Test Items and Long-Run Outcomes: How to Measure Achievement and What Drives It*

Jonathan Moreno-Medina
University of Texas at San Antonio

Eric R. Nielsen
Federal Reserve Board

Viviana Rodriguez
University of Texas at San Antonio

February 2026

PRELIMINARY AND INCOMPLETE

PLEASE DO NOT CIRCULATE

Abstract

Standard test scales aggregate item responses, obscuring the fact that some questions are far more predictive of economically relevant outcomes than others. We digitize item (question) content for $\sim$ 3,500 items and $\sim$1 billion student–item records linked to Texas administrative data. We document wide variation in how strongly different items predict later-life earnings and schooling outcomes, 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 given-score gaps while significantly reorder individual student rankings. To interpret what these prices capture, we bring modern machine learning and AI tools to item texts. We embed each item using state-of-the-art sentence encoders and estimate a flexible text-to-price mapping; we then use gradient-guided local comparative statics in embedding space with a embedding inverter to generate minimal variants of items, which we use to surface interpretable skill differences. 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 procedural and multi-step mathematics skills in driving outcomes.

*We would like to thank Samson Alva, Peter Arcidiacono, Elliot Ash, Bocar Ba, Pat Bayer, Stephen Billings, Aimee Chin, Ben Domingue, Elaine M. Liu, David Figlio, Josh Gilbert, Havisha Khurana, David Liebowitz, 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.

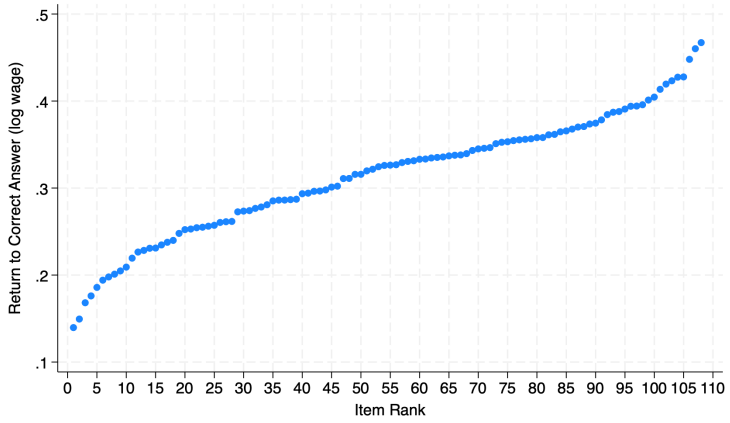
Keywords: large language models, machine learning, achievement, measurement, human capital, inequality, test items, psychometrics

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.

To this end, 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 generate interpretable skill groups without pre-specifying features. Throughout, we illustrate the framework with wages as the anchor outcome, while noting that the same approach applies to other outcomes such as college attendance, attainment, 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 ∼1 billion student–item records that are also linked to longitudinal administrative data (enrollment, completion, and earnings). Finally, we map individual test questions to detailed skill taxonomies of the CCSS, which codifies over 600 skills in English Language Arts (ELA) 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) on the vector of student item response to the math and ELA 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. In doing this, we update and extend the method developed in Nielsen (2019, forthcoming) to a new setting. 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 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" model-guided discovery approach which seeks to learn a flexible representation of what the item is testing and to generate interpretable hypotheses about skill dimensions that are not well captured by predefined taxonomies. 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. 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 implement a model-guided discovery approach that treats the text and structure of each item–its stem, text, response options, and supporting material– as data. Using the digitized items, we convert each question into a machine-readable representation and estimate a flexible prediction model that maps this representation to the item's estimated price. The goal is not to replace the standards-based mapping, but to complement it: while the CCSS provides an interpretable benchmark, the text may encode additional dimensions of what an item demands that are not well captured by a fixed list of skills. We show that models trained on item text and structure can explain meaningful variation in item prices beyond standard item metadata and psychometric characteristics, suggesting that the language of the question contains economically relevant information about what is being measured. Finally, we use this learned mapping as a hypothesis-generation tool: by comparing or generating "nearby" questions that the model predicts will have higher versus lower prices, we can surface potentially new candidate skill dimensions that can then be evaluated with targeted validation and human review.

The paper proceeds as follows. Section 2 introduces the two-stage item-anchored framework and defines anchored achievement and item prices; Section 3 covers the student–item linkage, and digitization of items; Section 4 formalizes estimation of the first stage prices; Section 5 revisits

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.

gaps and ranks under anchoring versus conventional scales; Section 6 interprets prices via CCSS and psychometric features; Section 7 models item text/structure and uses pseudo items to generate testable skill hypotheses; Section 8 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 argues for information loss from item aggregation using more recent cohorts of the Texas data than we do and focusing more on shorter-run outcomes such as school discipline, course performance, and school completion. Bruhn et al. (2025) further shows 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 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

[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.

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]

Regarding our model-guided discovery approach, our paper builds on recent work in machine learning and artificial intelligence (Ludwig and Mullainathan, 2024, BehnamGhader et al., 2024, Kapoor et al., 2025, BehnamGhader et al., 2024). Conceptually, our method adopts the gradient-guided morphing idea of Ludwig and Mullainathan (2024) to text rather than images. For this adaptation we need to overcome two challenges: (i) estimate a smooth, on-support map from sentence embeddings to item prices so that local gradients have an economic interpretation; and (ii) translate small representation-space moves into minimal, content-valid edits to the original item text using a constrained, embedding-conditioned editor, with embeddings kept compact enough that this inversion works reliably in practice. To do this, we build representations following the sequence-encoding setup in Du et al. (2024) and use LLM-based encoders in the spirit of LLM2Vec to obtain stable sentence vectors (see BehnamGhader et al. (2024)). Our method also works as an interpretable alternative to HotFlip-style gradient changes (Ebrahimi et al., 2018). In the space of using text and embeddings for predicting item-level characteristics, our paper fits along with Kapoor et al. (2025), which predicts difficulty of items with an embedding projected to a PCA and evaluated in a ridge model. We add to the idea of using embedding representation models for predicting item characteristics a method to translate features detected by the model.

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.

Finally our paper relates to the large literatures within economics that seek to understand the role that skills/human capital play in determining earnings,[8] health,[9], and many other outcomes.[10] 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

[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]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) among many others.

[9]See Cutler and Lleras-Muney (2010), Christopher Auld and Sidhu (2005), Kaestner and Callison (2011), Mocan and Altindag (2014).

[10]In particular, test scores predict 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.

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 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.[11]

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.[12]

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

[11] 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.

[12] 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.

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.

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.[13] 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.[14]

[13] 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.

[14] 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

As defined, the ω_m are population objects, and we must estimate them from our sample. We do this by first estimating Ψ, the parameter vector governing f. This allows us to compute, for each individual i, $\hat{A}_i[D_{im} = 1] = f(D_{i,m} = 1, \mathbf{D}_{i,-m}, X_i; \hat{\Psi})$ and $\hat{A}_i[D_{im} = 0] = f(D_{i,m} = 0, \mathbf{D}_{i,-m}, X_i; \hat{\Psi})$ where in both cases $\mathbf{D}_{i,-m}$ denotes the actual vector of non-m item responses for student i. We 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 T_m the actual text of the item or some mathematical representation of that text, with $\mathbf{T}$ representing the vector of these data across items. Then, we suppose that for some function g and an unknown error vector $\mathbf{\Xi}$

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

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 ($\mathbf{T}$), and (5) non-text item characteristics $\mathbf{R}$. To our knowledge, there are no extant data sets with all of these

(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.

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.[15]

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 (see Section 3.2 for more details on labor market outcomes.).[16]

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,

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

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

(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.[17]

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.

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) for the average item is 33%, while our conditional estimate, $\hat{\omega}$, indicates a 1% return.[18]

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 analo-

[17] 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.

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

gous 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 for any public school student who receives wage/salary income in Texas in adulthood.[19] 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.[20] As such, we are able to recover labor market information for approximately 72% of the universe of relevant test-takers.[21]

We take as our baseline "anchor" outcome adult earnings as earnings at age 25.[22] Ideally we would be able to measure student earnings at later ages in order to more accurately capture individual variation in lifetime earnings.[23] 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.[24]

To understand the implications of capturing earnings at different labor market ages, Figure 2 plots unconditional wage returns at ages 25, 30, and 35, to a correct answer for all items administered to Texas public school 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 and estimate returns for.[25] Panel (b) presents the correlation of item return rank across ages. This plot indicates a strong rank 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, however, we present robustness checks to our analyses using earnings at age 30.

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

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

[21] Panel (d) of Figure A.3 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%.

[22] 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.

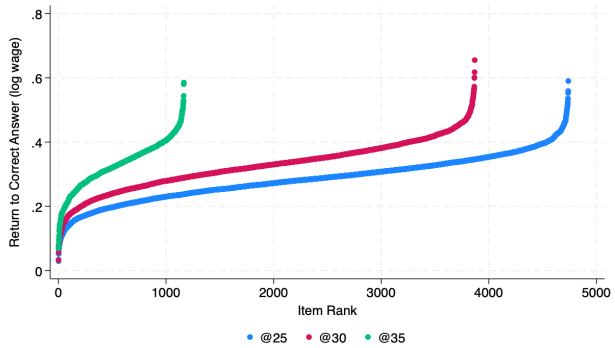
[23] Studies generally find that earnings around age 40 best proxy for lifetime earnings — see Haider and Solon (2006) and Mazumder (2005).

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

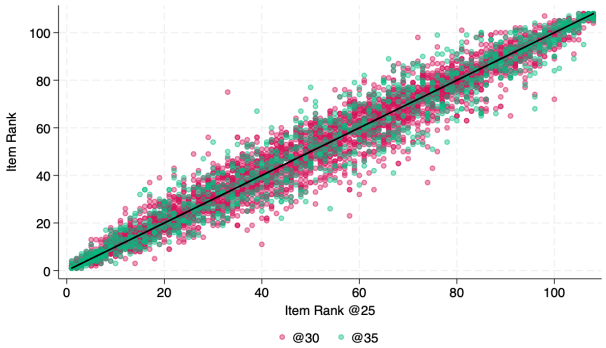
[25] At age 25, we have a larger coverage of wage data, with a match rate of 71.5%. If instead, we capture students at age 30, our average match rate lowers to 52.9%. Finally, at age 35, we loose a substantial amount of data. This is mostly driven by the fact that most students in our testing sample have yet to turn 35. See Figure A.3 for match rates for each grade-year combination in our sample.

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.

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 standard 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.[26]

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).[27] We then evaluate the quality of the digitization

[26] Table A.4 displays the availability of test booklets via Wayback Machine for the time period of this study.

[27] The output included item-level attributes such as: question number, question stem, question image indicator, and

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 corrected all questions with images in them. Thus, overall we obtain a digitization accuracy rate of 97.4%. [28] In the fourth 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 Section A.3.2.

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.[29] 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.[30] 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).[31] The final list describes 631 standards for both ELA (reading) and math across grades 3-12. Table F.1 presents the full list of standards, along with a short description of the skill used for exposition purposes.[32] To map each item to this skill taxonomy, we present the digitized item text together with the full list of CCSS standards to a large language model and prompt it to assign between one and three CCSS codes per question (see Appendix G for the coding prompt). Importantly, we do not tell the LLM what grade the question is from, nor do we restrict the possible matches to any grade-specific standards as defined by the CCSS. That means that we allow for the possibility that the standards set by

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.

[28] 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.1 for examples of question types and image position.

[29] 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.

[30] 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.

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

[32] Please refer to Association et al. (2010) for the full skill description.

CCSS for a 6th grader in the 2010s might match with a skill needed to answer a 4th grade or 8th grade question in Texas in the early 2000s.

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 (1) how we construct these psychometric measures, and (2) 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").[33] The resulting 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.[34] 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.[35]

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.

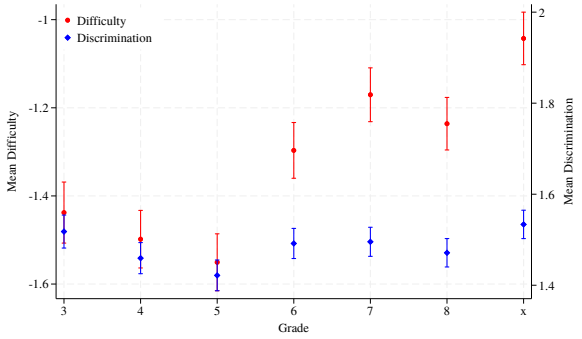
[33] The 3PL IRT model specifies the probability of a correct response as: $P(D = 1 \mid \theta) = c + (1 - c) \cdot \frac{1}{1+\exp[-a(\theta-b)]}$, where a is the discrimination parameter, b is the difficulty parameter, c is the guessing parameter, and student latent ability $\theta \sim \mathcal{N}(0, 1)$. 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 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.

[34] 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.

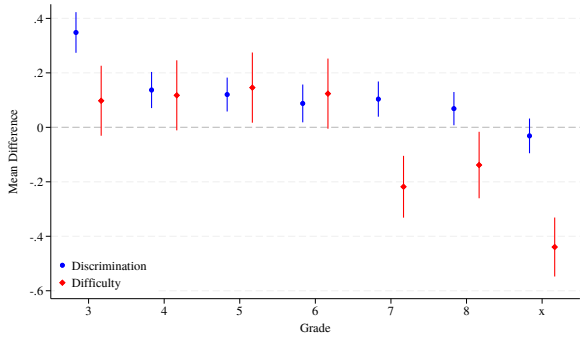
[35] To be clear, even an "easy" test can still differentiate between high performing students, albeit less efficiently than a test targeted towards the upper end of the achievement distribution. As an example, consider an exam where all items have difficulty = -1, discrimination = 1, and guessability = 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.

Figure 3: Item Difficulty and Discrimination by Grade and Subject

(a) Average Difficulty and Discrimination by Grade



(b) Average Difference: Reading - Math

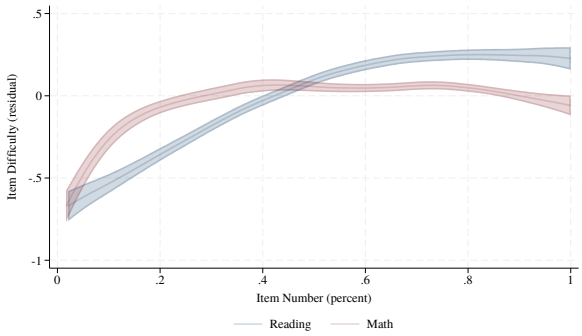


Notes: Panel (a) plots the average difficulty and discrimination, with 95% confidence intervals, of the items in our analysis sample after removing year fixed effects. Panel (b) plots the corresponding mean differences and 95% confidence intervals in difficulty and discrimination between reading and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

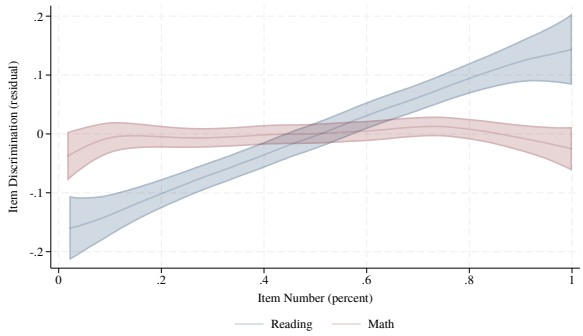
Figure 4 examines the relationship between an item's difficulty and its placement within the exam. Panel (a) shows that the items that come later in an exam tend to be more difficult. However, this relationship is not linear for either math or reading – after a certain point items do not become more 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. Taken together, these results show that the approach we adopt in Section 6 of adding the item's position in the test as an additional factor to explain $\widehat{\mathbf{\Omega}}$ is not redundant given that we also include item difficulty and discrimination.[36]

Figure 4: Item Difficulty and Discrimination and Question Number

(a) Difficulty



(b) Discrimination



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.

[36] 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 4 (not shown).

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.

4.1 Estimating Item Prices ($\widehat{\Omega}$) 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 $\mathbf{\Omega}$ is the OLS estimate of W:

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

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 B.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 D.1, suggests that linear models will do quite well in realistic scenarios without "too many" item-item interactions. In particular, these Monte Carlo experiments suggest that the OLS estimates are approximately unbiased across a wide range of data-generating processes, have very similar MSEs or 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).

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 $\widehat{\boldsymbol{\Omega}}$ can be extracted from the fitted model immediately as the coefficient vectors. Third, linear models allow for the straightforward estimation of $V(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega})$ because the sampling covariance matrix for OLS estimates is readily recoverable. Finally, the resulting estimates $\widehat{\boldsymbol{\Omega}}$ have the straightforward and familiar residual regression interpretation thanks to the Frisch–Waugh–Lovell theorem.

4.2 Controlling for Demographics

Our preferred interpretation of $\hat{\boldsymbol{\Omega}}$ 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).

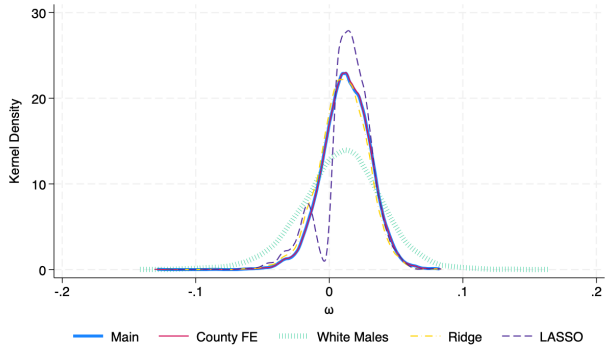
Because we consider data from a wide variety of years, and because earnings tend to grow over time, we include year fixed effects in X in all our specifications. In addition, we include in X commuting zone fixed effects in our preferred specification to account for differences in the local labor markets that our test takers have easy access to.[37] 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 $\widehat{\boldsymbol{\Omega}}$ 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 $\widehat{\boldsymbol{\Omega}}$s that result from this robustness check thus translate item responses into

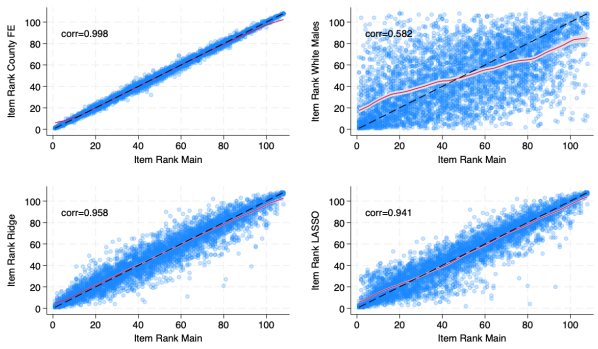
[37] Economic opportunities are quite different in large, growing, dynamic metro areas such as Austin, Houston, and Dallas compared to poorer and more rural parts of the state.

Figure 5: 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.

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 5 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 5 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 5 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 5 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 $H_0 : \omega^{m}_{j} = \omega^{m'}_{j}$. We present these results in Table B.1. Thus, the estimated item weights are quite similar to each

other across specifications.

5 Item-Anchoring, Achievement Gaps, and Student Ranks

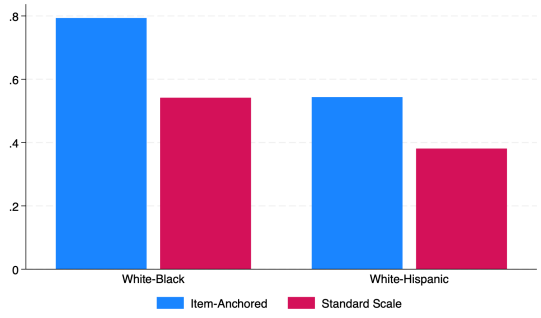
In the previous section, we showed that different, plausible anchor models and methods yield similar estimates $\widehat{\mathbf{\Omega}}$. Moreover, we showed that the individual estimated items prices, $\hat{\omega}_m$, display significant variation – different items predict outcomes very differently. 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.[38]

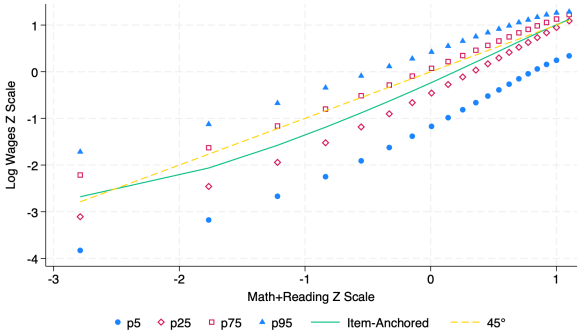
Stylized Fact 1: White-Minority gaps are larger than with standard scale: Item-anchored achievement gaps in general will differ from gaps estimated using given, psychometric scores. First, trivially, the units will differ – item-anchored gaps are in cardinally interpretable "outcome units," while psychometric scales will be in whatever scale the test-designers construct (which may or may not be cardinal).[39] 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.

Figure 6: Item-Anchored Achievement Differences

(a) Achievement Gaps



(b) Achievement Percentiles



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.

[38] 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.

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

Panel (a) of Figure 6 shows the average white-Black and white-Hispanic achievement gaps for 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.[40] 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.[41] 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 C discusses this point formally, adding details on how we deal with first stage noise in the estimates of the gaps.

Stylized Fact 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 6 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. 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 and Standards-Based 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 we explore these

[40] 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 C for details on the method and estimation.

[41] 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.

questions and develop the paper's central contribution. We divide this exploration in two parts: First, 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; second, we leverage information in the text and images in the questions as these are likely to encode information about the skills required to solve them.

6.1 Observable Characteristics

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 $\omega_k = \mathbf{R}'\boldsymbol{\Theta} + \varepsilon$. 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 7 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 associate with 0.2 p.p higher item wage returns. However, this is not the case for items that have a higher discrimination parameter.[42] 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.

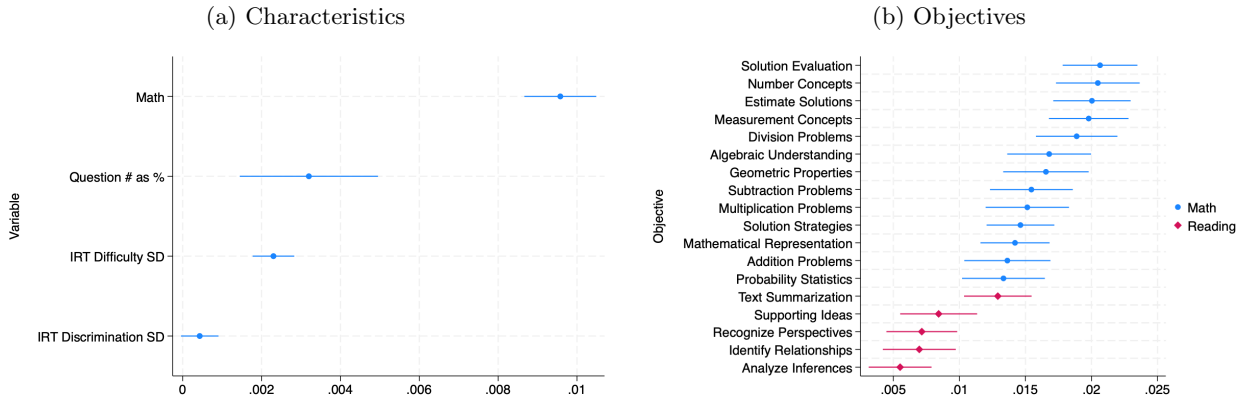
Panel b) of Figure 7 shows the results of an analogous regression where subjects (math and reading) are allowed to be subdivided into the learning objectives designed by test creators. Consistent with earlier evidence, math objectives are generally associated with higher returns on average than reading objectives (with "word meaning", a reading objective, as the omitted category). 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.

6.2 Standards-Based Skill Mapping

The central idea behind leveraging the text of questions to understand their returns is that the skills required to answer a question should be captured to a great extent by the text of the question itself. Thus, we consider a case in which the vector of item-level wage returns, $\widehat{\boldsymbol{\Omega}}$, is a function of

[42] Difficulty and discrimination IRT parameters can be approximated by percent-correct and the correlation of the item response to the total score. Figure A.2 presents the correlation of the IRT parameter estimates with these proxies.

Figure 7: 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 E.1 presents analogous results for the subset of items for which we were able to collect text data.

the text. We propose two novel frameworks to relate the item texts to $\widehat{\mathbf{\Omega}}$. The first one maps the text of the question into a highly detailed set of skills coded by education experts, so in this model $\widehat{\mathbf{\Omega}} = g(\text{skills}(\mathbf{text}))$. In the second framework, we are agnostic as to what is the possible set of skills that the text can encode, and thus directly model the returns as a function of the text. That is, $\widehat{\mathbf{\Omega}} = g(\mathbf{text})$.[43] This section explores results of the first model, while Section 7 develops and presents initial results for the second.

Our first approach relies on a taxonomy of skills defined by the Common Core State Standards (CCSS) (Association et al., 2010).[44] To map each item to this skill taxonomy, we present the digitized item text together with the full list of CCSS standards to a large language model and prompt it to assign between one and three CCSS codes per question (see Appendix G for the coding prompt). In total, we are able to identify 228 main skills from our 3,554 items with text data. Figure 8 shows the number of unique skills we can identify per grade-year with our method. Overall, the number of skills recovered per grade is fairly consistent within grades. Standardized assessments for earlier grades on average tend to test the smallest number of skills with third grade assessments covering around 30 different skills. The number of skills tested increases consistently with grade so that by the time students take the exit exam, the test contains items that measure around 50 different skills. The increase in the number of identified skills is consistent with the results in Nielsen (2025b) which finds that relatively more items have significant relationships with outcomes in later grades.

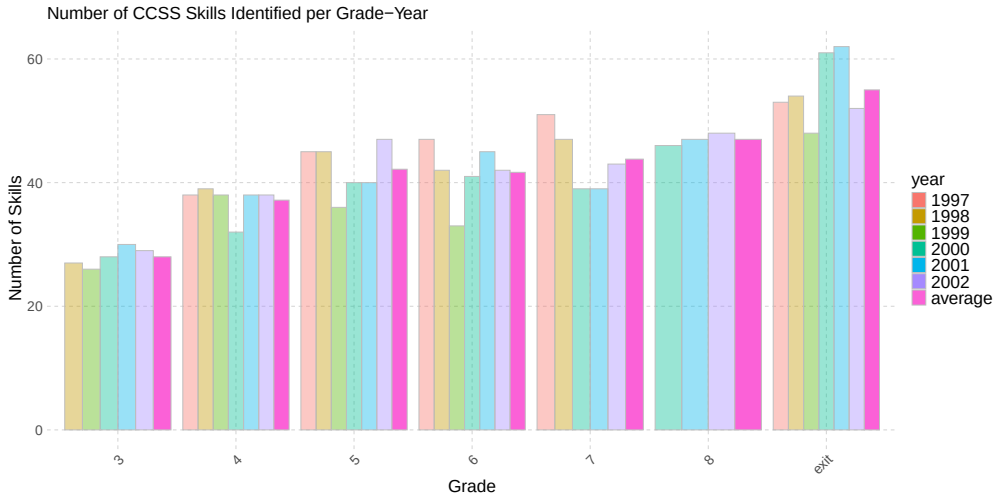
With this item-to-CCSS link at hand, we identify the skills that have the highest returns by regressing the estimated $\hat{\omega}$ as a function of skill-specific indicators, without intercept.[45] We restrict

[43] Although we can imagine $\mathbf{\Omega}$ as a function of the text of the item as well as other item characteristics (subject, objective, difficulty, etc.), the text itself captures most of those characteristics (see Figure 14).

[44] See Section 3.3.2 for a detailed description of these data.

[45] We weight this regression by the inverse of the variance $\hat{\omega}$. This analysis is a direct analogue of the one presented in panel b) of Figure 7. In this case, the CCSS allows us to explore a more comprehensive and hierarchically structure

Figure 8: Unique CCSS Skills by Grade



Notes: This figure plots the total number of skills by grade and year that we are able to recover using the digitized text of the items and the digitized text of the standards laid out by the CCSS.

our analysis to the skills that we find in at least 10 different items, which give us a total of 73 skills. In Figure 9 we present the full list of skill coefficients. To the left of each coefficient, we show a heatmap of the distribution of the item-level grades associated to each CCSS skill. Below we summarize the main results form this analysis.

Finding # 1: Math skills dominate the top rankings, but reading comprehension is also important: Figure 9 shows that almost all of the highest-ranked skills are math-related. While this result mirrors the one presented in panel b) of Figure 7, the granularity of skills provided by the CCSS allows us to recover two novel results. First, in contrast to the results presented in Figure 7, we find that not **all** math skills dominate reading skills. The highest-return reading skill — *'Identify main idea; summarize the text'* — appears in the top 20, in position 17, while the second — *'Identify main idea and supporting details'* — appears in position 21, both ranked higher than 30 other math skills.

Second, the highest return reading skills — basic reading comprehension — is a skill that is consistently tested throughout all of K-12 grades. Students work in reading comprehension (at different levels) throughout all K-12 grade. Unlike reading however, the highest return math skills are middle- and high-school-specific skills.

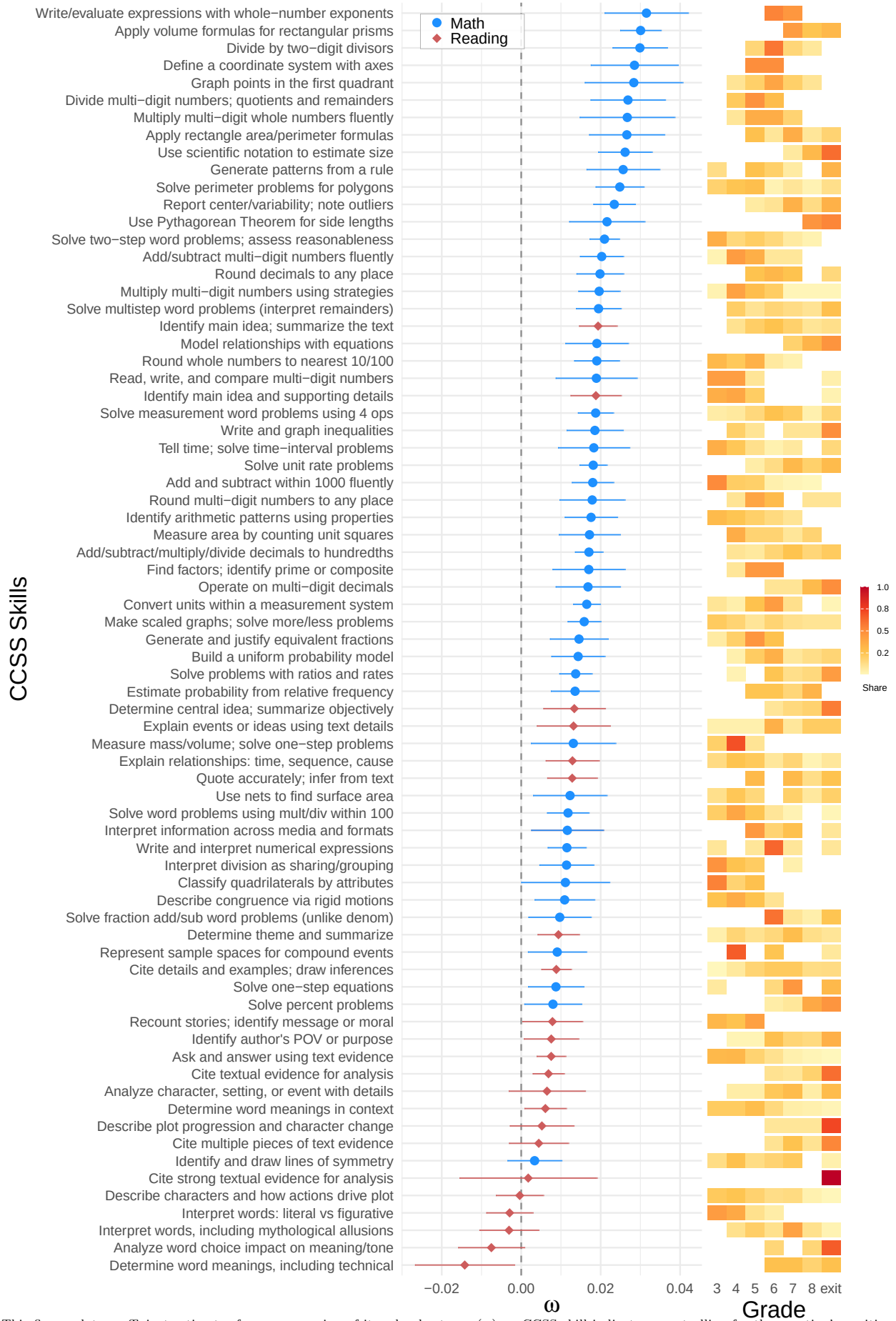
Finding # 2: Computation may matter more than conceptual understanding: Not all math skills are equal. The top math skills lean heavily toward what we might call computational or procedural fluency that require executing multi-step procedures accurately using formal notation to produce numerical results. Often these skills emphasize verbs such as *'apply', 'find', 'use'* and focus on whole-number computation, measurement, and coordinate-plane basics.

set of competencies by subject, which provides a richer understanding of the skills needed to answer each item.

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 (e.g., interpreting media, classifying quadrilaterals) or involve only basic operations without symbolic notation (e.g., solving one-step equations). Often these skills emphasize verbs such as *'understand', 'solve'* 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 end of the return distribution, their emphasis tends differ on either end. Geometry skills located at the top of the returns distribution favor procedural use of geometric tools (*'Define a coordinate system'; 'Graph points in the first quadrant'* — skills 4 and 5), while the bottom skills favor conceptual properties and transformations (*'Classify quadrilaterals by attributes'; 'Describe congruence via rigid motions'; 'Identify and draw line of symmetry'* — skills 51, 52, and 69).

Figure 9: Return by CCSS Skills



This figure plots coefficient estimates from a regression of item-level returns (ω) on CCSS skill indicators, controlling for the question's position within the test, discrimination and difficulty. To account for estimation precision, observations are weighted by the inverse of the variance of the estimates returns ($\hat{\omega}$).

Robustness: In principle, if a test has multiple questions for the same underlying skill, it is possible that when estimating $\mathbf{\Omega}$, by controlling for the full set of item responses we could bias the estimands for a specific question that requires the same unobserved skill. Our CCSS skill linking method allows to partially test for this possibility. In Figure H.2 we present equivalent results to those in Figure 9, but using the "unconditional" returns, $\mathbf{\Omega}^u$. These returns are estimated using a modified version of Equation 6 that controls for X and only one question d at a time (instead of controlling for the full vector D). We show the rank correlation between the CCSS skills with both Ω estimates in Figure H.1. Together, these results suggest that the implied CCSS skill prices are similar for $\widehat{\mathbf{\Omega}}^u$ and $\widehat{\mathbf{\Omega}}$.

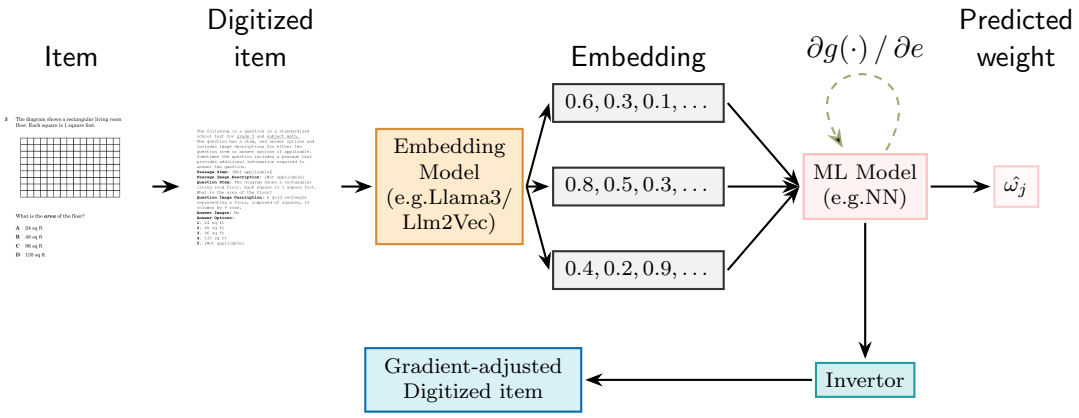
7 Inside the Item: Machine-Learned Representations

7.1 Conceptual framing for AI architecture

We have a dual purpose in this section. On the one hand we want to exploit the richness of information about the skills required to answered each item contained in the text (i.e. we want a flexible estimation of $g(\text{text})$). On the other, we also want to find a representation that is smooth (has a well defined gradient) and can be inverted back into text. The overall architecture we consider is summarized in Figure 10. We expand on each component below, but the sequence is this: i) transform items into text by digitizing the item; ii) embed the item text; iii) train a flexible ML model $g(\text{text})$; iv) construct an inverter from embedding back to text (a decoder) v) traverse the gradient of our model $g(\text{text})$ to generate from a base a higher-ω predicted text and a lower-ω predicted text; vi) compare those set of questions to extract features detected by $g(\text{text})$. We explain in more detail below.

We adapt the gradient-guided morphing idea from Ludwig and Mullainathan (2024) to work with text, which requires dealing with two wrinkles specific to language. First, we estimate a smooth, on-support mapping from sentence embeddings to item prices so that the implied local gradients behave like genuine comparative statics in text space-i.e., they actually tell us something economically meaningful about how wording changes predicted prices. Second, we build a constrained, embedding-aware editor that turns small steps in representation space into minimal, valid edits of the original sentence, using compact embeddings so that the inversion step is reliable in practice. To make all of this concrete, we follow the sequence-encoding setup used for economic text in Du et al. (2024) and rely on LLM-based sentence encoders in the spirit of BehnamGhader et al. (2024), which give us the kind of stable local geometry we need for both the price functional and its gradient. Our procedure is designed as an interpretable alternative to gradient-flip methods such as HotFlip, whose objective is adversarial rather than explanatory (Ebrahimi et al., 2018). Finally, within the literature that uses item text and embeddings to predict item-level characteristics, our approach complements Kapoor et al. (2025)—who predict difficulty via low-dimensional projections and penalized regression—by adding gradient-based local comparative statics and an inverter to return human-legible "up" versus "down" contrasts.

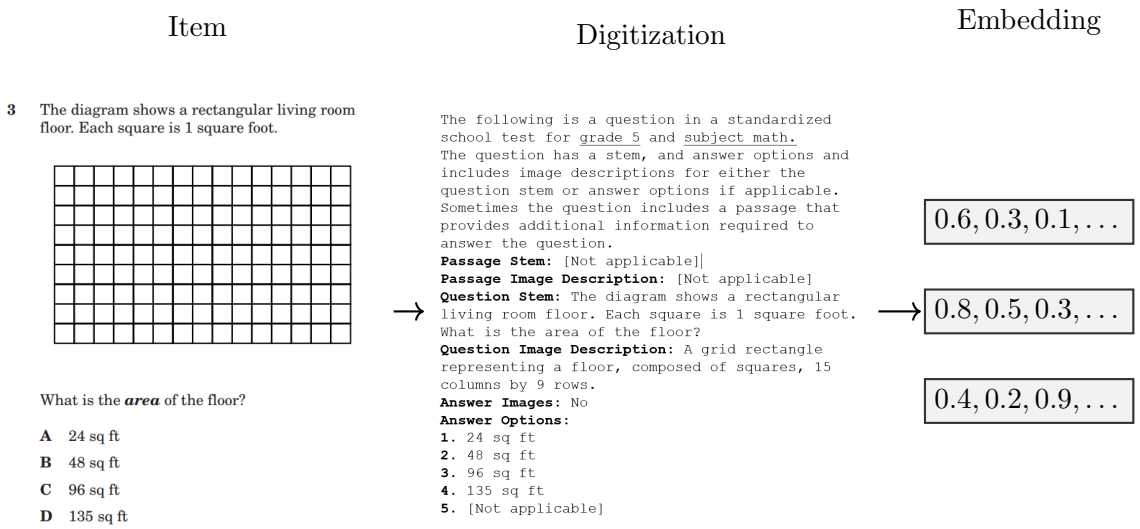
Figure 10: Prediction sequence



7.2 Item's text as embeddings

As is well known, language and unstructured text like the one in test items is a highly dimensional object. The way to get around this extreme high-dimensionality problem is to *embed* the text, usually with an already trained large-language embedding model. Although the model specifics vary somewhat, these models have been trained to complete a next-token prediction, and have been shown to represent the text in a space that captures its meaning (or semantic structure). For our application, we use **Llama 3** (**8 B**illion parameters version). This embedding occurs in a 4096-th dimensional space. Figure 11 presents the overall flow from the booklet information, passing to the digitization, to the final embedding.

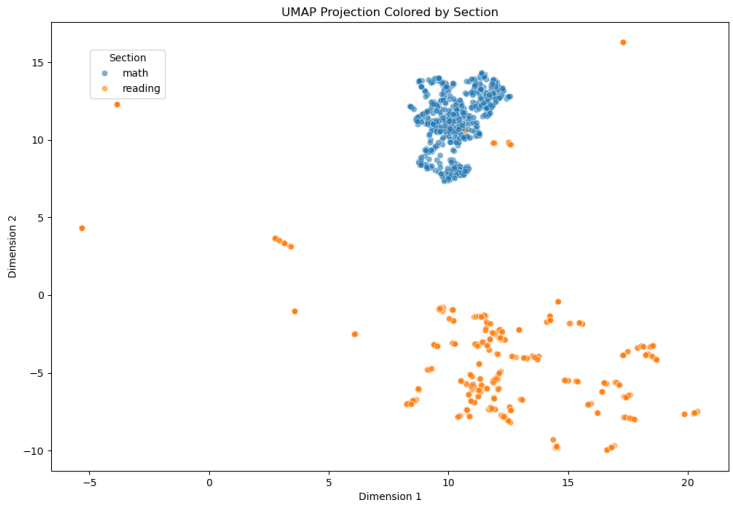
Figure 11: Item Digitization and Embedding



We find that the embeddings of the item texts are able to capture important features of the data.

Figure 12 shows the Uniform Manifold Approximation and Projection (UMAP) of the embedded item data into two dimensions (McInnes et al., 2018). Math (blue) and reading (orange) are very clearly differentiated in this projection, suggesting that they tend to occupy different locations in the embedding space.

Figure 12: Embeddings Differentiate Math and Reading



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

7.3 Predicting weights $\widehat{\Omega}$ as a function of text

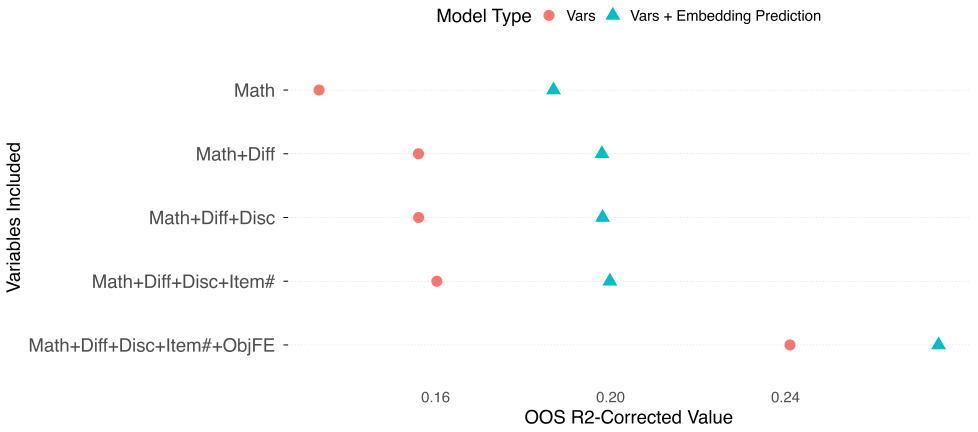
Once we have the embeddings, we adopt a Neural Network for the flexible estimation of $g(\text{text})$ (details are in Appendix J). Although we also considered an XGBoost, we preferred adopting a neural network as the performance was comparable, and the NN had the extra benefit of being fully differentiable for our next step. The model is estimated with a cross-fitting method, with 5 folds - each with a random sample of all questions **?**. We present below the out-of-sample (OOS) performance for each left-out fold, and the final prediction model is the ensemble of each of those.

ML model detects new features: In order to see if the ML model detects new features, we present the out-of-sample R^2 for the model. In thinking about how much variation in $\widehat{\mathbf{\Omega}}$ these featured can explain, it is important to adjust for the fact that $\widehat{\mathbf{\Omega}}$ is estimated; for each item m, we have $\hat{\omega}_m = \omega_m + e_m$, where e_m is a noise term that comes from the first-stage estimation of the anchor model. These errors will tend to bias down our R^2 estimates relative to what we would find if we could somehow observe $\mathbf{\Omega}$ perfectly. 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 $\mathbf{\Omega}$ (**?**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. Appendix I presents the derivation of this correction. We present the comparison between the raw R^2 and the corrected one in Figure I.1.

The resulting out-of-sample (OOS) R^2s corrected for the first-stage estimation error are presented in Figure 13. This figure demonstrates that the text embeddings recover new elements that help explain the item weights in $\mathbf{\Omega}$. The figure shows the OOS corrected R^2s for different sets of controls in $\mathbf{R}$ both with and without the text embeddings included as additional controls. [46]

The figure reveals two key findings. First, including the text embeddings always significantly increases the OOS R^2 – the boost is typically 20-60% of the baseline R^2. 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. Second, only a relatively small share of the variation in $\widehat{\mathbf{\Omega}}$ is explained by any of these factors, with the corrected R^2's never exceeding 0.25. Most of the variation in the item weights is not explained by any features we observe.

Figure 13: Embeddings Recover New Elements that Explain Returns



Notes: .

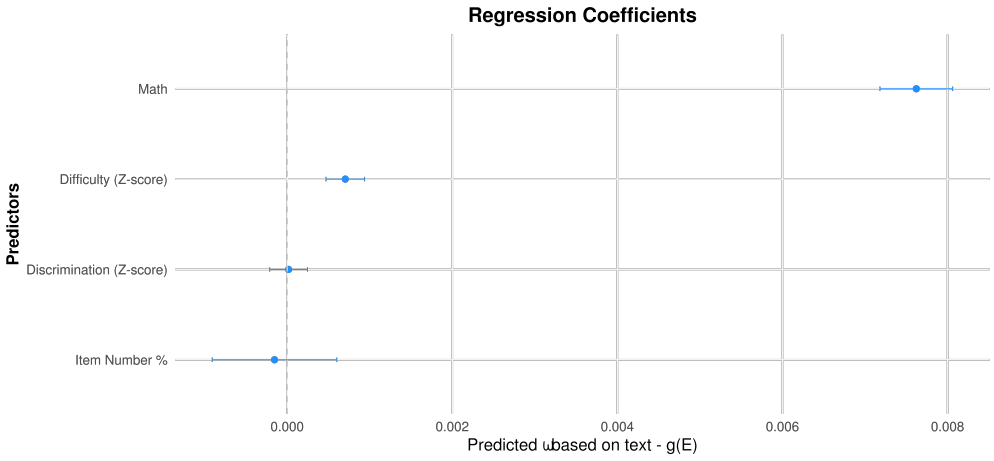
Inspecting the model further: Another way to probe what the model sees is to compare how the model prediction relates to other item-level observable characteristics, and to the correlations we already knew from Section 6.1. To test if this is the case with the predicted returns by $g(\cdot)$, we regress the prediction of the model on a series of characteristics. Figure 14 shows the extent to which the embeddings alone can predict item observables.

For example, as shown in panel (a) of Figure 7, math items tend to have higher returns than reading ones. When we look at the predictions of the model $(\cdot)$ we see something very similar. Overall, the embedding 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. This is important because, if g is extracting skill-relevant content from the text, nothing in the text should predict where the item appears; indeed, embeddings exhibit essentially no power to predict item position. Finally, the embeddings are also not able to predict whether a question is high discrimination or not - which we found are not critical to predict the

[46] We considered other ML approaches such as random forests, nearest neighbor, etc. but in every case, (i.e., with every control set we considered) XGBoost dominated other models.

actual ω, as shown in Figure 7.

Figure 14: Item-level Observables Captured by Text-Based Prediction



Notes: .

7.4 Understanding Skills Through Generated Pseudo-Tests

Because our predictor $g(\cdot)$ is differentiable in the item embedding e_j, we study *local* alterations to an item by moving the representation in the direction of the derivative of predicted price $\hat{\omega}$. Let $d_j \equiv \nabla_e g(e_j)$ denote the gradient at item j. We then consider the path $e_j(\tau) = e_j + \tau \cdot \tilde{v}_j$ where $\tilde{v}_j \equiv d_j / \|d_j\|$ and choose the step size τ by two criteria: (i) *support plausibility*—the morphed point remains close to the empirical manifold of observed items as measured by a k-NN or Mahalanobis radius; and (ii) *monotone improvement*—$g\big(e_j(\tau)\big) - g(e_j) > 0$ for "upward" morphs (and < 0 for "downward" morphs). To reduce sensitivity to local curvature and estimation noise in g, we average gradients across a small ensemble of networks. The resulting pair $\{e_j(+\tau), e_j(-\tau)\}$ defines a locally most-promising direction along which predicted prices change.

We then map each morphed point $e_j(\pm\tau)$ back into text using a embedding-conditioned generator (the "Invertor") that proposes minimal edits to the original stem, options, and exhibits. For each base item j, this yields a triad: the original and two pseudo-items targeted to lower and higher predicted prices. We screen candidates by semantic validity and content alignment (automatic checks plus human review on a small subset), so that the comparisons are based on sensible items.

Human annotators, blinded to direction and predicted prices, conduct paired comparisons of the high- vs. low-morph pseudo-items for each base item and *open-code* any salient textual, structural, or cognitive differences. The labeling exercise is ongoing; upon completion, we will report the aggregated difference patterns and evaluate to what extent the model-guided method produces new skill hypotheses not included in our Standards-Based Skill Mapping (see Section 6.2).

8 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 and computational 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 next 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. Finally, we use modified LLM tools to generate a large corpus of "pseudo-items" similar to those in the Texas data but which are predicted to have unusually low and high returns. A direct examination of this corpus gives additional insight into the drivers of item-level skill prices.

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. *The quarterly journal of economics*, 116(1):313–350.

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

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.

Brown, C., Kaur, S., Kingdon, G., and Schofield, H. (2025). Cognitive endurance as human capital. *The 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*.

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. *The 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.

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*.

Ebrahimi, J., Rao, A., Lowd, D., and Dou, D. (2018). Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36.

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*.

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.

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.

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.

Ludwig, J. and Mullainathan, S. (2024). Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827.

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.

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.

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.

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.

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.

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

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).

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.

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'.

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.[47] 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.[48]

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.

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

[48] 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 n with (unobserved) ability θ_n answers an item i with difficulty δ_i correctly, is defined as:

$$P(X_{n,i} = 1) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} \tag{8}$$

Under a logit formulation, the Rasch model yields a difficulty estimate for each test item ($\hat{\delta}_i$) 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:

$$score = \frac{R - R_{atstandard}}{\sigma_R} * 200 + 1500 \tag{9}$$

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.

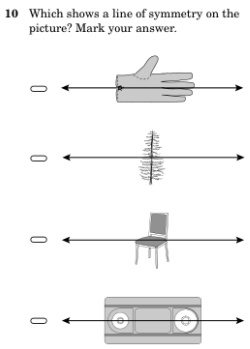
Items were then submitted to the Texas Education Agency (TEA) for review. For this review TEA's Student Assessment Division convened committees composed of teachers, curriculum directors, principals, superintendents, and administrators from regional education service centers to work with TEA staff in 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.[49]

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.[50]

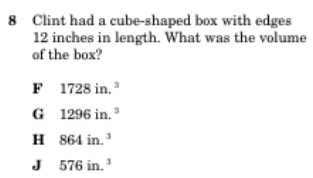
A.1.4 Other

Figure A.1: Question Examples By Type

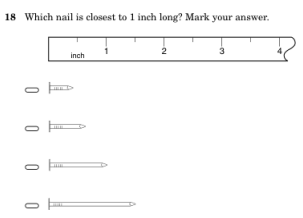
(a) No Image - Image



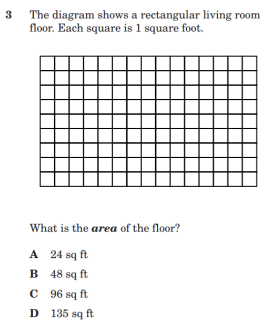
(b) No Image - No Image



(c) Image - Image



(d) Image - No Image

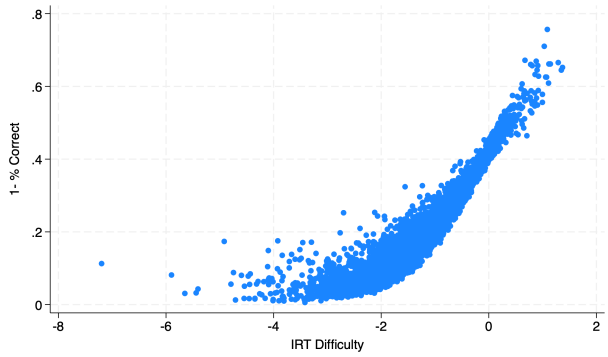


[49]The TEA review was exhaustive. For example in 1999 these committees met 75 times to review all newly developed test items and all new field test data.

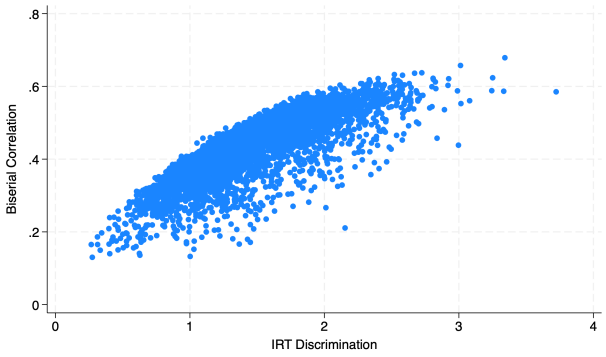
[50]Annual test releases to the public did not include these field testing items.

Figure A.2: Relationship Between IRT Estimates and Proxies

(a) Difficulty IRT & Percent Correct



(b) Discrimination IRT & Biserial Correlation



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.3 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.3 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.3 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.3 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.3 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)).

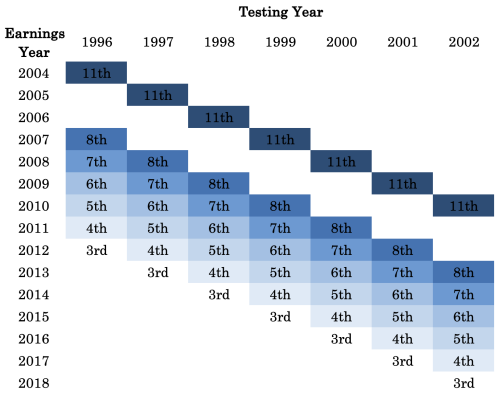
A.3 Booklet Data

A.3.1 Booklet Availability

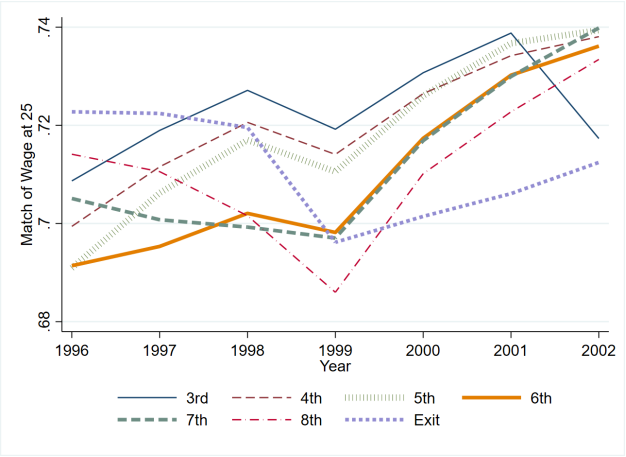
Table A.4 presents test booklet availability by grade-year. An 'X' indicates that the test booklet for that grade-year was recovered and digitized. All test booklets in 1996 are missing. For subsequent years, some grades are missing booklets. Overall, we were able to recover 88% of the test booklets

Figure A.3: Earnings Anchoring Timeline for Test-takers

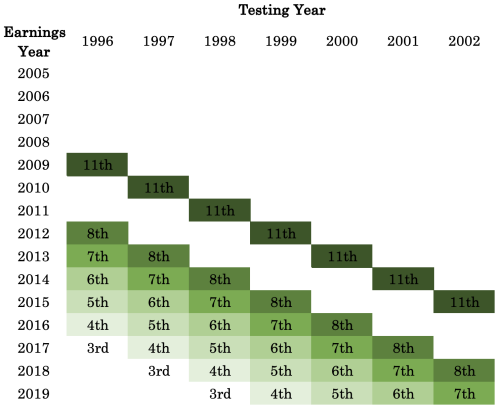
(a) Anchoring Timeline @ 25



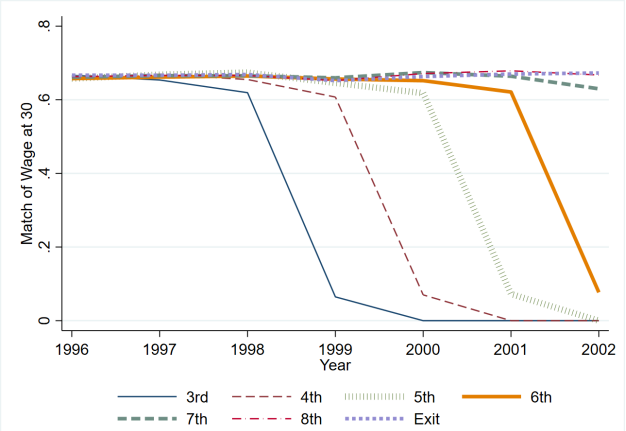
(b) Earnings Match Rate @ 25



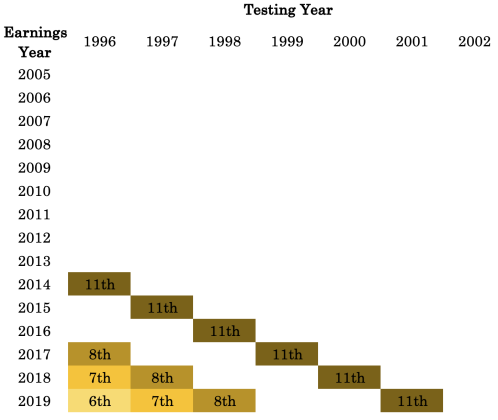
(c) Anchoring Timeline @ 30



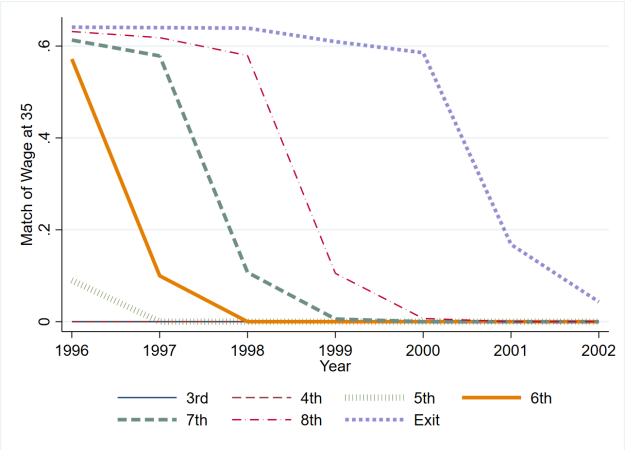
(d) Earnings Match Rate @30



(e) Anchoring Timeline @ 35



(f) Earnings Match Rate @35



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

A.3.2 Embedding format

Once the booklet data from each test-item has been digitized, we use a predefined template to then pass to the embedding model. Our first format, presented below, follows a similar structure to Du et al. (2024) who structure the job trajectory of survey respondents. We add more structure by adding tags to open and close specific parts of the test, such as the question options, though. This is called the 'Question' embedding (Listing 1). Next, we present the format for the expanded text, which takes as input the description from OpenAI on the skills required to answer the question, what are the steps required to solve the question, and what are the potential challenges in answering (Listing 2).

Listing 1: Question embedding

```
<QUESTION>
  <METADATA>
    <GRADE>{grade}</GRADE>
    <SUBJECT>{section}</SUBJECT>
  </METADATA>
  <PASSAGE>
    <STEM>{pass_stem}</STEM>
    <IMAGE_DESCRIPTION>{passage_image}</IMAGE_DESCRIPTION>
  </PASSAGE>
  <QUESTION_STEM>
    <TEXT>{q_stem}</TEXT>
    <IMAGE_DESCRIPTION>{q_image_description}</IMAGE_DESCRIPTION>
  </QUESTION_STEM>
  <ANSWER_OPTIONS image_flag="{a_image_yn}">
    <OPTION id="1">{a1}</OPTION>
    <OPTION id="2">{a2}</OPTION>
    <OPTION id="3">{a3}</OPTION>
    <OPTION id="4">{a4}</OPTION>
    <OPTION id="5">{a5}</OPTION>
  </ANSWER_OPTIONS>
</QUESTION>
```

Listing 2: Question+ Skills embedding

```
<QUESTION>
  <METADATA>
    <GRADE>{grade}</GRADE>
    <SUBJECT>{section}</SUBJECT>
  </METADATA>
  <PASSAGE>
    <STEM>{pass_stem}</STEM>
    <IMAGE_DESCRIPTION>{passage_image}</IMAGE_DESCRIPTION>
  </PASSAGE>
  <QUESTION_STEM>
    <TEXT>{q_stem}</TEXT>
    <IMAGE_DESCRIPTION>{q_image_description}</IMAGE_DESCRIPTION>
  </QUESTION_STEM>
  <ANSWER_OPTIONS image_flag="{a_image_yn}">
    <OPTION id="1">{a1}</OPTION>
    <OPTION id="2">{a2}</OPTION>
    <OPTION id="3">{a3}</OPTION>
    <OPTION id="4">{a4}</OPTION>
    <OPTION id="5">{a5}</OPTION>
  </ANSWER_OPTIONS>
  <QUESTION_SKILLS>
    <SKILL id="s1">{skill1}</SKILL>
    <SKILL id="s2">{skill2}</SKILL>
    <SKILL id="s3">{skill3}</SKILL>
  </QUESTION_SKILLS>
  <QUESTION_COMPLEX_STEPS>
    <COMSTEP id="c1">{step1}</COMSTEP>
    <COMSTEP id="c2">{step2}</COMSTEP>
    <COMSTEP id="c3">{step3}</COMSTEP>
    <COMSTEP id="c4">{step4}</COMSTEP>
  </QUESTION_COMPLEX_STEPS>
  <QUESTION_CHALLENGES>
    <CHALLENGE id="ch1">{challenge1}</CHALLENGE>
    <CHALLENGE id="ch2">{challenge2}</CHALLENGE>
    <CHALLENGE id="ch3">{challenge3}</CHALLENGE>
  </QUESTION_CHALLENGES>
</QUESTION>
```

B Testing and Validating the Item-Anchored Scales

Table B.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.

We first document that there is wide variation in how strongly individual test items relate to outcomes. For example, Figure B.1 and Figure B.2 present "unconditional" item returns. That is, for each item, we simply calculate the difference in mean earnings for students answering the question correctly versus incorrectly: $\hat{\Delta} = \hat{\mathbb{E}}[S|D_{ij} = 1] - \hat{\mathbb{E}}[S|D_{ij} = 0]$. While these mean differences are always positive, they vary tremendously, and some systematic differences by subject are evident.

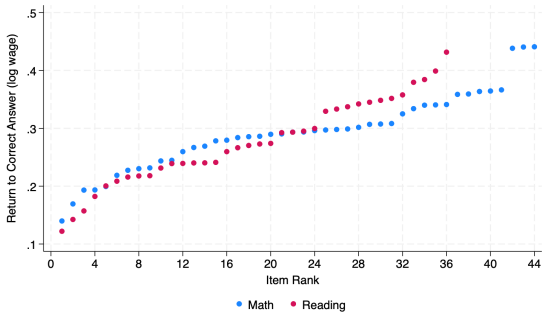


Figure B.1: Unconditional Item-level Returns for 3rd Grade in 1996

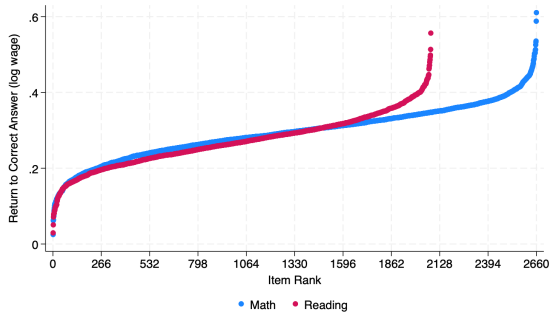


Figure B.2: Unconditional Item-level Returns - All grades & years

The actual OLS estimates of $\mathbf{\Omega}$, which are the baseline estimates we use for much of our analysis, assign negative weight to some items.[51] Figure B.3 shows the distributions of the math and reading weights for ln(wage), pooled across years and grades. Again, the spread in item weights for both subjects is evident. Moreover, this figure highlights one of the key findings of the paper: the math items, on average, have higher weights. This is a consistent finding across our many analyses.[52] Figure B.4 shows, moreover, that an item's outcome weight depends on the specific outcome used as the anchor. Looking across high school graduation and college enrollment, the item weights are

[51] As regards the item-anchored achievement scales, consistent with Nielsen (2019, forthcoming), we do not see notable differences if we instead use LASSO or if we allow for richer functional form specifications. As mentioned previously, we plan to explore Shapley values and other more general measures in future analysis.

[52] While these individual item weights are of course estimated with error, we can strongly reject equality of the item weights for all grades/subjects/years. Moreover, even accounting for multiple testing using the conservative Bonferroni correction, we are also able to reject equality in a significant share of pairwise item comparisons.

positively correlated but not identical.

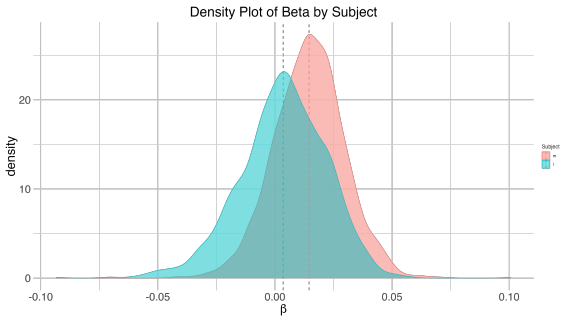


Figure B.3: Density of Item Weights for Log Wages

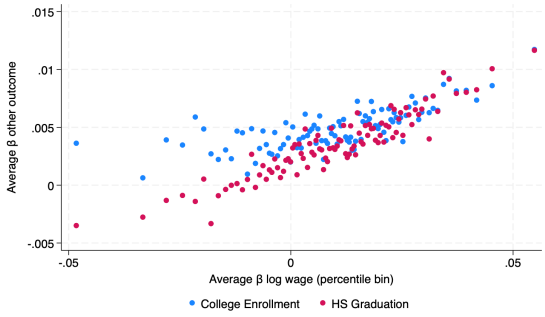


Figure B.4: Relationship of Item Weights Across Accounts

The wide range of item weights depicted in Figure B.1-Figure B.4 already provides substantial motivation to investigate what factors explain these differential returns. Nonetheless, we provide yet more motivation by showing that constructing alternative achievement measures using these item weights to aggregate has important implications for the measurement of achievement. The idea is to define student i's achievement to be $\mathbb{E}[S_i|\mathbf{D}_i = \mathbf{d}_i]$ and to then estimate this achievement using the estimated weights in $\mathbf{\Omega}$. This procedure produces alternative achievement measures better aligned with the outcome S.

?? and **??** exemplify two key general results from this analysis. First, the item-anchored achievement scores often rank individual students very differently from the given achievement scores. **??** plots for grade 3 students in 1996 the given math+reading score ventiles on the x-axis against the ventile-conditional means, 50 percent ranges, and 90 percent ranges of the ln(wage) item-anchored scores on the y-axis. There is 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. Other grades/years/outcomes show similarly wide conditional variation in the item-anchored scores. 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 cases. In fact, we can reject equality in almost all cases even within each percentile or half percentile of the given score distribution.

C 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:[53] $A_i = \sum_m \omega_m D_{i,m}$, $Z_i = \sum_m \alpha_m D_{i,m}$. 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

$$\sum_m (\omega_m - \alpha_m)(p_{m,H} - p_{m,L}). \tag{10}$$

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 $\hat{\Theta}$ means that $\hat{A}_i = A_i + \nu_i$ for some error term ν_i.[54]

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

$$\mathbb{E}[S|\hat{A}_i] = R_{A,\nu}\hat{A}_i + (1 - R_{A,\nu})\bar{A}, \text{ where } R_{A,\nu} \equiv \frac{\sigma_A^2}{\sigma_A^2 + \sigma_\nu^2}. \tag{11}$$

Equation (11) 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

$$\text{plim}_{N\to\infty}(\hat{A}_H - \hat{A}_L) = R_{A,\nu}(\bar{A}_H - \bar{A}_L) < (\bar{A}_H - \bar{A}_L). \tag{12}$$

Equation (12) 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

[53] 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.

[54] If f is correctly specified, which is the assumption we maintain throughout this analysis, ν_i will come only through sampling variability in $\hat{\Theta}$. 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 $D_i^{(1)}$ and $D_i^{(2)}$ denote the odd- and even-numbered item responses for student i. We then use the item and outcome data to estimate $\hat{A}_i^{(1)} = f(D_i^{(1)}, X_i; \hat{\Theta}^{(1)})$ and $\hat{A}_i^{(2)} = f(D_i^{(2)}, X_i; \hat{\Theta}^{(2)})$. 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,

$$Z_i^{(1)} = N_i^{-1} \sum_{j \neq i : \hat{A}_j^{(1)} = \hat{A}_i^{(1)}} S_j, \tag{13}$$

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 $\hat{A}_i^{(1)} = f(D_i^{(1)}, X_i; \hat{\Theta}^{(1)})$ and $\hat{A}_i^{(2)} = f(D_i^{(2)}, X_i; \hat{\Theta}^{(2)})$.

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 (13). 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 C.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	

D 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.

D.1 Anchor Results Under Different Model Specifications

D.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 D.1 and D.4 below.

Result 2: OLS and LASSO perform very similarly in all cases considered.

Figure D.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 and RMSEs as the lasso models. Typically, the lasso models that are correctly specified (i.e. two-way interactions when that is the true dgp) perform best, but the OLS estimates are never far behind.[55]

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

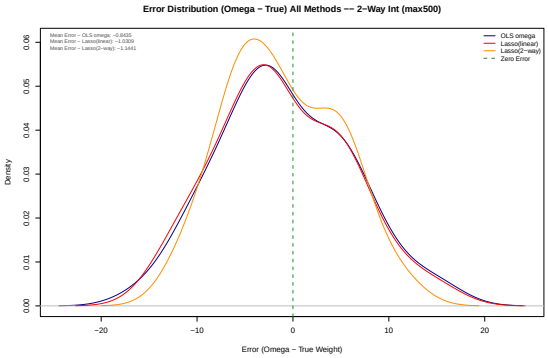


Figure D.2: Two-Way (N =20k; $M = 50$)

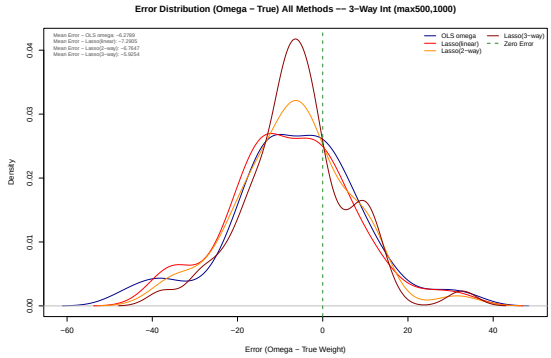


Figure D.3: Three-Way (N =20k; $M = 50$)

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 often performs better than random forests.

Figure D.4 provides a representative picture of the OLS vs. random forest comparisons in our Monte Carlo experiments. The left panels compare the error distributions for OLS and various random forest models for data generating processes that feature some two-way item interactions but no higher-order interactions. The right panels show analogous error distributions for cases with both two- and three-way item interactions. In all panels, it is clear that the OLS estimates have lower mean errors than any of the random forest estimates. When there are a large number of items, the random forest models do show lower RMSEs, but they also show clear bias.

[55] We do not show results for $N = 200,000$ and $M = 250$ as we do for random forests below due to computational limitations.

Figure D.4: OLS versus Random Forest Estimates of $\boldsymbol{\Omega}$

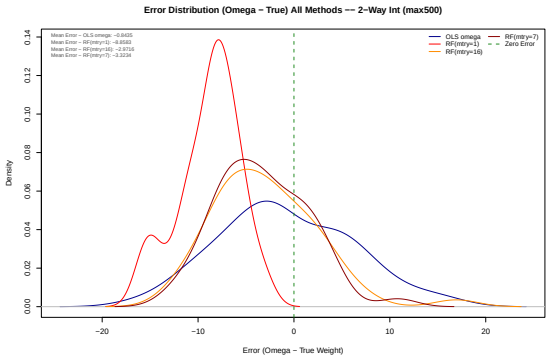


Figure D.5: Two-Way (N =20k; $M = 50$)

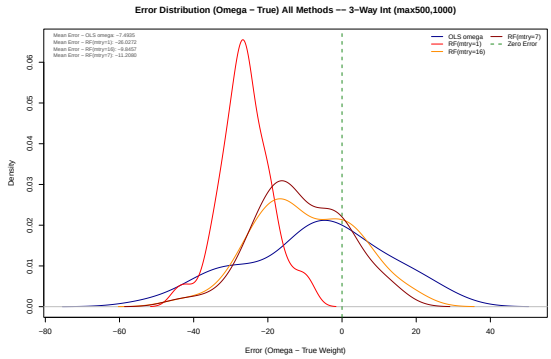


Figure D.6: Three-Way (N =20k; $M = 50$)

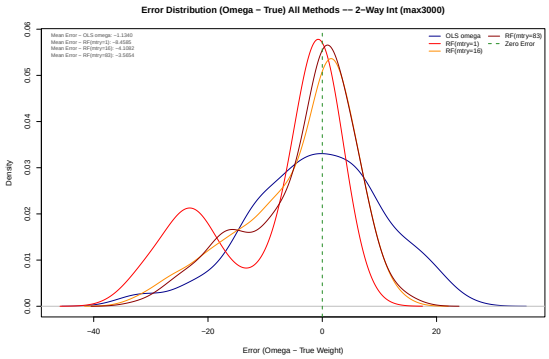


Figure D.7: Two-Way (N =200k; $M = 250$)

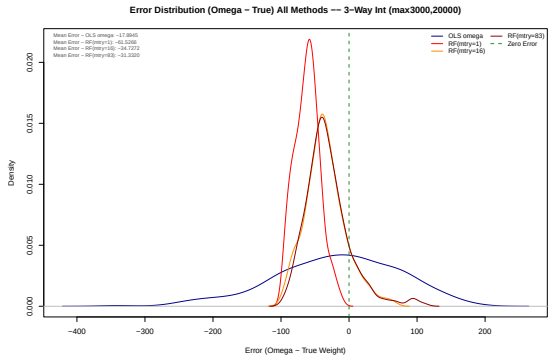


Figure D.8: Three-Way (N =200k; $M = 250$)

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

The data used in Figure D.4 feature interactions that are not too "concentrated" in the sense that, conditional on having an interaction, a given item is likely involved in only a few total interactions. Figure D.9 shows that the outperformance of OLS relative to the random forest models is even greater for data generating processes with a high degree of interaction concentration.

[gg]When each interacted item is involved in many interactions, the random forests often apparently miss some very high-value items (those that happen to have many interactions with other items).

Figure D.9: OLS versus Random Forest Estimates of $\mathbf{\Omega}$ with High Interaction Concentration

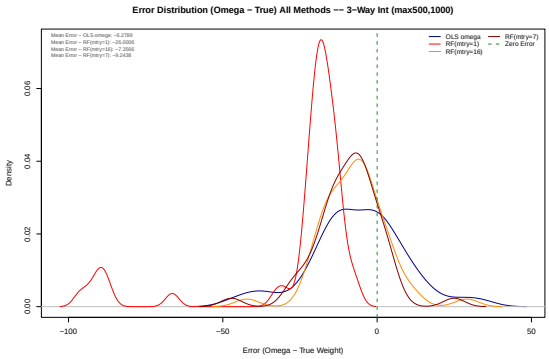


Figure D.10: Two-Way (N =20k; $M = 50$)

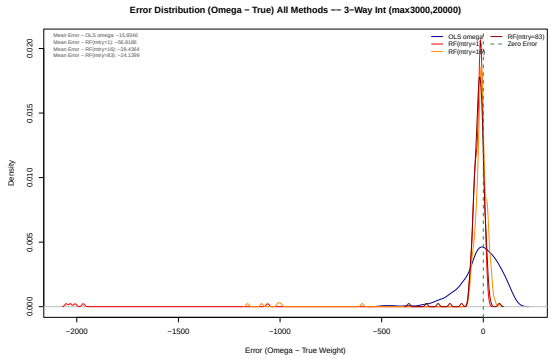


Figure D.11: Three-Way (N =20k; $M = 50$)

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

D.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 D.1 - D.9. 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.[56] 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 lasso or random forests as the case may be. The propensity scores we estimate via linear lasso. Then, the per-observation score contribution is

$$\hat{\xi}_{im} = \hat{\mu}_{m,1}(d_{i,-m}) - \hat{\mu}_{m,0}(d_{i,-m}) + \frac{d_{im}}{\hat{p}_m(d_{i,-m})}\left(s_i - \hat{\mu}_{m,1}(d_{i,-m})\right) - \frac{1 - d_{im}}{1 - \hat{p}_m(d_{i,-m})}\left(s_i - \hat{\mu}_{m,0}(d_{i,-m})\right). \tag{14}$$

[56] 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).

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).

Figure D.12: OLS versus LASSO DML Estimates of $\boldsymbol{\Omega}$

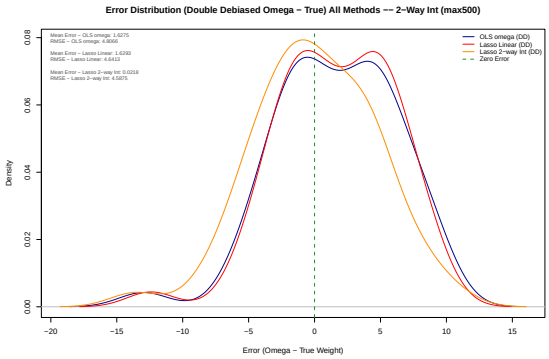


Figure D.13: Two-Way (N =20k; $M = 50$)

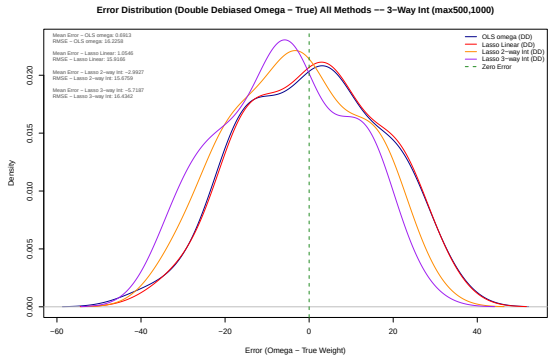


Figure D.14: Three-Way (N =20k; $M = 50$)

Notes: The left panel shows the error distributions for $\widehat{\boldsymbol{\Omega}}^{\text{DML}}$ 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.

Figure D.15: OLS versus Random Forest DML Estimates of $\boldsymbol{\Omega}$

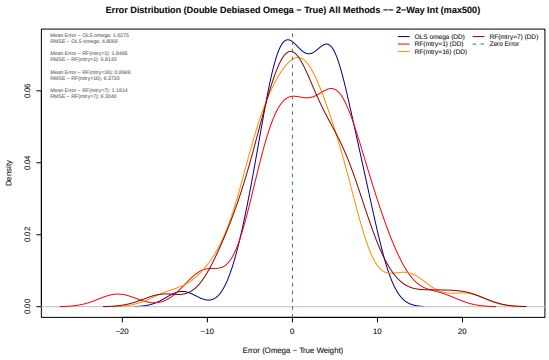


Figure D.16: Two-Way (N =20k; $M = 50$)

Figure D.17: Three-Way (N =20k; $M = 50$)

Figure D.18: Two-Way (N =200k; $M = 250$)

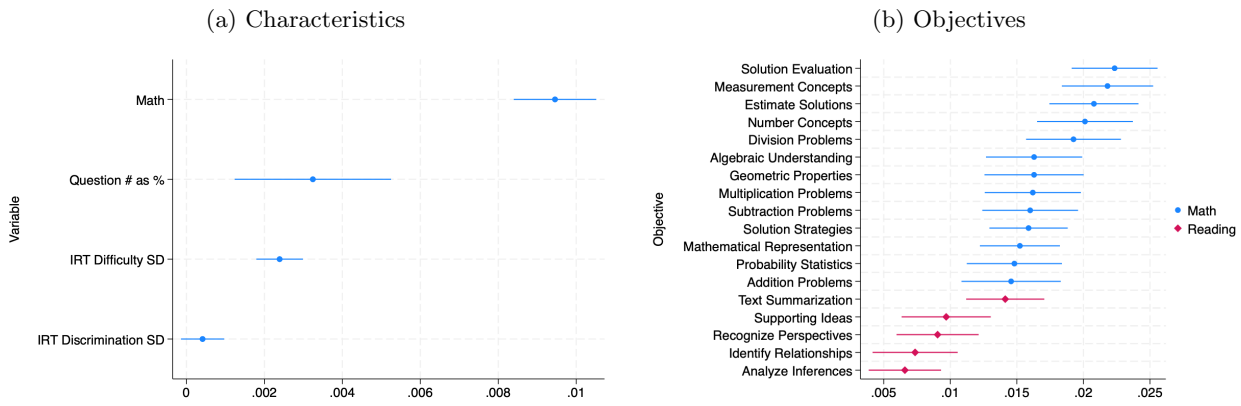
Figure D.19: Three-Way (N =200k; $M = 250$)

The overarching conclusion we draw from this analysis is that OLS is acceptable for use in estimating $\boldsymbol{\Omega}$. 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.

E Robustness Checks

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



Notes: This figure plots analogous results to Figure 7 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".

F Common Core State Standards List

Table F.1: CCSS Codes and Short Description

ID	Content Type	Category	Description
CCSS.ELA-LITERACY.RF.3.3	ELA	Reading: Foundational Skills	Apply phonics and word analysis
CCSS.ELA-LITERACY.RF.3.3.a	ELA	Reading: Foundational Skills	Identify common prefixes and suffixes
CCSS.ELA-LITERACY.RF.3.3.b	ELA	Reading: Foundational Skills	Decode words with Latin suffixes
CCSS.ELA-LITERACY.RF.3.3.c	ELA	Reading: Foundational Skills	Decode multisyllable words
CCSS.ELA-LITERACY.RF.3.3.d	ELA	Reading: Foundational Skills	Read irregularly spelled words
CCSS.ELA-LITERACY.RF.3.4	ELA	Reading: Foundational Skills	Read accurately and fluently
CCSS.ELA-LITERACY.RF.3.4.a	ELA	Reading: Foundational Skills	Read with purpose and understanding
CCSS.ELA-LITERACY.RF.3.4.b	ELA	Reading: Foundational Skills	Read aloud with accuracy and expression
CCSS.ELA-LITERACY.RF.3.4.c	ELA	Reading: Foundational Skills	Use context to self-correct while reading
CCSS.ELA-LITERACY.RF.4.3	ELA	Reading: Foundational Skills	Apply phonics and word analysis
CCSS.ELA-LITERACY.RF.4.3.a	ELA	Reading: Foundational Skills	Use phonics and morphology to decode words
CCSS.ELA-LITERACY.RF.4.4	ELA	Reading: Foundational Skills	Read accurately and fluently
CCSS.ELA-LITERACY.RF.4.4.a	ELA	Reading: Foundational Skills	Read with purpose and understanding
CCSS.ELA-LITERACY.RF.4.4.b	ELA	Reading: Foundational Skills	Read aloud with accuracy and expression
CCSS.ELA-LITERACY.RF.4.4.c	ELA	Reading: Foundational Skills	Use context to self-correct while reading
CCSS.ELA-LITERACY.RF.5.3	ELA	Reading: Foundational Skills	Apply phonics and word analysis
CCSS.ELA-LITERACY.RF.5.3.a	ELA	Reading: Foundational Skills	Use phonics and morphology to decode words
CCSS.ELA-LITERACY.RF.5.4	ELA	Reading: Foundational Skills	Read accurately and fluently
CCSS.ELA-LITERACY.RF.5.4.a	ELA	Reading: Foundational Skills	Read with purpose and understanding
CCSS.ELA-LITERACY.RF.5.4.b	ELA	Reading: Foundational Skills	Read aloud with accuracy and expression
CCSS.ELA-LITERACY.RF.5.4.c	ELA	Reading: Foundational Skills	Use context to self-correct while reading
CCSS.ELA-LITERACY.RI.11-12.1	ELA	Reading: Informational Text	Cite evidence; note textual uncertainty
CCSS.ELA-LITERACY.RI.11-12.10	ELA	Reading: Informational Text	Read complex literary nonfiction (11–12 band)
CCSS.ELA-LITERACY.RI.11-12.2	ELA	Reading: Informational Text	Analyze multiple central ideas and interactions
CCSS.ELA-LITERACY.RI.11-12.3	ELA	Reading: Informational Text	Analyze interactions in complex ideas/events
CCSS.ELA-LITERACY.RI.11-12.4	ELA	Reading: Informational Text	Track and analyze refinement of key terms
CCSS.ELA-LITERACY.RI.11-12.5	ELA	Reading: Informational Text	Evaluate effectiveness of argument structure
CCSS.ELA-LITERACY.RI.11-12.6	ELA	Reading: Informational Text	Analyze how rhetoric and style create impact
CCSS.ELA-LITERACY.RI.11-12.7	ELA	Reading: Informational Text	Integrate and evaluate sources across media
CCSS.ELA-LITERACY.RI.11-12.8	ELA	Reading: Informational Text	Evaluate reasoning in seminal U.S. texts
CCSS.ELA-LITERACY.RI.11-12.9	ELA	Reading: Informational Text	Analyze themes and rhetoric in foundational U.S. docs
CCSS.ELA-LITERACY.RI.3.1	ELA	Reading: Informational Text	Ask and answer using text evidence
CCSS.ELA-LITERACY.RI.3.10	ELA	Reading: Informational Text	Read complex informational texts independently
CCSS.ELA-LITERACY.RI.3.2	ELA	Reading: Informational Text	Identify main idea and supporting details

Continued on next page

Table F.1 – continued from previous page

ID	Content Type	Category	Description
CCSS.ELA-LITERACY.RI.3.3	ELA	Reading: Informational Text	Explain relationships: time, sequence, cause
CCSS.ELA-LITERACY.RI.3.4	ELA	Reading: Informational Text	Determine word meanings in context
CCSS.ELA-LITERACY.RI.3.5	ELA	Reading: Informational Text	Use text features to locate information
CCSS.ELA-LITERACY.RI.3.6	ELA	Reading: Informational Text	Distinguish own and author point of view
CCSS.ELA-LITERACY.RI.3.7	ELA	Reading: Informational Text	Use visuals and text to understand events
CCSS.ELA-LITERACY.RI.3.8	ELA	Reading: Informational Text	Describe links among sentences and paragraphs
CCSS.ELA-LITERACY.RI.3.9	ELA	Reading: Informational Text	Compare two texts on key points and details
CCSS.ELA-LITERACY.RI.4.1	ELA	Reading: Informational Text	Cite details and examples; draw inferences
CCSS.ELA-LITERACY.RI.4.10	ELA	Reading: Informational Text	Read complex informational texts proficiently
CCSS.ELA-LITERACY.RI.4.2	ELA	Reading: Informational Text	Identify main idea; summarize the text
CCSS.ELA-LITERACY.RI.4.3	ELA	Reading: Informational Text	Explain events or ideas using text details
CCSS.ELA-LITERACY.RI.4.4	ELA	Reading: Informational Text	Determine word meanings in context
CCSS.ELA-LITERACY.RI.4.5	ELA	Reading: Informational Text	Describe text structure and organization
CCSS.ELA-LITERACY.RI.4.6	ELA	Reading: Informational Text	Compare firsthand and secondhand accounts
CCSS.ELA-LITERACY.RI.4.7	ELA	Reading: Informational Text	Interpret information across media and formats
CCSS.ELA-LITERACY.RI.4.8	ELA	Reading: Informational Text	Explain how reasons and evidence support claims
CCSS.ELA-LITERACY.RI.4.9	ELA	Reading: Informational Text	Integrate information from multiple texts
CCSS.ELA-LITERACY.RI.5.1	ELA	Reading: Informational Text	Quote accurately; infer from text
CCSS.ELA-LITERACY.RI.5.10	ELA	Reading: Informational Text	Read complex informational texts independently
CCSS.ELA-LITERACY.RI.5.2	ELA	Reading: Informational Text	Identify multiple main ideas; summarize
CCSS.ELA-LITERACY.RI.5.3	ELA	Reading: Informational Text	Explain relationships among individuals/ideas
CCSS.ELA-LITERACY.RI.5.4	ELA	Reading: Informational Text	Determine word meanings in context
CCSS.ELA-LITERACY.RI.5.5	ELA	Reading: Informational Text	Compare structures across texts
CCSS.ELA-LITERACY.RI.5.6	ELA	Reading: Informational Text	Analyze multiple accounts; compare POV
CCSS.ELA-LITERACY.RI.5.7	ELA	Reading: Informational Text	Use multiple sources to answer efficiently
CCSS.ELA-LITERACY.RI.5.8	ELA	Reading: Informational Text	Link reasons and evidence to points
CCSS.ELA-LITERACY.RI.5.9	ELA	Reading: Informational Text	Integrate information from several texts
CCSS.ELA-LITERACY.RI.6.1	ELA	Reading: Informational Text	Cite textual evidence for analysis
CCSS.ELA-LITERACY.RI.6.10	ELA	Reading: Informational Text	Read complex literary nonfiction
CCSS.ELA-LITERACY.RI.6.2	ELA	Reading: Informational Text	Determine central idea; summarize objectively
CCSS.ELA-LITERACY.RI.6.3	ELA	Reading: Informational Text	Analyze how key ideas/events are developed
CCSS.ELA-LITERACY.RI.6.4	ELA	Reading: Informational Text	Determine word meanings, including technical
CCSS.ELA-LITERACY.RI.6.5	ELA	Reading: Informational Text	Analyze how text parts build ideas
CCSS.ELA-LITERACY.RI.6.6	ELA	Reading: Informational Text	Identify author's POV or purpose
CCSS.ELA-LITERACY.RI.6.7	ELA	Reading: Informational Text	Integrate information across media formats
CCSS.ELA-LITERACY.RI.6.8	ELA	Reading: Informational Text	Evaluate claims and supporting evidence
CCSS.ELA-LITERACY.RI.6.9	ELA	Reading: Informational Text	Compare authors' presentations of events
CCSS.ELA-LITERACY.RI.7.1	ELA	Reading: Informational Text	Cite multiple pieces of text evidence
CCSS.ELA-LITERACY.RI.7.10	ELA	Reading: Informational Text	Read complex literary nonfiction

Continued on next page

Table F.1 – continued from previous page

ID	Content Type	Category	Description
CCSS.ELA-LITERACY.RI.7.2	ELA	Reading: Informational Text	Identify multiple central ideas; summarize
CCSS.ELA-LITERACY.RI.7.3	ELA	Reading: Informational Text	Analyze interactions among individuals/events/ideas
CCSS.ELA-LITERACY.RI.7.4	ELA	Reading: Informational Text	Analyze word meanings and tone
CCSS.ELA-LITERACY.RI.7.5	ELA	Reading: Informational Text	Analyze author's organizational structure
CCSS.ELA-LITERACY.RI.7.6	ELA	Reading: Informational Text	Analyze author's POV vs others
CCSS.ELA-LITERACY.RI.7.7	ELA	Reading: Informational Text	Compare text to multimedia version
CCSS.ELA-LITERACY.RI.7.8	ELA	Reading: Informational Text	Assess reasoning and evidence sufficiency
CCSS.ELA-LITERACY.RI.7.9	ELA	Reading: Informational Text	Analyze how authors frame information
CCSS.ELA-LITERACY.RI.8.1	ELA	Reading: Informational Text	Cite strongest textual evidence
CCSS.ELA-LITERACY.RI.8.10	ELA	Reading: Informational Text	Read complex literary nonfiction independently
CCSS.ELA-LITERACY.RI.8.2	ELA	Reading: Informational Text	Analyze central idea development; summarize
CCSS.ELA-LITERACY.RI.8.3	ELA	Reading: Informational Text	Analyze connections and distinctions in text
CCSS.ELA-LITERACY.RI.8.4	ELA	Reading: Informational Text	Analyze word meaning and tone; allusions
CCSS.ELA-LITERACY.RI.8.5	ELA	Reading: Informational Text	Analyze paragraph structure and sentence roles
CCSS.ELA-LITERACY.RI.8.6	ELA	Reading: Informational Text	Analyze author POV; address conflicting evidence
CCSS.ELA-LITERACY.RI.8.7	ELA	Reading: Informational Text	Evaluate media choices for presenting ideas
CCSS.ELA-LITERACY.RI.8.8	ELA	Reading: Informational Text	Evaluate argument, evidence, and relevance
CCSS.ELA-LITERACY.RI.8.9	ELA	Reading: Informational Text	Identify conflicts across texts
CCSS.ELA-LITERACY.RI.9-10.1	ELA	Reading: Informational Text	Cite strong textual evidence for analysis
CCSS.ELA-LITERACY.RI.9-10.10	ELA	Reading: Informational Text	Read complex literary nonfiction (9–10 band)
CCSS.ELA-LITERACY.RI.9-10.2	ELA	Reading: Informational Text	Analyze central idea refinement; summarize
CCSS.ELA-LITERACY.RI.9-10.3	ELA	Reading: Informational Text	Analyze how author develops analysis and sequence
CCSS.ELA-LITERACY.RI.9-10.4	ELA	Reading: Informational Text	Analyze cumulative impact of word choice
CCSS.ELA-LITERACY.RI.9-10.5	ELA	Reading: Informational Text	Analyze how claims are developed and refined
CCSS.ELA-LITERACY.RI.9-10.6	ELA	Reading: Informational Text	Analyze rhetoric advancing author purpose
CCSS.ELA-LITERACY.RI.9-10.7	ELA	Reading: Informational Text	Compare accounts across media for emphasis
CCSS.ELA-LITERACY.RI.9-10.8	ELA	Reading: Informational Text	Evaluate argument validity and fallacies
CCSS.ELA-LITERACY.RI.9-10.9	ELA	Reading: Informational Text	Analyze themes in seminal U.S. documents
CCSS.ELA-LITERACY.RL.11-12.1	ELA	Reading: Literature	Cite evidence; note textual uncertainty
CCSS.ELA-LITERACY.RL.11-12.10	ELA	Reading: Literature	Read complex literature (11-12 band)
CCSS.ELA-LITERACY.RL.11-12.2	ELA	Reading: Literature	Analyze multiple themes/ideas and interactions
CCSS.ELA-LITERACY.RL.11-12.3	ELA	Reading: Literature	Analyze impact of author's development choices
CCSS.ELA-LITERACY.RL.11-12.4	ELA	Reading: Literature	Analyze word choice effects on meaning and tone
CCSS.ELA-LITERACY.RL.11-12.5	ELA	Reading: Literature	Analyze structural choices and aesthetic impact
CCSS.ELA-LITERACY.RL.11-12.6	ELA	Reading: Literature	Analyze implied meaning (irony/satire)
CCSS.ELA-LITERACY.RL.11-12.7	ELA	Reading: Literature	Evaluate interpretations across productions
CCSS.ELA-LITERACY.RL.11-12.8	ELA	Reading: Literature	Not applicable
CCSS.ELA-LITERACY.RL.11-12.9	ELA	Reading: Literature	Compare themes across foundational American works
CCSS.ELA-LITERACY.RL.3.1	ELA	Reading: Literature	Ask and answer using text evidence

Continued on next page

Table F.1 – continued from previous page

ID	Content Type	Category	Description
CCSS.ELA-LITERACY.RL.3.10	ELA	Reading: Literature	Read complex literature independently
CCSS.ELA-LITERACY.RL.3.2	ELA	Reading: Literature	Recount stories; identify message or moral
CCSS.ELA-LITERACY.RL.3.3	ELA	Reading: Literature	Describe characters and how actions drive plot
CCSS.ELA-LITERACY.RL.3.4	ELA	Reading: Literature	Interpret words: literal vs figurative
CCSS.ELA-LITERACY.RL.3.5	ELA	Reading: Literature	Refer to parts and how they build the whole
CCSS.ELA-LITERACY.RL.3.6	ELA	Reading: Literature	Distinguish own and narrator point of view
CCSS.ELA-LITERACY.RL.3.7	ELA	Reading: Literature	Explain how illustrations shape meaning or mood
CCSS.ELA-LITERACY.RL.3.8	ELA	Reading: Literature	Not applicable
CCSS.ELA-LITERACY.RL.3.9	ELA	Reading: Literature	Compare themes, settings, plots by same author
CCSS.ELA-LITERACY.RL.4.1	ELA	Reading: Literature	Cite details and examples; draw inferences
CCSS.ELA-LITERACY.RL.4.10	ELA	Reading: Literature	Read complex literature proficiently
CCSS.ELA-LITERACY.RL.4.2	ELA	Reading: Literature	Determine theme and summarize
CCSS.ELA-LITERACY.RL.4.3	ELA	Reading: Literature	Analyze character, setting, or event with details
CCSS.ELA-LITERACY.RL.4.4	ELA	Reading: Literature	Interpret words, including mythological allusions
CCSS.ELA-LITERACY.RL.4.5	ELA	Reading: Literature	Compare genres; reference structural elements
CCSS.ELA-LITERACY.RL.4.6	ELA	Reading: Literature	Compare narrative point of view
CCSS.ELA-LITERACY.RL.4.7	ELA	Reading: Literature	Connect written text to visual or oral version
CCSS.ELA-LITERACY.RL.4.8	ELA	Reading: Literature	Not applicable
CCSS.ELA-LITERACY.RL.4.9	ELA	Reading: Literature	Compare themes and patterns across cultures
CCSS.ELA-LITERACY.RL.5.1	ELA	Reading: Literature	Quote accurately; infer from text
CCSS.ELA-LITERACY.RL.5.10	ELA	Reading: Literature	Read complex literature independently
CCSS.ELA-LITERACY.RL.5.2	ELA	Reading: Literature	Determine theme; summarize with details
CCSS.ELA-LITERACY.RL.5.3	ELA	Reading: Literature	Compare characters, settings, and events
CCSS.ELA-LITERACY.RL.5.4	ELA	Reading: Literature	Interpret figurative language
CCSS.ELA-LITERACY.RL.5.5	ELA	Reading: Literature	Explain how parts create overall structure
CCSS.ELA-LITERACY.RL.5.6	ELA	Reading: Literature	Explain how narrator POV shapes description
CCSS.ELA-LITERACY.RL.5.7	ELA	Reading: Literature	Analyze effects of visual/multimedia elements
CCSS.ELA-LITERACY.RL.5.8	ELA	Reading: Literature	Not applicable
CCSS.ELA-LITERACY.RL.5.9	ELA	Reading: Literature	Compare stories within the same genre
CCSS.ELA-LITERACY.RL.6.1	ELA	Reading: Literature	Cite textual evidence for analysis
CCSS.ELA-LITERACY.RL.6.10	ELA	Reading: Literature	Read complex literature (6-8 band)
CCSS.ELA-LITERACY.RL.6.2	ELA	Reading: Literature	Determine theme/central idea; summarize objectively
CCSS.ELA-LITERACY.RL.6.3	ELA	Reading: Literature	Describe plot progression and character change
CCSS.ELA-LITERACY.RL.6.4	ELA	Reading: Literature	Analyze word choice impact on meaning/tone
CCSS.ELA-LITERACY.RL.6.5	ELA	Reading: Literature	Analyze how parts develop theme, setting, plot
CCSS.ELA-LITERACY.RL.6.6	ELA	Reading: Literature	Explain development of narrator POV
CCSS.ELA-LITERACY.RL.6.7	ELA	Reading: Literature	Compare reading to listening/viewing
CCSS.ELA-LITERACY.RL.6.8	ELA	Reading: Literature	Not applicable
CCSS.ELA-LITERACY.RL.6.9	ELA	Reading: Literature	Compare texts across forms or genres

Continued on next page

Table F.1 – continued from previous page

ID	Content Type	Category	Description
CCSS.ELA-LITERACY.RL.7.1	ELA	Reading: Literature	Cite multiple pieces of text evidence
CCSS.ELA-LITERACY.RL.7.10	ELA	Reading: Literature	Read complex literature (6-8 band)
CCSS.ELA-LITERACY.RL.7.2	ELA	Reading: Literature	Analyze theme/central idea development
CCSS.ELA-LITERACY.RL.7.3	ELA	Reading: Literature	Analyze interaction of story elements
CCSS.ELA-LITERACY.RL.7.4	ELA	Reading: Literature	Analyze sound devices' effect
CCSS.ELA-LITERACY.RL.7.5	ELA	Reading: Literature	Analyze form/structure and meaning
CCSS.ELA-LITERACY.RL.7.6	ELA	Reading: Literature	Analyze contrasting points of view
CCSS.ELA-LITERACY.RL.7.7	ELA	Reading: Literature	Compare written to multimedia; analyze techniques
CCSS.ELA-LITERACY.RL.7.8	ELA	Reading: Literature	Not applicable
CCSS.ELA-LITERACY.RL.7.9	ELA	Reading: Literature	Compare fiction and historical accounts
CCSS.ELA-LITERACY.RL.8.1	ELA	Reading: Literature	Cite strongest textual evidence
CCSS.ELA-LITERACY.RL.8.10	ELA	Reading: Literature	Read complex literature independently
CCSS.ELA-LITERACY.RL.8.2	ELA	Reading: Literature	Analyze theme development in narrative; summarize
CCSS.ELA-LITERACY.RL.8.3	ELA	Reading: Literature	Analyze dialogue/incidents that advance plot
CCSS.ELA-LITERACY.RL.8.4	ELA	Reading: Literature	Analyze word meaning and tone; allusions
CCSS.ELA-LITERACY.RL.8.5	ELA	Reading: Literature	Compare text structures and their effects
CCSS.ELA-LITERACY.RL.8.6	ELA	Reading: Literature	Analyze POV differences and dramatic irony effects
CCSS.ELA-LITERACY.RL.8.7	ELA	Reading: Literature	Evaluate adaptation fidelity and choices
CCSS.ELA-LITERACY.RL.8.8	ELA	Reading: Literature	Not applicable
CCSS.ELA-LITERACY.RL.8.9	ELA	Reading: Literature	Analyze modern fiction's use of mythic sources
CCSS.ELA-LITERACY.RL.9-10.1	ELA	Reading: Literature	Cite strong textual evidence for analysis
CCSS.ELA-LITERACY.RL.9-10.10	ELA	Reading: Literature	Read complex literature (9–10 band)
CCSS.ELA-LITERACY.RL.9-10.2	ELA	Reading: Literature	Analyze theme development; summarize objectively
CCSS.ELA-LITERACY.RL.9-10.3	ELA	Reading: Literature	Analyze complex character development
CCSS.ELA-LITERACY.RL.9-10.4	ELA	Reading: Literature	Analyze cumulative word choice and tone
CCSS.ELA-LITERACY.RL.9-10.5	ELA	Reading: Literature	Analyze structure and pacing effects
CCSS.ELA-LITERACY.RL.9-10.6	ELA	Reading: Literature	Analyze cultural perspective in world literature
CCSS.ELA-LITERACY.RL.9-10.7	ELA	Reading: Literature	Compare artistic representations across media
CCSS.ELA-LITERACY.RL.9-10.8	ELA	Reading: Literature	Not applicable
CCSS.ELA-LITERACY.RL.9-10.9	ELA	Reading: Literature	Analyze transformation of source material
CCSS.MATH.CONTENT.3.G.A.1	Math	Geometry	Classify quadrilaterals by attributes
CCSS.MATH.CONTENT.3.G.A.2	Math	Geometry	Partition shapes; express area as unit fraction
CCSS.MATH.CONTENT.3.MD.A.1	Math	Measurement & Data	Tell time; solve time-interval problems
CCSS.MATH.CONTENT.3.MD.A.2	Math	Measurement & Data	Measure mass/volume; solve one-step problems
CCSS.MATH.CONTENT.3.MD.B.3	Math	Measurement & Data	Make scaled graphs; solve more/less problems
CCSS.MATH.CONTENT.3.MD.B.4	Math	Measurement & Data	Measure lengths; make a line plot
CCSS.MATH.CONTENT.3.MD.C.5	Math	Measurement & Data	Recognize area and basic measurement concepts
CCSS.MATH.CONTENT.3.MD.C.5.a	Math	Measurement & Data	Use unit squares to measure area
CCSS.MATH.CONTENT.3.MD.C.5.b	Math	Measurement & Data	Define area by covering with unit squares

Continued on next page

Table F.1 – continued from previous page

ID	Content Type	Category	Description
CCSS.MATH.CONTENT.3.MD.C.6	Math	Measurement & Data	Measure area by counting unit squares
CCSS.MATH.CONTENT.3.MD.C.7	Math	Measurement & Data	Relate area to multiplication and addition
CCSS.MATH.CONTENT.3.MD.C.7.a	Math	Measurement & Data	Find rectangle area by tiling/multiplying
CCSS.MATH.CONTENT.3.MD.C.7.b	Math	Measurement & Data	Multiply side lengths to find rectangle area
CCSS.MATH.CONTENT.3.MD.C.7.c	Math	Measurement & Data	Use area models for distributive property
CCSS.MATH.CONTENT.3.MD.C.7.d	Math	Measurement & Data	Decompose figures; add areas
CCSS.MATH.CONTENT.3.MD.D.8	Math	Measurement & Data	Solve perimeter problems for polygons
CCSS.MATH.CONTENT.3.NBT.A.1	Math	Number & Operations in Base Ten	Round whole numbers to nearest 10/100
CCSS.MATH.CONTENT.3.NBT.A.2	Math	Number & Operations in Base Ten	Add and subtract within 1000 fluently
CCSS.MATH.CONTENT.3.NBT.A.3	Math	Number & Operations in Base Ten	Multiply 1-digit by multiples of 10
CCSS.MATH.CONTENT.3.NF.A.1	Math	Numbers & Operations - Fractions	Define fractions as parts of a whole
CCSS.MATH.CONTENT.3.NF.A.2	Math	Numbers & Operations - Fractions	Represent fractions on a number line
CCSS.MATH.CONTENT.3.NF.A.2.a	Math	Numbers & Operations - Fractions	Locate 1/b on a number line
CCSS.MATH.CONTENT.3.NF.A.2.b	Math	Numbers & Operations - Fractions	Locate a/b on a number line
CCSS.MATH.CONTENT.3.NF.A.3	Math	Numbers & Operations - Fractions	Explain and compare equivalent fractions
CCSS.MATH.CONTENT.3.NF.A.3.a	Math	Numbers & Operations - Fractions	Understand equivalence as same size/point
CCSS.MATH.CONTENT.3.NF.A.3.b	Math	Numbers & Operations - Fractions	Generate and justify equivalent fractions
CCSS.MATH.CONTENT.3.NF.A.3.c	Math	Numbers & Operations - Fractions	Express whole numbers as fractions
CCSS.MATH.CONTENT.3.NF.A.3.d	Math	Numbers & Operations - Fractions	Compare fractions with same numerator/denominator
CCSS.MATH.CONTENT.3.OA.A.1	Math	Operations & Algebraic Thinking	Interpret multiplication as equal groups
CCSS.MATH.CONTENT.3.OA.A.2	Math	Operations & Algebraic Thinking	Interpret division as sharing/grouping
CCSS.MATH.CONTENT.3.OA.A.3	Math	Operations & Algebraic Thinking	Solve word problems using mult/div within 100
CCSS.MATH.CONTENT.3.OA.A.4	Math	Operations & Algebraic Thinking	Find unknown in mult/div equations
CCSS.MATH.CONTENT.3.OA.B.5	Math	Operations & Algebraic Thinking	Use properties to multiply and divide
CCSS.MATH.CONTENT.3.OA.B.6	Math	Operations & Algebraic Thinking	Treat division as unknown-factor
CCSS.MATH.CONTENT.3.OA.C.7	Math	Operations & Algebraic Thinking	Fluently multiply and divide within 100
CCSS.MATH.CONTENT.3.OA.D.8	Math	Operations & Algebraic Thinking	Solve two-step word problems; assess reasonableness
CCSS.MATH.CONTENT.3.OA.D.9	Math	Operations & Algebraic Thinking	Identify arithmetic patterns using properties
CCSS.MATH.CONTENT.4.G.A.1	Math	Geometry	Draw and identify basic geometric objects
CCSS.MATH.CONTENT.4.G.A.2	Math	Geometry	Classify figures by lines and angles
CCSS.MATH.CONTENT.4.G.A.3	Math	Geometry	Identify and draw lines of symmetry
CCSS.MATH.CONTENT.4.MD.A.1	Math	Measurement & Data	Convert units within a measurement system
CCSS.MATH.CONTENT.4.MD.A.2	Math	Measurement & Data	Solve measurement word problems using 4 ops
CCSS.MATH.CONTENT.4.MD.A.3	Math	Measurement & Data	Apply rectangle area/perimeter formulas
CCSS.MATH.CONTENT.4.MD.B.4	Math	Measurement & Data	Use line plots to add/subtract fractions
CCSS.MATH.CONTENT.4.MD.C.5	Math	Measurement & Data	Recognize angles and angle measurement
CCSS.MATH.CONTENT.4.MD.C.5.a	Math	Measurement & Data	Define degree measure using a circle
CCSS.MATH.CONTENT.4.MD.C.5.b	Math	Measurement & Data	Relate angle measure to degrees
CCSS.MATH.CONTENT.4.MD.C.6	Math	Measurement & Data	Measure and draw angles with a protractor

Continued on next page

Table F.1 – continued from previous page

ID	Content Type	Category	Description
CCSS.MATH.CONTENT.4.MD.C.7	Math	Measurement & Data	Use additive angle measures to find unknowns
CCSS.MATH.CONTENT.4.NBT.A.1	Math	Number & Operations in Base Ten	Recognize place value as powers of ten
CCSS.MATH.CONTENT.4.NBT.A.2	Math	Number & Operations in Base Ten	Read, write, and compare multi-digit numbers
CCSS.MATH.CONTENT.4.NBT.A.3	Math	Number & Operations in Base Ten	Round multi-digit numbers to any place
CCSS.MATH.CONTENT.4.NBT.B.4	Math	Number & Operations in Base Ten	Add/subtract multi-digit numbers fluently
CCSS.MATH.CONTENT.4.NBT.B.5	Math	Number & Operations in Base Ten	Multiply multi-digit numbers using strategies
CCSS.MATH.CONTENT.4.NBT.B.6	Math	Number & Operations in Base Ten	Divide multi-digit numbers; quotients and remainders
CCSS.MATH.CONTENT.4.NF.A.1	Math	Numbers & Operations - Fractions	Explain equivalent fractions using models
CCSS.MATH.CONTENT.4.NF.A.2	Math	Numbers & Operations - Fractions	Compare fractions using benchmarks/common units
CCSS.MATH.CONTENT.4.NF.B.3	Math	Numbers & Operations - Fractions	Express improper fractions as sums of unit fractions
CCSS.MATH.CONTENT.4.NF.B.3.a	Math	Numbers & Operations - Fractions	Add/subtract fractions as joining/separating
CCSS.MATH.CONTENT.4.NF.B.3.b	Math	Numbers & Operations - Fractions	Decompose fractions with same denominator
CCSS.MATH.CONTENT.4.NF.B.3.c	Math	Numbers & Operations - Fractions	Add/subtract mixed numbers (like denominators)
CCSS.MATH.CONTENT.4.NF.B.3.d	Math	Numbers & Operations - Fractions	Solve fraction add/sub word problems (like denom)
CCSS.MATH.CONTENT.4.NF.B.4	Math	Numbers & Operations - Fractions	Multiply a fraction by a whole number
CCSS.MATH.CONTENT.4.NF.B.4.a	Math	Numbers & Operations - Fractions	Interpret a/b as a multiple of 1/b
CCSS.MATH.CONTENT.4.NF.B.4.b	Math	Numbers & Operations - Fractions	Use multiples to multiply fractions by whole numbers
CCSS.MATH.CONTENT.4.NF.B.4.c	Math	Numbers & Operations - Fractions	Solve fraction-by-whole word problems
CCSS.MATH.CONTENT.4.NF.C.5	Math	Numbers & Operations - Fractions	Convert tenths to hundredths; add
CCSS.MATH.CONTENT.4.NF.C.6	Math	Numbers & Operations - Fractions	Write decimals for tenths and hundredths
CCSS.MATH.CONTENT.4.NF.C.7	Math	Numbers & Operations - Fractions	Compare decimals to hundredths
CCSS.MATH.CONTENT.4.OA.A.1	Math	Operations & Algebraic Thinking	Interpret multiplication as a comparison
CCSS.MATH.CONTENT.4.OA.A.2	Math	Operations & Algebraic Thinking	Solve multiplicative comparison word problems
CCSS.MATH.CONTENT.4.OA.A.3	Math	Operations & Algebraic Thinking	Solve multistep word problems (interpret remainders)
CCSS.MATH.CONTENT.4.OA.B.4	Math	Operations & Algebraic Thinking	Find factors; identify prime or composite
CCSS.MATH.CONTENT.4.OA.C.5	Math	Operations & Algebraic Thinking	Generate patterns from a rule
CCSS.MATH.CONTENT.5.G.A.1	Math	Geometry	Define a coordinate system with axes
CCSS.MATH.CONTENT.5.G.A.2	Math	Geometry	Graph points in the first quadrant
CCSS.MATH.CONTENT.5.G.B.3	Math	Geometry	Recognize attributes in shape hierarchies
CCSS.MATH.CONTENT.5.G.B.4	Math	Geometry	Classify 2D figures in a hierarchy
CCSS.MATH.CONTENT.5.MD.A.1	Math	Measurement & Data	Convert measurement units; solve problems
CCSS.MATH.CONTENT.5.MD.B.2	Math	Measurement & Data	Use line plots to solve fraction problems
CCSS.MATH.CONTENT.5.MD.C.3	Math	Measurement & Data	Recognize volume and measurement concepts
CCSS.MATH.CONTENT.5.MD.C.3.a	Math	Measurement & Data	Use unit cubes to measure volume
CCSS.MATH.CONTENT.5.MD.C.3.b	Math	Measurement & Data	Define volume by packing unit cubes
CCSS.MATH.CONTENT.5.MD.C.4	Math	Measurement & Data	Measure volume by counting unit cubes
CCSS.MATH.CONTENT.5.MD.C.5	Math	Measurement & Data	Relate volume to mult/add; solve problems
CCSS.MATH.CONTENT.5.MD.C.5.a	Math	Measurement & Data	Find volume of a rectangular prism
CCSS.MATH.CONTENT.5.MD.C.5.b	Math	Measurement & Data	Apply volume formulas for rectangular prisms

Continued on next page

Table F.1 – continued from previous page

ID	Content Type	Category	Description
CCSS.MATH.CONTENT.5.MD.C.5.c	Math	Measurement & Data	Add volumes of composite prisms
CCSS.MATH.CONTENT.5.NBT.A.1	Math	Number & Operations in Base Ten	Use place value (10x and 1/10)
CCSS.MATH.CONTENT.5.NBT.A.2	Math	Number & Operations in Base Ten	Explain patterns with powers of 10
CCSS.MATH.CONTENT.5.NBT.A.3	Math	Number & Operations in Base Ten	Read, write, and compare thousandths
CCSS.MATH.CONTENT.5.NBT.A.3.a	Math	Number & Operations in Base Ten	Write thousandths in expanded form
CCSS.MATH.CONTENT.5.NBT.A.3.b	Math	Number & Operations in Base Ten	Compare decimals to thousandths
CCSS.MATH.CONTENT.5.NBT.A.4	Math	Number & Operations in Base Ten	Round decimals to any place
CCSS.MATH.CONTENT.5.NBT.B.5	Math	Number & Operations in Base Ten	Multiply multi-digit whole numbers fluently
CCSS.MATH.CONTENT.5.NBT.B.6	Math	Number & Operations in Base Ten	Divide by two-digit divisors
CCSS.MATH.CONTENT.5.NBT.B.7	Math	Number & Operations in Base Ten	Add/subtract/multiply/divide decimals to hundredths
CCSS.MATH.CONTENT.5.NF.A.1	Math	Numbers & Operations - Fractions	Add/subtract fractions with unlike denominators
CCSS.MATH.CONTENT.5.NF.A.2	Math	Numbers & Operations - Fractions	Solve fraction add/sub word problems (unlike denom)
CCSS.MATH.CONTENT.5.NF.B.3	Math	Numbers & Operations - Fractions	Interpret fractions as division; solve problems
CCSS.MATH.CONTENT.5.NF.B.4	Math	Numbers & Operations - Fractions	Multiply fractions and whole numbers by fractions
CCSS.MATH.CONTENT.5.NF.B.4.a	Math	Numbers & Operations - Fractions	Interpret fraction products (a/b) x q
CCSS.MATH.CONTENT.5.NF.B.4.b	Math	Numbers & Operations - Fractions	Find rectangle area with fractional side lengths
CCSS.MATH.CONTENT.5.NF.B.5	Math	Numbers & Operations - Fractions	Interpret multiplication as scaling
CCSS.MATH.CONTENT.5.NF.B.5.a	Math	Numbers & Operations - Fractions	Compare product size to factors
CCSS.MATH.CONTENT.5.NF.B.5.b	Math	Numbers & Operations - Fractions	Explain scaling effects of fractions
CCSS.MATH.CONTENT.5.NF.B.6	Math	Numbers & Operations - Fractions	Solve real-world fraction/mixed multiplication
CCSS.MATH.CONTENT.5.NF.B.7	Math	Numbers & Operations - Fractions	Divide unit fractions and whole numbers
CCSS.MATH.CONTENT.5.NF.B.7.a	Math	Numbers & Operations - Fractions	Compute unit fraction divided by whole number
CCSS.MATH.CONTENT.5.NF.B.7.b	Math	Numbers & Operations - Fractions	Compute whole number divided by unit fraction
CCSS.MATH.CONTENT.5.NF.B.7.c	Math	Numbers & Operations - Fractions	Solve real-world unit fraction division problems
CCSS.MATH.CONTENT.5.OA.A.1	Math	Operations & Algebraic Thinking	Evaluate expressions with grouping symbols
CCSS.MATH.CONTENT.5.OA.A.2	Math	Operations & Algebraic Thinking	Write and interpret numerical expressions
CCSS.MATH.CONTENT.5.OA.B.3	Math	Operations & Algebraic Thinking	Generate two patterns and graph ordered pairs
CCSS.MATH.CONTENT.6.EE.A.1	Math	Expressions & Equations	Write/evaluate expressions with whole-number exponents
CCSS.MATH.CONTENT.6.EE.A.2	Math	Expressions & Equations	Write/read/evaluate expressions with variables
CCSS.MATH.CONTENT.6.EE.A.2.a	Math	Expressions & Equations	Write expressions with variables for operations
CCSS.MATH.CONTENT.6.EE.A.2.b	Math	Expressions & Equations	Identify parts of an expression (terms, factors)
CCSS.MATH.CONTENT.6.EE.A.2.c	Math	Expressions & Equations	Evaluate expressions using values and formulas
CCSS.MATH.CONTENT.6.EE.A.3	Math	Expressions & Equations	Generate equivalent expressions
CCSS.MATH.CONTENT.6.EE.A.4	Math	Expressions & Equations	Identify equivalent expressions
CCSS.MATH.CONTENT.6.EE.B.5	Math	Expressions & Equations	Test solutions by substitution
CCSS.MATH.CONTENT.6.EE.B.6	Math	Expressions & Equations	Use variables to represent numbers
CCSS.MATH.CONTENT.6.EE.B.7	Math	Expressions & Equations	Solve one-step equations
CCSS.MATH.CONTENT.6.EE.B.8	Math	Expressions & Equations	Write and graph inequalities

Continued on next page

Table F.1 – continued from previous page

ID	Content Type	Category	Description
CCSS.MATH.CONTENT.6.EE.C.9	Math	Expressions & Equations	Model relationships with equations
CCSS.MATH.CONTENT.6.G.A.1	Math	Geometry	Find area by composing/decomposing
CCSS.MATH.CONTENT.6.G.A.2	Math	Geometry	Find prism volume with fractional edges
CCSS.MATH.CONTENT.6.G.A.3	Math	Geometry	Use coordinates to find side lengths
CCSS.MATH.CONTENT.6.G.A.4	Math	Geometry	Use nets to find surface area
CCSS.MATH.CONTENT.6.NS.A.1	Math	The Number System	Divide fractions and solve problems
CCSS.MATH.CONTENT.6.NS.B.2	Math	The Number System	Divide multi-digit numbers (standard)
CCSS.MATH.CONTENT.6.NS.B.3	Math	The Number System	Operate on multi-digit decimals
CCSS.MATH.CONTENT.6.NS.B.4	Math	The Number System	Find GCF and LCM
CCSS.MATH.CONTENT.6.NS.C.5	Math	The Number System	Use positive/negative numbers in context
CCSS.MATH.CONTENT.6.NS.C.6	Math	The Number System	Plot rational numbers on number lines
CCSS.MATH.CONTENT.6.NS.C.6.a	Math	The Number System	Interpret opposites on number line
CCSS.MATH.CONTENT.6.NS.C.6.b	Math	The Number System	Interpret signs in ordered pairs
CCSS.MATH.CONTENT.6.NS.C.6.c	Math	The Number System	Locate integers and rationals on plane
CCSS.MATH.CONTENT.6.NS.C.7	Math	The Number System	Order rational numbers; use absolute value
CCSS.MATH.CONTENT.6.NS.C.7.a	Math	The Number System	Interpret inequalities on number line
CCSS.MATH.CONTENT.6.NS.C.7.b	Math	The Number System	Write order comparisons in context
CCSS.MATH.CONTENT.6.NS.C.7.c	Math	The Number System	Interpret absolute value as distance
CCSS.MATH.CONTENT.6.NS.C.7.d	Math	The Number System	Distinguish absolute value vs order
CCSS.MATH.CONTENT.6.NS.C.8	Math	The Number System	Graph points; find distances using abs
CCSS.MATH.CONTENT.6.RP.A.1	Math	Ratios & Proportional Relationships	Describe ratio relationships
CCSS.MATH.CONTENT.6.RP.A.2	Math	Ratios & Proportional Relationships	Compute and interpret unit rates
CCSS.MATH.CONTENT.6.RP.A.3	Math	Ratios & Proportional Relationships	Solve problems with ratios and rates
CCSS.MATH.CONTENT.6.RP.A.3.a	Math	Ratios & Proportional Relationships	Make ratio tables and plot pairs
CCSS.MATH.CONTENT.6.RP.A.3.b	Math	Ratios & Proportional Relationships	Solve unit rate problems
CCSS.MATH.CONTENT.6.RP.A.3.c	Math	Ratios & Proportional Relationships	Solve percent problems
CCSS.MATH.CONTENT.6.RP.A.3.d	Math	Ratios & Proportional Relationships	Convert units using ratios
CCSS.MATH.CONTENT.6.SP.A.1	Math	Statistics & Probability	Identify statistical questions
CCSS.MATH.CONTENT.6.SP.A.2	Math	Statistics & Probability	Describe distributions (center/spread/shape)
CCSS.MATH.CONTENT.6.SP.A.3	Math	Statistics & Probability	Use measures of center and variability
CCSS.MATH.CONTENT.6.SP.B.4	Math	Statistics & Probability	Plot data (dot/hist/box)
CCSS.MATH.CONTENT.6.SP.B.5	Math	Statistics & Probability	Summarize numerical data in context
CCSS.MATH.CONTENT.6.SP.B.5.a	Math	Statistics & Probability	Report number of observations
CCSS.MATH.CONTENT.6.SP.B.5.b	Math	Statistics & Probability	Describe variable and units measured
CCSS.MATH.CONTENT.6.SP.B.5.c	Math	Statistics & Probability	Report center/variability; note outliers
CCSS.MATH.CONTENT.6.SP.B.5.d	Math	Statistics & Probability	Choose measures based on distribution
CCSS.MATH.CONTENT.7.EE.A.1	Math	Expressions & Equations	Manipulate linear expressions
CCSS.MATH.CONTENT.7.EE.A.2	Math	Expressions & Equations	Rewrite expressions to reveal relationships
CCSS.MATH.CONTENT.7.EE.B.3	Math	Expressions & Equations	Solve multi-step problems with rational numbers

Continued on next page

Table F.1 – continued from previous page

ID	Content Type	Category	Description
CCSS.MATH.CONTENT.7.EE.B.4	Math	Expressions & Equations	Model with variables and inequalities
CCSS.MATH.CONTENT.7.EE.B.4.a	Math	Expressions & Equations	Solve two-step equations
CCSS.MATH.CONTENT.7.EE.B.4.b	Math	Expressions & Equations	Solve and graph inequalities
CCSS.MATH.CONTENT.7.G.A.1	Math	Geometry	Solve problems with scale drawings
CCSS.MATH.CONTENT.7.G.A.2	Math	Geometry	Construct geometric shapes from conditions
CCSS.MATH.CONTENT.7.G.A.3	Math	Geometry	Describe cross-sections of solids
CCSS.MATH.CONTENT.7.G.B.4	Math	Geometry	Use circle area and circumference formulas
CCSS.MATH.CONTENT.7.G.B.5	Math	Geometry	Use angle relationships to solve equations
CCSS.MATH.CONTENT.7.G.B.6	Math	Geometry	Solve area/volume/surface area problems
CCSS.MATH.CONTENT.7.NS.A.1	Math	The Number System	Add and subtract rational numbers
CCSS.MATH.CONTENT.7.NS.A.1.a	Math	The Number System	Describe opposites that sum to zero
CCSS.MATH.CONTENT.7.NS.A.1.b	Math	The Number System	Interpret addition as movement on number line
CCSS.MATH.CONTENT.7.NS.A.1.c	Math	The Number System	Interpret subtraction and distance on number line
CCSS.MATH.CONTENT.7.NS.A.1.d	Math	The Number System	Use properties to add/subtract rationals
CCSS.MATH.CONTENT.7.NS.A.2	Math	The Number System	Multiply and divide rational numbers
CCSS.MATH.CONTENT.7.NS.A.2.a	Math	The Number System	Extend multiplication to rational numbers
CCSS.MATH.CONTENT.7.NS.A.2.b	Math	The Number System	Extend division to rational numbers
CCSS.MATH.CONTENT.7.NS.A.2.c	Math	The Number System	Use properties to multiply/divide rationals
CCSS.MATH.CONTENT.7.NS.A.2.d	Math	The Number System	Convert rationals to decimals (long division)
CCSS.MATH.CONTENT.7.NS.A.3	Math	The Number System	Solve problems with rational numbers
CCSS.MATH.CONTENT.7.RP.A.1	Math	Ratios & Proportional Relationships	Compute unit rates for fractional ratios
CCSS.MATH.CONTENT.7.RP.A.2	Math	Ratios & Proportional Relationships	Recognize proportional relationships
CCSS.MATH.CONTENT.7.RP.A.2.a	Math	Ratios & Proportional Relationships	Test whether quantities are proportional
CCSS.MATH.CONTENT.7.RP.A.2.b	Math	Ratios & Proportional Relationships	Identify constant of proportionality
CCSS.MATH.CONTENT.7.RP.A.2.c	Math	Ratios & Proportional Relationships	Write equations for proportional relationships
CCSS.MATH.CONTENT.7.RP.A.2.d	Math	Ratios & Proportional Relationships	Interpret points on proportional graphs
CCSS.MATH.CONTENT.7.RP.A.3	Math	Ratios & Proportional Relationships	Solve multistep ratio and percent problems
CCSS.MATH.CONTENT.7.SP.A.1	Math	Statistics & Probability	Use random samples to infer about populations
CCSS.MATH.CONTENT.7.SP.A.2	Math	Statistics & Probability	Use samples to estimate and gauge variation
CCSS.MATH.CONTENT.7.SP.B.3	Math	Statistics & Probability	Compare distributions using center and spread
CCSS.MATH.CONTENT.7.SP.B.4	Math	Statistics & Probability	Compare populations using samples (center/spread)
CCSS.MATH.CONTENT.7.SP.C.5	Math	Statistics & Probability	Interpret probability as likelihood (0 to 1)
CCSS.MATH.CONTENT.7.SP.C.6	Math	Statistics & Probability	Estimate probability from relative frequency
CCSS.MATH.CONTENT.7.SP.C.7	Math	Statistics & Probability	Use probability models; compare to data
CCSS.MATH.CONTENT.7.SP.C.7.a	Math	Statistics & Probability	Build a uniform probability model
CCSS.MATH.CONTENT.7.SP.C.7.b	Math	Statistics & Probability	Build probability models from data
CCSS.MATH.CONTENT.7.SP.C.8	Math	Statistics & Probability	Find probabilities of compound events
CCSS.MATH.CONTENT.7.SP.C.8.a	Math	Statistics & Probability	Define probability of compound events
CCSS.MATH.CONTENT.7.SP.C.8.b	Math	Statistics & Probability	Represent sample spaces for compound events

Continued on next page

Table F.1 – continued from previous page

ID	Content Type	Category	Description
CCSS.MATH.CONTENT.7.SP.C.8.c	Math	Statistics & Probability	Use simulation for compound probabilities
CCSS.MATH.CONTENT.8.EE.A.1	Math	Expressions & Equations	Apply integer exponent rules
CCSS.MATH.CONTENT.8.EE.A.2	Math	Expressions & Equations	Use square/cube roots; recognize irrationals
CCSS.MATH.CONTENT.8.EE.A.3	Math	Expressions & Equations	Use scientific notation to estimate size
CCSS.MATH.CONTENT.8.EE.A.4	Math	Expressions & Equations	Operate with scientific notation
CCSS.MATH.CONTENT.8.EE.B.5	Math	Expressions & Equations	Graph proportional relationships; interpret slope
CCSS.MATH.CONTENT.8.EE.B.6	Math	Expressions & Equations	Use similar triangles to derive slope and y=mx+b
CCSS.MATH.CONTENT.8.EE.C.7	Math	Expressions & Equations	Solve linear equations (one variable)
CCSS.MATH.CONTENT.8.EE.C.7.a	Math	Expressions & Equations	Classify solutions to linear equations
CCSS.MATH.CONTENT.8.EE.C.7.b	Math	Expressions & Equations	Solve linear equations with rational coefficients
CCSS.MATH.CONTENT.8.EE.C.8	Math	Expressions & Equations	Solve systems of linear equations
CCSS.MATH.CONTENT.8.EE.C.8.a	Math	Expressions & Equations	Interpret system solutions as intersections
CCSS.MATH.CONTENT.8.EE.C.8.b	Math	Expressions & Equations	Solve systems algebraically and graphically
CCSS.MATH.CONTENT.8.EE.C.8.c	Math	Expressions & Equations	Model problems with linear systems
CCSS.MATH.CONTENT.8.F.A.1	Math	Functions	Define functions and their graphs
CCSS.MATH.CONTENT.8.F.A.2	Math	Functions	Compare functions across representations
CCSS.MATH.CONTENT.8.F.A.3	Math	Functions	Identify linear vs non-linear functions
CCSS.MATH.CONTENT.8.F.B.4	Math	Functions	Build linear functions from situations
CCSS.MATH.CONTENT.8.F.B.5	Math	Functions	Describe and sketch function behavior
CCSS.MATH.CONTENT.8.G.A.1	Math	Geometry	Verify rigid motion properties
CCSS.MATH.CONTENT.8.G.A.1.a	Math	Geometry	Verify lines/segments map with same length
CCSS.MATH.CONTENT.8.G.A.1.b	Math	Geometry	Verify angles preserve measure
CCSS.MATH.CONTENT.8.G.A.1.c	Math	Geometry	Verify parallel lines stay parallel
CCSS.MATH.CONTENT.8.G.A.2	Math	Geometry	Describe congruence via rigid motions
CCSS.MATH.CONTENT.8.G.A.3	Math	Geometry	Describe transformations using coordinates
CCSS.MATH.CONTENT.8.G.A.4	Math	Geometry	Describe similarity via transformations
CCSS.MATH.CONTENT.8.G.A.5	Math	Geometry	Use informal proofs about angles and similarity
CCSS.MATH.CONTENT.8.G.B.6	Math	Geometry	Explain Pythagorean Theorem and converse
CCSS.MATH.CONTENT.8.G.B.7	Math	Geometry	Use Pythagorean Theorem for side lengths
CCSS.MATH.CONTENT.8.G.B.8	Math	Geometry	Use Pythagorean Theorem for distance
CCSS.MATH.CONTENT.8.G.C.9	Math	Geometry	Use volume formulas for cones/cylinders/spheres
CCSS.MATH.CONTENT.8.NS.A.1	Math	The Number System	Identify irrationals; convert repeating decimals
CCSS.MATH.CONTENT.8.NS.A.2	Math	The Number System	Approximate irrationals and locate on number line
CCSS.MATH.CONTENT.8.SP.A.1	Math	Statistics & Probability	Create scatter plots; describe association
CCSS.MATH.CONTENT.8.SP.A.2	Math	Statistics & Probability	Fit and assess linear models from scatter plots
CCSS.MATH.CONTENT.8.SP.A.3	Math	Statistics & Probability	Use linear models; interpret slope/intercept
CCSS.MATH.CONTENT.8.SP.A.4	Math	Statistics & Probability	Analyze association in two-way tables
CCSS.MATH.CONTENT.HSA.APR.A.1	Math	Arithmetic with Polynomials & Rational Expressions	Add, subtract, and multiply polynomials
CCSS.MATH.CONTENT.HSA.APR.B.2	Math	Arithmetic with Polynomials & Rational Expressions	Apply the Remainder Theorem

Continued on next page

Table F.1 – continued from previous page

ID	Content Type	Category	Description
CCSS.MATH.CONTENT.HSA.APR.B.3	Math	Arithmetic with Polynomials & Rational Expressions	Use zeros to sketch polynomial graphs
CCSS.MATH.CONTENT.HSA.APR.C.4	Math	Arithmetic with Polynomials & Rational Expressions	Prove and use polynomial identities
CCSS.MATH.CONTENT.HSA.APR.C.5	Math	Arithmetic with Polynomials & Rational Expressions	Apply the Binomial Theorem
CCSS.MATH.CONTENT.HSA.APR.D.6	Math	Arithmetic with Polynomials & Rational Expressions	Rewrite rational expressions via division
CCSS.MATH.CONTENT.HSA.APR.D.7	Math	Arithmetic with Polynomials & Rational Expressions	Operate on rational expressions
CCSS.MATH.CONTENT.HSA.CED.A.1	Math	Creating Equations	Create and solve one-variable equations/inequalities
CCSS.MATH.CONTENT.HSA.CED.A.2	Math	Creating Equations	Create and graph equations in two variables
CCSS.MATH.CONTENT.HSA.CED.A.3	Math	Creating Equations	Model constraints with equations/inequalities
CCSS.MATH.CONTENT.HSA.CED.A.4	Math	Creating Equations	Rearrange formulas to solve for a variable
CCSS.MATH.CONTENT.HSA.REI.A.1	Math	Reasoning with Equations & Inequalities	Justify steps when solving equations
CCSS.MATH.CONTENT.HSA.REI.A.2	Math	Reasoning with Equations & Inequalities	Solve rational and radical equations; check extraneous
CCSS.MATH.CONTENT.HSA.REI.B.3	Math	Reasoning with Equations & Inequalities	Solve linear equations and inequalities (1 variable)
CCSS.MATH.CONTENT.HSA.REI.B.4	Math	Reasoning with Equations & Inequalities	Solve quadratic equations (1 variable)
CCSS.MATH.CONTENT.HSA.REI.B.4.a	Math	Reasoning with Equations & Inequalities	Complete the square; derive quadratic formula
CCSS.MATH.CONTENT.HSA.REI.B.4.b	Math	Reasoning with Equations & Inequalities	Solve quadratics using multiple methods
CCSS.MATH.CONTENT.HSA.REI.C.5	Math	Reasoning with Equations & Inequalities	Prove row operation preserves solutions
CCSS.MATH.CONTENT.HSA.REI.C.6	Math	Reasoning with Equations & Inequalities	Solve systems of two linear equations
CCSS.MATH.CONTENT.HSA.REI.C.7	Math	Reasoning with Equations & Inequalities	Solve linear-quadratic systems
CCSS.MATH.CONTENT.HSA.REI.C.8	Math	Reasoning with Equations & Inequalities	Represent linear systems with matrices
CCSS.MATH.CONTENT.HSA.REI.C.9	Math	Reasoning with Equations & Inequalities	Use matrix inverses to solve systems
CCSS.MATH.CONTENT.HSA.REI.D.10	Math	Reasoning with Equations & Inequalities	Interpret graphs as solution sets
CCSS.MATH.CONTENT.HSA.REI.D.11	Math	Reasoning with Equations & Inequalities	Find intersections by solving f(x)=g(x)
CCSS.MATH.CONTENT.HSA.REI.D.12	Math	Reasoning with Equations & Inequalities	Graph linear inequalities in two variables
CCSS.MATH.CONTENT.HSA.SSE.A.1	Math	Seeing Structure in Expressions	Interpret expressions in context
CCSS.MATH.CONTENT.HSA.SSE.A.1.a	Math	Seeing Structure in Expressions	Identify terms, factors, and coefficients
CCSS.MATH.CONTENT.HSA.SSE.A.1.b	Math	Seeing Structure in Expressions	View parts of expressions as single entities
CCSS.MATH.CONTENT.HSA.SSE.A.2	Math	Seeing Structure in Expressions	Rewrite expressions using structure
CCSS.MATH.CONTENT.HSA.SSE.B.3	Math	Seeing Structure in Expressions	Choose equivalent forms to reveal properties
CCSS.MATH.CONTENT.HSA.SSE.B.3.a	Math	Seeing Structure in Expressions	Factor quadratics to find zeros
CCSS.MATH.CONTENT.HSA.SSE.B.3.b	Math	Seeing Structure in Expressions	Complete the square to find extrema
CCSS.MATH.CONTENT.HSA.SSE.B.3.c	Math	Seeing Structure in Expressions	Use exponent rules to transform exponentials
CCSS.MATH.CONTENT.HSA.SSE.B.4	Math	Seeing Structure in Expressions	Use finite geometric series formula
CCSS.MATH.CONTENT.HSF.BF.A.1	Math	Building Functions	Write functions from relationships
CCSS.MATH.CONTENT.HSF.BF.A.1.a	Math	Building Functions	Write explicit or recursive rules from context
CCSS.MATH.CONTENT.HSF.BF.A.1.b	Math	Building Functions	Combine functions to build new models
CCSS.MATH.CONTENT.HSF.BF.A.1.c	Math	Building Functions	Compose functions
CCSS.MATH.CONTENT.HSF.BF.A.2	Math	Building Functions	Write sequences recursively and explicitly
CCSS.MATH.CONTENT.HSF.BF.B.3	Math	Building Functions	Analyze graph transformations of functions
CCSS.MATH.CONTENT.HSF.BF.B.4	Math	Building Functions	Find inverse functions

Continued on next page

Table F.1 – continued from previous page

ID	Content Type	Category	Description
CCSS.MATH.CONTENT.HSF.BF.B.4.a	Math	Building Functions	Solve using inverses; write inverse expressions
CCSS.MATH.CONTENT.HSF.BF.B.4.b	Math	Building Functions	Verify inverses by composition
CCSS.MATH.CONTENT.HSF.BF.B.4.c	Math	Building Functions	Read inverse values from graphs/tables
CCSS.MATH.CONTENT.HSF.BF.B.4.d	Math	Building Functions	Restrict domain to make a function invertible
CCSS.MATH.CONTENT.HSF.BF.B.5	Math	Building Functions	Use log-exponent inverse relationship
CCSS.MATH.CONTENT.HSF.IF.A.1	Math	Interpreting Functions	Understand functions, domain, and range
CCSS.MATH.CONTENT.HSF.IF.A.2	Math	Interpreting Functions	Use function notation and evaluate
CCSS.MATH.CONTENT.HSF.IF.A.3	Math	Interpreting Functions	Recognize sequences as functions
CCSS.MATH.CONTENT.HSF.IF.B.4	Math	Interpreting Functions	Interpret and sketch graphs from contexts
CCSS.MATH.CONTENT.HSF.IF.B.5	Math	Interpreting Functions	Choose appropriate domains from context
CCSS.MATH.CONTENT.HSF.IF.B.6	Math	Interpreting Functions	Compute average rate of change
CCSS.MATH.CONTENT.HSF.IF.C.7	Math	Interpreting Functions	Graph functions from symbolic forms
CCSS.MATH.CONTENT.HSF.IF.C.7.a	Math	Interpreting Functions	Graph linear and quadratic functions
CCSS.MATH.CONTENT.HSF.IF.C.7.b	Math	Interpreting Functions	Graph root and piecewise functions
CCSS.MATH.CONTENT.HSF.IF.C.7.c	Math	Interpreting Functions	Graph polynomials; zeros and end behavior
CCSS.MATH.CONTENT.HSF.IF.C.7.d	Math	Interpreting Functions	Graph rational functions; zeros and asymptotes
CCSS.MATH.CONTENT.HSF.IF.C.7.e	Math	Interpreting Functions	Graph exponential/log and trig functions
CCSS.MATH.CONTENT.HSF.IF.C.8	Math	Interpreting Functions	Rewrite functions to reveal properties
CCSS.MATH.CONTENT.HSF.IF.C.8.a	Math	Interpreting Functions	Use factoring/completing square on quadratics
CCSS.MATH.CONTENT.HSF.IF.C.8.b	Math	Interpreting Functions	Interpret exponential expressions (growth/decay)
CCSS.MATH.CONTENT.HSF.IF.C.9	Math	Interpreting Functions	Compare function properties across representations
CCSS.MATH.CONTENT.HSF.LE.A.1	Math	Linear, Quadratic, & Exponential Models	Distinguish linear vs exponential models
CCSS.MATH.CONTENT.HSF.LE.A.1.a	Math	Linear, Quadratic, & Exponential Models	Prove linear vs exponential growth patterns
CCSS.MATH.CONTENT.HSF.LE.A.1.b	Math	Linear, Quadratic, & Exponential Models	Identify constant rate situations
CCSS.MATH.CONTENT.HSF.LE.A.1.c	Math	Linear, Quadratic, & Exponential Models	Identify constant percent-rate situations
CCSS.MATH.CONTENT.HSF.LE.A.2	Math	Linear, Quadratic, & Exponential Models	Construct linear/exponential functions from data
CCSS.MATH.CONTENT.HSF.LE.A.3	Math	Linear, Quadratic, & Exponential Models	Observe exponential outgrows polynomial growth
CCSS.MATH.CONTENT.HSF.LE.A.4	Math	Linear, Quadratic, & Exponential Models	Solve exponential equations using logarithms
CCSS.MATH.CONTENT.HSF.LE.B.5	Math	Linear, Quadratic, & Exponential Models	Interpret parameters in linear/exponential models
CCSS.MATH.CONTENT.HSF.TF.A.1	Math	Trigonometric Functions	Interpret radians as arc length
CCSS.MATH.CONTENT.HSF.TF.A.2	Math	Trigonometric Functions	Use unit circle to extend trig to real angles
CCSS.MATH.CONTENT.HSF.TF.A.3	Math	Trigonometric Functions	Use special triangles/unit circle for trig values
CCSS.MATH.CONTENT.HSF.TF.A.4	Math	Trigonometric Functions	Explain trig symmetry and periodicity
CCSS.MATH.CONTENT.HSF.TF.B.5	Math	Trigonometric Functions	Model periodic phenomena with trig functions
CCSS.MATH.CONTENT.HSF.TF.B.6	Math	Trigonometric Functions	Restrict trig domains to define inverses
CCSS.MATH.CONTENT.HSF.TF.B.7	Math	Trigonometric Functions	Solve trig equations with inverse functions
CCSS.MATH.CONTENT.HSF.TF.C.8	Math	Trigonometric Functions	Use the Pythagorean trig identity
CCSS.MATH.CONTENT.HSF.TF.C.9	Math	Trigonometric Functions	Use trig addition/subtraction formulas
CCSS.MATH.CONTENT.HSG.C.A.1	Math	Circles	Prove all circles are similar

Continued on next page

Table F.1 – continued from previous page

ID	Content Type	Category	Description
CCSS.MATH.CONTENT.HSG.C.A.2	Math	Circles	Analyze inscribed angle and chord relationships
CCSS.MATH.CONTENT.HSG.C.A.3	Math	Circles	Construct inscribed/circumscribed circles; prove cyclic angles
CCSS.MATH.CONTENT.HSG.C.A.4	Math	Circles	Construct a tangent from an external point
CCSS.MATH.CONTENT.HSG.C.B.5	Math	Circles	Derive arc length and sector area formulas
CCSS.MATH.CONTENT.HSG.CO.A.1	Math	Congruence	Use precise geometric definitions
CCSS.MATH.CONTENT.HSG.CO.A.2	Math	Congruence	Represent and compare geometric transformations
CCSS.MATH.CONTENT.HSG.CO.A.3	Math	Congruence	Describe rotational/reflection symmetries of figures
CCSS.MATH.CONTENT.HSG.CO.A.4	Math	Congruence	Define rotations, reflections, and translations
CCSS.MATH.CONTENT.HSG.CO.A.5	Math	Congruence	Perform transformations and specify sequences
CCSS.MATH.CONTENT.HSG.CO.B.6	Math	Congruence	Use rigid motions to decide congruence
CCSS.MATH.CONTENT.HSG.CO.B.7	Math	Congruence	Link triangle congruence to matching sides/angles
CCSS.MATH.CONTENT.HSG.CO.B.8	Math	Congruence	Justify ASA/SAS/SSS congruence criteria
CCSS.MATH.CONTENT.HSG.CO.C.10	Math	Congruence	Prove triangle theorems (angles, midsegments, medians)
CCSS.MATH.CONTENT.HSG.CO.C.11	Math	Congruence	Prove parallelogram theorems
CCSS.MATH.CONTENT.HSG.CO.C.9	Math	Congruence	Prove theorems about lines and angles
CCSS.MATH.CONTENT.HSG.CO.D.12	Math	Congruence	Perform formal geometric constructions
CCSS.MATH.CONTENT.HSG.CO.D.13	Math	Congruence	Construct regular figures inscribed in a circle
CCSS.MATH.CONTENT.HSG.GMD.A.1	Math	Geometric Measurement & Dimension	Derive circle/solid formulas informally
CCSS.MATH.CONTENT.HSG.GMD.A.2	Math	Geometric Measurement & Dimension	Use Cavalieri's principle for sphere volume
CCSS.MATH.CONTENT.HSG.GMD.A.3	Math	Geometric Measurement & Dimension	Apply volume formulas for solids
CCSS.MATH.CONTENT.HSG.GMD.B.4	Math	Geometric Measurement & Dimension	Identify cross-sections and solids of revolution
CCSS.MATH.CONTENT.HSG.GPE.A.1	Math	Expressing Geometric Properties with Equations	Derive circle equations; complete the square
CCSS.MATH.CONTENT.HSG.GPE.A.2	Math	Expressing Geometric Properties with Equations	Derive parabola equation from focus/directrix
CCSS.MATH.CONTENT.HSG.GPE.A.3	Math	Expressing Geometric Properties with Equations	Derive ellipse/hyperbola equations from foci
CCSS.MATH.CONTENT.HSG.GPE.B.4	Math	Expressing Geometric Properties with Equations	Prove geometry theorems using coordinates
CCSS.MATH.CONTENT.HSG.GPE.B.5	Math	Expressing Geometric Properties with Equations	Use slope to prove parallel/perpendicular lines
CCSS.MATH.CONTENT.HSG.GPE.B.6	Math	Expressing Geometric Properties with Equations	Find point dividing a segment in a ratio
CCSS.MATH.CONTENT.HSG.GPE.B.7	Math	Expressing Geometric Properties with Equations	Compute perimeter/area with coordinates
CCSS.MATH.CONTENT.HSG.MG.A.1	Math	Modeling with Geometry	Model real objects with geometric shapes
CCSS.MATH.CONTENT.HSG.MG.A.2	Math	Modeling with Geometry	Apply density concepts in modeling
CCSS.MATH.CONTENT.HSG.MG.A.3	Math	Modeling with Geometry	Solve design problems with geometry
CCSS.MATH.CONTENT.HSG.SRT.A.1	Math	Similarity, Right Triangles, & Trigonometry	Verify properties of dilations
CCSS.MATH.CONTENT.HSG.SRT.A.1.a	Math	Similarity, Right Triangles, & Trigonometry	Describe dilation effects on lines
CCSS.MATH.CONTENT.HSG.SRT.A.1.b	Math	Similarity, Right Triangles, & Trigonometry	Scale segments by a dilation factor
CCSS.MATH.CONTENT.HSG.SRT.A.2	Math	Similarity, Right Triangles, & Trigonometry	Decide similarity using transformations
CCSS.MATH.CONTENT.HSG.SRT.A.3	Math	Similarity, Right Triangles, & Trigonometry	Establish AA similarity criterion
CCSS.MATH.CONTENT.HSG.SRT.B.4	Math	Similarity, Right Triangles, & Trigonometry	Prove triangle theorems via similarity
CCSS.MATH.CONTENT.HSG.SRT.B.5	Math	Similarity, Right Triangles, & Trigonometry	Use triangle criteria to solve and prove

Continued on next page

Table F.1 – continued from previous page

ID	Content Type	Category	Description
CCSS.MATH.CONTENT.HSG.SRT.C.6	Math	Similarity, Right Triangles, & Trigonometry	Define trig ratios using similar right triangles
CCSS.MATH.CONTENT.HSG.SRT.C.7	Math	Similarity, Right Triangles, & Trigonometry	Relate sine and cosine of complementary angles
CCSS.MATH.CONTENT.HSG.SRT.C.8	Math	Similarity, Right Triangles, & Trigonometry	Solve right triangles using trig and Pythagorean
CCSS.MATH.CONTENT.HSG.SRT.D.10	Math	Similarity, Right Triangles, & Trigonometry	Prove Laws of Sines and Cosines
CCSS.MATH.CONTENT.HSG.SRT.D.11	Math	Similarity, Right Triangles, & Trigonometry	Apply Law of Sines/Cosines to solve triangles
CCSS.MATH.CONTENT.HSG.SRT.D.9	Math	Similarity, Right Triangles, & Trigonometry	Derive triangle area formula (1/2 ab sin C)
CCSS.MATH.CONTENT.HSN.CN.A.1	Math	The Complex Number System	Define i and complex numbers
CCSS.MATH.CONTENT.HSN.CN.A.2	Math	The Complex Number System	Add/subtract/multiply complex numbers
CCSS.MATH.CONTENT.HSN.CN.A.3	Math	The Complex Number System	Use conjugates for modulus and division
CCSS.MATH.CONTENT.HSN.CN.B.4	Math	The Complex Number System	Represent complex numbers in rectangular/polar form
CCSS.MATH.CONTENT.HSN.CN.B.5	Math	The Complex Number System	Use geometric representation of complex operations
CCSS.MATH.CONTENT.HSN.CN.B.6	Math	The Complex Number System	Compute distance and midpoint in complex plane
CCSS.MATH.CONTENT.HSN.CN.C.7	Math	The Complex Number System	Solve quadratics with complex solutions
CCSS.MATH.CONTENT.HSN.CN.C.8	Math	The Complex Number System	Extend polynomial identities to complex numbers
CCSS.MATH.CONTENT.HSN.CN.C.9	Math	The Complex Number System	State Fundamental Theorem of Algebra
CCSS.MATH.CONTENT.HSN.Q.A.1	Math	Quantities	Use units and scales in modeling
CCSS.MATH.CONTENT.HSN.Q.A.2	Math	Quantities	Define quantities for modeling
CCSS.MATH.CONTENT.HSN.Q.A.3	Math	Quantities	Choose appropriate precision
CCSS.MATH.CONTENT.HSN.RN.A.1	Math	The Real Number System	Define rational exponents and radicals
CCSS.MATH.CONTENT.HSN.RN.A.2	Math	The Real Number System	Rewrite radicals using exponent rules
CCSS.MATH.CONTENT.HSN.RN.B.3	Math	The Real Number System	Explain rational/irrational sums and products
CCSS.MATH.CONTENT.HSN.VM.A.1	Math	Vector & Matrix Quantities	Represent vector quantities (magnitude/direction)
CCSS.MATH.CONTENT.HSN.VM.A.2	Math	Vector & Matrix Quantities	Compute vector components from coordinates
CCSS.MATH.CONTENT.HSN.VM.A.3	Math	Vector & Matrix Quantities	Solve problems with vectors (e.g., velocity)
CCSS.MATH.CONTENT.HSN.VM.B.4	Math	Vector & Matrix Quantities	Add and subtract vectors
CCSS.MATH.CONTENT.HSN.VM.B.4.a	Math	Vector & Matrix Quantities	Add vectors using parallelogram rule
CCSS.MATH.CONTENT.HSN.VM.B.4.b	Math	Vector & Matrix Quantities	Find magnitude and direction of vector sums
CCSS.MATH.CONTENT.HSN.VM.B.4.c	Math	Vector & Matrix Quantities	Represent and compute vector subtraction
CCSS.MATH.CONTENT.HSN.VM.B.5	Math	Vector & Matrix Quantities	Multiply vectors by scalars
CCSS.MATH.CONTENT.HSN.VM.B.5.a	Math	Vector & Matrix Quantities	Represent scalar multiplication graphically
CCSS.MATH.CONTENT.HSN.VM.B.5.b	Math	Vector & Matrix Quantities	Compute magnitude/direction of scalar multiples
CCSS.MATH.CONTENT.HSN.VM.C.10	Math	Vector & Matrix Quantities	Use zero/identity; link determinant to inverse
CCSS.MATH.CONTENT.HSN.VM.C.11	Math	Vector & Matrix Quantities	Multiply matrices and vectors; view as transforms
CCSS.MATH.CONTENT.HSN.VM.C.12	Math	Vector & Matrix Quantities	Interpret 2x2 matrices as plane transforms
CCSS.MATH.CONTENT.HSN.VM.C.6	Math	Vector & Matrix Quantities	Use matrices to represent data
CCSS.MATH.CONTENT.HSN.VM.C.7	Math	Vector & Matrix Quantities	Scale matrices by scalars
CCSS.MATH.CONTENT.HSN.VM.C.8	Math	Vector & Matrix Quantities	Add/subtract/multiply matrices
CCSS.MATH.CONTENT.HSN.VM.C.9	Math	Vector & Matrix Quantities	Recognize non-commutativity of matrix multiplication
CCSS.MATH.CONTENT.HSS.CP.A.1	Math	Conditional Probability & the Rules of Probability	Describe events with set operations

Continued on next page

Table F.1 – continued from previous page

ID	Content Type	Category	Description
CCSS.MATH.CONTENT.HSS.CP.A.2	Math	Conditional Probability & the Rules of Probability	Determine independence via product rule
CCSS.MATH.CONTENT.HSS.CP.A.3	Math	Conditional Probability & the Rules of Probability	Interpret conditional probability and independence
CCSS.MATH.CONTENT.HSS.CP.A.4	Math	Conditional Probability & the Rules of Probability	Use two-way tables for conditional probability
CCSS.MATH.CONTENT.HSS.CP.A.5	Math	Conditional Probability & the Rules of Probability	Explain conditional probability in context
CCSS.MATH.CONTENT.HSS.CP.B.6	Math	Conditional Probability & the Rules of Probability	Compute conditional probability
CCSS.MATH.CONTENT.HSS.CP.B.7	Math	Conditional Probability & the Rules of Probability	Apply addition rule for probabilities
CCSS.MATH.CONTENT.HSS.CP.B.8	Math	Conditional Probability & the Rules of Probability	Apply general multiplication rule
CCSS.MATH.CONTENT.HSS.CP.B.9	Math	Conditional Probability & the Rules of Probability	Use permutations/combinations for probability
CCSS.MATH.CONTENT.HSS.IC.A.1	Math	Making Inferences & Justifying Conclusions	View statistics as inference from samples
CCSS.MATH.CONTENT.HSS.IC.A.2	Math	Making Inferences & Justifying Conclusions	Assess models using simulation
CCSS.MATH.CONTENT.HSS.IC.B.3	Math	Making Inferences & Justifying Conclusions	Differentiate surveys, experiments, and observational studies
CCSS.MATH.CONTENT.HSS.IC.B.4	Math	Making Inferences & Justifying Conclusions	Estimate mean/proportion; develop margin of error
CCSS.MATH.CONTENT.HSS.IC.B.5	Math	Making Inferences & Justifying Conclusions	Compare treatments using experiments
CCSS.MATH.CONTENT.HSS.IC.B.6	Math	Making Inferences & Justifying Conclusions	Evaluate statistical reports
CCSS.MATH.CONTENT.HSS.ID.A.1	Math	Interpreting Categorical & Quantitative Data	Represent data with dot/hist/box plots
CCSS.MATH.CONTENT.HSS.ID.A.2	Math	Interpreting Categorical & Quantitative Data	Compare distributions by center and spread
CCSS.MATH.CONTENT.HSS.ID.A.3	Math	Interpreting Categorical & Quantitative Data	Interpret differences including outliers
CCSS.MATH.CONTENT.HSS.ID.A.4	Math	Interpreting Categorical & Quantitative Data	Use normal model with mean and SD
CCSS.MATH.CONTENT.HSS.ID.B.5	Math	Interpreting Categorical & Quantitative Data	Summarize categorical data in two-way tables
CCSS.MATH.CONTENT.HSS.ID.B.6	Math	Interpreting Categorical & Quantitative Data	Use scatter plots to describe relationships
CCSS.MATH.CONTENT.HSS.ID.B.6.a	Math	Interpreting Categorical & Quantitative Data	Fit models to data and use them
CCSS.MATH.CONTENT.HSS.ID.B.6.b	Math	Interpreting Categorical & Quantitative Data	Assess fit using residuals
CCSS.MATH.CONTENT.HSS.ID.B.6.c	Math	Interpreting Categorical & Quantitative Data	Fit a linear function to data
CCSS.MATH.CONTENT.HSS.ID.C.7	Math	Interpreting Categorical & Quantitative Data	Interpret slope and intercept in context
CCSS.MATH.CONTENT.HSS.ID.C.8	Math	Interpreting Categorical & Quantitative Data	Compute and interpret correlation coefficient
CCSS.MATH.CONTENT.HSS.ID.C.9	Math	Interpreting Categorical & Quantitative Data	Distinguish correlation from causation
CCSS.MATH.CONTENT.HSS.MD.A.1	Math	Measurement & Data	Define random variables and distributions
CCSS.MATH.CONTENT.HSS.MD.A.2	Math	Measurement & Data	Calculate expected value
CCSS.MATH.CONTENT.HSS.MD.A.3	Math	Measurement & Data	Develop theoretical distributions; find expected value
CCSS.MATH.CONTENT.HSS.MD.A.4	Math	Measurement & Data	Develop empirical distributions; find expected value
CCSS.MATH.CONTENT.HSS.MD.B.5	Math	Measurement & Data	Evaluate decisions using expected value
CCSS.MATH.CONTENT.HSS.MD.B.5.a	Math	Measurement & Data	Compute expected payoff in games
CCSS.MATH.CONTENT.HSS.MD.B.5.b	Math	Measurement & Data	Compare strategies using expected value
CCSS.MATH.CONTENT.HSS.MD.B.6	Math	Measurement & Data	Use probability to make fair decisions
CCSS.MATH.CONTENT.HSS.MD.B.7	Math	Measurement & Data	Analyze decisions with probability concepts

G Common Core State Standard Coding Protocol

Listing 3: Prompt for assigning CCSS code to digitized items

```
SYSTEM:
You are a meticulous data-coder. You follow classification protocols
   precisely and document your reasoning with traceable references.

TASK
------
Read these instructions carefully and follow the protocol exactly.

You are an education specialist tasked with evaluating standardized exam
   questions. Your goal is to classify each question according to the
   Common Core State Standards (CCSS) for either Mathematics or English
   Language Arts (ELA).

Each question includes:
- `grade`: target grade level
- `subject`: either "math" or "reading"
- `passage`: optional reading passage (can be "not-applicable")
- `question_stem`: the main question
- `answer_options`: list of 4 answer choices, labeled `a_id=1` to `a_id=4`

For each question:
1. Determine the most appropriate CCSS code(s) from the given CCSS coding
   instruction section.
2. Provide up to three (3) CCSS code(s). It is not necessary to fill all
   three if fewer are relevant.
3. Justify your selection using direct quotes from the coding section.
4. Assign a confidence score between 0 and 1 (rounded to two decimals) to
   every CCSS code provided.
5. Cite only the reference IDs you actually use.

PITFALL CHECKLIST
------
Before finalizing each classification, check for:
- Ambiguity in subject (e.g., reading comprehension vs. vocabulary)
- Grade mismatch (e.g., question complexity vs. stated grade)
- Multiple plausible standards (choose the most specific and relevant)
- Misleading distractors in answer options
- Missing passage context (if "not-applicable", adjust expectations)

OUTPUT
------
Return one UTF-8 encoded JSON object with two keys:

```json
{
  "data": [
    {
      "ccss_code": "CCSS.ELA-LITERACY.RL.3.1",
      "justification": "This question asks students to refer explicitly to
          the text to support their answer, which aligns with [1].",
      "confidence": 0.92
    },
    ...
  ],
  "references": [
```
```

```
    {
      "id": 1,
      "text": "Ask and answer questions to demonstrate understanding of a
         text, referring explicitly to the text as the basis for the
         answers.",
    },
    ...
  ]
}
```

```
## NOTES
- Use only standards from the instructions present in the coding section
- Do not hallucinate codes or guidelines.
- Justifications must be concise and reference actual text.
- Confidence should reflect your certainty in the match (e.g., 0.95 for
   exact match, 0.70 for partial alignment).
```

H Robustness to Ranked CCSS Skills

In this section, we present alternative ranking of skills using ω^u, the unconditional item weights, instead of ω, the conditional on items weights. We first present the whole set of weights by CCSS skills, and the rank correlation between skills between the two methods.

As can be seen, although the rank correlation differs somewhat, the overall Spearman correlation is 0.7, which is fairly strong. That is to say, under both methods, the ranking by CCSS skill is relatively similar. This is particularly true for the top 10 and bottom 10 CCSS skills.

Figure H.1: Rank Correlation by Skill (Cond vs Uncond Weights)

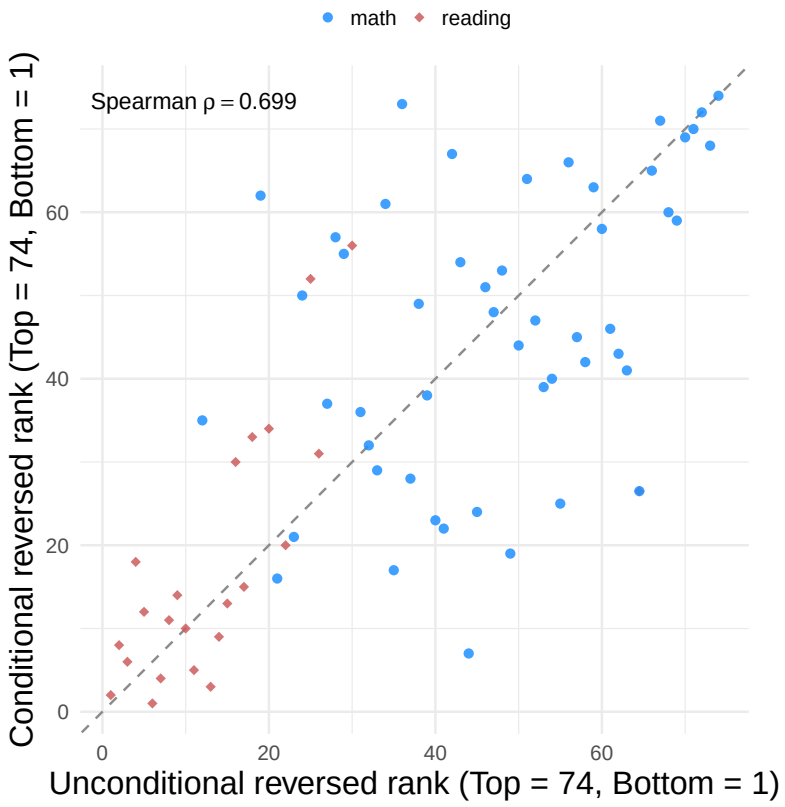
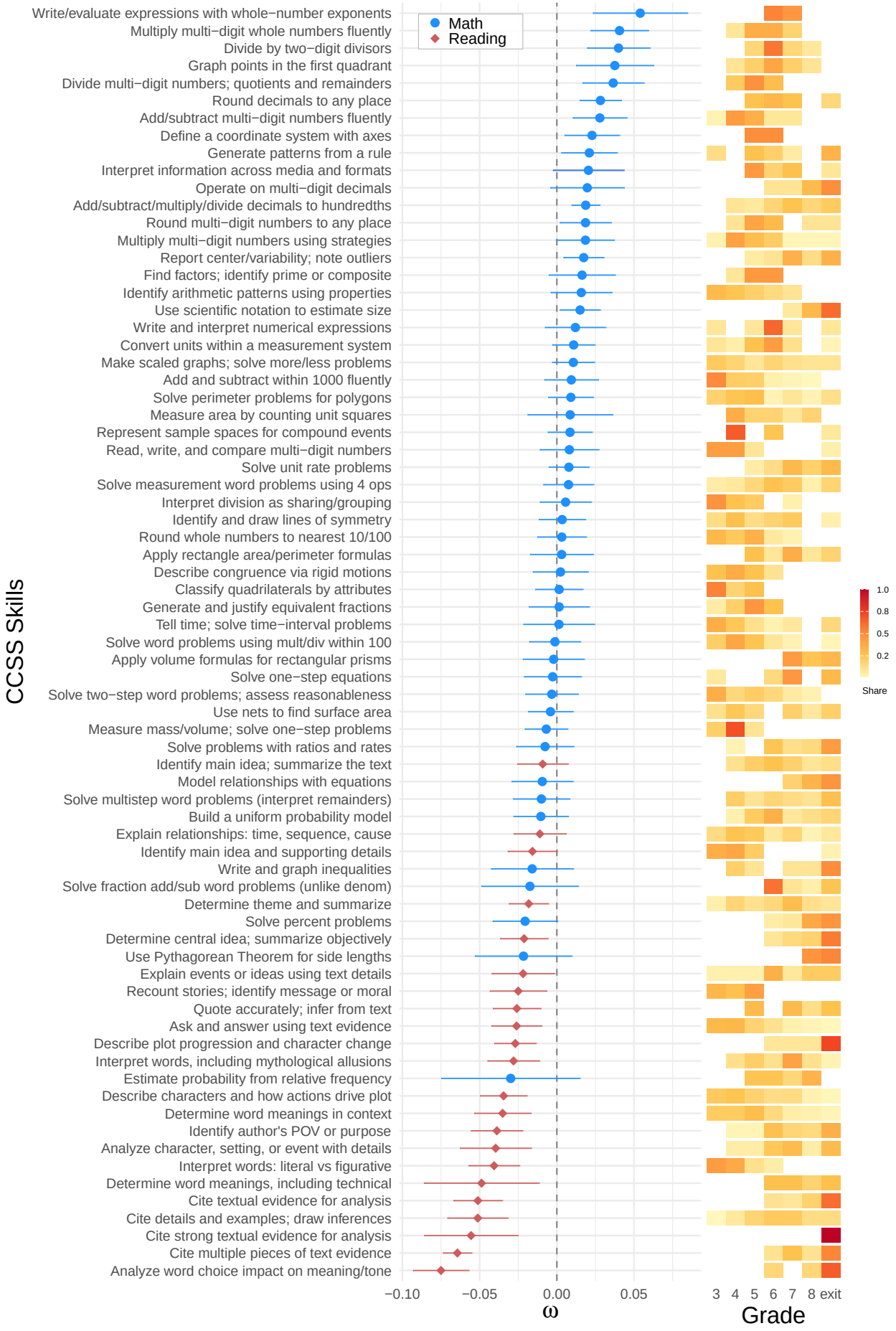


Figure H.2: Return by CCSS Skills



This figure plots coefficient estimates from a regression of unconditional item-level returns (ω^u) on Common Core skill indicators, controlling for the question's position within the test, discrimination and difficulty parameters. To account for estimation precision, observations are weighted by the inverse of the variance of the returns.

I 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 **?**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 $C \equiv I - \frac{1}{N} 1 1'$.[57]

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

$$V(\widehat{\mathbf{\Omega}}) = \frac{1}{N} \widehat{\mathbf{\Omega}}' C \widehat{\mathbf{\Omega}} \tag{15}$$

Now, let's consider the expected value of 15 conditional on the unobserved $\mathbf{\Omega}$: [58]

$$\begin{aligned} E[V(\widehat{\mathbf{\Omega}})|\mathbf{\Omega}] &= E\left[\frac{1}{N}\widehat{\mathbf{\Omega}}' C \widehat{\mathbf{\Omega}} | \mathbf{\Omega}\right] \\ &= \frac{1}{N}\left(\mathbf{\Omega}' C \mathbf{\Omega} + tr(C\Sigma)\right) \\ &= \underbrace{\frac{1}{N}\mathbf{\Omega}' C \mathbf{\Omega}}_{V_{\text{signal}}} + \underbrace{\frac{1}{N} tr(C\Sigma)}_{V_{\text{exp.noise}}} \qquad (16) \\ &= \underbrace{\frac{1}{N}\mathbf{\Omega}' C \mathbf{\Omega}}_{V_{\text{signal}}} + \underbrace{\frac{1_b}{N}\sum_{b=1}^{B}\left(tr(\Sigma_b) - \frac{1}{N} 1_b' \Sigma_b 1_b\right)}_{V_{\text{exp.noise}}} \qquad (17) \end{aligned}$$

That is, $V_{\text{exp.noise}}$ is the bias from estimating the variance of the expected observed coefficients

[57] 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.

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

($\widehat{\mathbf{\Omega}}$) 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:

$$R^2(h) = 1 - \frac{(\widehat{\mathbf{\Omega}} - h)'(\widehat{\mathbf{\Omega}} - h)}{\widehat{\mathbf{\Omega}}'C\widehat{\mathbf{\Omega}}} \tag{18}$$

Consider the expected value of Equation 18. Then we have

$$\begin{aligned} E[R^2(h)|\mathbf{\Omega}] &= 1 - E\left[\frac{(\widehat{\mathbf{\Omega}} - h)'(\widehat{\mathbf{\Omega}} - h)}{\widehat{\mathbf{\Omega}}'C\widehat{\mathbf{\Omega}}}|\mathbf{\Omega}\right] \\ &\approx 1 - \frac{E[(\widehat{\mathbf{\Omega}} - h)'(\widehat{\mathbf{\Omega}} - h)|\mathbf{\Omega}]}{E[\widehat{\mathbf{\Omega}}'C\widehat{\mathbf{\Omega}}|\mathbf{\Omega}]} \qquad (19) \\ &= 1 - \frac{(\mathbf{\Omega} - h)'(\mathbf{\Omega} - h) + tr(\mathbf{\Sigma})}{\mathbf{\Omega}'C\mathbf{\Omega} + tr(C\Sigma)} \qquad (20) \end{aligned}$$

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:

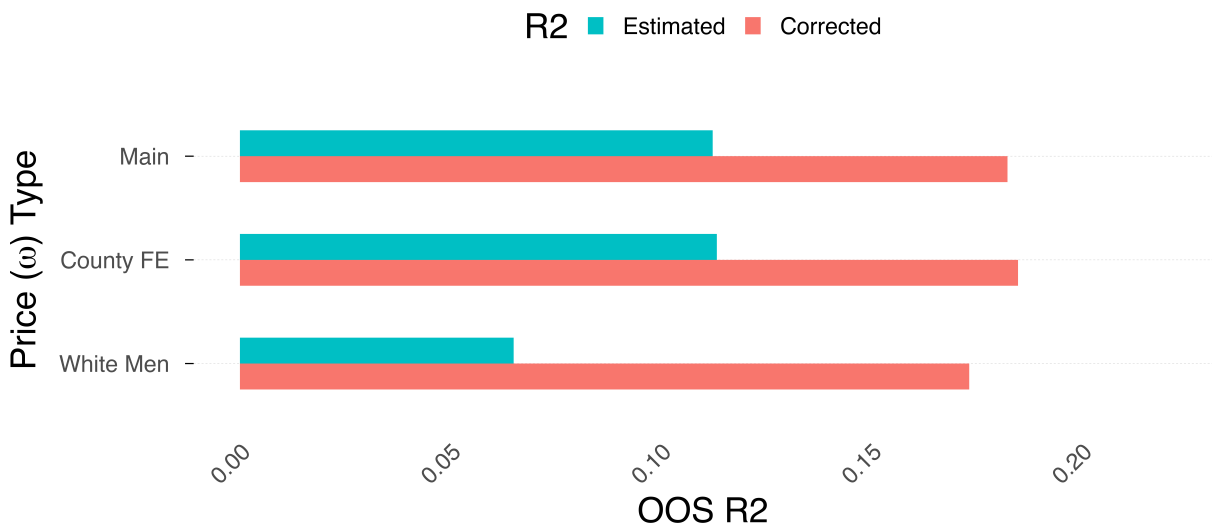
$$\widehat{V}_{\text{signal}} = \widehat{\mathbf{\Omega}}'C\widehat{\mathbf{\Omega}} - tr(C\widehat{\Sigma})$$

Finally,

$$R^2_{\text{corrected}} = 1 - \frac{\widehat{\text{SSE}}_{\text{true}}}{\widehat{V}_{\text{signal}}} = 1 - \frac{(\widehat{\mathbf{\Omega}} - h)'(\widehat{\mathbf{\Omega}} - h) - tr(\widehat{\Sigma})}{\widehat{\mathbf{\Omega}}'C\widehat{\mathbf{\Omega}} - tr(C\widehat{\Sigma})}. \tag{21}$$

We report $R^2_{\text{corrected}}$ in Section 7.3. Figure 13 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 I.1: R2 Correction by ω



Notes: .

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

J.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.

J.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.