skills provided by the CCSS allows us to recover two novel results. First, in contrast to the results presented in Figure 6, we find that not all math skills dominate reading skills. The highest-return reading skills – *'Compare text structures and their effects'* – rank higher than approximately 300 lower-return math skills. Nonetheless, the highest return math skills have much higher estimated returns than the highest ranking reading skills; the point estimates at the top end for math are about double those for reading.

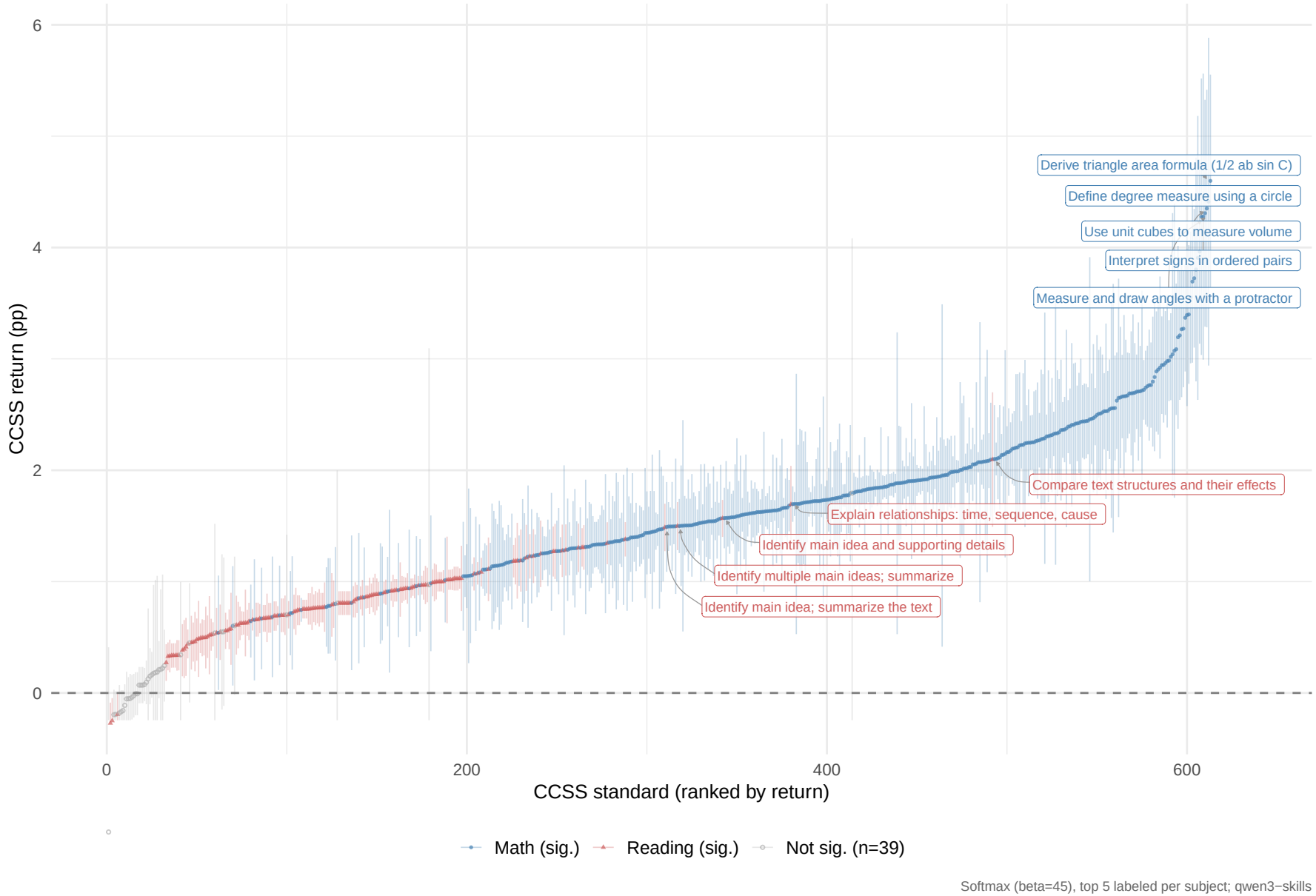
Second, consistent with the results presented in Figure 6, the highest return reading skills often involve basic reading comprehension and text summarization. Twelve of the top 20 reading skills require students to *'summarize'* or *'identify main idea'* of the text. In contrast, none of the bottom reading skills require any text summarization. Instead, they focus on *'word meaning', 'word choice'* or *'word tone'*.

Finding #2: Computation may matter more than conceptual understanding: Within math, differences across the distribution are more nuanced than for reading. From a simple comparison of skills on either end of the distribution, we can nonetheless see a few key differences. First, the top math skills lean heavily toward what we might call computational or procedural fluency that require executing multi-step procedures accurately using formal notation to produce numerical results. Often these skills emphasize verbs such as *'derive', 'measure', 'find'*, and they tend to focus on whole-number computation, measurement, and coordinate-plane basics.

By contrast, the bottom math skills lean towards conceptual definitions and reasoning focused on understanding meanings, building representations, and recognizing patterns, often with minimal computation. These tasks either avoid formulas entirely or involve only basic operations without symbolic notation. Often these skills emphasize verbs such as *'compare', 'describe', 'explain'*, and focus on fractions/ratios/percent, probability, and conceptual geometry—areas that typically require more abstraction and representational flexibility.

For example, while geometry-related skills are located on both ends of the returns distribution, their emphasis tends differ on either end. Geometry skills located at the top of the returns distribution favor procedural use of geometric tools (*'Derive triangle area formula'; 'Define degree measure using a circle'*), while the bottom skills favor conceptual properties and transformations (*'Prove parallelogram theorems'; 'Identify and draw line of symmetry'*).

Figure 11: Distribution of CCSS Skill Returns



Notes: Each point represents one of 613 CCSS standards, ranked by estimated return ($\tilde{\omega}_j$). Vertical bars show 95% confidence intervals. Returns computed following Equation 10 using a softmax kernel ($\beta = 45$) with precision weighting ($1/\text{SE}^2$). Skills with insignificant returns (CI includes zero) are shown in gray. Top and bottom 5 skills per subject are labeled.

Table 2: Top and Bottom 20 CCSS by Subject

Top 20 Skills		Bottom 20 Skills	
Rank	CCSS	Rank	CCSS
Panel A. Math			
1	Derive triangle area formula (1/2 ab sin C)	514	Link triangle congruence to matching sides/angles
2	Define degree measure using a circle	518	Use multiples to multiply fractions by whole numbers
3	Use unit cubes to measure volume	522	Describe congruence via rigid motions
4	Interpret signs in ordered pairs	528	Interpret a/b as a multiple of 1/b
5	Measure and draw angles with a protractor	532	Define fractions as parts of a whole
6	Recognize angles and angle measurement	534	Justify steps when solving equations
7	Write order comparisons in context	543	Use matrix inverses to solve systems
8	Use additive angle measures to find unknowns	544	Use rigid motions to decide congruence
9	Relate angle measure to degrees	549	Find inverse functions
10	Define volume by packing unit cubes	550	Use zero/identity; link determinant to inverse
11	Order rational numbers; use absolute value	552	Understand equivalence as same size/point
12	Generate equivalent expressions	554	Report number of observations
13	Interpret graphs as solution sets	568	Explain and compare equivalent fractions
14	Define a coordinate system with axes	573	Explain equivalent fractions using models
15	Derive circle/solid formulas informally	583	Solve using inverses; write inverse expressions
16	Identify linear vs non-linear functions	584	Generate and justify equivalent fractions
17	Evaluate expressions using values and formulas	586	Compare fractions using benchmarks/common units
18	Graph points; find distances using abs	587	Identify and draw lines of symmetry
19	Draw and identify basic geometric objects	588	Compare fractions with same numerator/denominator
20	Fit a linear function to data	613	Prove parallelogram theorems
Panel B. Reading			
122	Compare text structures and their effects	593	Analyze cumulative impact of word choice
234	Explain relationships: time, sequence, cause	594	Use context to self-correct while reading
272	Identify main idea and supporting details	595	Use context to self-correct while reading
297	Identify multiple main ideas; summarize	596	Use context to self-correct while reading
304	Identify main idea; summarize the text	597	Use phonics and morphology to decode words
326	Analyze author's organizational structure	598	Use phonics and morphology to decode words
336	Describe text structure and organization	599	Determine word meanings in context
349	Identify multiple central ideas; summarize	600	Determine word meanings in context
350	Analyze multiple central ideas and interactions	601	Analyze sound devices' effect
352	Analyze paragraph structure and sentence roles	602	Interpret words: literal vs figurative
359	Analyze central idea development; summarize	603	Interpret figurative language
361	Determine central idea; summarize objectively	604	Analyze word meaning and tone; allusions
365	Analyze theme development; summarize objectively	605	Analyze cumulative word choice and tone
366	Analyze central idea refinement; summarize	606	Analyze word meaning and tone; allusions
373	Analyze theme development in narrative; summarize	607	Determine word meanings in context
377	Analyze theme/central idea development	608	Analyze word choice effects on meaning and tone
379	Analyze multiple themes/ideas and interactions	609	Determine word meanings, including technical
381	Determine theme/central idea; summarize objectively	610	Analyze word choice impact on meaning/tone
384	Analyze structure and pacing effects	611	Analyze word meanings and tone
386	Determine theme and summarize	612	Interpret words, including mythological allusions

Notes: This table displays the top and bottom 20 CCSS skills alongside its rank. Panel A. presents CCSS top and bottom CCSS for math, Panel B. presents an analogous list for reading.

8.6 What Predicts High-Return Skills?

The results in Section 8.4 show that different CCSS skills have very different wage returns. Although Table 2 presents a clearer picture for reading (i.e. the most important reading skills seem directly related to basic reading comprehension and text summarization), math skill rankings require a more nuanced analysis. In this section, we classify each of the 613 CCSS standards along three dimensions that capture cognitive and labor-market-relevant features of each skill: complexity, exposure to automation, and spatial content.

Complexity – Given the patterns observed in Figure 11 and Table 2, we assess how cognitive

complexity is associated with skill returns. To assess how systematic these patterns are across the full CCSS return distribution, we categorize each skill using the Depth of Knowledge (DOK) framework, which captures the cognitive complexity of academic tasks and classifies them into three ordered categories.[51]

Automation exposure – In addition to cognitive complexity, a related measure to computational/procedural fluency is its automation exposure. As defined in Autor et al. (2003)'s task-based framework, cognitive activities easily governed by explicit rules are more susceptible to computer/capital substitution. By contrast, non-routine tasks that require creative problem-solving and abstract reasoning act as complements to this technological change. Thus, we pull from this framework to classify each skill based on its susceptibility to replacement by 2010-era technology - the time for which we are measuring wages. It is worth emphasizing that this construct does not measure exposure to automation by current, cutting-edge AI technology.

Spatial reasoning – Finally, we also classify the extent to which the skill requires spatial reasoning. Spatial reasoning has been found to be highly predictive of academic success, particularly in STEM and high-earning fields (Uttal et al., 2013a,b).

To characterize each CCSS standard, we provide a classification protocol to an LLM to code each of these three dimensions, summarized in Table 3.[52] Each standard is presented to the model with its full text description, stripped of grade identifiers to prevent anchoring. Importantly, the classification is performed independently for each standard, and the model receives no information about the standard's estimated return. The full classification protocol, dimension distributions, and inter-dimension correlations are reported in Appendix G. Particularly for math, these three measures are only modestly correlated, suggesting that they are picking up distinct skill dimensions.

8.6.1 Skill Dimensions as Predictors of Skill Returns

We begin by examining each dimension separately. Figure 12 plots coefficients from bivariate regressions of CCSS returns on each skill dimension, estimated separately for Math and Reading. Both the outcome and the regressors are standardized (the outcome in pooled SD of CCSS returns across all standards; regressors in within-subject SD, except spatial reasoning which enters as a raw 0/1 indicator). Each coefficient comes from its own separate regression, so they are not conditional on the other dimensions.

[51] DOK 1: Recall & Reproduction: involves basic recall of facts, definitions, or simple procedures. In math, these involve straightforward actions such as recalling multiplication facts, identifying the formula for the area of a rectangle, or performing a routine computation. DOK 2: Skills & Concepts: requires cognitive processing such as comparing, classifying, interpreting, or applying concepts in familiar situations. For example, comparing two fractions with unlike denominators, selecting an appropriate operation to solve a word problem, or interpreting information from a simple table or graph. DOK 3: Strategic Thinking: demands reasoning, justification, planning, and evidence-based decision-making, often involving abstract or non-routine problems. For example, analyzing the structure of an algebraic expression to determine equivalence, or determining the most efficient strategy for solving a multi-step percentage problem. We do not include DOK 4: Extended Thinking, as there are no skills in the CCSS that match this level. This level encompasses complex, sustained work such as designing investigations, synthesizing information across sources, or applying concepts in novel contexts over time. See Webb (2002, 2007) for more details.

[52] We use Anthropic's Claude 4.6 for this task.

Table 3: Skill Classification Dimensions

Dimension	Scale	Subjects	Definition	Literature
Automation exposure	1–5	Both	How susceptible the skill is to replacement by 2010-era technology. 1 = non-automatable (e.g., "analyze how characters develop"); 5 = fully automatable (e.g., "compute perimeter from coordinates").	Autor et al. (2003) Autor (2013)
Spatial reasoning	0/1	Math	Whether the skill requires mental manipulation of shapes, positions, or coordinate systems. 1 = spatial (e.g., "graph points on a coordinate plane"); 0 = non-spatial (e.g., "solve linear equations").	Uttal et al. (2013b) Uttal et al. (2024)
Depth of Knowledge	1–3	Both	Level of mathematical thinking/Reading: 1 = recall/reproduction; 2 = concepts/multi-step; 3 = justification/complex problems.	Webb (1997), Hess (2009)

Notes: Each of 613 CCSS standards is classified independently by Claude (Opus 4). Grade identifiers are stripped from the standard text before classification to prevent anchoring. "Both" = dimension applies to math and reading standards. Automation exposure is coded so that higher values indicate *greater* susceptibility to automation. Examples in the Definition column are illustrative CCSS standards (paraphrased).

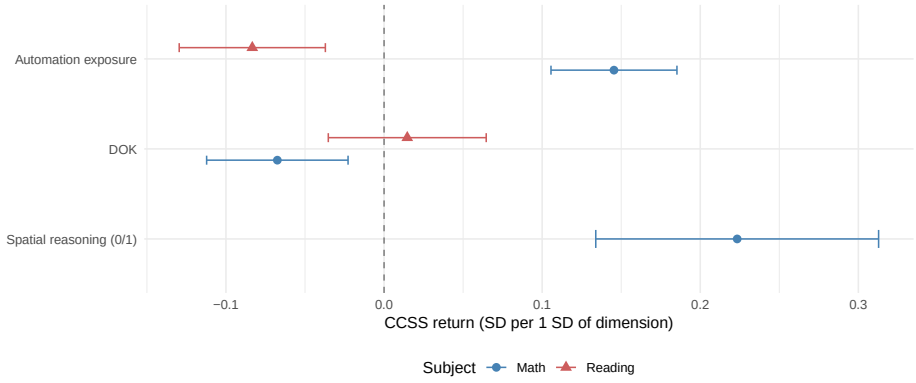
For mathematics, spatial reasoning is the strongest single predictor: math skills that involve spatial content have substantially higher returns. Automation exposure is the next strongest—skills that are more exposed to automation have *higher* returns, consistent with the finding that procedural, computational skills outperform more conceptual ones. DOK level shows a negative association: less complex math skills tend to have higher returns. For Reading, the patterns differ. Automation exposure is negatively associated associated with returns, while DOK level shows no apparent relationship (and spatial reasoning clearly does not apply to reading skills).

Given that our classification is discrete, panel (a) of Figure 13 repeats the analysis using indicators for each level of each dimension, with the modal category as the reference group. In panel (b) of Figure 13, we present the same coefficients but including all three dimensions simultaneously. For math, automation exposure seems to have a robust monotonic pattern, where higher exposure predicts higher returns even controlling for the other dimensions. The DOK dimension in math shows that a lot of the initial negative association between complexity and returns gets attenuated when controlling for the other two dimensions. Lastly, spatial reasoning has almost the same, if not a larger, association with skill returns when controlling for the other two dimensions. For reading, however, higher automation exposure implies lower returns, particularly when controlling for DOK. Regarding DOK, the bivariate analysis shows a noisy relationship with returns, but in the joint analysis we do see that lower complexity tasks have higher returns in reading.

8.6.2 Visualizing Skill Returns by Dimension

Finally, Figure 14 provides a more intuitive summary by organizing math CCSS returns along the three dimensions in form of a heat map. Each cell shows the precision-weighted mean return together with a representative example skill (we select the standard with the highest cosine similarity

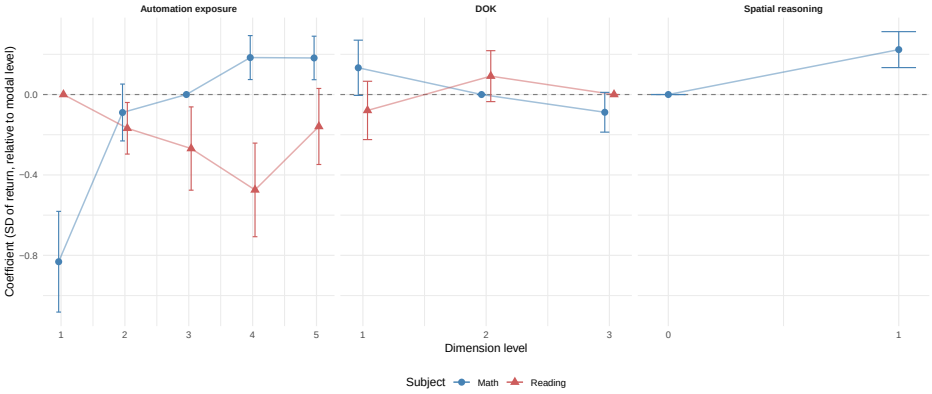
Figure 12: Continuous Measure Coefficients



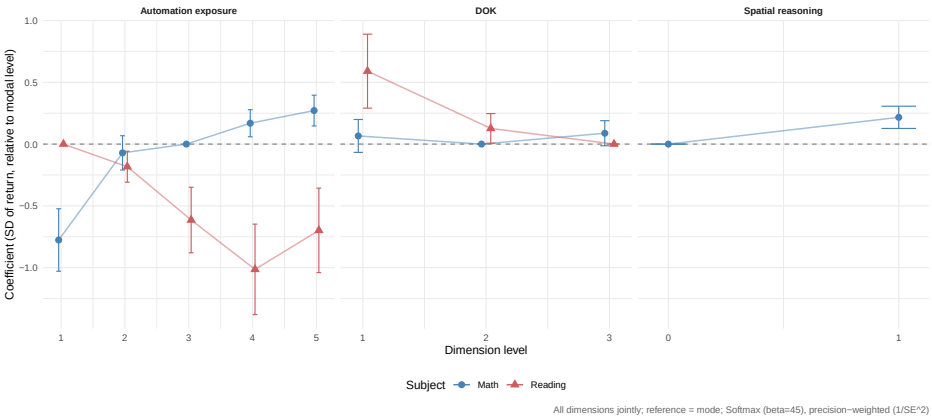
Notes: Each panel shows coefficients from a bivariate regression of CCSS returns on continuous measure for the dimension, with the modal level as reference (coefficient = 0). Precision-weighted.

Figure 13: Categorical Coefficients

(a) Individual Regressions



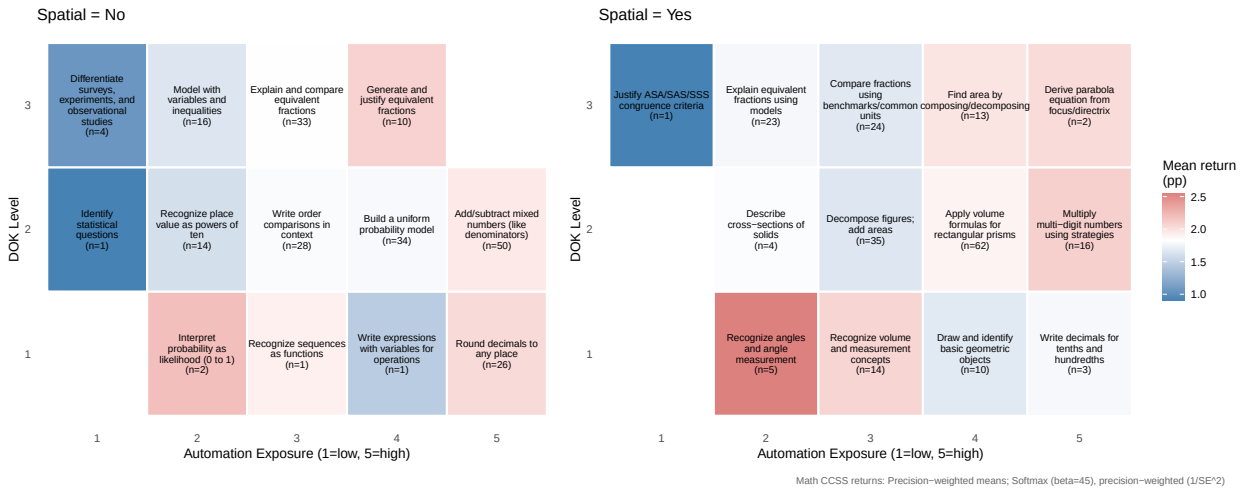
(b) Joint Regressions



Notes: This figure presents estimates of CCSS regressions per subject on on factor indicators for the skill dimensions. Panel a) presents estimates for each dimension estimated individually. Panel b) presents estimates when all dimensions are included. For each dimension, the modal level is set as reference (coefficient = 0). Precision-weighted.

to any item in the data, so that each example is well-identified). The heatmap exemplifies the results from the previous analysis: skills with lower complexity and that are more procedural (high automation exposure) tend to have higher returns. The higher average returns for skills involving spatial reasoning are also apparent in this figure.

Figure 14: Math CCSS Returns by Transferability and Automation Exposure



Notes: Precision-weighted mean return (pp) in each transferability × automation exposure cell. One example skill per cell (highest max cosine similarity to any item). Color scale: blue = below average, red = above average. Reading equivalent is presented in Figure G.1.

8.7 Robustness

Alternative models to extract skills: The main results use skills extracted by Qwen3-8B ("qwen3-skill"). As an alternative, we extract skills using o3 ("o3-skill"), a different model with substantially different architecture. The correlation between CCSS returns under the two extraction methods is essentially 1 ($r > 0.99$), and the point estimates for the skill dimension regressions are nearly identical (see Appendix G.3 for the corresponding coefficient and factor-level plots).

Robustness to weighting kernel: As noted above, Appendix G.3 documents that the CCSS returns under the softmax kernel ($\beta = 45$) are highly correlated with those from the power kernel ($p = 30$, selected by the same CV procedure). The skill dimension regression results are substantively unchanged.

8.8 Discussion and Interpretation

The empirical results in this section are striking and suggest a number of interesting (and not mutually exclusive) explanations. We examine these possibilities in this section, and we provide additional discussion of the method and data to help put them in context. This discussion is speculative and not exhaustive; we leave a detailed analysis of these and other explanations for future work.

High-Weight CCSS Skills Have Higher Labor Market Returns: This is just a "straight-read" of our results. It could be that the types of math skills most rewarded in the labor market are procedural, computational skills (i.e., those that have low DOK or are particularly exposed to automation) as well those that require spatial reasoning. That is, employers might directly value such skills and pay more for them accordingly.

High-Weight CCSS Skills Predict the Future Acquisition of Valuable Skills: Alternatively, it could be that these types of skills, assessed during childhood and young adulthood, predict the future acquisition of economically skills. In this case, the skills valued directly by the labor market need not be low-DOK, exposed to automation, or spatial in nature – all that matters is that their presence is predicted by the high-weight CCSS skill categories. Thus, for example, it could be that the types of higher-level math skills used intensively in STEM careers are best predicted in earlier schooling by mastery of basic computation and spatial tasks. Indeed, prior research, including some item level analysis (Nielsen, 2025b) provides support for this "accumulative" model of skill acquisition.[53]

Soft Skills: Our results indicate that math skills involving procedural, multi-step computation are associated with higher returns than more conceptual or interpretative tasks. One possible interpretation is that these patterns partially reflect returns to soft skills rather than purely academic competencies. If procedural skills require more repeated practice and reinforcement than do more conceptual skills, then they might correlate more strongly with traits such as grit or perseverance.

Unfortunately, our ability to speak directly to soft skills in this context is limited. Even with rich item response and text data, reconstructing meaningful measures of soft skills is difficult because these constructs differ fundamentally from the academic competencies typically assessed in reading or mathematics.[54] The best approximation for such traits in our setting is the position of the item within the test. Items appearing later in the test plausibly load more heavily on cognitive endurance or sustained effort, which could be interpreted as a behavioral manifestation of grit.[55] Indeed, we find evidence consistent with this interpretation. As shown in panel a) of Figure 6, items located toward the end of the test are associated with systematically higher estimated returns. Importantly, this finding persists even after controlling for key item characteristics such as subject, difficulty, and discrimination.[56]

[53] See also Duncan et al. (2007), Jordan et al. (2009) for research connecting early academic skills to later academic skills. Ritchie and Bates (2013) finds that early math skills are strong predictors of later SES, a closely related question to ours.

[54] For one, soft skills are not domain-bound — grit or persistence displayed in mathematics does not necessarily translate to similar behavior in reading. Furthermore, soft skills generally manifest as patterns of behavior over time. A persistent student is one that routinely sticks with challenging tasks or returns to a problem after an initial failure. Isolated responses to individual test items do not allow us to observe these behavioral patterns.

[55] See Brown et al. (2025), Brunello et al. (2021), Borgonovi et al. (2021), Debeer et al. (2014), Weirich et al. (2017), Reyes (2023, 2025), Reyes et al. (2024).

[56] See Section A.1.4 for a detailed discussion of the relationship between an item's position in the test and IRT parameters.

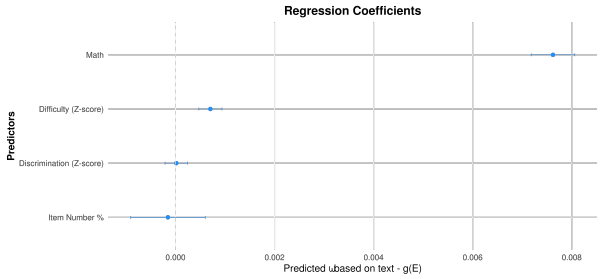
Furthermore, we find that higher returns for later items are not driven by differences in the the language of these questions. Panel a) of Figure 15 shows the extent to which the embeddings alone can predict item observables. Overall, the embeddings are able to predict whether a question is a math or reading question and to some extent the difficulty of the question. However, the embedding alone cannot predict the position of the question in the test or whether it is a high discrimination question or not.

Given that the item's position in the test is not related to discrimination, difficulty, or language, we consider it to be a reasonable proxy for a measure of student's cognitive endurance and control for it in every specification. To the extent that the item's position in the test captures variation in student endurance or grit, these controls absorb the influence of soft skills that might otherwise confound our estimates of CCSS returns. In a final check, we assess whether high- and low-return CCSS skills are associated with the item's position in the test. Panel b) of Figure 15 shows that high skill-level returns are not disproportionately concentrated on late-test item response.

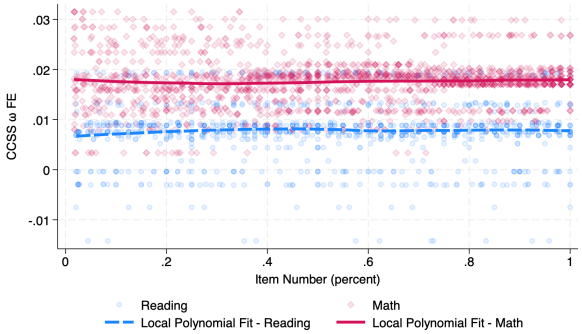
Together, these results suggest that while soft skills —particularly cognitive endurance — may play some role in test performance, our empirical strategy adjusts for these factors. The observed heterogeneity in returns across CCSS skills is therefore unlikely to be an artifact of unmeasured non-cognitive traits.

Figure 15: Item Question Number

(a) Text-Based Prediction



(b) CCSS Returns



Notes: Panel a) of this figure presents regression coefficients on the embeddings-based prediction of ω on item-level observables. The estimates suggest that embeddings are able to capture the item's subject (math/reading) and difficulty, but the embeddings alone are not able to capture the item's position in the test or it's discrimination. Panel b) presents the average position in the test for items associated to each CCSS return level. Overall, we observe no relationship between the CCSS estiamted return and the position in the test of the average item linked to it.

The TAAS is an Easy Assessment: The typical question on the TAAS exams is quite easy – the test was designed to assess whether students are meeting broad, basic learning objectives. The typical IRT-estimated difficulty is well below -1, meaning that the items tend to be most effective at differentiating students with low to very low achievement (a standard deviation or more below the mean). Consistent with this, Table 1 shows that the items are answered correctly about 80% of the time pooling across grades and years. These items were simply not designed to differentiate very high performing students. The items may therefore not cover rarer, higher-level skills that may have significant labor market returns. Put differently, the results are *conditional on the pool*

of TAAS items and are thus most informative about the value of the skills spanned by this pool, that is, the skills covered by the Texas state curriculum.

9 Conclusion

This paper demonstrates that standardized test scores, which aggregate item responses according to psychometric principles, discard economically valuable information. Using administrative data for the universe of Texas public school students linking nearly 1 billion student-item records for 3,555 items to longitudinal earnings records for 1996-2002, we show that different test items exhibit dramatically different predictive power for wage earnings at age 25. We show that this variation in item "skill prices" is consequential: achievement scales that weight items by these skill prices yield white-minority achievement gaps that are roughly 50% larger than what standard achievement scales indicate. Moreover, these item-anchored scales substantially reorder individual student rankings.

Our primary contribution lies in explaining what drives variation in item skill prices. To do this, we use AI tools to digitize item texts from images of old test booklets. We then develop a novel algorithmic method using LLMs to map test items to the Common Core State Standards, producing the first systematic evidence on which academic skills best predict labor market outcomes. Math skills dominate the highest-return academic skills. Within mathematics, procedural, computational, and spatial skills outperform conceptual and interpretive tasks. Within reading, summarizing a text and identifying its main idea outperform word meaning, character analysis, tone analysis, and other more fine-grained tasks. We also show that machine learning models applied to the embeddings associated with the item texts explain additional variation in skill prices beyond readily observable item characteristics – item content and structure contain economically relevant information not captured by standard psychometric parameters.

Our results have wide-ranging implications for education research and policy. Some test items predict economically relevant outcomes much more strongly than others, and this paper pioneers multiple tools that allow one to understand what is driving this variation. This understanding, in turn, could inform almost every educational policy and decision imaginable. For example, knowing which math skills have high returns could inform the design of math curricula. More generally, this paper highlights the value of using item-level data and the power of AI/ML tools in analyzing such data.

References

Ahmed, I., Bertling, M., Zhang, L., Ho, A. D., Loyalka, P., Xue, H., Rozelle, S., and Domingue, B. W. (2025). Heterogeneity of item-treatment interactions masks complexity and generalizability in randomized controlled trials. *Journal of Research on Educational Effectiveness*, 18(4):854–877.

Aisch, G., Gebeloff, R., and Quealy, K. (2014). Where We Came From and Where We Went, State by State. *New York Times*, August 19. https://www.nytimes.com/interactive/2014/08/13/upshot/where-people-in-each-state-were-born.html.

Altonji, J. G. and Pierret, C. R. (2001). Employer learning and statistical discrimination. *Quarterly Journal of Economics*, 116(1):313–350.

Association, N. G. et al. (2010). Common core state standards. *Washington, DC*.

Autor, D. H. (2013). The "task approach" to labor markets: an overview. *Journal for Labour Market Research*, 46(3):185–199.

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, 118(4):1279–1333.

Bach, P., Chernozhukov, V., Kurz, M. S., and Spindler, M. (2021). Doubleml-an object-oriented implementation of double machine learning in r. arxiv.

BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., and Reddy, S. (2024). Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.

Blazar, D., Heller, B., Kane, T. J., Polikoff, M., Staiger, D. O., Carrell, S., Goldhaber, D., Harris, D. N., Hitch, R., Holden, K. L., and Kurlaender, M. (2020). Curriculum reform in the common core era: Evaluating elementary math textbooks across six u.s. states. *Journal of Policy Analysis and Management*, 39(4):pp. 966–1019.

Bond, T. N. and Lang, K. (2013). The evolution of the black–white test score gap in grades k–3: The fragility of results. *The Review of Economics and Statistics*, 95(5):1468–1479.

Bond, T. N. and Lang, K. (2018). The black–white education scaled test-score gap in grades k-7. *Journal of Human Resources*, 53(4):891–917.

Bond, T. N. and Lang, K. (2019). The sad truth about happiness scales. *Journal of Political Economy*, 127(4):1629–1640.

Borgonovi, F., Ferrara, A., and Piacentini, M. (2021). Performance decline in a low-stakes test at age 15 and educational attainment at age 25: Cross-country longitudinal evidence. *Journal of Adolescence*, 92:114–125.

Brown, C., Kaur, S., Kingdon, G., and Schofield, H. (2025). Cognitive endurance as human capital. *Quarterly Journal of Economics*, 140(2):943–1002.

Bruhn, J., Gilraine, M., Ludwig, J., and Mullainathan, S. (2025). Do test scores misrepresent test results? an item-by-item analysis. *National Bureau of Economic Research Working Paper*.

Brunello, G., Crema, A., and Rocco, L. (2021). Some unpleasant consequences of testing at length. *Oxford Bulletin of Economics and Statistics*, 83(4):1002–1023.

Cawley, J., Heckman, J., and Vytlacil, E. (2001). Three observations on wages and measured cognitive ability. *Labour Economics*, 8(4):419–442.

Chen, L., Lin, J., Wang, Z., and Wu, G. (2025). The impact of student's ordinal cognitive ability rank on school violence: Evidence from china. *Economic Modelling*, 143:106967.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? evidence from project star. *Quarterly Journal of Economics*, 126(4):1593–1660.

Christopher Auld, M. and Sidhu, N. (2005). Schooling, cognitive ability and health. *Health Economics*, 14(10):1019–1034.

Cobb, P. and Jackson, K. (2011). Assessing the quality of the common core state standards for mathematics. *Educational Researcher*, 40(4):183–185.

Costrell, R. M. (1997). Can centralized educational standards raise welfare? *Journal of Public Economics*, 65(3):271–293.

Cunha, F., Heckman, J. J., and Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931.

Cunha, F., Nielsen, E., and Williams, B. (2021). The econometrics of early childhood human capital and investments. *Annual Review of Economics*, 13(Volume 13, 2021):487–513.

Cutler, D. M. and Lleras-Muney, A. (2010). Understanding differences in health behaviors by education. *Journal of Health Economics*, 29(1):1–28.

Debeer, D., Buchholz, J., Hartig, J., and Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 pisa reading assessment. *Journal of educational and behavioral statistics*, 39(6):502–523.

Deming, D. and Silliman, M. (2025). Skills and human capital in the labor market. In *Handbook of Labor Economics*, volume 6, pages 115–157. Elsevier.

Deming, D. J. (2023). Multidimensional human capital and the wage structure. *Handbook of the Economics of Education*, 7:469–504.

Du, T., Kanodia, A., Brunborg, H., Vafa, K., and Athey, S. (2024). Labor-llm: Language-based occupational representations with large language models. *arXiv preprint arXiv:2406.17972*.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., and Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, 43(6):1428–1446.

Fryer, Roland G., J. and Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *The Review of Economics and Statistics*, 86(2):447–464.

Fryer, Roland G., J. and Levitt, S. D. (2006). The black-white test score gap through third grade. *American Law and Economics Review*, 8(2):249–281.

Gilbert, J. B., Hieronymus, F., Eriksson, E., and Domingue, B. W. (2024). Item-level heterogeneous treatment effects of selective serotonin reuptake inhibitors (ssris) on depression: Implications for inference, generalizability, and identification. *Epidemiologic Methods*, 13(s2):20240006.

Gilbert, J. B., Himmelsbach, Z., Soland, J., Joshi, M., and Domingue, B. W. (2025). Estimating heterogeneous treatment effects with item-level outcome data: Insights from item response theory. *Journal of Policy Analysis and Management*, 44(4):1417–1449.

Hahm, D. W. (2026). From curriculum to career: Early-career labor market effects of the common core. *Economics of Education Review*, 110:102758.

Haider, S. and Solon, G. (2006). Life-cycle variation in the association between current and lifetime earnings. *American Economic Review*, 96(4):1308–1320.

Hanushek, E. A. (1974). Efficient estimators for regressing regression coefficients. *The American Statistician*, 28(2):66–67.

Hanushek, E. A. and Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature*, 46(3):607–68.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). Random forests. In *The elements of statistical learning: Data mining, inference, and prediction*, pages 587–604. Springer.

Heckman, J. J. and Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, 19(4):451–464.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, 24(3):411–482.

Hedges, L. V. and Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic press.

Hemphill, F. C. and Vanneman, A. (2011). Achievement gaps: How hispanic and white students in public schools perform in mathematics and reading on the national assessment of educational progress. statistical analysis report. nces 2011-459. *National Center for Education Statistics*.

Hess, K. (2009). Hess' cognitive rigor matrix & curricular examples: Applying webb's depth-of-knowledge levels to bloom's cognitive process dimensions – math/science.

Hiebert, E. H. and Mesmer, H. A. E. (2013). Upping the ante of text complexity in the common core state standards: Examining its potential impact on young readers. *Educational Researcher*, 42(1):44–51.

Jacob, B. and Rothstein, J. (2016). The measurement of student ability in modern assessment systems. *Journal of Economic Perspectives*, 30(3):85–108.

Jordan, N. C., Kaplan, D., Ramineni, C., and Locuniak, M. N. (2009). Early math matters: kindergarten number competence and later mathematics outcomes. *Developmental psychology*, 45(3):850.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2):183–202.

Kaestner, R. and Callison, K. (2011). Adolescent cognitive and noncognitive correlates of adult health. *Journal of Human Capital*, 5(1):29–69.

Kapoor, R., Truong, S. T., Haber, N., Ruiz-Primo, M. A., and Domingue, B. W. (2025). Prediction of item difficulty for reading comprehension items by creation of annotated item repository. *arXiv preprint arXiv:2502.20663*.

Kolk, M. and Barclay, K. (2019). Cognitive ability and fertility among swedish men born 1951–1967: evidence from military conscription registers. *Proceedings of the Royal Society B: Biological Sciences*, 286(1902):20190359.

Kusupati, A., Bhatt, G., Rber, A., et al. (2022). Matryoshka representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Lee, S. and Schaelling, M. (2025). Content relatability and standardized testing: Evidence from texas. *Working Paper*.

Lord, F. M. (1975). The 'ability'scale in item characteristic curve theory. *Psychometrika*, 40(2):205–217.

Mansour, H. and McKinnish, T. (2014). Who marries differently aged spouses? ability, education, occupation, earnings, and appearance. *The Review of Economics and Statistics*, 96(3):577–580.

Mazumder, B. (2005). Fortunate sons: New estimates of intergenerational mobility in the united states using social security earnings data. *Review of Economics and Statistics*, 87(2):235–255.

McFadden, D. (1974). The measurement of urban travel demand. *Journal of Public Economics*, 3(4):303–328.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mears, D. P. and Cochran, J. C. (2013). What is the effect of iq on offending? *Criminal Justice and Behavior*, 40(11):1280–1300.

Mocan, N. and Altindag, D. T. (2014). Education, cognition, health knowledge, and health behavior. *The European Journal of Health Economics*, 15:265–279.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. (2023). MTEB: Massive text embedding benchmark. In *Proceedings of EACL*.

Murnane, R. J., Willett, J. B., and Levy, F. (1995). The growing importance of cognitive skills in wage determination. *The Review of Economics and Statistics*, 77(2):251–266.

Murphy, K. M. and Topel, R. H. (2002). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, 20(1):88–97.

Neal, D. (2006). Why has black–white skill convergence stopped? *Handbook of the Economics of Education*, 1:511–576.

Neal, D. A. and Johnson, W. R. (1996). The role of premarket factors in black-white wage differences. *Journal of political Economy*, 104(5):869–895.

Nielsen, E. (2019). Test Questions, Economic Outcomes, and Inequality. *Finance and Economics Discussion Series 2019-013, Federal Reserve Board*.

Nielsen, E. (2023a). How sensitive are standard statistics to the choice of scale? *Unpublished Working Paper*.

Nielsen, E. (2023b). Is the greater variability in achievement for males a psychometric artifact? *Working Paper*.

Nielsen, E. (2025a). The income–achievement gap and adult outcome inequality. *Journal of Human Resources*, 60(4):1217–1252.

Nielsen, E. (2025b). The variance of achievement increases during childhood. *Working Paper*.

Nielsen, E. (2026). Test Questions, Economic Outcomes, and Inequality. *Journal of Political Economy Microeconomics*. Forthcoming.

Opfer, V. D., Kaufman, J. H., and Thompson, L. E. (2016). Implementation of k–12 state standards for mathematics and english language arts and literacy. *Santa Monica, CA: RAND*.

Porter, A., McMaken, J., Hwang, J., and Yang, R. (2011). Common core standards: The new us intended curriculum. *Educational researcher*, 40(3):103–116.

Quinn, D. M. (2015). Kindergarten black–white test score gaps: Re-examining the roles of socioeconomic status and school quality with new data. *Sociology of Education*, 88(2):120–139.

Ramanujam, S. S., Alonso, A., Kataria, S., Dangi, S., Gupta, A., Tiwana, B. S., Somaiya, M., Simon, L., Byrne, D., Ha, S., Zhou, S., Akterskii, A., Liu, Z., Sriram, S., Xiong, C., Pei, Z., Shao, A., Li, A., Xiao, A., Kolb, C., Kistler, T., Moore, Z., and Firooz, H. (2025). Large scale retrieval for the linkedin feed using causal language models. arXiv.

Reardon, S. F. and Galindo, C. (2009). The hispanic-white achievement gap in math and reading in the elementary grades. *American educational research journal*, 46(3):853–891.

Reardon, S. F., Kalogrides, D., and Shores, K. (2019). The geography of racial/ethnic test score gaps. *American Journal of Sociology*, 124(4):1164–1221.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied psychological measurement*, 9(4):401–412.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate behavioral research*, 47(5):667–696.

Reyes, G. (2023). Cognitive endurance, talent selection, and the labor market returns to human capital. *arXiv preprint arXiv:2301.02575*.

Reyes, G. (2025). Cognitive endurance, talent selection, and the labor market returns to human capital. *arXiv preprint arXiv:2301.02575*.

Reyes, G., Riehl, E., and Xu, R. (2024). Stakes and signals: An empirical investigation of muddled information in standardized testing. *NBER Working Paper*, (w32608).

Ritchie, S. J. and Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological science*, 24(7):1301–1308.

Schakel, A. M. J. and Wilson, B. J. (2015). Measuring word significance using distributed representations of words. *arXiv preprint arXiv:1508.02297*.

Schennach, S. (2022). Measurement systems. *Journal of Economic Literature*, 60(4):1223–1263.

Schmidt, W. H. and Houang, R. T. (2012). Curricular coherence and the common core state standards for mathematics. *Educational Researcher*, 41(8):294–308.

Schröder, C. and Yitzhaki, S. (2017). Revisiting the evidence for cardinal treatment of ordinal variables. *European Economic Review*, 92:337–358.

Shearer, D. L., Mulvihill, B. A., Klerman, L. V., Wallander, J. L., Hovinga, M. E., and Redden, D. T. (2002). Association of early childbearing and low cognitive ability. *Perspectives on sexual and reproductive health*, pages 236–243.

Stanford Center for Education Policy Analysis (2012). The educational opportunity monitoring project: Racial and ethnic achievement gaps. https://cepa.stanford.edu/educational-opportunity-monitoring-project/achievement-gaps/race/. Accessed 2025-11-14.

Su, H., Shi, W., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, W.-t., Smith, N. A., Zettlemoyer, L., and Yu, T. (2023). One embedder, any task: Instruction-finetuned text embeddings. In *Findings of ACL*.

Thirukovalluru, R. and Dhingra, B. (2025). GenEOL: Harnessing the generative power of LLMs for training-free sentence embeddings. In Chiruzzo, L., Ritter, A., and Wang, L., editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2295–2308, Albuquerque, New Mexico. Association for Computational Linguistics.

Train, K. E. (2009). Discrete choice methods with simulation.

Ttofi, M. M., Farrington, D. P., Piquero, A. R., Lösel, F., DeLisi, M., and Murray, J. (2016). Intelligence as a protective factor against offending: A meta-analytic review of prospective longitudinal studies. *Journal of Criminal Justice*, 45:4–18. 'Protective Factors against Youth Offending and Violence: Results from Prospective Longitudinal Studies'.

Uttal, D. H., McKee, K., Simms, N., Hegarty, M., and Newcombe, N. S. (2024). How can we best assess spatial skills? practical and conceptual challenges. *Journal of Intelligence*, 12(1):8.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., and Newcombe, N. S. (2013a). The malleability of spatial skills: a meta-analysis of training studies. *Psychological Bulletin*, 139(2):352.

Uttal, D. H., Miller, D. I., and Newcombe, N. S. (2013b). Exploring and enhancing spatial thinking: Links to achievement in science, technology, engineering, and mathematics? *Current Directions in Psychological Science*, 22(5):367–373.

Webb, N. L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. research monograph no. 6.

Webb, N. L. (2002). Depth-of-knowledge levels for four content areas. Unpublished paper, March 28, 2002.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, 20(1):7–25.

Weirich, S., Hecht, M., Penk, C., Roppelt, A., and Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied psychological measurement*, 41(2):115–129.

Zhang, Y. et al. (2025). Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Online Appendices

A Data Appendix

A.1 Texas Assessment of Academic Skills

A.1.1 Subjects and Objectives

Since 1996, the state of Texas testing program has consistently measured student learning across subjects and grades every academic year. Table A.1 displays the subjects that are tested across all grades for the Texas Assessment of Academic Skills (TAAS).

Table A.1: Test Subjects

		Grades								
		3rd	4th	5th	6th	7th	8th	9th	10th	11th
	Reading	X	X	X	X	X	X		X*	
TAAS	Mathematics	X	X	X	X	X	X		X*	
1996-2002	Writing		X				X			

Notes: * Exit exam. Typically administered in 10th grade, and ocasionally in 11th grade.

Table A.2 provides and overview of the test structure across grades for both reading and math standardized test administrations. On average, reading tests were designed with fewer items than math tests.[57] Students tend to respond correctly to reading items at a higher rate than math ones. Finally, the number of items increase as grade level increases.[58]

Table A.2: Item-level Statistics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Reading				Mathematics					
	No.	Correct	IRT Diff.	IRT Disc.	No.	Correct	IRT Diff.	IRT Disc.	Obs.	Booklets
Grade 3	36	0.82	-1.38	1.71	44	0.79	-1.48	1.36	1,716,485	0.71
Grade 4	40	0.80	-1.43	1.54	50	0.79	-1.55	1.40	1,779,431	0.86
Grade 5	40	0.82	-1.47	1.49	52	0.81	-1.61	1.37	1,764,271	0.86
Grade 6	40	0.79	-1.22	1.54	56	0.79	-1.35	1.46	1,814,889	0.86
Grade 7	45	0.80	-1.29	1.55	58	0.74	-1.08	1.45	1,827,132	0.71
Grade 8	48	0.79	-1.31	1.51	60	0.74	-1.17	1.44	1,855,899	0.43
Exit	48	0.75	-1.29	1.52	60	0.69	-0.85	1.55	1,974,388	0.86

The TAAS was designed to measure broad learning objectives consistently over time. For each testing item, we observe which learning objective it corresponds to. Table A.3 presents a list of these objective for Reading and Mathematics tests.

[57] The number of items for each grade-year combination did not change throughout the period of our study (1996-2002).

[58] See Section A.1.3 for more details on the selection process of test items.

Table A.3: Summary of Reading and Mathematics Objectives

No.	Reading	Math
1	Word Meaning	Number Concepts
2	Supporting Ideas	Algebraic Understanding
3	Text Summarization	Geometric Properties
4	Identify Relationships	Measurement Concepts
5	Analyze Inferences	Probability Statistics
6	Recognize Perspectives	Addition Problems
7		Subtraction Problems
8		Multiplication Problems
9		Division Problems
10		Estimate Solutions
11		Solution Strategies
12		Mathematical Representation
13		Solution Evaluation

A.1.2 Scaling

The TAAS scale scores were derived using a one-parameter Item Response Theory (IRT) or Rasch model. Under the Rasch model, the probability that a person i with (unobserved) ability θ_i answers an item m with difficulty δ_m correctly, is defined as:

$$P(X_{i,m} = 1) = \frac{\exp(\theta_i - \delta_m)}{1 + \exp(\theta_i - \delta_m)} \tag{11}$$

Under a logit formulation, the Rasch model yields a difficulty estimate for each test item ($\hat{\delta}_m$) and latent ability estimate for each student ($\hat{\theta}_i$).

The TAAS scale scores are calibrated under a 70%-correct standard, which means that a $\theta_{standard}$ is defined to be the student ability level that on average leads to a 70% correct rate. $\theta_{standard}$ is also know as the *R at standard* and can be interpreted as the logit-scale ability estimate for a typical test taker who earns 70% correct on the test, under the Rasch model.

The scale score at this standard is then set at 1500 — the passing standard. All other deviations on the logit scale as transformed following:

$$score = \frac{R - R_{atstandard}}{\sigma_R} * 200 + 1500 \tag{12}$$

Under this transformation, scale scores range from 400-2400 with a passing standard set at 1500 (corresponding to 70% correct standard). This scale transformation also ensures that the passing standard is maintained at the same level of difficulty across administrations. However, note that the passing standards are set independently at each grade. Thus, direct comparisons of performance across grades should not be made.

A.1.3 Item Selection

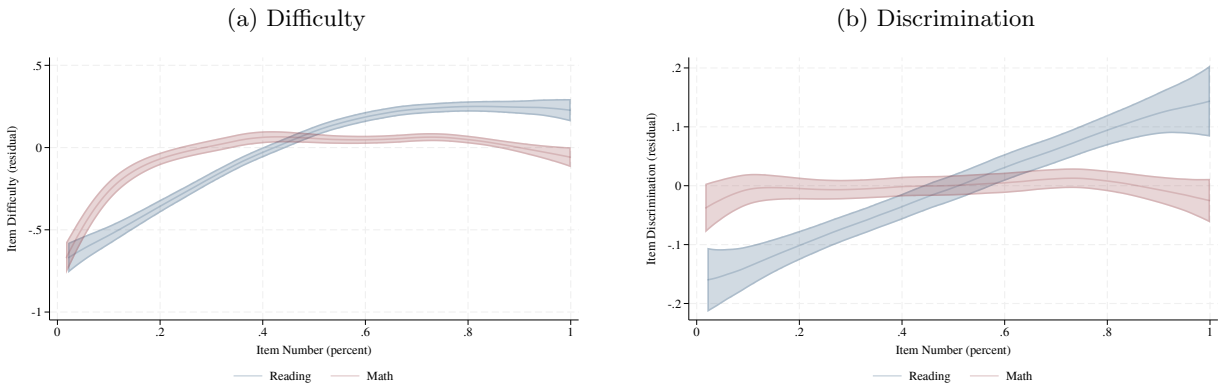
The development, publication, and distribution of TAAS was contracted out to Harcourt Education Measurement (HEM — now Pearson PLC). HEM item writers developed items for TAAS that fell under their specific content-area knowledge or their teaching/curriculum development experience — many item writers were current or former Texas teachers. HEM provided training to item writers that highlighted the scope of the testing program, security issues, adherence to the measurement specifications, and avoidance of possible economic, regional, cultural, gender, and ethnic bias. Items were reviewed annually by HEM to check the appropriateness of the items to the test objectives, difficulty range, clarity of the items, correctness of answer choices, and plausibility of the distractors. As well as their depiction of minority, gender, and other demographic groups.

Items were then submitted to the Texas Education Agency (TEA) for review. For this review TEA's Student Assessment Division convened committees composed of teachers, curriculum directors, principals, superintendents, and administrators from regional education service centers to work with TEA staff in reviewing test items developed by HEM. Under this review, the committees scrutinized each item for content-to-specification match, item difficulty, plausibility of the distractors, and any potential ethnic, gender, economic, or cultural bias.[59]

At the end of this vetting process by TEA's Student Assessment Division, items were then field tested. Newly developed items were embedded in regular Spring test administrations for representative samples of students from across the test. Student responses for these items were not included in their test score calculation. Rather, these data were reviewed to determine whether new items would be included in the following testing cycle.[60]

A.1.4 Item Difficulty, Discrimination, and Position in the Test

Figure A.1: Item Question Number



Note: Each panel shows a local linear polynomial fit (bandwidth = 0.15) of the relevant residualized psychometric characteristic on the item's number (as a percent of the relevant exam), along with the associated 95% confidence intervals. Both difficulty and discrimination are residualized using fully interacted fixed effects for subject, grade, year, and learning objective.

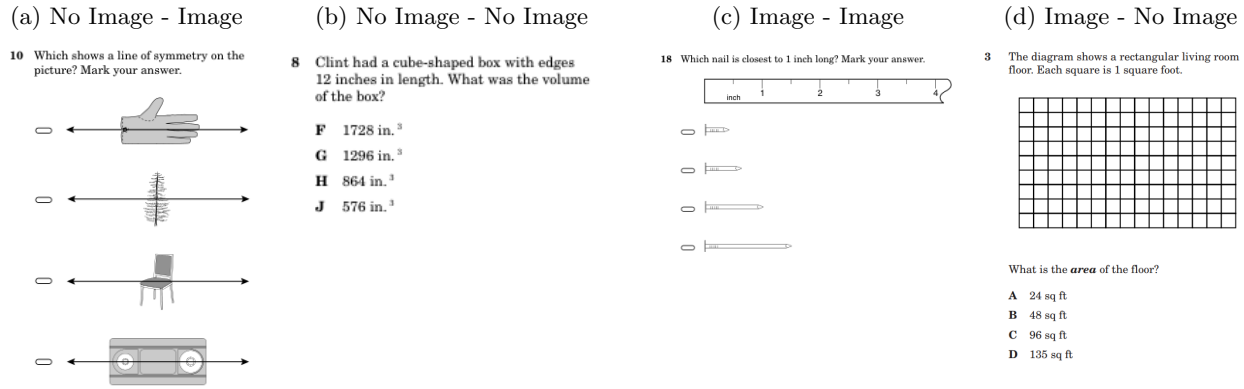
[59] The TEA review was exhaustive. For example in 1999 these committees met 75 times to review all newly developed test items and all new field test data.

[60] Annual test releases to the public did not include these field testing items.

Panels a) and b) of Figure A.1 examines the relationship between an item's difficulty and discrimination and its placement within the exam. Panel (a) shows that the items that come later in an exam tend to be more difficult. However, this relationship is not linear for either math or reading – after a certain point items do not become more difficult as one progresses through the exam. Panel (b) shows that there is no relationship between an item's discrimination and its placement in the exam for math, but a strong positive relationship for reading. [61]

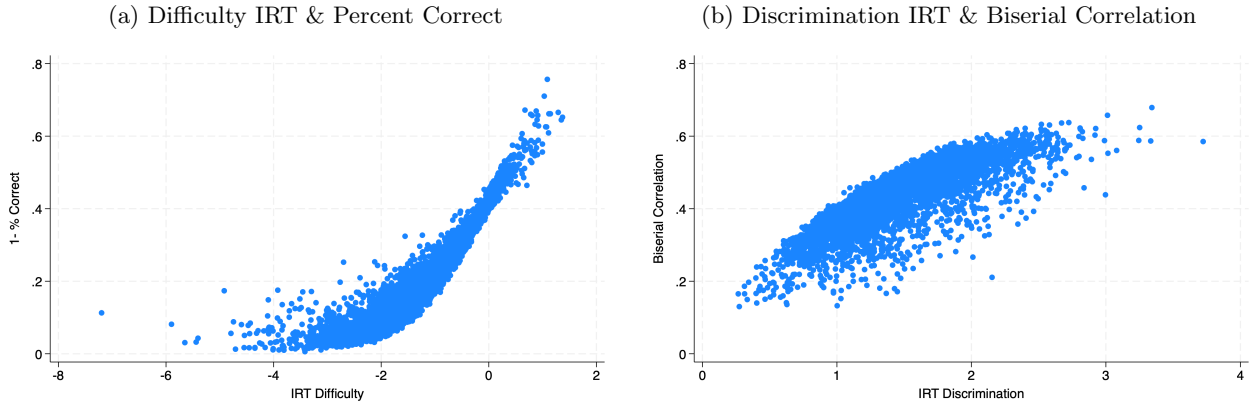
A.1.5 Other

Figure A.2: Question Examples By Type



[61] Even in the case of discrimination and reading, where the relationship is positively and linear, the item's position in the test explains only about 20% of the variation in discrimination. Note also that the patterns within each grade are quite similar to what is presented in Figure 15 (not shown).

Figure A.3: Relationship Between IRT Estimates and Proxies



A.2 Anchoring Timelines for Earnings

We focus on exploring the anchoring timeline for student wages at ages 25, 30, and 35. Panel (a) of Figure A.4 presents the mapping between testing year (horizontal axis) and earnings at ages 25 year (vertical axis). This plot shows that for all grades between 1996-2002 we expect to have complete and reliable wage information for students when they turn 25. For example, a third grader that is tested in 2002, will, on average, turn 25 years of age in the year 2018. Because we have earnings data up until 2019, we can expect to have good quality earning matches for this group of students.

To examine empirically whether this is the case, panel (b) of Figure A.4 displays the share of students for which we observe wages at age 25 across grades and time. This figure shows that, on average, we observe stable earnings data at age 25 for all grades and years, with a match rate of around 72% across grade-year combinations.

Panels (c) and (d) of Figure A.4 show analogous plots for earnings captured at age 30. At this age, panel c) shows that some early grades tested at later years (e.g. third graders in 2002) would not have feasible matches as they have not yet turned 30 by 2019. This is corroborated by panel d) of Figure A.4 that shows low earnings match rates for these grade-year combinations. The average match rate for *feasible* grade-year combinations is 66% for earnings at age 30. Finally, panels and (e) and (f) of Figure A.4 show analogous plots for earnings captured at age 35. At this age, we are unable to capture reliable earnings information for most grades and years (as shown by low earnings match rates in panel (f)).

A.3 Booklet Data

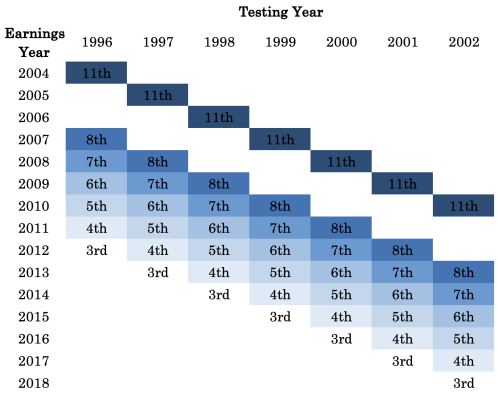
Table A.4 presents test booklet availability by grade-year. An 'X' indicates that the test booklet for that grade-year was recovered and digitized. All test booklets in 1996 are missing. For subsequent years, some grades are missing booklets. Overall, we were able to recover 88% of the test booklets for the years 1997-2002.

Table A.4: Booklet Availability

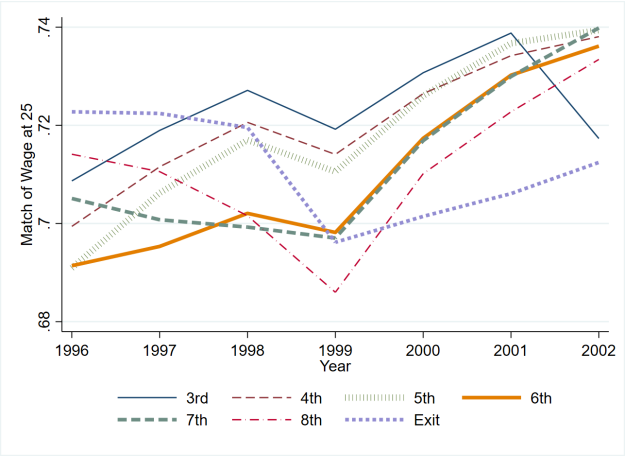
School Year	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Exit
1995-1996							
1996-1997		X	X	X	X		X
1997-1998	X	X	X	X	X		X
1998-1999	X	X	X	X			X
1999-2000	X	X	X	X	X	X	X
2000-2001	X	X	X	X	X	X	X
2001-2002	X	X	X	X	X	X	X

Figure A.4: Earnings Anchoring Timeline for Test-takers

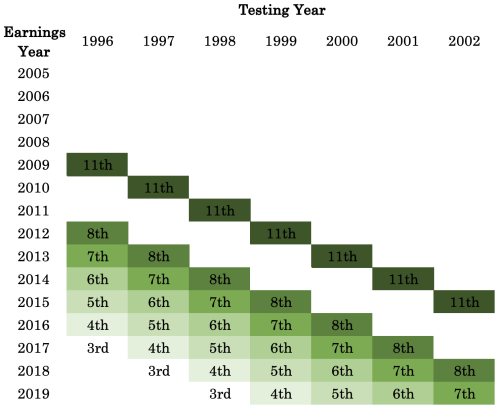
(a) Anchoring Timeline @ 25



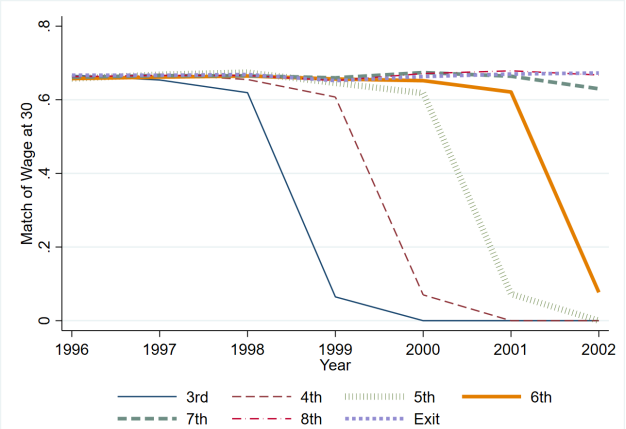
(b) Earnings Match Rate @ 25



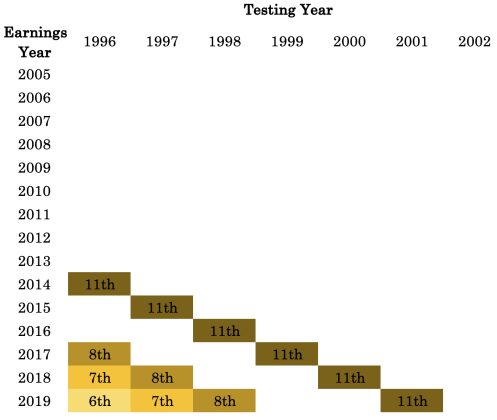
(c) Anchoring Timeline @ 30



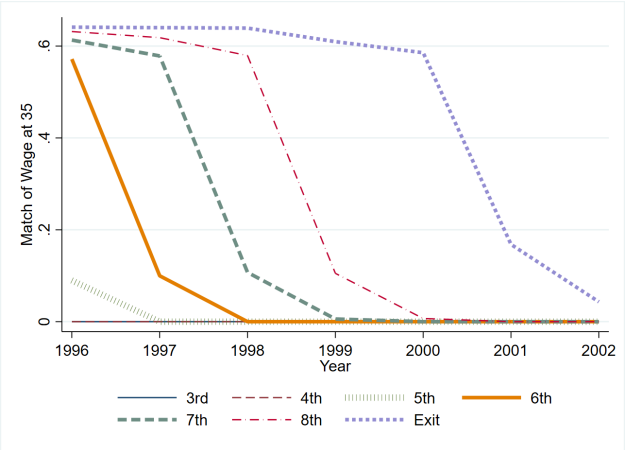
(d) Earnings Match Rate @30



(e) Anchoring Timeline @ 35



(f) Earnings Match Rate @35



B Achievement Gaps and Attenuation Bias

To illustrate how item-anchored and given achievement gaps can differ, even when they are on the same scale, suppose that the (population) item-anchored (A_i) and given scores (Z_i) are both given by weighted sums of the items, where both sets of weights sum to one:[62] $A_i = \sum_m \omega_m D_{i,m}$, $Z_i = \sum_m \alpha_m D_{i,m}$. Then, consider the mean achievement difference between students from groups H and L, and let $p_{m,H} = \mathbb{E}[D_{i,m}|i \in H]$ and $p_{m,L} = \mathbb{E}[D_{i,m}|i \in L]$ be the probabilities of correct answers to item m for groups H and L respectively. Then the difference in the item-anchored and given $H - L$ achievement gaps is given by

$$\sum_m (\omega_m - \alpha_m)(p_{m,H} - p_{m,L}). \tag{13}$$

This equation shows that, abstracting from scaling differences, item-anchored and given achievement gaps will different if the two scales weight items differently that are answered by the two groups under comparison at different rates.

A naive estimate of the achievement gap between groups H and L would be the in-sample mean of $\hat{A}$ for group-H students less then mean for group-L students. However, this estimator will be biased – measurement error in $\hat{A}$ will tend to attenuate this naive estimator towards zero. That is, sampling variability in $\widehat{\Psi}$ means that $\hat{A}_i = A_i + \nu_i$ for some error term ν_i.[63]

Supposing that $A \sim N(\bar{A}, \sigma_A^2)$ and $\nu \sim N(0, \sigma_\nu^2)$,

$$\mathbb{E}[S|\hat{A}_i] = R_{A,\nu}\hat{A}_i + (1 - R_{A,\nu})\bar{A}, \text{ where } R_{A,\nu} \equiv \frac{\sigma_A^2}{\sigma_A^2 + \sigma_\nu^2}. \tag{14}$$

Equation (14) is intuitive: because the anchored scale is a noisy estimate of true (anchored) achievement at the individual level, the best guess of student i's achievement gives weight to both the observed, noisy score for i and the population mean score. Student i's estimated score is given more weight in this sum the less noisy a measure it is (that higher is $R_{A,\nu}$). Thus, letting $\hat{A}_H$ and $\hat{A}_L$ denote the group-level averages of $\hat{A}$, we get

$$\text{plim}_{N\to\infty}(\hat{A}_H - \hat{A}_L) = R_{A,\nu}(\bar{A}_H - \bar{A}_L) < (\bar{A}_H - \bar{A}_L). \tag{15}$$

Equation (15) shows that, while "shading" towards the population mean is optimal at the individual level, this shading is not needed for group mean achievement differences because measurement error in the estimated anchored scales is immaterial at the group level.

Thus, the consistent estimation of $(\bar{A}_H - \bar{A}_L)$ requires a consistent estimate of $R_{A,\nu}$. Consider the ordinal least squares estimate $\hat{\gamma}$ from $\hat{A}_i = \kappa + \gamma S_i + \epsilon_i$. Because S_i is a noisy estimate of A_i, via equation (**??**), $\hat{\gamma}$ will be attenuated towards zero. To solve this errors-in-variables problem, we

[62] For given achievement scales based on item response theory, the weighted sum formula below will be an approximation. This approximation will be more accurate the more items the test has.

[63] If f is correctly specified, which is the assumption we maintain throughout this analysis, ν_i will come only through sampling variability in $\widehat{\Psi}$. More generally, ν_i would also pick up any misspecification in the anchoring relationship f.

seek an instrument for S_i, some Z_i that is correlated with A_i and uncorrelated with η_i.

The richness of our item-level data and the large size of our year-grade samples allows for the construction of many such instruments by estimating models separately using disjoint subsets of the test items. In particular, let $\mathbf{D}_i^{(1)}$ and $\mathbf{D}_i^{(2)}$ denote the odd- and even-numbered item responses for student i. We then use the item and outcome data to estimate $\hat{A}_i^{(1)} = f(\mathbf{D}_i^{(1)}, \mathbf{X}_i; \widehat{\mathbf{\Psi}}^{(1)})$ and $\hat{A}_i^{(2)} = f(\mathbf{D}_i^{(2)}, \mathbf{X}_i; \widehat{\mathbf{\Psi}}^{(2)})$. In words, $\hat{A}_1^{(1)}$ and $\hat{A}_1^{(2)}$ are the estimated item-anchored achievement measures using only the odd and even items. Each of these scores is a noisy measure of A_i. We thus take $\hat{A}_i^{(1)}$ as our "base" measure of item-anchored achievement. We then use the even-anchored scores $\{\hat{A}_i{}^{(2)}\}$ to construct the necessary instruments for the odd-item achievement estimates.

An instrument for S_i when $\hat{A}_i^{(1)}$ is the base achievement measure is the average S among test takers other than i (to avoid a mechanical correlation) but who nevertheless have the same value of $\hat{A}^{(2)}$ as i. That is,

$$Z_i^{(1)} = N_i^{-1} \sum_{j \neq i : \hat{A}_j^{(1)} = \hat{A}_i^{(1)}} S_j, \tag{16}$$

where N_i is the number of students other than i satisfying the condition $\hat{A}_j^{(1)} = \hat{A}_i^{(1)}$. This condition ensures the relevance of the instrument, and exogeneity is guaranteed by the leave-on-out construction.

In broad outline, then, our approach consists of the following steps:

1. For a particular choice of f and estimation approach, estimate $\hat{A}_i^{(1)} = f(\mathbf{D}_i^{(1)}, \mathbf{X}_i; \widehat{\mathbf{\Psi}}^{(1)})$ and $\hat{A}_i^{(2)} = f(\mathbf{D}_i^{(2)}, \mathbf{X}_i; \widehat{\mathbf{\Psi}}^{(2)})$.

2. Estimate the biased achievement gap between student groups H and L using the sample averages of $\hat{A}_i^{(1)}$:
$$\hat{\Delta}_{HL} = \frac{1}{N_H} \sum_{i \in H} \hat{A}_i^{(1)} - \frac{1}{N_L} \sum_{i \in L} \hat{A}_i^{(1)}.$$

3. Construct $Z_i^{(1)}$ according to equation (16). Regress $\hat{A}_i^{(1)}$ on S_i, instrumenting S_i with $Z_i^{(1)}$. Denote the resulting IV coefficient by $\hat{\gamma}_{IV}^{(1)}$. This regression coefficient estimates $R_{A,\nu}$.

4. Estimate the corrected HL item-anchored achievement gap by $\hat{\Delta}_{HL}/\hat{\gamma}_{IV}^{(1)}$.

Table B.1: Balance Across Observables for Odd and Even Items

	(1)	(2)	(3)
	Even	Odd	Difference
IRT Difficulty	-1.285	-1.325	0.040
IRT Discrimination	1.481	1.489	-0.009
Percent Correct	0.798	0.804	-0.006
Biserial Correlation	0.435	0.435	-0.000
Math	0.562	0.560	0.002
Question # as %	0.520	0.501	0.019*
Grade: 3	0.118	0.118	0.000
Grade: 4	0.133	0.133	0.000
Grade: 5	0.136	0.136	0.000
Grade: 6	0.142	0.142	0.000
Grade: 7	0.151	0.153	-0.003
Grade: 8	0.160	0.159	0.000
Grade: Exit	0.160	0.159	0.000
Observations		4739	

C Assessing Different Anchor Models

This appendix presents evidence justifying our reliance on simple linear-in-items anchor models estimated via OLS. We provide two lines of evidence. First, we show that different, plausible anchor model specifications and estimation approaches yield similar $\hat{A}_i$ and $\widehat{\mathbf{\Omega}}$ estimates. We then discuss the results of a number of Monte Carlo experiments which demonstrate that linear OLS works "well" even in settings where the true data generating process is not linear.

C.1 Anchor Results Under Different Model Specifications

Table C.1: Item-level Tests for Statistical Differences Across Specifications

	(1)	(2)
Rejected	Main v. County FE	Main v. White Males
5%-level	0.13%	1.16%
10%-level	0.19%	1.50%

Notes: This table presents the results of item-level statistical tests for differences in estimates between our chosen specification and an alternative one. Column (1) presents results of differences between our main specification and one that uses county FE instead on commuting zone FE. Column (2) presents results of differences between our main specification and one that restricts the sample to white males one. p-values were adjusted using a Bonferroni correction at the grade-year level.

C.2 Monte Carlo Analysis

In this section, we report the results of a number of Monte Carlo experiments designed to assess the performance of linear-in-items OLS models in situations where the data generating process is known to contain nonlinearities (interactions). Across a wide variety of true data-generating models with different numbers of two- and three-way interactions, we find that OLS performs quite well.

Data Generating Process

We generate data according to the following process:

1. We fix N as the total number of students and M as the total number of items. We set $\bar{R}^2$ as the desired share of outcome variation due to the items.
2. Each student i's academic ability θ_i is drawn independently from $N(0,1)$.
3. For each item m, we set the IRT parameters and item response probabilities as follows:
   (a) Guessing: $c_m = 0.25,\ \forall m$
   (b) Discrimination: $a_m = 1,\ \forall m$
   (c) Difficulty: $b_m \sim N(0,1)$ drawn independently $\forall m$

(d) For each (i, m), the probability of a correct response is given by the three parameter logistic IRT model:

$$p_{i,m} = c_m + \frac{1 - c_m}{1 + e^{-a_m(\theta_i - b_m)}}.$$

4. For each i, we construct the vector of item responses $\mathbf{d}_i$ by drawing each $d_{i,m}$ from a Bernoulli($p_{i,m}$) distribution.

5. For each item m, we construct the "linear" item outcome weights according to

$$\omega_m^{(1)} = \mu + \gamma * b_m + \xi_m$$

where $\xi_m \sim N(0, \sigma_1^2)$. In this formulation, μ gives the average outcome weight, and γ controls how strongly related are item difficulty and outcome weight.

6. Achievement under linearity is then defined by

$$A_i^{(1)} = \sum_m \omega_m^{(1)} * d_{i,m}.$$

7. The observed outcome under linearity is

$$S_i^{(1)} = A_i^{(1)} + \upsilon_i^{(1)}$$

where $\upsilon_i^{(1)}$ in an iid draw from $N(0, \tilde{\sigma}_1^2)$. Here, $\tilde{\sigma}_1^2 = \widehat{Var}(A_i)(1/\bar{R}^2 - 1)$ is set so that items explain a fraction $\bar{R}^2$ of the variation in S.

8. We generate achievement and outcomes with two-way interactions by supposing that, for each pair of distinct items m and m', there is a non-zero interaction between them with probability $p^{(2)}$. That is, for each pair (m, m'), we draw $\psi_{m,m'}^{(2)}$ from a Bernoulli($p^{(2)}$) distribution. We generate interactions until we reach a fixed number $T^{(2)} < \binom{M}{2}$ set as a parameter of the simulation.

9. If $\psi_{m,m'}^{(2)} = 1$, we generate an interaction weight for (m, m') according to $\omega_{m,m'}^{(2)} \sim N(\mu_2, \sigma_2^2)$. In practice, throughout we set $\mu_2 = 0$ and $\sigma_2^2 = 1$. If $\psi_{m,m'}^{(2)} = 1$, we set $\omega_{m,m'}^{(2)} = 0$.

10. The true achievement with interactions is then given by

$$A_i^{(2)} = A_i^{(1)} + \sum_m \sum_{m' \neq m} \psi_{m,m'}^{(2)} \omega_{m,m'}^{(2)} d_{i,m} d_{i,m'}.$$

11. Then, as in the linear case, we construct the observed outcomes $S_i^{(2)} = A_i^{(2)} + \upsilon_i^{(2)}$ with $\upsilon_i^{(2)}$ distributed normally with a variance chosen as in the linear case so that items (with the relevant interactions) account for a share $\bar{R}^2$ of the variation in $S_i^{(2)}$.

12. The three-way interaction achievements $\{A_i^{(3)}\}$ and outcomes $\{S_i^{(3)}\}$ are constructed following an analogous process. Thus, for both the three-way and two-way interactions, there is no systematic relationship between the "kinds" of items that are interacted. We generate three-way interactions up to $T^{(3)} < \binom{M}{3}$.

13. The end result for a given choice of parameters, sample sizes, etc., is a data set for N individuals where each $i \in N$ is defined by $(A_i^{(1)}, S_i^{(1)}, A_i^{(2)}, S_i^{(2)}, A_i^{(3)}, S_i^{(3)}, \mathbf{d}_i)$. We also retain $\{\omega_m^{(1)}, \omega_{m,m'}^{(2)}, \omega_{m,m',m''}^{(3)}, \psi_{m,m'}^{(2)}, \psi_{m,m',m''}^{(3)}, T^{(2)}, T^{(3)}\}$.

Assessing Different Anchor Models

For each combination of parameters and each interaction order (1-way, 2-way, 3-way) we estimate the following models:

1. *Linear OLS:* This corresponds to the baseline specification from which we calculate $\widehat{\mathbf{\Omega}}$ in the main body of the paper. We simply run a linear regression of $S^{(k)}$ on D, where $k \in \{1, 2, 3\}$.

2. *Linear LASSO:* We estimate the exact same specification as the linear OLS case, but instead we fit a LASSO model with the penalty parameter set by cross-validation.

3. *Multi-Way LASSO:* We estimate LASSO models that consider all possible item interactions of all orders up to whatever order interaction actually generated the data. Thus, for example, if we are considering $S^{(2)}$, we fit a LASSO model where the right-hand side consists of item indicators and indicators for each possible item interaction ($\binom{M}{2} + M$ total indicators).

4. *Random Forest:* We estimate random forests for each model using the 'ranger' package in R with the number of trees capped at 500. We consider multiple different values of 'mtry', which is the parameter that governs the size of the random subset of items considered for each split of a node. We first fit random forests with mtry=1, which constitutes extreme feature subsampling. The trees tend to have low correlation in this case. We also assess random forest models where mtry is set to either the order of the interaction being considered, $M/3$, or $\sqrt{M}$, with the last two being commonly-employed rules-of-thumb (Hastie et al., 2008).

Our simulations proceed through the following steps:

1. Select a sample size N and number of items M, and then generate data according to the process outlined above.

   (a) We set $\mu = 4$, $\mu_2 = \mu_3 = 2$, $\gamma = 1$, $\sigma_1 = \sigma_2 = \sigma_3 = 0.1$, and $\bar{R}^2 = 0.2$.

   (b) We consider $N \in \{20,000;\ 200,000\}$ and $M \in \{50, 250\}$.

2. Randomly split the sample in half. Estimate the models listed above on one of the random subsamples. Then, compute $\widehat{\mathbf{\Omega}}$ using each fitted model using the other random subsample (the holdout).

3. For each model, compare the elements of $\widehat{\mathbf{\Omega}}$ to the estimated "true" weights given by

$$\begin{aligned}
\hat{\tau}_m^{(1)} &= \omega_m^{(1)} \\
\hat{\tau}_m^{(2)} &= \tau_m^{(1)} + \sum_m \sum_{m' \neq m} \psi_{m,m'}^{(2)} \omega_{m,m'}^{(2)} \widehat{\mathbb{E}}[D_{i,m'} | D_{i,m} = 1] \\
\hat{\tau}_m^{(3)} &= \hat{\tau}_m^{(2)} + \sum_m \sum_{m' \neq m} \sum_{m'' \neq m'} \psi_{m,m',m''}^{(3)} \omega_{m,m',m''}^{(3)} \widehat{\mathbb{E}}[D_{i,m'} D_{i,m''} | D_{i,m} = 1],
\end{aligned}$$

where

$$\begin{aligned}
\widehat{\mathbb{E}}[D_{i,m'} | D_{i,m} = 1] &= \frac{\sum_{i \in H} d_{im} d_{im'}}{\sum_{i \in H} d_{im}} \\
\widehat{\mathbb{E}}[D_{i,m'} D_{i,m''} | D_{i,m} = 1] &= \frac{\sum_{i \in H} d_{im} d_{im'} d_{im''}}{\sum_{i \in H} d_{im}}.
\end{aligned}$$

That is, we compare errors defined by $e_m^{(k)} = \hat{\omega}_m^{(k)} - \hat{\tau}_m^{(k)}$ for $k \in \{1, 2, 3\}$.

Result 1: OLS has low bias in all cases considered.

Notably, OLS performs well, in the sense of being approximately unbiased, even in the face of two-way and three-way interactions. This result is evident in Figures C.1 and C.4 below.

Result 2: OLS and LASSO have similar bias.

Figure C.1 provides a representative picture of the OLS vs. LASSO comparisons in our Monte Carlo experiments. The left panel compares the error distributions for OLS and various lasso models for a data generating process that features two-way item interactions but no higher-order interactions. The right panel shows analogous error distributions for a case with both two- and three-way item interactions. In both panels, it is clear that the OLS estimates have very similar mean errors as the lasso models. The LASSO models with interactions do have lower RMSEs; typically, the lasso models that are correctly specified (i.e. two-way interactions when that is the true dgp) perform best.

Figure C.1: OLS versus LASSO Estimates of $\mathbf{\Omega}$

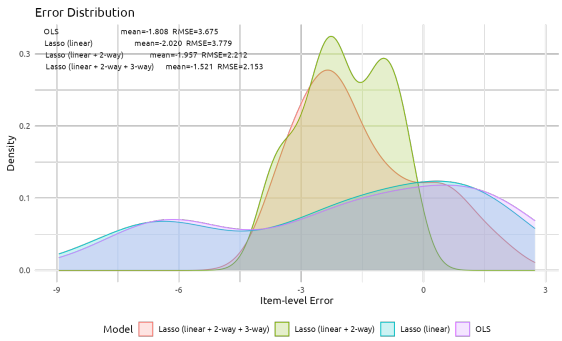


Figure C.2: Two-Way (N =20k; $M = 50$)

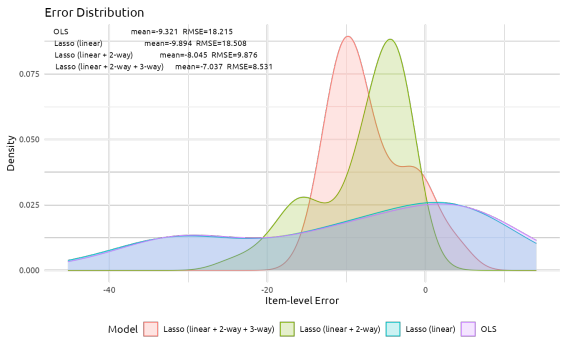


Figure C.3: Three-Way (N =20k; $M = 50$)

Notes: The left panel shows the error distributions for $\widehat{\mathbf{\Omega}}$ for the case with two-way interactions. The right panel show the same but for data-generating processes featuring both two-way and three-way interactions.

Result 3: OLS typically performs better than random forests.

Figure C.4 provides a representative picture of the OLS vs. random forest comparisons in our Monte Carlo experiments. The left panel compares the error distributions for OLS and various random forest models for a data generating process that features some two-way item interactions but no higher-order interactions. The right panel shows analogous error distributions for a case with both two- and three-way item interactions. In both panels, it is clear that the OLS estimates have lower mean errors than any of the random forest estimates. Moreover, the OLS estimates also have lower RMSEs.

Figure C.4: OLS versus Random Forest Estimates of $\mathbf{\Omega}$

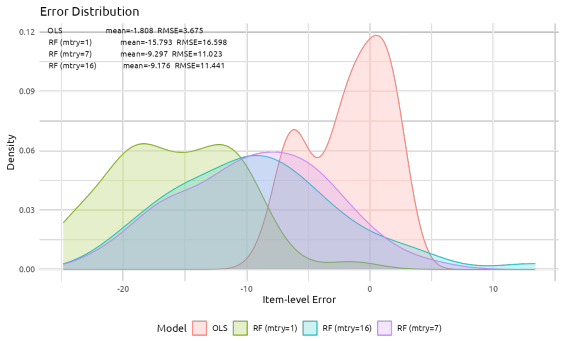


Figure C.5: Two-Way (N =20k; $M = 50$)

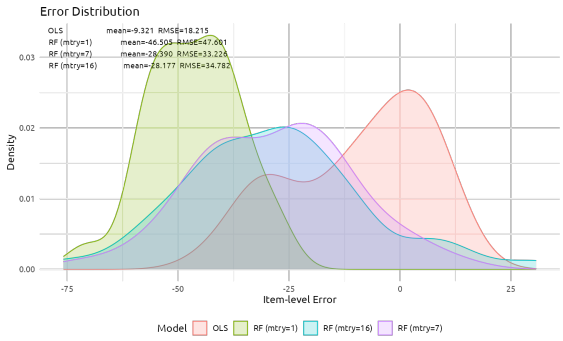


Figure C.6: Three-Way (N =20k; $M = 50$)

To show that these results are not unusual, we plot in Figure C.7 the distributions of the OLS and random forests mean errors and RMSEs for 250 iterations of the above analysis. Both panels make clear that OLS dominates the random forest models for both 2-way and 3-way interacted DGPs both in terms of mean error and in terms of RMSE.

Figure C.7: Bootstrapped Error Distributions: OLS vs. Random Forests

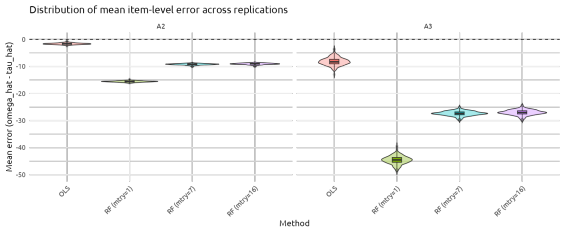


Figure C.8: Mean Errors (N =20k; $M = 50$)

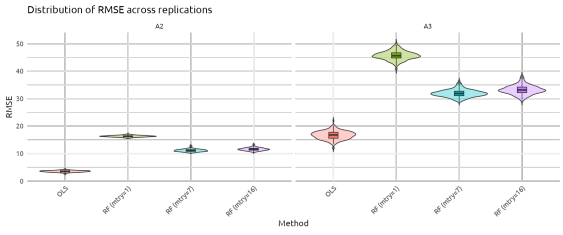


Figure C.9: RMSEs (N =20k; $M = 50$)

C.3 Double-Debiased Methods

The estimates for $\mathbf{\Omega}$ assessed above suffer from a number of well-known problems inherited from the ML models used to estimate f. First, because $\hat{f}$ is obtained from regularized/flexible ML models, the plug in estimator for $\hat{\omega}$ can have non-negligible finite sample bias. Indeed, this is evident by the non-zero locations of the error distributions in Figures C.1 - C.7. Second, $\hat{\omega_m}$ averages over the

marginal distribution of $\mathbf{D}_{-m}$, but the conditional distributions $\mathbf{D}_{-m}|D_m = 1$ and $\mathbf{D}_{-m}|D_m = 0$ might differ substantially. When the overlap in these distributions is limited, the plug-in estimator can be sensitive to extrapolation into regions where either $D_m = 1$ or $D_m = 0$ is rare.

Chernozhukov et al. (2018) develop a double/debiased machine learning (DML) approach that corrects for model bias and reweights to address imbalance-induced extrapolation bias.[64] The DML estimator adds to the simple plug-in estimator $\hat{\omega}$ propensity-reweighted residual corrections with cross-fitting to evaluate nuisance estimates out-of-sample.

To adapt the DML method to our setting define the following nuisance objects:

$$\begin{aligned} \mu_{m,1}(d_{-m}) &= \mathbb{E}[Y|D_m = 1, \mathbf{D}_{-m} = d_{-m}] \\ \mu_{m,0}(d_{-m}) &= \mathbb{E}[Y|D_m = 0, \mathbf{D}_{-m} = d_{-m}] \\ p_m(d_{-m}) &= Pr(D_m = 1|\mathbf{D}_{-m} = d_{-m}). \end{aligned}$$

We estimate the nuisance functions $\mu_{m,1}(x)$ and $\mu_{m,0}(x)$ via OLS, LASSO, or random forests as the case may be. The propensity scores we estimate via lasso allowing for 2-way interactions.[65] Then, the per-observation score contribution is

$$\hat{\xi}_{im} = \hat{\mu}_{m,1}(d_{i,-m}) - \hat{\mu}_{m,0}(d_{i,-m}) + \frac{d_{im}}{\hat{p}_m(d_{i,-m})}\left(s_i - \hat{\mu}_{m,1}(d_{i,-m})\right) - \frac{1 - d_{im}}{1 - \hat{p}_m(d_{i,-m})}\left(s_i - \hat{\mu}_{m,0}(d_{i,-m})\right). \tag{17}$$

The DML estimate for ω_m is then given by

$$\hat{\omega}_m^{\text{DML}} = \frac{1}{N}\sum_i \hat{\xi}_{im}.$$

Intuitively, the residual terms correct errors in $\hat{f}(1,\cdot)$ ($\hat{f}(0,\cdot)$), while the propensity weights re-target those corrections to the marginal distribution of D_{-m} relevant for ω_m. We implement this approach using our same test-train samples as in the non-DML case (that is, we do not employ $K-$fold crossfitting). If we were interested in conducting inference on the resulting estimates $\hat{\omega}_m^{\text{DML}}$, then we would use $K-$fold crossfitting. Then, the Neyman orthogonality of the DML scores helps – the first-stage estimation errors are second-order, and by assumption the OLS estimates are $O_p(n^{-0.5})$. Thus, the propensity scores need only be $o_p(1)$. Stacking vertically across m, the full variance-covariance matrix is then

$$\widehat{\mathbf{\Sigma}}^{\text{DML}} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{\xi}_i - \hat{\omega}^{\text{DML}}\right)\left(\hat{\xi}_i - \hat{\omega}^{\text{DML}}\right)'. \tag{18}$$

Figure C.10 shows the error distributions for the OLS and random forest models estimated using the above DML procedure. Compared to Figure C.4, the DML estimates are less biased and have

[64] It also handles overfitting concerns through cross validation. However, as the analysis above already used a test-train split, this is not as much an issue (although cross-fitting could improve the efficiency of the estimates).

[65] The DML estimates when the propensities are estimated as linear functions of the items are slightly more biased and noisy.

lower RMSEs for every model specification. However, OLS continues to perform very well against the random forest models – OLS has the lowest bias and RMSE in the 2-way case and a still-low bias with the lowest RMSE in the 3-way case.

Figure C.10: OLS versus RF DML Estimates of $\mathbf{\Omega}$

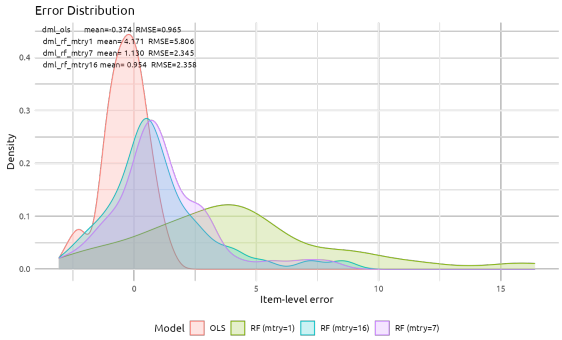


Figure C.11: Two-Way (N =20k; $M = 50$)

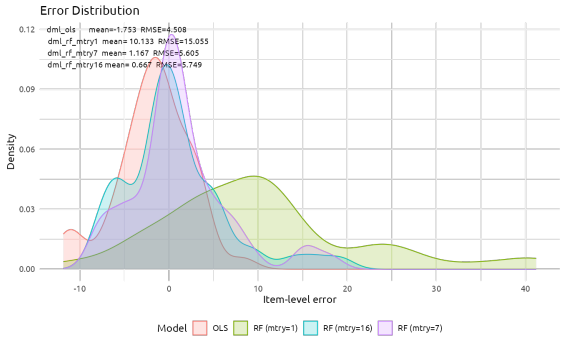


Figure C.12: Three-Way (N =20k; $M = 50$)

Figure C.13 likewise shows the error distributions for the OLS and LASSO models estimated using the above DML procedure. OLS continues to perform very well – the OLS estimates have lower bias than the higher-order LASSO estimates while still achieving low RMSEs.

Figures C.10 and C.13 are not exceptional. Across many bootstrap iterations, we find that the OLS models perform similarly or better than the random forest models in terms of mean bias, and they perform better on RMSE in most casess. Similarly, while the DML LASSO models often have lower RMSEs than the OLS models, the differences are modest, and the OLS estimates are typically less biased.

Figure C.13: OLS versus LASSO DML Estimates of $\mathbf{\Omega}$

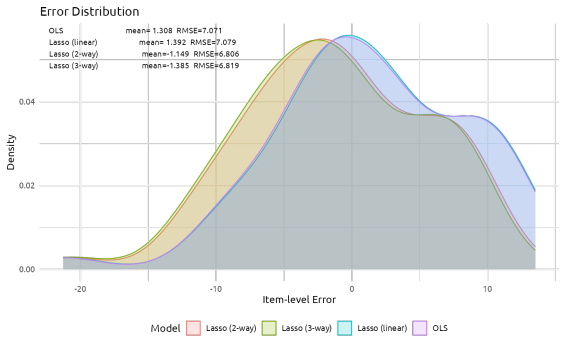


Figure C.14: Two-Way (N =20k; $M = 50$)

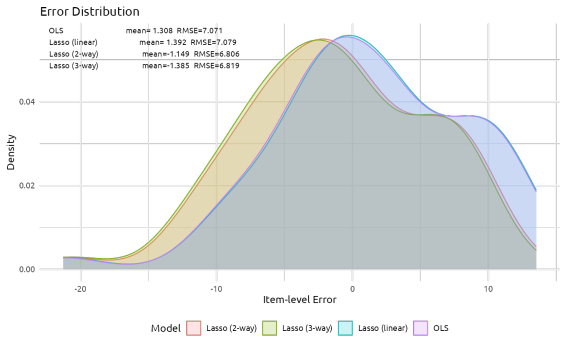


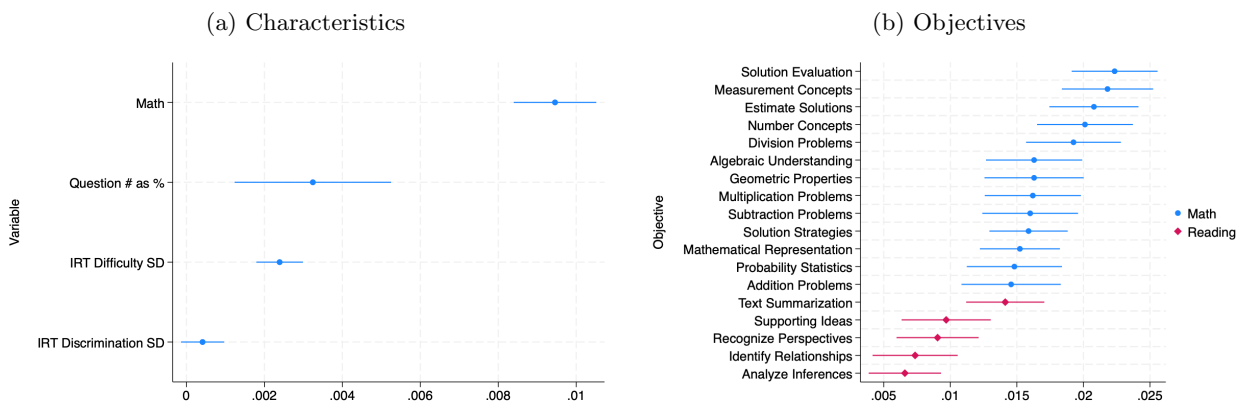
Figure C.15: Three-Way (N =20k; $M = 50$)

The overarching conclusion we draw from this analysis is that OLS is acceptable for use in estimating $\mathbf{\Omega}$. This conclusion will hold so long as (1) the Monte Carlo dgp accurately enough represents the true dgp in our analysis sample and (2) the number of two- and three-way interactions is modest relative to $\binom{M}{2}$ and $\binom{M}{3}$. To be clear, OLS could well work also in the case that there

are very many interactions; this case was simply not covered in our Monte Carlo analysis due to computational limitations.

D Robustness Checks

Figure D.1: Relationship of Item Characteristics to $\hat{\mathbf{\Omega}}$



Notes: This figure plots analogous results to Figure 6 for the subset of items for which we were able to collect text data. Panel (a) of this figure shows estimates of regression coefficients of estimated item-level returns (ω) on observable item-level characteristics. All regressions have grade and year fixed effects and are weighted by the inverse of the square of the SE of ω_K to account for estimate precision (Hedges and Olkin, 2014). Panel (b) presents analogous estimates when the subject is further split into subject objectives as designed by test creators, using the base objective as "Word Meaning - Reading".

E How much of the Variation in $\widehat{\Omega}$ is Sampling Error?

In this section we want to understand how to judge the quality in the prediction of $\widehat{\mathbf{\Omega}}$ as generated by machine learning model in Section 4. This implies considering the sampling expected variation in $\widehat{\mathbf{\Omega}}$, and correcting our observed R^2 by the maximum expected R^2. The discussion below focuses in **out-of-sample predictions** unless otherwise stated, but we remove any extra notation to simplify the exposition. That is, throughout, we impose the standard cross-fitting independence: for each index m, the ML predictor h_m is computed on a fold that excludes the observations used to estimate $\widehat{\omega}_m$; conditional on the true vector $\mathbf{\Omega}$, we treat h as fixed with respect to the first-stage estimation error $e = \widehat{\mathbf{\Omega}} - \mathbf{\Omega}$, so $E[e \mid \mathbf{\Omega}] = 0$ and $\mathrm{Cov}(e, h \mid \mathbf{\Omega}) = 0$ (see Chernozhukov et al. (2018), Bach et al. (2021)).

Notationally, let b denote a given grade-year combination, and let $\widehat{\mathbf{\Omega}}_b$ and $\widehat{\mathbf{\Sigma}}_b$ the coefficient and variance-covariance estimates from anchor model b (for a fixed anchor outcome), both of which are unbiased estimators of $\mathbf{\Omega}_b$ and $\mathbf{\Sigma}_b$, respectively. The latter are the true, unobserved coefficient vector, and unobserved variance-covariance matrix. The pooled coefficient estimates and variance-covariance matrix are given by:

$$\widehat{\mathbf{\Omega}} = \begin{bmatrix} \widehat{\mathbf{\Omega}}_1 \\ \vdots \\ \widehat{\mathbf{\Omega}}_B \end{bmatrix}, \qquad \widehat{\Sigma} = \mathrm{blockdiag}\big(\widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_B\big).$$

We use an equivalent notation for the true $\mathbf{\Omega}$ and $\mathbf{\Sigma}$. Let N be the overall number of elements in $N = \sum_b N_b = \sum_b |\widehat{\mathbf{\Omega}}_b|$. Also define the centering matrix $C \equiv I - \frac{1}{N} 11'$.[66]

So, the observed variance of $\widehat{\mathbf{\Omega}}$ will be:

$$V(\widehat{\mathbf{\Omega}}) = \frac{1}{N} \widehat{\mathbf{\Omega}}' C \widehat{\mathbf{\Omega}} \tag{19}$$

Now, let's consider the expected value of 19 conditional on the unobserved $\mathbf{\Omega}$: [67]

[66] Note that the expressions that follow assume equal weighting of all terms. If one wanted to apply different non-negative weights $\varphi \in \mathbb{R}^N$ with $\sum_i \varphi_i = 1$, then one could replace C for P such that $P = \mathrm{diag}(\varphi) - \varphi\varphi'$. One such matrix would be the inverse-variance weighting matrix where $\varphi_i \propto 1/\Sigma_i i$, such that ω's that are more precisely estimated get more weight. Another would be to use a Mahalanobis metric, such that $P = blkdiag(\Sigma_1^{-1}, ..., \Sigma_B^{-1})$. In this case the metric penalizes errors according to the full covariance of estimation uncertainty, not just the diagonal.

[67] The second step comes from the fact that for any RV X with mean m and covariance S and any symmetric matrix A, $E[X'AX] = tr(AS) + m'Am$.

$$
\begin{aligned}
E[V(\widehat{\mathbf{\Omega}})|\mathbf{\Omega}] &= E\left[\frac{1}{N}\widehat{\mathbf{\Omega}}'C\widehat{\mathbf{\Omega}}|\mathbf{\Omega}\right] \\
&= \frac{1}{N}\left(\mathbf{\Omega}'C\mathbf{\Omega} + tr(C\Sigma)\right) \\
&= \underbrace{\frac{1}{N}\mathbf{\Omega}'C\mathbf{\Omega}}_{V_{\text{signal}}} + \underbrace{\frac{1}{N}tr(C\Sigma)}_{V_{\text{exp.noise}}} \qquad (20) \\
&= \underbrace{\frac{1}{N}\mathbf{\Omega}'C\mathbf{\Omega}}_{V_{\text{signal}}} + \underbrace{\frac{1_b}{N}\sum_{b=1}^{B}\left(tr(\Sigma_b) - \frac{1}{N}1_b'\Sigma_b 1_b\right)}_{V_{\text{exp.noise}}} \qquad (21)
\end{aligned}
$$

That is, $V_{\text{exp.noise}}$ is the bias from estimating the variance of the expected observed coefficients ($\widehat{\mathbf{\Omega}}$) with the actual observed variance of the coefficients ($\frac{1}{N}\mathbf{\Omega}'C\mathbf{\Omega}$).

Now let's consider the expected value of R^2 for any h prediction model of $\widehat{\mathbf{\Omega}}$. In order to correct R^2, we need to correct both the total sum of squares and the residual sum of squares. The population level R^2 for any predictor h is:

$$
R^2(h) = 1 - \frac{(\widehat{\mathbf{\Omega}} - h)'(\widehat{\mathbf{\Omega}} - h)}{\widehat{\mathbf{\Omega}}'C\widehat{\mathbf{\Omega}}} \qquad (22)
$$

Consider the expected value of Equation 22. Then we have

$$
\begin{aligned}
E[R^2(h)|\mathbf{\Omega}] &= 1 - E\left[\frac{(\widehat{\mathbf{\Omega}} - h)'(\widehat{\mathbf{\Omega}} - h)}{\widehat{\mathbf{\Omega}}'C\widehat{\mathbf{\Omega}}}|\mathbf{\Omega}\right] \\
&\approx 1 - \frac{E[(\widehat{\mathbf{\Omega}} - h)'(\widehat{\mathbf{\Omega}} - h)|\mathbf{\Omega}]}{E[\widehat{\mathbf{\Omega}}'C\widehat{\mathbf{\Omega}}|\mathbf{\Omega}]} \qquad (23) \\
&= 1 - \frac{(\mathbf{\Omega} - h)'(\mathbf{\Omega} - h) + tr(\mathbf{\Sigma})}{\mathbf{\Omega}'C\mathbf{\Omega} + tr(C\Sigma)} \qquad (24)
\end{aligned}
$$

where the approximation is the first order approximation.

Thus, to estimate the true performance, we correct both the numerator and denominator. For the expected true SSE, $\widehat{\text{SSE}}_{\text{true}}$, we have:

$$
\widehat{\text{SSE}}_{\text{true}} = (\widehat{\mathbf{\Omega}} - h)'(\widehat{\mathbf{\Omega}} - h) - tr(\widehat{\Sigma})
$$

And for the expected variance of the signal, $\widehat{V}_{\text{signal}}$, we have:

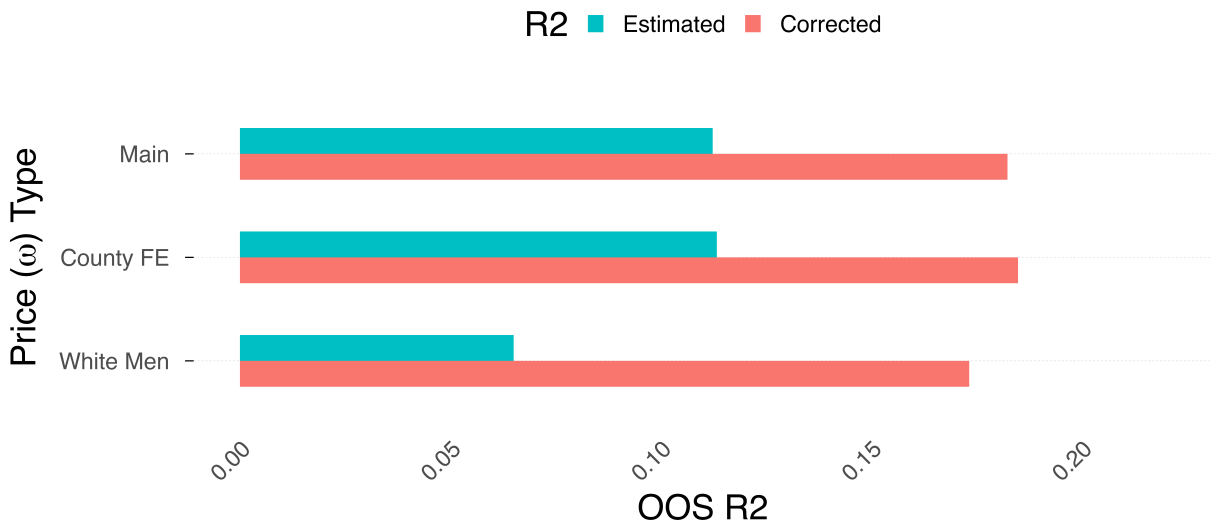
$$
\widehat{V}_{\text{signal}} = \widehat{\mathbf{\Omega}}'C\widehat{\mathbf{\Omega}} - tr(C\widehat{\Sigma})
$$

Finally,

$$R^2_{\text{corrected}} = 1 - \frac{\widehat{\text{SSE}}_{\text{true}}}{\widehat{V}_{\text{signal}}} = 1 - \frac{(\widehat{\mathbf{\Omega}} - h)'(\widehat{\mathbf{\Omega}} - h) - tr(\widehat{\Sigma})}{\widehat{\mathbf{\Omega}}' C \widehat{\mathbf{\Omega}} - tr(C\widehat{\Sigma})}. \tag{25}$$

We report $R^2_{\text{corrected}}$ in **??**. Figure 7 presents a comparison of the "raw" (uncorrected) and corrected R^2s for various models for $\widehat{\mathbf{\Omega}}$. This figure shows the importance of correcting for the first-stage estimation error in $\widehat{\mathbf{\Omega}}$–the corrected R^2s are 50-200% larger than their uncorrected counterparts. Not surprisingly, the correction makes the largest difference for the item weights estimated on white men only, a smaller subset of the full analysis sample. Interestingly, the corrected R^2s are quite similar across different anchor models, unlike for the uncorrected R^2s.

Figure E.1: R2 Correction by ω



Notes: .

F Details on architecture of $g(\text{text})$

F.1 Neural Network Architecture

For the NN model we use a fully connected feed-forward neural network to predict a single continuous outcome - our $\widehat{\omega}$. As is standard in neural network models, both the outcomes and the input covariates are standardized to zero mean and unit variance. For the outcome, after we fit the model on a standardized version of the target we invert the transformation for evaluation and reporting, so metrics are in the original units.

The architecture consists of a stack of dense (ReLU) layers followed by a linear output layer with one unit. The depth and width are tuned. Concretely, the first hidden layer contains between 32 and 512 units, and we allow an additional 1–3 hidden layers, each with 64–512 units. Every hidden layer is followed by batch normalization and dropout. We also apply $L2$ (ridge) penalties

to all dense layers. Both the dropout rates and the $L2$ coefficients are selected by the tuner. The final layer is a single linear neuron producing the scalar prediction.

Hyperparameters are chosen with 'KerasTuner' using random search. The search space includes: the number of hidden layers (2–4 total, counting the first), the units per layer (as above, in steps of 32), the per-layer $L2$ penalty (log-uniform between 10^{-4} and 10^{-2}), the dropout rate after each hidden layer (0.20–0.50), and the Adam learning rate (chosen from 10^{-3}, 10^{-4}, and 10^{-5}). Each trial is trained for up to 100 epochs with early stopping on validation loss (patience 10, restoring the best weights), and we average performance across three executions per trial to reduce training noise. The selected model is the one with the lowest validation mean squared error over the search.

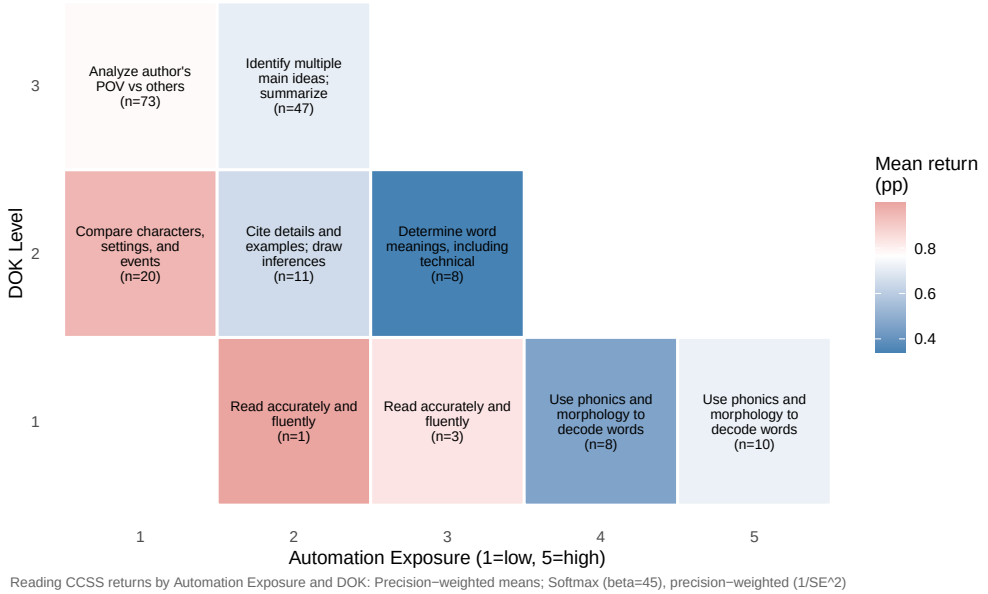
F.2 Computational Resources

All ML models were estimated on TACC's Lonestar6 (LS6) GPU nodes. Each LS6 A100 node comprises dual AMD EPYC 7763 ('Milan') CPUs with 128 physical cores, 256 GB RAM, and three NVIDIA A100 (40 GB HBM2) GPUs per node; GPU-accelerated jobs were scheduled to the A100 partitions accordingly. We relied on the system CUDA-enabled XGBoost builds and TensorFlow with cuDNN.

G Skill Classification Details

G.1 Reading CCSS Returns by DOK and Automation Exposure

Figure G.1: Reading CCSS Returns by DOK and Automation Exposure



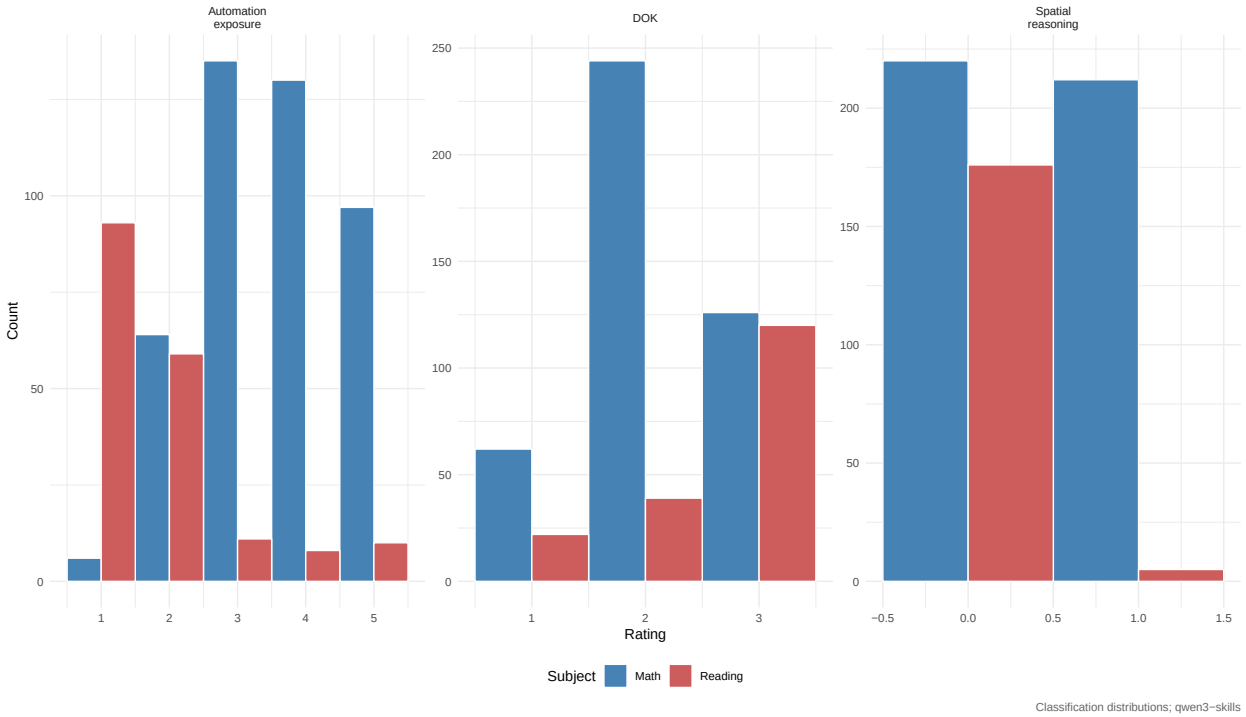
Notes: Precision-weighted mean return (pp) in each DOK × automation exposure cell. One example skill per cell (highest max cosine similarity to any item). Color scale: blue = below average, red = above average.

G.2 Dimension Distributions

Figure G.2 shows the distribution of all three classification dimensions across math and ELA standards.

Figure G.3 shows precision-weighted correlation matrices for the relevant skill dimensions.

Figure G.2: Distribution of Skill Classifications

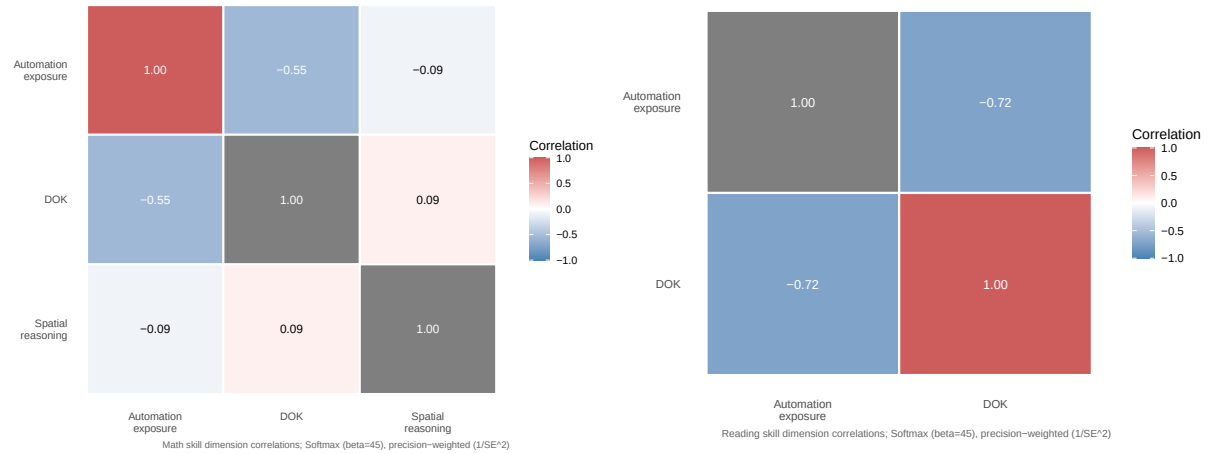


Notes: Distribution of LLM-classified skill dimensions across 613 CCSS standards. Math: blue; ELA: red. Transferability and automation exposure are shared across subjects (1–5). Spatial reasoning is binary (0/1, math only). Representation translation (1–4) and cognitive demand (1–4 math, 1–3 ELA) are subject-specific.

Figure G.3: Skill Dimension Correlations

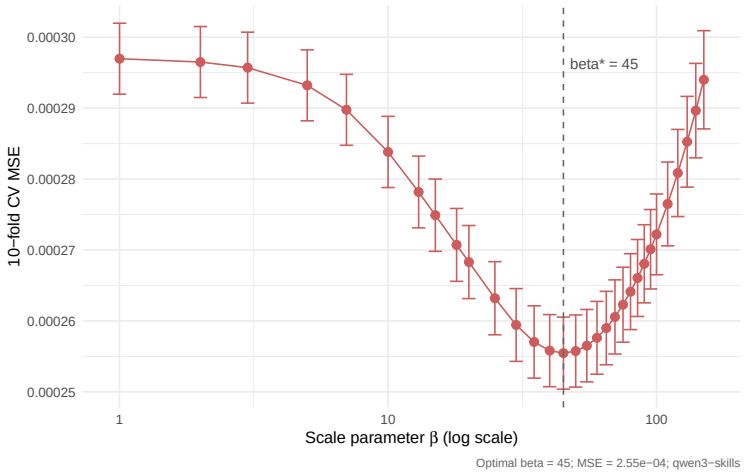
(a) Math

(b) Reading



G.3 Softmax vs. Power Kernel Comparison

Figure G.4: Cross-Validation for Softmax Scale Parameter



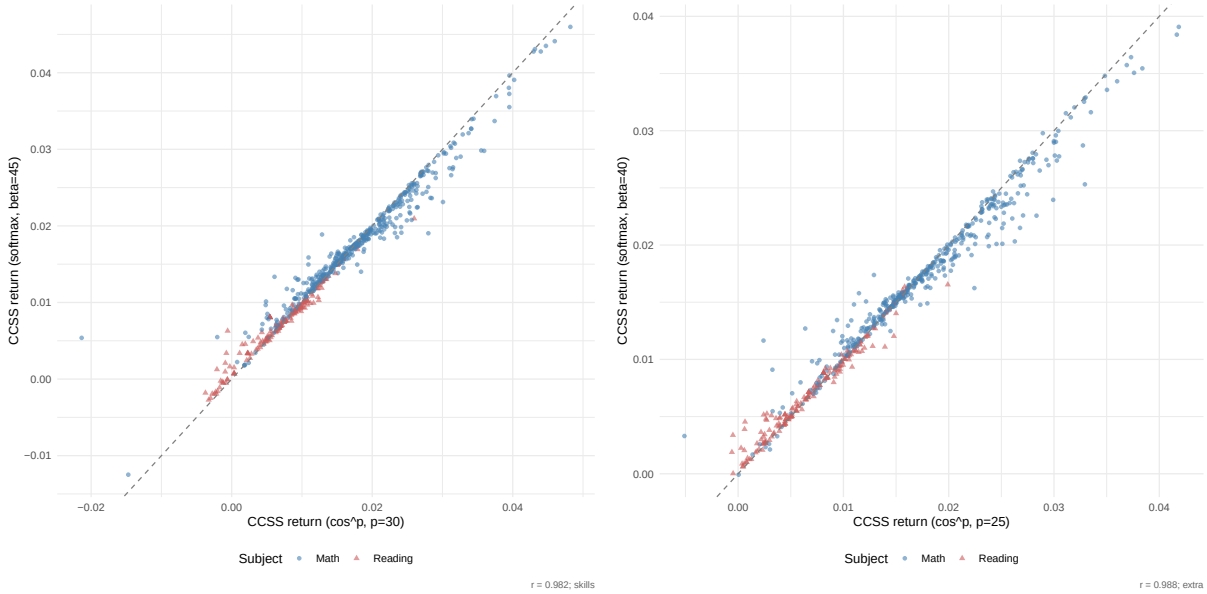
Notes: This figure presents the 10-fold cross-validation MSE as a function of β. Each fold holds out a stratified sample of items, computes CCSS returns from the training set, and predicts held-out item returns. The optimal $\beta^* = 45$ minimizes prediction error. Bars represent +/- 1 SD of mean MSE across folds.

Figure G.5 shows the relationship between CCSS returns estimated under the softmax kernel and the power kernel. The correlation between the returns to CCSS skills across both methods is about 0.94.

Figure G.5: Softmax vs. Power Kernel CCSS Returns

(a) qwen3-skill

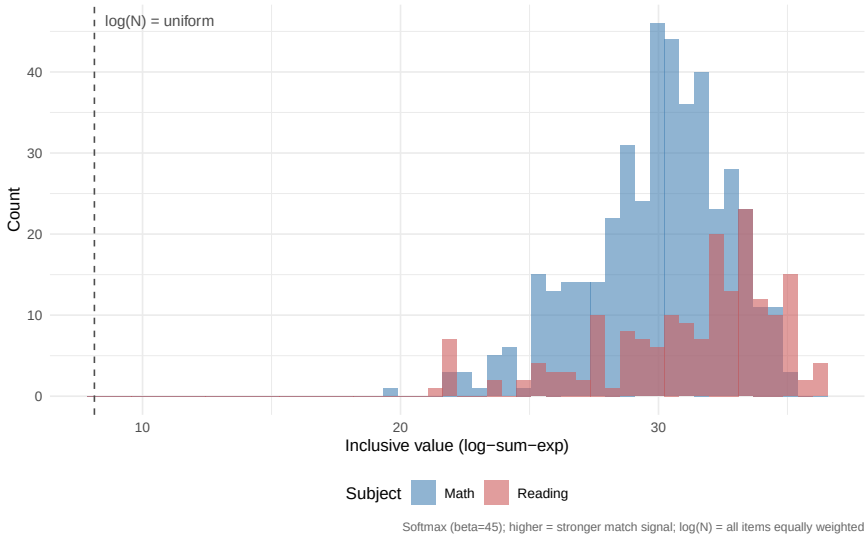
(b) o3-skill



G.4 Inclusive Value Distribution

The inclusive value $IV_j = \log \sum_m \exp(\beta \cdot C_{m,j})$ measures how well each CCSS standard is matched to the item pool. Standards with higher IV have at least some items that match well; those with low IV are effectively unidentified in the data. Figure G.6 shows the distribution. Most standards have IV values concentrated in a narrow range, suggesting that the embedding space provides reasonable matches for the majority of standards.

Figure G.6: Inclusive Value Distribution



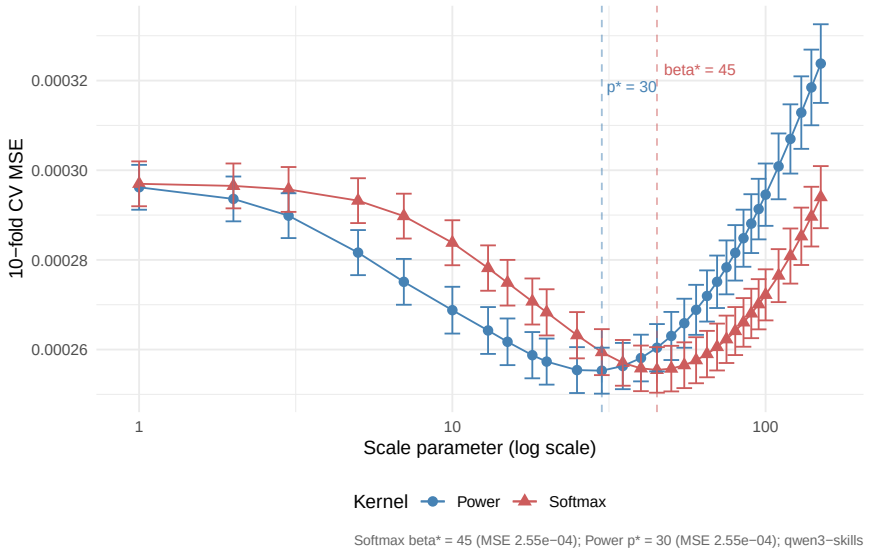
G.5 Cross-Validation: Full Comparison

Figure G.7 shows the full CV curve comparing both the softmax (β) and power (p) kernels.

G.6 CCSS Returns Distribution by Subject

Figure G.8 shows the distribution of CCSS returns by subject under the softmax kernel.

Figure G.7: Cross-Validation: Softmax vs. Power Kernel



Notes: This figure presents the 10-fold cross-validation MSE as a function of β and p. Each fold holds out a stratified sample of items, computes CCSS returns from the training set, and predicts held-out item returns. Bars represent +/- 1 SD of mean MSE across folds.

Figure G.8: CCSS Returns Distribution by Subject (Softmax)

