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

Jonathan Moreno-Medina ⓡ Eric R. Nielsen ⓡ Viviana Rodriguez

UT San Antonio Federal Reserve Board UT San Antonio

June 2026

Abstract

Standardized test scores aggregate item (question) responses into a single scalar, collapsing distinct skills into an undifferentiated measure of proficiency. Which of these component skills is most predictive of long-run economic outcomes is a question that aggregate scores cannot answer. We develop a framework that looks both *inside the score*—re-weighting items by their predictive power for a chosen outcome ("item-level prices")—and *inside the item*—using the digitized text of each question to identify what skills drive the variation in these prices. We apply this framework to over 3,500 items linked to approximately 1 billion student-by-item-response records and adult earnings from Texas administrative data. Achievement scales that weight items by their estimated economic prices yield white–minority gaps that are 31%–58% larger than conventional scales and substantially reorder individual student rankings. To interpret these prices, we show that item text carries economically relevant information beyond standard psychometric characteristics, and we develop a novel text-based mapping of items to the over 600 skills comprising the Common Core State Standards. The mapping reveals that math skills involving procedural fluency, spatial reasoning, and codifiable, rule-based tasks have the highest estimated prices, while basic reading comprehension dominates more fine-grained reading skills. To our knowledge, this provides the first granular evidence on which specific K–12 curricular standards are most predictive of long-run labor-market outcomes.

Keywords: large language models, machine learning, achievement, measurement, human capital, inequality
JEL Codes: I26, J24, C53

*We would like to thank Samson Alva, Peter Arcidiacono, Elliot Ash, Bocar Ba, Pat Bayer, Stephen Billings, Aimee Chin, Michael Chrzan, Ben Domingue, David Figlio, Josh Gilbert, Havisha Khurana, David Liebowitz, Elaine M. Liu, Nolan Pope, Anthony Rios, Ed Rubin, Yona Rubinstein, Yotam Shem-Tov, David Slichter, and Sandy Student for helpful comments and suggestions. Similarly for the participants of the seminars at 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, University of Wisconsin-Madison, Educational Measurement Workshop at Brown University, Association of Education Policy and Finance, the Federal Reserve Board, Association for Public Policy Analysis and Management, and the Monash-Paris-Warwick-Zurich Text-as-Data Workshop. All remaining errors are our own. Jaidheer Sirigineedi, Margot Duque, Hannah Landel, Brendan Elliott, 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. The author order has been randomized and recorded on the AEA Author Randomization Tool, with confirmation "AnvneGJTJaQ4." The randomization of order is indicated by the symbol ⓡ. Contact email: Moreno-Medina: jonmorenomedina@gmail.com; Nielsen: Eric.R.Nielsen@frb.gov; Rodriguez: viviana.rodriguez@utsa.edu.

1 Introduction

Standardized test scores are constructed to measure academic proficiency—a goal that shapes the choices psychometricians make about how to aggregate individual item (question) responses into a single scalar. Researchers across the social sciences then commonly re-purpose these scores for their own, often very different, objectives: estimating returns to education, understanding the determinants of intergenerational mobility, among others. Such uses come with two limitations. First, the aggregation embedded in the score was not designed to measure the other latent constructs of interest, e.g., human capital. In psychometric terms, test scores can be valid measures of academic proficiency and yet may lack validity as measures of economically valuable skills.[1] Second, the aggregation makes it impossible to distinguish component skills whose relevance for social and economic outcomes may differ sharply. A student's math score, for instance, collapses spatial reasoning, algebraic manipulation, and data interpretation into a single number, implicitly weighting each as it pertains to the original psychometric goal of the scale. Which of these skills matter most for long-run outcomes is a question that aggregate test scores, by construction, cannot answer.

This paper argues that two related but distinct steps are needed to bridge the gap between psychometric test-score construction and economic skill measurement. The first is to look *inside the score*: because different test items predict economic outcomes to very different degrees, reweighting items by their predictive power for a chosen outcome yields achievement measures that can differ markedly from conventional scores. The second is to look *inside the item*: if we observe the content of each question – its text, structure, and cognitive demands – we can move from knowing *which questions* predict outcomes to identifying *which skills* do. We show that both steps are essential, and that together they open a path to credibly re-purpose test score data for skill measurement relevant for long-run economic outcomes.

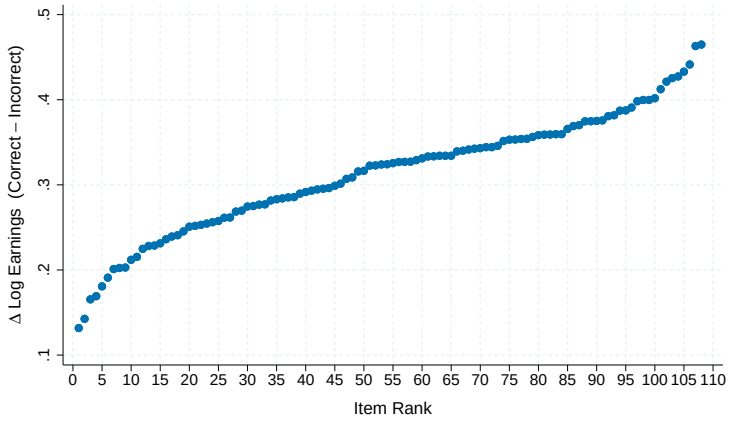
Figure 1 illustrates the kind of variation that standard, psychometrically-derived aggregations of item responses conceal. Each point plots the raw log-earnings difference, roughly one decade after testing, between students who answered a given item correctly and those who did not for the universe of Texas public-school eighth graders in 1996. Some items are associated with earnings differences of roughly 10%, while others exceed 50%. This wide variation suggests that item responses contain potentially valuable information for understanding human capital. However, to make full use of these data we must move beyond documenting *which questions* predict outcomes to identifying *what skills* predict outcomes. That is the central question of this paper.

We develop a two-stage framework to tackle this question. In the first stage, adapting and extending Nielsen (2019, accepted), we estimate the "price" of each test item, defined as the expected difference in an anchor outcome (e.g., log earnings) between answering the item correctly and incorrectly, conditional on the other item responses. Although these prices need not equal causal

[1] As defined in the *Standards for Educational and Psychological Testing*, the gold standard guide to testing,"[v]alidity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests" (American Educational Research Association, 2014, p. 11).

returns to skill, we nonetheless argue that these prices likely reflect the causal return to skills more so than the presence of confounders.[2] The item prices naturally generate an *item-anchored* achievement scale consisting of the price-weighted sum of the item responses. This scale is in interpretable "outcome" units and reflects the item-outcome relationships conventional psychometric scales ignore. In the second stage, we use the *content* of the items to explain why some have high prices and others do not. We do this through two complementary approaches. First, we convert each digitized item into a numerical representation using state-of-the-art text embeddings, and we estimate a flexible machine-learning model mapping these representations to item prices. We show that the language of a question contains economically relevant information above and beyond standard psychometric characteristics. Second, to uncover the information in the text, we develop a novel text-based mapping from items to the detailed skill taxonomy of the Common Core State Standards (CCSS), producing an estimated "price," analogous to the item prices, for each of the over 600 skills (standards) in the CCSS. Throughout, we apply the framework with wages as the anchor outcome, noting that the same approach can be applied to other outcomes such as college attendance or high-school graduation.

Figure 1: Item-level Earnings Differences for 8th Graders in 1996



Notes: This figure plots for each item (test question) administered to 8th graders in 1996, the difference in log earnings at age 25 between students who answered each item correctly versus those who did not. Items are ordered by 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 earnings differences.

Implementing this framework requires data that, to our knowledge, have not previously been assembled. We collect and digitize over 3,500 standardized-test items from scanned test booklets administered across roughly 12 million student-grade records of Texas public-school students in grades 3-8 and an exit exam administered in grade 10 or 11. We link these digitized items to approximately 1 billion student item-response records, which are in turn linked to adult earnings via state unemployment-insurance records. Finally, we map each item to the CCSS (over 600 standards in reading and mathematics) using a novel algorithmic procedure based on the semantic

[2]See Section 4, particularly 4.3, for details of this argument.

similarity of item and CCSS standard texts.[3] To our knowledge, this is the first general, replicable method for mapping any test item to any skill taxonomy based on the textual content of the items and taxonomy entries.

We report three sets of results. First, we show that achievement scales constructed from estimated item prices substantially alter two objects that are central to applied work: achievement gaps across racial and ethnic groups, and achievement ranks across individuals. In relative terms, white–Hispanic and white–Black gaps measured on the item-anchored scale are 31% and 58% larger, respectively, than those measured on the conventional psychometric scale.[4] These differences arise because Black and Hispanic students perform disproportionately worse on items that strongly predict earnings yet receive relatively little implicit weight under the conventional scale. The item-anchored scores also sharply reorder individual student rankings: within a given conventional-score ventile, the 90% range of item-anchored scores often exceeds two standard deviations. These findings on gaps and ranks echo Nielsen (2019, accepted) in a new context with substantially more data and policy relevance.[5]

Second, we find that the text of the item carries economically relevant information beyond standard psychometric and test-metadata characteristics. Using state-of-the-art sentence encoders, we embed each digitized item and estimate a flexible neural-network mapping from embeddings, psychometric characteristics (e.g., difficulty and discrimination), and test metadata (e.g., subject and item placement in the test) to item prices. Notably, while more difficult and more discriminating items do have higher prices on average, these psychometric characteristics explain only a modest share of the observed item-price variation. Including text embeddings raises the out-of-sample R^2 by 20–50% relative to models that use only metadata and psychometric characteristics, confirming that the language of a question encodes novel information about its power to predict economic outcomes.

Third, to unpack *what* in an item's language explains its predictiveness, we develop a standards-based approach that maps each item to a detailed set of CCSS standards based on the semantic similarity of their texts.[6] Because any given item may draw on multiple skills, we define the price of each CCSS standard as a weighted average of the item prices mapped to it, where the weights follow a softmax (multinomial logit) kernel over the text similarities, so that each standard's estimated price is concentrated on the items most closely aligned with it in the embedding space. We find a striking concentration of high-price skills in mathematics relative to reading, and, within math, a

[3] Note that the tests we assess pre-date the CCSS, which was also never adopted by Texas. We thus use the CCSS as an ex post interpretive taxonomy, not as the standards these tests were designed to adhere to.

[4] We correct for attenuation due to measurement error using a split-half IV reliability adjustment (odd/even items) as in Nielsen (2019, accepted). See Section 5.

[5] Nielsen (2019, accepted) uses the National Longitudinal Survey of Youth 1979, a panel survey of roughly 12,000 youth who were teenagers in 1980. The test items come from the Armed Forces Qualifying Test, administered to each respondent only once.

[6] While the CCSS developed clear and consistent grade-level standards in English Language Arts (ELA) and Mathematics, they were written as outcome expectations rather than a prescribed curriculum, which makes them useful as a common language for describing assessed skills. Thus, throughout the paper we refer to individual CCSS standards as *skills*.

clear tilt toward procedural, multi-step, rule-based computation (e.g., formula application, ordering numbers), rather than more conceptual or interpretive tasks. We also find that spatial reasoning tasks have notably higher prices than non-spatial tasks in mathematics.

While the majority of high-price standards are concentrated in math, we find that basic reading comprehension ranks higher than half of all math standards. This is in contrast with the lowest-ranked reading skills, which focus on analyzing tone, citing textual evidence, or determining word meanings. These results provide the first granular, item-level, standards-based account of which K–12 curricular skills are most predictive of long-run labor-market outcomes.

We assess the sensitivity of both stages of the analysis to specification and modeling choices. For the first-stage item prices, we show that estimates are stable across alternative geographic controls (county versus commuting-zone or school fixed effects), demographic samples (restricting to white males), and estimation methods (ridge and LASSO), with pairwise tests of equality between baseline and alternative item-price estimates rejecting for fewer than 2% of items at the 5% level. Because our preferred anchor model specification is linear in items, we also show through a series of Monte Carlo experiments that the linear-in-items specification recovers item prices with low bias even when the true data-generating process contains higher-order item interactions. For the second-stage skill mapping, the CCSS prices are nearly identical whether the individual CCSS skills are extracted from the full CCSS taxonomy by an open-source model or a closed-source model, with correlations exceeding 0.9. Moreover, the CCSS prices are stable across alternative weighting kernels for the item-to-skill similarity/distance, and re-estimating the full pipeline on the white-male subsample yields a CCSS-level rank correlation of 0.88 with qualitatively identical patterns across skill dimensions.

Prior Literature

Our results connect to several literatures in economics, psychometrics, and education policy. We organize the discussion around the paper's three main contributions.

We develop a general, replicable method for mapping any test item to any skill taxonomy based on the semantic similarity between the texts of the items and the skills, and we apply this method to map Texas items to the Common Core State Standards. This produces, to our knowledge, the first evidence on which curricular skills measured at the granularity of the CCSS predict long-run earnings. This contribution speaks directly to the large literature on the role of skills and human capital in determining earnings and other outcomes.[7] That literature has generally treated "achievement" either as a unidimensional or as a coarsely partitioned object (e.g., cognitive vs. non-cognitive, math vs. reading, etc.). By contrast, our framework, which assumes that different outcomes correspond *by construction* to different achievement scales, echoes Deming (2023), Dem-

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

ing and Silliman (2025), and Nielsen (2026, 2025b, 2023b) in arguing that human capital should be understood as multidimensional. Although multidimensional item response theory (MIRT) shares the premise that achievement is multidimensional (Reckase, 1985, 2009), it differs fundamentally in how the dimensions are defined: it recovers the latent vector of skills from the covariance matrix of the item responses rather than from the relationships between the item responses and different economic outcomes.[8] Our CCSS-based mapping also speaks to the extensive debate surrounding curriculum construction and standards-based reform.[9] By revealing that the CCSS standards have widely varying earnings prices, our paper highlights the context-dependent nature of curricular choices and suggests that a reorientation toward high-price standards may be beneficial. This analysis also builds on prior work demonstrating that different school subjects and different subsections of standardized achievement tests may have varying associations with outcomes (Cawley et al., 1999, Rose and Betts, 2004, Bettinger et al., 2013). In contemporaneous work, Conrad et al. (2026) finds that math scores predict early-career earnings much more strongly than ELA scores, with especially large gradients for middle-school assessments. Although our setting is different and our data considerably more granular, our results likewise point to a larger role for math, particularly assessed around grade 8, in predicting earnings. Finally, our classification of high- and low-price skills draws on prior work on the cognitive "depth of knowledge" of academic tasks (Webb, 1997, 2007) and on routine/nonroutine tasks (Simon, 1960, Autor et al., 2003, Autor, 2013, Spitz-Oener, 2006, Polanyi, 1966).

Our second contribution is to bring AI and machine learning tools to the economics of achievement measurement by digitizing, embedding, and analyzing item content at scale, showing that item text contains information about economic value that standard psychometric parameters and test metadata do not capture. This contribution connects to the growing literature within psychometrics showing that such tools can be used to predict item difficulty. In recent work, Kapoor et al. (2025) shows that LLM embeddings of item text can accurately predict item difficulty, providing a proof of concept for the approach we take here to relate embeddings to item-level economic prices. Benedetto et al. (2023) and Al-Khuzaey et al. (2024) survey recent efforts at text-based estimation of question difficulty. Relative to this new literature within psychometrics, our approach differs both in objective — predicting/explaining economic prices rather than psychometric difficulty — and in methodological details. Moreover, the use of text embeddings to generate a map between a curricular skill taxonomy and items is, to our knowledge, novel.

We also contribute to a growing literature within both psychometrics and economics that seeks to leverage item-response data in novel ways. In psychometrics, Gilbert et al. (2025), Ahmed et al.

[8] As with single-dimension IRT, MIRT factor scales are not cardinal and moreover are identified only up to rotation. Anchoring such factors to an outcome (Cunha et al., 2010) can yield cardinal achievement measures. We instead anchor items to the outcome directly, avoiding any commitment to a particular latent-factor structure.

[9] 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), Betts and Grogger (2003). This prior literature has mostly studied the effects of exposure to CCSS or some other curricular standard on either earnings or some measure of academic achievement. The extent to which individual CCSS standards predict earnings is not addressed.

(2025), and Gilbert et al. (2024) show that item-level heterogeneous treatment effects are pervasive: interventions often differentially impact specific items or subdomains rather than affecting the latent construct uniformly. Gilbert et al. (2025) proposes using item-level covariates to explain residual variation in these effects. In economics, recent work has used item-level data to estimate returns to cognitive endurance (Reyes, 2023), identify coachable items (Reyes et al., 2024), and study the effects of differential representation on test performance (Lee and Schaelling, 2025). Relative to these literatures, our paper differs in its focus on explaining the *economic* importance of items, its more extensive use of LLMs and machine learning for digitizing and analyzing item content, and its focus on achievement measurement rather than treatment-effect heterogeneity.[10]

Third, we confirm (with 12 million students and 3,500 items) that item-anchored achievement scales meaningfully alter estimated racial achievement gaps and individual student rankings, extending prior work to a large-scale administrative setting. This contribution builds most directly on Nielsen (2019, accepted), the first paper to argue that psychometric aggregation of items discards economically valuable information and that outcome-anchored aggregation can paint a very different picture of both individual and group-level achievement.[11] Bruhn et al. (2025) likewise document information loss from item aggregation using data from more recent cohorts of Texas students and focusing on shorter-run outcomes such as school discipline, course performance, and school completion. Bruhn et al. (2025) further show that item-level analyses significantly alter teacher value-added estimates and that, absent item content, item-specific teacher value-added patterns can link substantively similar items across years and cohorts.

The rest of the paper proceeds as follows. Section 2 introduces the item-anchored framework and defines both anchored achievement and item prices. Section 3 describes the student–item response and earnings data, as well as the collection and digitization of item texts. Section 4 formalizes the estimation of item prices. Section 5 revisits estimated achievement gaps and individual student ranks under item-anchoring versus conventional scales. Section 6 examines item prices through observable psychometric and test-metadata features. Section 7 relates item prices to text embeddings via flexible prediction models. Section 8 interprets prices via the item-to-CCSS skill mapping and presents the skill-price estimates. Section 9 concludes.

[10] Our paper also contributes to broader literatures on the measurement of achievement and the properties of standardized test scores. Past research has argued that psychometric scales lack a cardinal interpretation (Lord, 1975, Jacob and Rothstein, 2016, Cunha et al., 2021) and that standard results in economics can be sensitive to economically arbitrary scaling choices (Cunha et al., 2010, Bond and Lang, 2018, 2013, 2019, Schröder and Yitzhaki, 2017, Nielsen, 2025a, 2023a).

[11] One result in Nielsen (2019, accepted) is that item-predicted white–Black differences in labor-market outcomes equal observed gaps, though this does not hold for household income. A second result is that individual achievement ranks shift notably under 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 item-anchored achievement variability increases dramatically as children progress through school, with clear implications for analyses that standardize scores to unit variance at each grade.

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 a price 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 is concerned with understanding which item-level features are associated with low or high estimated prices. This section lays out the basic definitions and conceptual framework underlying our approach, which builds on the frameworks in Nielsen (2019, accepted) and Bond and Lang (2018).

We seek to define the notion of a "price" for each test item. 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 $\mathbf{D}_i = [D_{i,m}, \mathbf{D}_{i,-m}]$ where $\mathbf{D}_{i,-m}$ denotes the length $M-1$ vector of all i's item responses other than to item m.[12]

Implicit in 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 responses ($\mathbf{D}_i$) 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 correctly. 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.[13]

In this paper, we instead seek a definition of $\mathbf{\Omega}$ based on the economic usefulness of each item. Thus, we use "prices" to describe the elements of $\mathbf{\Omega}$. We adapt the item-anchoring approach in Nielsen (2019, accepted) which defines achievement A_i as the component of some outcome S_i that is predictable from ideal psychometric data:

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

Here, η_i represents determinants of the outcome S_i that are not predictable from psychometric data. Thus, $\mathbb{E}[\eta_i A_i] = 0$ holds by construction. In practice, no test has complete and perfect psychometric data, and so the best we can hope to identify given the observed item data is $\tilde{A}_i \equiv \mathbb{E}[S_i|\mathbf{D}_i, \mathbf{X}_i]$. Thus, $\tilde{A}_i$ is defined relative to the particular items used to assess achievement. $\tilde{A}_i$ will by necessity

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

[13] 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 for achievement.

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

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

Data on outcomes, item responses, 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 according to the structure imposed by f.[14] Section 4 presents the empirical details of our estimation approach.

Our definition of the item-price vector $\mathbf{\Omega}$ follows from our definition of item-anchored achievement. The price of item m, ω_m, is defined as the average change in $\tilde{A}_i$ from toggling item m from incorrect to correct, holding fixed the remaining item responses and covariates and averaging over their population distribution:

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

Equivalently, ω_m is the average difference in the expected anchor outcome S between students who answer item m correctly and those who do not, holding everything else fixed. Like conventional test scales, our approach assigns a weight to each item; unlike equal-weight or psychometric scales, however, it chooses these weights according to each item's predictive relevance for a long-run economic outcome of interest.

Importantly, ω_m is not a causal or structural parameter. It is a predictive price: any factor that is correlated with both the item response and the anchor outcome can load on ω_m. There are therefore several channels through which an item can acquire a high estimated price. First, the academic skill assessed by the item may have a direct labor-market return; many jobs require reading comprehension, numeracy, and other basic academic skills. Second, the skill assessed by item m may be an input into the future acquisition of higher-level skills that themselves have labor-market returns. Basic arithmetic, for example, may be valuable partly because it is a prerequisite for higher-level quantitative tasks which are themselves directly rewarded in the labor market. In a different context, Nielsen (2025b) finds empirical support for this type of cumulative skill-production framework. Third, the item might have a high price for reasons unrelated to the academic skill being assessed, either through noncognitive skills helpful for solving the item or through sociodemographic and other confounders.

These additional channels do not all have the same interpretation. In particular, noncognitive

[14] 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, accepted) for more details.

components of item performance are not necessarily a threat to our framework. Item responses will generally depend on skills such as conscientiousness, grit, etc., in addition to the particular academic skill being assessed. A correct response to an item thus indicates both that the student has the necessary academic skill *and* is willing and able to put in the care and effort needed to actually complete the item. In this case, ω_m should be interpreted not as the price of a narrowly defined academic skill, but as the price of the bundle of academic and nonacademic attributes that contribute to a correct response on item m.[15]

By contrast, other channels are more naturally viewed as confounds. An item might predict adult earnings not because it captures a valuable cognitive or noncognitive skill, but rather because it is correlated with family background or other non-skill determinants of later outcomes. For example, an item involving yachting terminology might predict earnings even if knowledge of yachting vocabulary has no independent value in the labor market. Our research design cannot definitively rule out such channels. However, in Section 3 and Section 4, we provide substantial evidence that the estimates of $\mathbf{\Omega}$ reflect economically relevant skills, broadly construed to include noncognitive components of item performance, more so than sociodemographic and other confounds.

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

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

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.

With $\widehat{\mathbf{\Omega}}$ in hand, the main part of our analysis seeks to understand how observable characteristics of the items explain (in the statistical sense) these item prices. In particular, we consider two conceptually distinct types of item-level data. First, we denote by $\mathbf{R}_m$ 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 stack of these characteristics across the items m. Second, we denote by $\mathbf{E}_m$ a mathematical

[15] $\mathbf{\Omega}$ can be interpreted also as defining an outcome-relevant direction in the M-dimensional item-response space. If item responses are driven by a latent vector of skills $\mathbf{L}$, our approach is to use the full vector of item responses for student i to proxy for $\mathbf{L}_i$. Prior literatures have generally employed factor model approaches to recover low-dimensional latent skills in similar contexts. In psychometrics, Item Response Theory (IRT) relies on assumed factor structure for skills (e.g., Jöreskog (1969), Reise (2012), Reckase (1985)). In economics, item-level response data remain relatively uncommon, and the dominant approach treats test scores and behavioral outcomes as noisy measures of latent 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. By contrast, our method does not require specifying the dimension of the latent skill space in advance, nor does it require a parametric measurement model linking latent skills to item responses.

representation of the item text $\mathbf{T}_m$, with $\mathbf{E}$ representing the matrix of these data across items.[16] Then, we suppose that for some function g and a residual component $\boldsymbol{\Xi}$,

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

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

3 Data

Constructing estimates of $\boldsymbol{\Omega}$ 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 covariates ($\mathbf{X}_i$). Furthermore, understanding what characterizes high-price 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 ingredients. 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-run economic outcomes, covering the universe of public school students in Texas from the 1995-96 to 2001-02 school years.[17]

However, 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, its subject (math or reading), and its position in the test. Thus, we supplement these data in two ways. First, we collect and digitize the test booklets used in statewide testing, which allows us to recover image and text information for each test item. Second, we use the content of the items to link each item to the skill taxonomy defined by the Common Core State Standards (CCSS) to obtain a richer picture of the skills assessed by each item.

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

[16] We use "$\mathbf{E}_m$" because our empirical work will use embedding space representations of the text. $\mathbf{E}_m$ will thus be a high-dimensional real vector.

[17] Throughout, we use the year of the spring term to refer to each school year.

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 (typically grade 10 or 11). The purpose of these assessments is to measure the level and growth of student proficiency and learning in Texas public schools. In this paper we link student item-response patterns to subsequent adult earnings. Thus, we focus our analysis on the Texas Assessment of Academic Skills (TAAS) (1990–2002), the earliest test available in the ERC data, which allows us to observe wages at age 25 for all test-takers while also preserving the same assessment design.[18]

A key feature of the state's data collection efforts for our purposes 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 selected, (2) whether the answer was correct, and (3) whether the student skipped the item. In addition to the student item responses, we also observe the test subject (math or reading) and learning objective associated with each question.[19]

Panel B of Table 1 presents item-level descriptive statistics for our analysis sample. Overall, the average item is answered correctly by 80% of students. Importantly, the percent correct variable ranges from 24% to 99% across our 4,739 items, with the 1st and 99th percentiles at 46% and 98%, respectively. Thus, the items retain meaningful information content – the are answered neither correctly nor incorrectly by all students. Our sample contains slightly more math than reading questions. The raw log earnings difference between a correct versus incorrect answer (as plotted in Figure 1) for the average item is 0.28 log points (approximately 32%), while our conditional estimate, $\hat{\omega}$, indicates a 0.01 log point (approximately 1%) average price.[20]

We are able to recover the text of 76% of the items administered between 1996–2002 (see Section 3.3.1 for more details). Thus, in columns (4) through (6) we present analogous statistics for the subsample with item text data available. The descriptive statistics for this subsample do not differ meaningfully from the full sample. In total, our analysis draws on more than 1.2 billion student-item responses, approximately 940 million of which we can link to recovered item text.

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 link item responses to earnings for any public school student who receives wage/salary income in Texas in adulthood.[21] 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.[22] As such, we

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

[19] The standardized tests we study were designed to measure broad learning objectives consistently over time. See Table A.3 for the list of TAAS testing objectives in reading and mathematics.

[20] See Section 2 and Section 4 for a discussion of the estimation of these prices.

[21] This excludes self-employed workers and independent contractors.

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

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						
Earnings at 25	$29,693	$29,010	9,111,275	$29,932	$29,452	6,996,713
Earnings at 30	$42,466	$44,479	6,717,698	$42,775	$44,295	4,731,630
Earnings 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	0.76	0.43	4,739	1.00	0.00	3,555
Raw Log Earnings Difference	0.28	0.07	4,739	0.28	0.07	3,555
$\hat{\omega}_m$	0.01	0.02	4,739	0.01	0.02	3,555
Discrimination IRT	1.49	0.44	4,739	1.48	0.44	3,555
Difficulty IRT	-1.31	0.87	4,739	-1.35	0.87	3,555
Student-Grade-Item Responses		1,235,875,456			937,619,520	

Notes: This table presents descriptive statistics of student-by-grade demographics and long-run outcomes in Panel A, and item-level characteristics in Panel B. All earnings data are converted to 2019 dollars using the U.S. Bureau of Labor Statistics Historical Consumer Price Index for all urban consumers (CPI-U). Columns (1)–(3) present statistics for the universe of students in Texas from 1996–2002 who took standardized tests (grades 3–8 and the exit exam). Columns (4)–(6) present analogous statistics for the grade-year combinations for which we recovered test booklets.

are able to recover labor market information for approximately 72% of the universe of test-takers.

We take as our baseline "anchor" outcome adult earnings at age 25.[23] Ideally we would be able to measure student earnings at later ages in order to more accurately capture individual variation in lifetime earnings.[24] To avoid pandemic-era labor-market distortions, we restrict to cohorts whose earnings are observed *pre–COVID*, which prevents us from observing most of the

[23]TWC data are reported quarterly. We aggregate quarterly earnings for all students to generate yearly earnings. To reduce the importance of measurement error and transitory earnings fluctuations, we measure earnings at age 25 by averaging the labor market earnings observed at ages 24, 25, and 26. We then convert earnings to 2019 dollars based on the CPI.

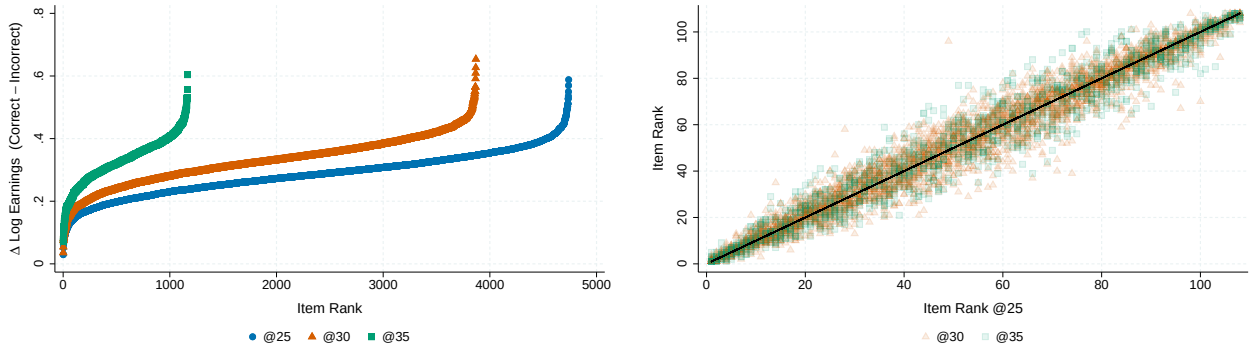
[24]Studies generally find that earnings around age 40 best proxy for lifetime earnings; see Haider and Solon (2006) and Mazumder (2005).

sample at later career ages. Accordingly, we use earnings at age 25, which we view as a reasonable early-career measure: by that age most individuals have completed schooling and have entered the labor market.[25]

Figure 2: Unconditional Item-Level Earnings Prices for All Grades and Years

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

(b) Rank Correlations Across Ages



Notes: Panel (a) of this figure plots for each item (test question) administered in any grade-year the difference in log earnings (at age 25, 30, and 35) between students who answered the item correctly and those who did not. Items are ordered by the value of this difference. We are missing earnings data for some grade-year combinations in which students have yet to reach 30 and 35 years of age by 2019. This explains differences in sample size between item-level prices at age 25, 30, and 35. Panel (b) shows the correlation between the item percentile ranks at age 25 and the corresponding percentile ranks at age 30 and 35.

To understand the implications of capturing earnings at different labor-market ages, Figure 2 plots unconditional item-level earnings differences at ages 25, 30, and 35 for all items administered to Texas public-school students during our period of study. Panel (a) of Figure 2 shows that: i) the range of prices is very similar for all ages; 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.[26] Panel (b) presents the correlation between the item price ranks across ages. This plot indicates a strong rank correlation of prices estimated at different labor market ages ($\rho = 0.95$). Thus, items that have high (low) prices for earnings at 25 continue to have high (low) prices at ages 30 and 35. Given these results, throughout the paper, our preferred specification uses earnings data at age 25.

3.3 Item Characteristics Data

Item-level information included in the Texas ERC data is limited to the item's broad learning objective, its subject (math or reading), and its position in the test. Thus, we supplement these data in three different ways. First, we collect and digitize documentation on the test booklets used for each grade-year in our sample to recover information on the actual content of each item. Second, we use the content of each item to map it to the taxonomy of skills specified by the Common

[25] See Appendix A.2 for additional discussion of our earnings definition and choice of earnings age.

[26] At age 25, we have a broader coverage of earnings data, with a match rate of 71.5%. If instead, we capture students at age 30, our average match rate falls to 52.9%. Finally, at age 35, we lose a substantial amount of data. This is mostly driven by the fact that most students in our testing sample had not yet turned 35 by 2019.

Core State Standards (CCSS). Finally, we recover psychometric parameters (e.g., item difficulty and discrimination) for each question via the estimation of a standard 3PL 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 online following each year's test administration cycle — which were subsequently removed from its website. Using the Internet Archive's Wayback Machine, we are able to recover 76% of the test booklets for the years 1996–2002.[27]

With the booklets in hand, we extract item-level information from the PDFs in a multi-step process. First, each assessment PDF is systematically split into its respective math and reading sections using a document segmentation approach. Following this, each section is segmented at the page level and converted to PNG format to ensure compatibility with downstream processing tasks. These images are then used as inputs for OpenAI's GPT-o3 model to perform Optical Character Recognition (OCR) and extract textual content. Through an API call with a predefined prompt, GPT-o3 generates structured JSON outputs containing two distinct data types: items and passages (for reading).[28]

We then evaluate the quality of the digitization separately for items without images (86.5% of all items) and items with images (13.5% of all items). We found very high accuracy for the imageless group (97% accuracy), but we found lower accuracy for items with images. To address this, two research assistants manually checked and corrected all items with images. This procedure yields a digitization accuracy rate of 97.4%.[29] 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 the Online Appendix.

3.3.2 Common Core State Standards Initiative

We use the digitized content of items 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

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

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

[29] Accuracy rates for items without images are estimated from a review of a random sample of 100 questions out of 3,077 total such items. Accuracy rates for items with images are estimated from the full set of questions with images in our sample. See Figure A.2 for examples of question types and image positions.

educators in K–12 and higher education.[30]

Crucially 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 the skill demands of each item. 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. Our final dataset of digitized CCSS standards includes 613 standards for both reading and math across grades 3–12.[31]

3.3.3 Psychometric Properties of the Items

In addition to the text of the items, we also consider whether and how traditional psychometric item characteristics relate to $\widehat{\mathbf{\Omega}}$. Using the IRT routines built into Stata, we recover item-level estimates of difficulty and discrimination assuming a three-parameter logistic IRT model ("3PL model").[32] The average item on these TAAS exams is fairly easy, with most items having negative estimated difficulties.[33] This is also reflected in the correct response rates in Table 1, 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 students and are thus not targeted toward the top of the achievement distribution.[34]

4 Estimation of the 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 vector $\mathbf{\Omega}$. We

[30]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.

[31]The math standards can be found here. The reading-ELA standards can be found here. The Online Appendix presents the full list of standards, along with a short description of the skill used for exposition purposes. See NGA Center and CCSSO (2010) for the full standard/skill description.

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

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

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

then show the robustness of our estimates of $\mathbf{\Omega}$ to alternative model specifications.

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

We focus in this paper on log earnings of individual i at age 25 as our long-run outcome of interest (S_i), although the methodology could be straightforwardly applied to other outcomes. In our baseline approach, we suppose that log earnings 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

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

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

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

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

Price 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 $\mathbf{\Omega}$ within different demographic groups. Specifically, we consider alternative models, including a fully interacted model with race through sample splitting. For these alternative models, Table C.1 shows that we can rarely reject the null hypothesis that the individual item prices estimated controlling for demographics in different ways are equal. 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 Appendix 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.

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

In detail, the Monte Carlo experiments in Appendix C.1 generate item-response data in a realistic way by assuming a 3PL IRT model but allowing for outcomes to be determined by items and item interactions. These experiments reveal that the OLS estimates of $\mathbf{\Omega}$ are approximately unbiased across a wide range of data-generating processes (different IRT parameters, item prices, and

interaction specifications). Moreover, the OLS estimates have very similar RMSEs as lasso models that include the correct order of item interactions in the feature set (e.g., two-way or three-way), and generally outperform random forest models in terms of bias (though not always in terms of RMSE).

In interpreting the strong performance of linear OLS in these Monte Carlo experiments, recall from equation (3) that ω_m is the increment in the expected outcome from a correct response to item m averaging over the population distribution of responses to the other items. When item-item interactions are present, ω_m will not equal the coefficient on m in the correctly specified model. Instead, it will combine that "main" effect along with "secondary" effects operating through the other items interacted with it. Crucially, $\widehat{w}_m$, the OLS coefficient for item m in the linear model, will absorb the same secondary effects. Thus, although $\widehat{w}_m$ will be asymptotically biased for the main effect, this "bias" can push the estimate towards our target estimand ω_m rather than away from it. Indeed, in a simple case with two independent items, the misspecified linear coefficient identifies ω_m exactly.[35]

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

4.2 Controlling for Demographics

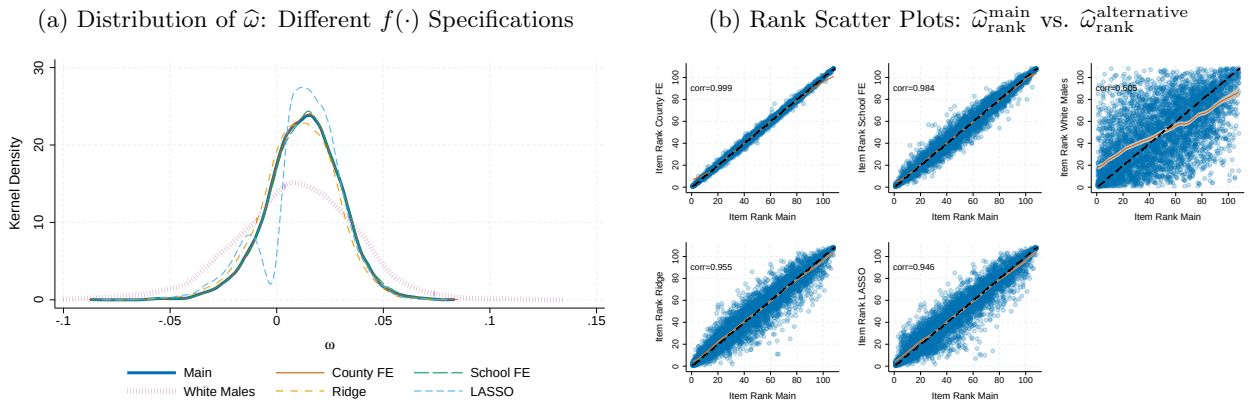
Our preferred interpretation of $\widehat{\mathbf{\Omega}}$ is that it is a vector of item prices that themselves reflect "skill prices" – the labor market value of the skills assessed by the items. We thus want to control for non-skill factors that affect the level and distribution of earnings. We also want to control for confounders – factors that are correlated with both 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).

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.[36] Commuting zone fixed effects allow us to compare students with similar labor market opportunities who have different

[35] Suppose the true model is $S = \psi_0 + \psi_1 D_1 + \psi_2 D_2 + \psi_{12} D_1 D_2 + \varepsilon$ with $D_1 \perp D_2$. If the researcher erroneously estimates a linear model in D_1 and D_2 alone, independence implies that the linear projection of the omitted regressor $D_1 D_2$ onto $(1, D_1, D_2)$ places coefficient $\mathbb{E}[D_2]$ on D_1. Thus, the omitted-variable formula gives plim $\widehat{w}_1 = \psi_1 + \psi_{12}\mathbb{E}[D_2]$. But this is exactly ω_1 under equation (3): the expected return to answering item 1 correctly, averaged over the distribution of D_2. This exact equivalence relies on the independence of the item responses. With many, correlated items, plim $\widehat{w}_m$ and ω_m will generally not coincide exactly, but the divergence will be small when pairwise correlations among item responses are modest and response rates are broadly similar across items—consistent with the Monte Carlo results in Appendix C.1.

[36] 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 3: Estimated Item Prices for Different Anchor Model Specifications



Notes: Panel (a) of this figure plots the distribution of $\{\widehat{\omega}_m\}$ for different anchor model specifications. Panel (b) shows the correlation between the grade-year item-price ranks of our main specification and the grade-year item-price ranks of various alternative specifications. A local polynomial fit is added in an orange solid line with 95% confidence intervals in gray. The black dashed line represents the 45-degree line. Item-price rank-rank correlations are reported in the top left corner.

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.

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

Before delving into our primary analysis of what item features explain $\widehat{\mathbf{\Omega}}$, we first explore the stability of the estimated item-price vectors across alternative estimation approaches. Panel (a) of Figure 3 shows the distribution across m of the estimated item prices for several different anchor model specifications. Our baseline regression estimates of the item weights are robust to differences in how we control for geography and race. In particular, the "County FE" estimates, $\{\omega_m^{\text{CF}}\}$, control for county rather than commuting zone fixed effects in order to consider a different, typically smaller, local labor market. The "School FE" estimates, $\{\omega_m^{\text{SF}}\}$, include school fixed effects instead to control for differences in school inputs, social networks, and any other school-level unobserved determinant of adult earnings. Finally, the "White Males" estimates, $\{\omega_m^{\text{WM}}\}$, are derived from the subsample of white males only — a population that is less likely to experience labor-market discrimination and which has high labor-force participation. The resulting $\widehat{\mathbf{\Omega}}$ therefore translates item responses into predicted outcomes as they are priced for white males only (Nielsen, 2019, accepted). All three specifications yield similar item-price distributions. Figure 3 also plots estimates obtained from ridge and lasso specifications with the relevant penalty parameter for each model selected via cross-validation. For both of these regularized models, the distribution of estimated item-level prices mirrors that of our main specification.

Because the item prices depicted in panel (a) of Figure 3 are estimated, some of the appar-

ent dispersion will be due to estimation error. We assess the stability of the estimates across specifications in two ways. First, panel (b) of Figure 3 plots the rank correlations between the item prices estimated under the main specification, $\{\hat{\omega}_m^b\}$, and under the alternative specifications, $\{\hat{\omega}_m^{b'}\}_{b' \in \{\text{CF,SF,WM}\}}$. Overall, these rank correlations are positive and large. However, the comparison with the white-male estimates is noticeably noisier. This is likely because we estimate these prices with only a quarter of the observations used in every other specification. In our second exercise, we run item-level statistical difference tests for the estimated prices across specifications. We find that for 98.5% of items m we cannot reject $H_0 : \omega_m^b = \omega_m^{\text{WM}}$ at 10%. The rejection rate is even lower comparing the county fixed effects estimates to our main specification – we cannot reject coefficient equality at 10% for about 99.8% of items. We likewise fail to reject equality for about 92.5% of items comparing our main specification to the school fixed effects specification. We present these results in Table C.1. Thus, the estimated item prices are quite similar to each other across specifications. Furthermore, a central exercise of this paper is to aggregate these item-level prices into prices for each CCSS standard, as we detail in Section 8. Previewing those results, the aggregation averages out item-level estimation noise, and the rank correlation between full-sample and white-male-only CCSS prices is 0.88, with substantively identical skill-dimension regression patterns (Section 8.7).

5 Item-Anchoring, Achievement Gaps, and Student Ranks

In the previous section, we showed that different plausible anchor models and methods yield similar item-price estimates. Moreover, we showed that the individual estimated item prices, $\hat{\omega}_m$, display significant variation: different items predict adult earnings 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 item-anchored scores rather than 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, accepted) to a new context.[37]

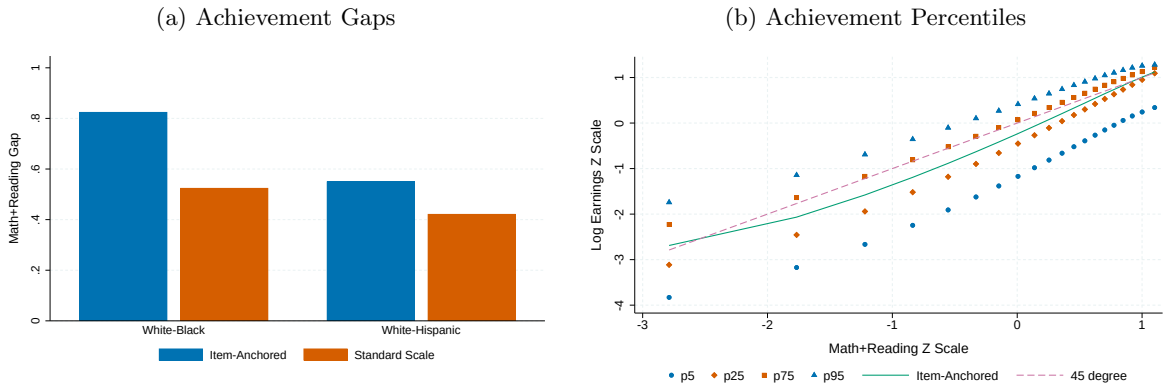
Finding 1: White-minority item-anchored gaps are larger than standard-scale gaps. Item-anchored achievement gaps in general will differ from gaps estimated using conventional psychometric scores. First, trivially, the units will differ – item-anchored gaps are in cardinally interpretable "outcome units," while psychometric scores will be in whatever scale the test designers construct (which may or may not be cardinal).[38] Second, and more fundamentally, a mean achieve-

[37] Nielsen (2019, accepted) 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, accepted), our analysis covers more recent cohorts over a wider range of grades and ages.

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

ment gap could differ across scales because students from one group may do particularly well or poorly on items that are emphasized differently by the item-anchored and conventional psychometric scales.

Figure 4: Item-Anchored Achievement Differences



Notes: Panel (a) of this figure shows white-Black and white-Hispanic achievement gaps measured using log-earnings item-anchored scores (blue bars), alongside analogous gaps using the conventional math+reading test scale (orange bars). Gaps for both scales are reported in standard deviation units. Panel (b) plots average given-scale math+reading score ventiles on the x-axis against the ventile-conditional means, 50% ranges, and 90% ranges of the log-earnings item-anchored scores on the y-axis.

Panel (a) of Figure 4 shows the average white-Black and white-Hispanic achievement gaps for all grades and years in our sample. Standard-scale gaps are presented in orange bars, while log-earnings item-anchored achievement gaps are presented in blue bars.[39] For comparability, both sets of estimates are presented in SD units. In our sample, achievement gaps estimated using traditional test scales yield an average white-Black gap of approximately 0.52 SD and a white-Hispanic gap of 0.42 SD. These findings align with extensive research documenting large racial achievement gaps, with Black students showing particularly pronounced disadvantages.[40] Yet, achievement gaps estimated using log-earnings item-anchored scores are substantially larger, at around 0.82 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 are less likely to answer high-price questions correctly that are relatively less emphasized by the TAAS scoring rules. These achievement deficits are thus partly obscured by the aggregation inherent to the TAAS scale scores. Appendix B discusses this point formally, adding details on how we deal with first-stage noise in the estimates of the gaps.

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

[40] 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.

Finding 2: Anchored achievement scales rank students differently from given scales. In light of the achievement gap results, a natural question is whether item-anchored scales also reorder students relative to conventional test scores. In other words, is our achievement measure A_i just a rescaled version of the conventional standardized score? To explore this possibility, Panel (b) of Figure 4 plots average given math+reading score ventiles on the x-axis against the ventile-conditional means, 50% ranges, and 90% ranges of the log-earnings item-anchored scores on the y-axis. The estimates plotted in this figure suggest substantial variation in the item-anchored scores even among students with very similar given scores – the ventile-conditional 90% ranges often span two 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 bin of the given score distribution.

6 Inside the Item: Observable Features

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

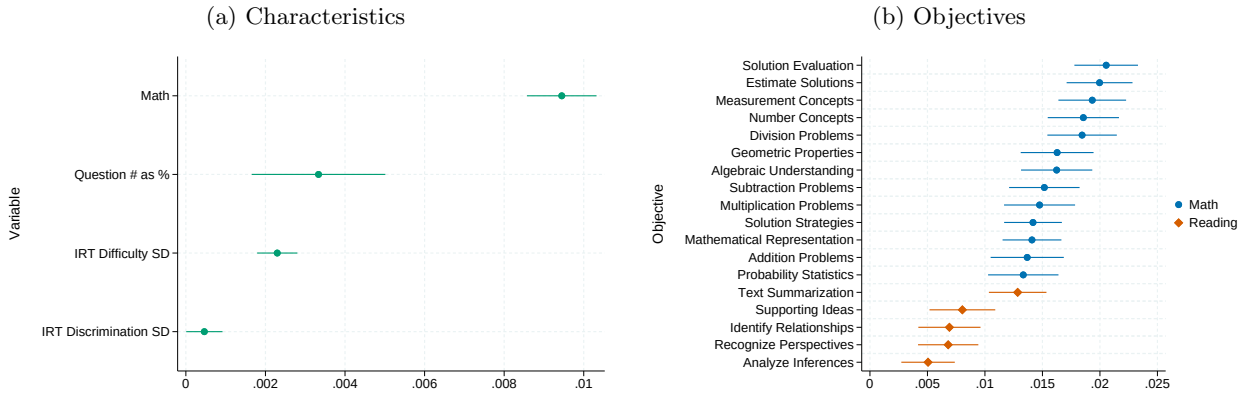
In this section, we show how much variation in item prices can be explained by readily available item-level characteristics, such as the subject, difficulty, and broad objectives. To assess the relationship between our estimated item-level prices and standard item-level characteristics we estimate simple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

Figure 5 presents the estimated coefficients from these regressions. The results presented in Panel (a) indicate that math items are associated with higher log-earnings item prices. On average, math item prices are about 1 percent higher than reading item prices. Similarly, items that occur later in the test (have a higher question number) have higher prices. Conditional on all other observable characteristics, the last question of a 100-question test has a predicted item price about 0.3 percent higher than the first. This result is in line with the research finding high academic and labor-market returns to student cognitive endurance (Brown et al., 2025, Reyes, 2023).

Finally, items that are more difficult, as estimated by a 3PL IRT model, also tend to be associated with higher log-earnings item prices.[41] A one-SD increase in item difficulty is associated with about a 0.2 percentage point higher log-earnings item price. However, this is not the case for items that have a higher discrimination parameter. Recall that the discrimination parameter can be approximated by the correlation between answering the item correctly and the overall test score. Given that for each item m, the estimation of the item price conditions on the full vector of student

[41] We control for the information content of each item (Fisher information) in a 3PL IRT model, which is approximately quadratic in the item's difficulty relative to the population mean. Figure D.2 presents estimates of this regression. Including the quadratic-in-difficulty term does not meaningfully change our estimates.

Figure 5: Relationship of Item Characteristics to $\hat{\mathbf{\Omega}}$



Notes: Panel (a) shows coefficient estimates from regressions of estimated item-level prices ($\hat{\omega}_m$) on observable item-level characteristics. All regressions include grade and year fixed effects and are weighted by the inverse of the estimated variance of $\hat{\omega}_m$. Panel (b) presents the same regression as Panel (a), but replaces the subject indicator for math with subject-specific learning-objective indicators as designated by the State of Texas, using "Word Meaning – Reading" as the omitted objective. Figure D.1 presents analogous results for the subset of items for which we were able to collect text data.

item responses, $\mathbf{D}_{i,-m}$, it is not surprising that the discrimination parameter does little to explain log-earnings item prices.

Panel (b) of Figure 5 shows the results of an analogous regression in which the math indicator is replaced by learning-objective indicators, with *"word meaning,"* a reading objective, as the omitted category. These learning objectives were designated by the State of Texas. Consistent with earlier evidence, math objectives are generally associated with higher prices on average than reading objectives. Among reading objectives, *"text summarization"* has the highest estimated price, similar in magnitude to the lowest-price math objectives. Moreover, different objectives are often starkly different from each other in average price. However, high prices are not concentrated in a single objective.

7 Inside the Item: Item Text Embeddings

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

Formally, we estimate the sample analogue of equation (5) by relating $\widehat{\mathbf{\Omega}}$ to the non-text item characteristics $\mathbf{R}$ and the matrix of item-text embeddings $\mathbf{E}$. Our basic approach is to compare the fit of this model to that of a restricted model that uses only psychometric item characteristics and

test metadata — that is, a model in which we use only $\mathbf{R}$ (or subsets of $\mathbf{R}$) and exclude $\mathbf{E}$.

Text representation. In order to implement this procedure, we must first convert the item texts, which are high-dimensional, into lower-dimensional numeric data amenable to quantitative analysis. We accomplish this through a *text embedding*: a function, implemented by a pretrained neural network, that maps an arbitrary piece of text to a fixed-length numerical vector. These vectors are trained so that texts with similar meaning are mapped to nearby points in the embedding space, providing a continuous measure of semantic similarity between any two pieces of text. We provide further details about the embedding model in Section 8.2.[42] We denote the embedding for item m by $\mathbf{E}_m$.

Prediction model and evaluation. We estimate $\widehat{\mathbf{\Omega}}^{ML} = \hat{g}(\mathbf{E})$ flexibly using five-fold cross-fitting and a neural network (NN).[43] To illustrate the incremental information contained in item text, we compare and show two sets of regressions of our estimated vector $\widehat{\mathbf{\Omega}}$ on: i) only non-text observables, $\mathbf{R}$; and ii) $\mathbf{R}$ and $\hat{g}(\mathbf{E})$, thereby including all the information on the item text embeddings, $\mathbf{E}$.[44] Figure 6 presents the out-of-sample (OOS) R^2s for our NN model for different sets of predictors in $\mathbf{R}$ both with and without the text embeddings included as additional predictors. Because $\widehat{\mathbf{\Omega}}$ is estimated in a first stage, we adjust the reported R^2 values to subtract variation due solely to first-stage estimation error; see Appendix E for the derivation and Figure E.1 for implementation details.[45]

Results. Figure 6 reports out-of-sample R^2 for several predictor sets drawn from $\mathbf{R}$ (subject, learning objective, item position in the exam, IRT difficulty, and IRT discrimination) with and without embeddings. It reveals two key findings.

- **Incremental explanatory power of text.** Including the text embeddings $\mathbf{E}_m$ consistently increases the OOS R^2 – roughly 20-50% of the baseline that uses only $\mathbf{R}$. The boost is similar for different non-text predictor sets, suggesting further that the information captured by the text embeddings is not well explained by standard psychometric item characteristics such as

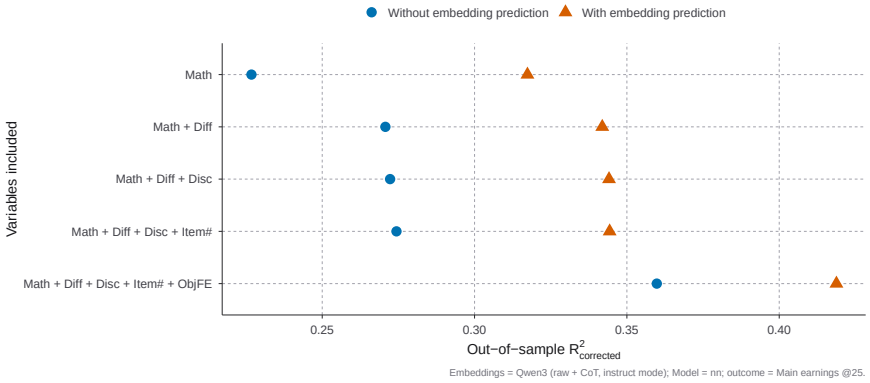
[42] See Reimers and Gurevych (2019) for a general introduction to text embeddings and Muennighoff et al. (2023) for the Massive Text Embedding Benchmark (MTEB) used to evaluate embedding models.

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

[44] Notice that we are essentially modeling $\widehat{\mathbf{\Omega}}^{ML}$ as a semiparametric model, with $\widehat{\mathbf{\Omega}}^{ML} = \hat{g}(\mathbf{R}, \mathbf{E}) = \theta'\mathbf{R} + \hat{g}(\mathbf{E})$. This assumption is not very important for the substantive conclusion here presented, so we choose it for ease of exposition.

[45] In summary, because each item weight is estimated via cross-fitting, we treat the ML predictor for item m as fixed with respect to the first-stage estimation error, conditional on $\mathbf{\Omega}$ (Chernozhukov et al., 2018, Bach et al., 2024). 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.

Figure 6: Item Text Embeddings Add Explanatory Power for Item Prices



Notes: Out-of-sample $R^2_{\text{corrected}}$ (sampling-error corrected; see Appendix E) for nested OLS specifications predicting the age-25 log-earnings item prices. Blue circles are model fit without the neural-network embedding prediction; orange triangles include it. Embeddings are Qwen3 (raw question + chain-of-thought-extracted skills, instruct mode); the prediction model is a precision-weighted neural network with hyperparameters selected by random search. Across all five predictor sets, adding the embedding prediction increases explanatory power. The embedding prediction is never shown on its own in this figure—each specification adds it to a set of item observables. The standalone out-of-sample $R^2_{\text{corrected}} = 0.28$ for our main-specification $\widehat{\mathbf{\Omega}}$. See Figure E.1 for this and other embedding-only R^2s. Figure F.1 reports the same comparison using ridge + PCA as an alternative ML estimator, yielding qualitatively identical patterns.

difficulty, discrimination, and learning objective.[46]

- **Most variation remains unexplained.** Only a relatively small share of the variation in $\widehat{\mathbf{\Omega}}$ is explained by any of these factors, with the corrected R^2s never exceeding 0.42. Moreover, the non-embedding features of the items (difficulty, subject, etc.) explain even less—about 20-35% of the total variation depending on the control set. Thus, high-price items are not simply difficult and discriminating items. Most of the variation in the item weights is not explained by any features we observe.

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

8 Inside the Item: Standards-Based Skill Mapping

The embedding results in Section 7 show that item language helps explain item prices, but the mapping from text embeddings to item prices is difficult to interpret. This interpretability problem is compounded by the fact that the variables analyzed in Section 6 are either psychometric/metadata characteristics that are difficult to interpret as skills, such as difficulty and discrimination, or are too coarse to identify specific skills, such as broad learning objectives like "Number Concepts."

[46] Regressing the NN text-based predictions $\widehat{g}(\mathbf{E}_m)$ on observable item characteristics confirms this: the largest loading is on the math indicator, with item difficulty and item position carrying smaller but statistically distinguishable coefficients, while item discrimination is indistinguishable from zero. The magnitudes of the difficulty, position, and discrimination loadings are roughly an order of magnitude smaller than the math loading. Figure F.2 reports the full set of coefficients.

To obtain a more interpretable skill-level decomposition of the item prices, we develop and implement a *text-based mapping* from items to the Common Core State Standards (CCSS) and estimate "skill prices" that summarize the price associated with each CCSS skill. As mentioned in Section 3.3.2, the CCSS provides a fine-grained skill taxonomy that is actionable and granular enough to identify the specific skills that are predictive of later earnings. Our mapping proceeds in three steps, which we detail below.

8.1 Step 1: Skill Extraction

For each digitized test item, we prompt two large language models, one open-source chain-of-thought (Qwen 3-8B) and one closed-source (OpenAI's o3), to return a structured description of the item's cognitive demands.[47] In particular, we ask for the skills required to answer the item, the reasoning steps a student would follow, and needed prerequisite knowledge. We ensure that the model receives only the digitized item text in the prompt – no grade label, no curriculum reference, and no CCSS skills – so the extracted skills are agnostic to any standards framework. This yields a short text summary per item consisting typically of about three required skills, four reasoning steps, and three prerequisite skills, which characterizes *what the item tests* rather than *what the item says*.[48]

8.2 Step 2: Text Embeddings

We next embed the skill descriptions from Step 1 and the full text of each CCSS skill using Qwen3-Embedding, a dedicated open-source embedding model based on a chain-of-thought equivalent generation model.[49,50] By embedding the LLM-extracted skill descriptions rather than the raw question text, we focus the comparison on cognitive demands rather than surface-level question language. This is important because the CCSS entries are skill-based; we thus translate the items into the kinds of skills that can be plausibly matched to the CCSS taxonomy.

Following Su et al. (2023) and Muennighoff et al. (2023), we add the instruction "*represent the cognitive skills and knowledge a student needs to answer it correctly*," so that the resulting vectors emphasize skill demands rather than surface-level question language. CCSS skill descriptions

[47] We use OpenAI's `o3-2025-04-16` with high reasoning effort and Qwen3-8B with chain-of-thought reasoning. Both prompts are reproduced verbatim in Section 2 of the Online Appendix.

[48] This skill-extraction step is distinct from the digitization described in Section 3.3.1. Digitization recovers the item text from scanned booklet images; skill extraction takes that text as input and produces a structured cognitive profile of each item. The digitized item template fed into Qwen3 is documented in Section 1.1 of the Online Appendix.

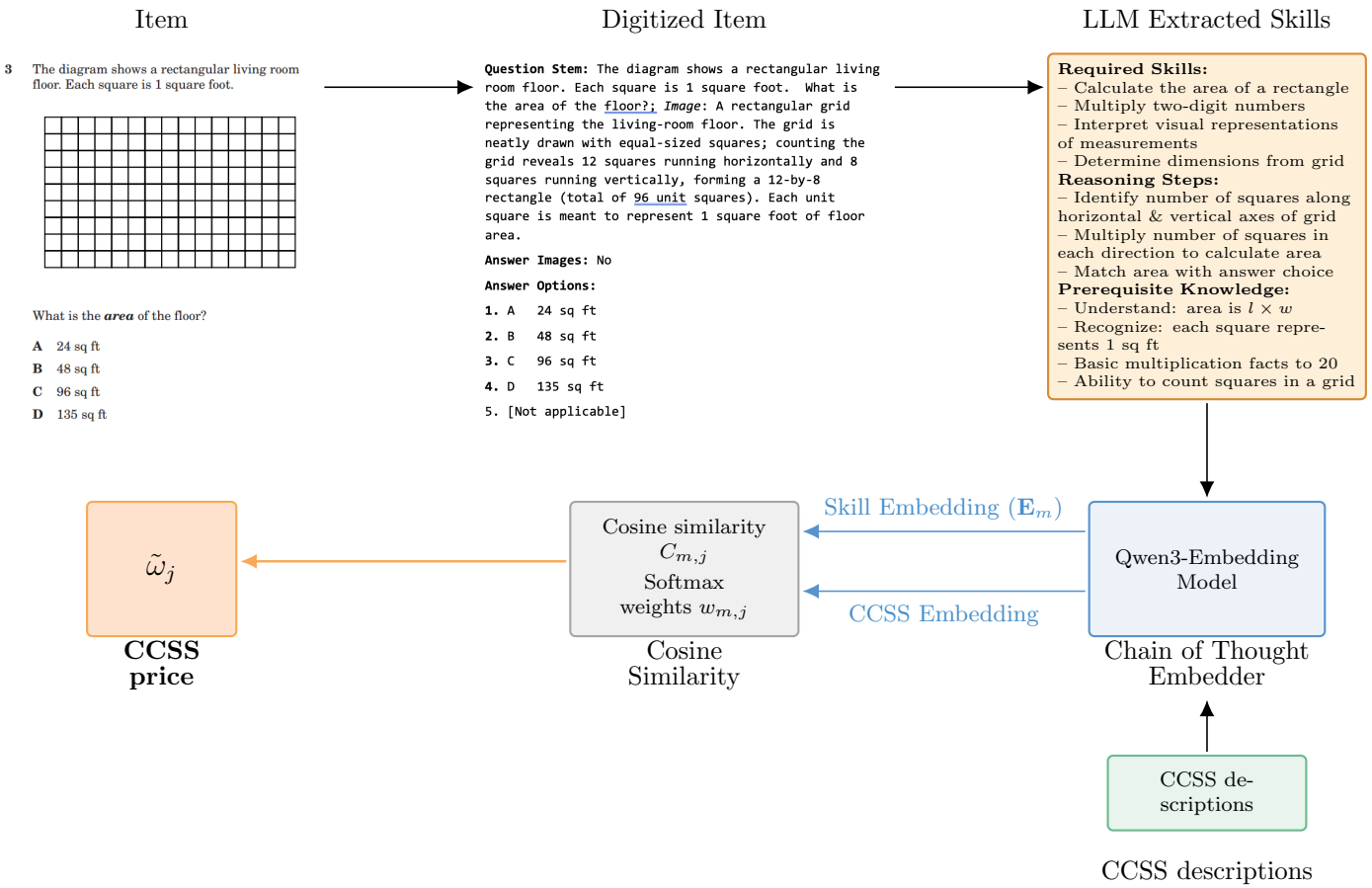
[49] The Qwen3-Embedding (Zhang et al., 2025) is an 8-billion-parameter model purpose-built for text embedding, which maps each input text to a vector of 1,024 real numbers. The model supports Matryoshka Representation Learning (Kusupati et al., 2022), which trains the model so that the leading components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

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

are embedded without a prefix, following the asymmetric query-document convention standard in information retrieval.[51]

Figure 7 illustrates the overall flow from a scanned test booklet, through digitization of the item text, to the final numerical embedding. Figure 8 projects the item embeddings into two dimensions via the Uniform Manifold Approximation and Projection (UMAP) method (McInnes et al., 2018). Math (blue) and reading (orange) items occupy clearly distinct regions of the embedding space, illustrating that the embeddings capture subject-level semantic structure.

Figure 7: From Test Item to Skill Price



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

8.3 Step 3: Cosine Similarity

With embeddings for both the item-level skills and the CCSS skills in hand, we measure the semantic proximity of each item to each CCSS skill via cosine similarity. Let $\mathbf{E}_m$ be the embedding of the

[51] Instruction-prefixed embedding prepends a short natural-language directive to the input before encoding, directing the model to foreground task-relevant information. Models fine-tuned with instruction prefixes substantially outperform their non-instruction counterparts across nearly all MTEB task categories, with especially large gains on retrieval tasks. See Su et al. (2023), Muennighoff et al. (2023).

Figure 8: Embeddings Differentiate Math and Reading

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

skill description for item m, and let $\mathbf{V}_j$ be the embedding of CCSS skill j. The cosine similarity between these embeddings is

$$C_{m,j} = \frac{\mathbf{E}_m \cdot \mathbf{V}_j}{\|\mathbf{E}_m\| \, \|\mathbf{V}_j\|}. \tag{8}$$

That is, cosine similarity is defined as the normalized dot product of the embedding vectors. Geometrically, $C_{m,j} = \cos(\theta_{m,j})$, where $\theta_{m,j}$ is the angle between the two vectors in the embedding space. Two embeddings have high similarity whenever they "point" in the same direction. Cosine similarity has been shown to provide a good summary of semantic similarity between the texts associated with the compared embeddings.[52] Let $\mathbf{C}$ denote the $M \times J$ matrix of cosine similarities – $\mathbf{C}$ thus connects each item to each CCSS skill, with the magnitude and sign of $C_{m,j}$ determining the "strength" of the item-to-CCSS match.[53] Figure 9 illustrates the item-to-CCSS match with the sample example item used in Figure 7. The rightmost panel lists the seven most similar CCSS skills to this item according to the estimated cosine similarity, $C_{m,j}$.

8.4 Step 4: Defining Skill Prices

For a given CCSS skill, we define

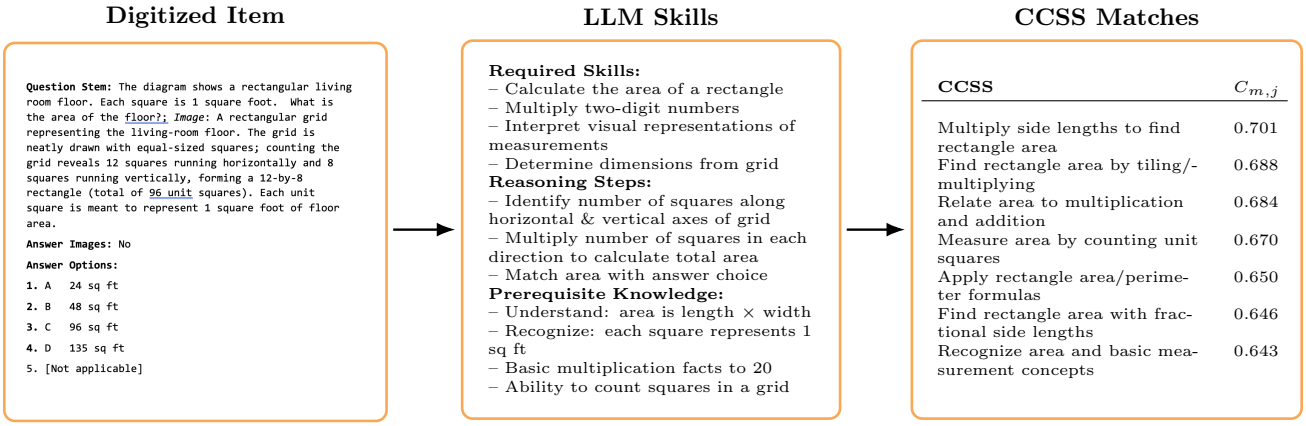
$$\tilde{\omega}_j = \sum_m w_{m,j} \times \hat{\omega}_m, \quad \text{where } w_{m,j} = \frac{\exp(\beta \cdot C_{m,j})}{\sum_{m'} \exp(\beta \cdot C_{m',j})}. \tag{9}$$

In words, $\tilde{\omega}_j$ attributes to CCSS skill j a weighted average of all item-level prices, where the weights follow a multinomial logit (or "softmax") kernel (McFadden, 1974). The scale parameter β controls how sharply the weighting discriminates between good and poor matches: a higher β

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

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

Figure 9: Item-to-CCSS Match via Cosine Similarity Example



means that items with low similarity to skill j are discounted more aggressively relative to better-matched items. In the language of discrete choice, β is inversely proportional to the scale of the error term in a random utility model (Train, 2009). This formulation has two useful properties. First, it reduces the weighting to a single free parameter β. Second, the log of the denominator—the *inclusive value* from discrete choice theory—provides a natural diagnostic for the quality of the item-to-skill match: standards with low inclusive values have no well-matched items in the data (see Appendix H). We select β by 10-fold cross-validation, stratified by grade and subject, which yields an optimal value $\beta^* = 45$ that we use throughout the remainder of the paper.[54]

8.5 Results

Figure 10 presents the full distribution of estimated CCSS prices following equation (9) with the associated 95% confidence intervals. The standard error of each skill price is derived from the item-price standard errors (Section 4) via the variance propagation formula, treating the softmax weights as fixed. We weight all subsequent regressions by $1/\text{SE}(\tilde{\omega}_j)^2$ to account for the fact that some skill prices are estimated more precisely than others (Hanushek, 1974, Hedges and Olkin, 2014). While Figure 10 labels the five highest-price skills for both math and reading, Table 2 expands this list to present the 20 top and bottom CCSS skills for each. Below, we summarize the main takeaways from this analysis.

Finding 1: Most skills have positive prices. Estimated skill prices vary substantially across CCSS skills, ranging from (essentially) zero to about 4.5 percentage points. Only about 6% of CCSS skills have prices that are not statistically distinguishable from zero.

Finding 2: Math skills dominate the top rankings, but reading comprehension is also important. Consistent with the results presented in Section 6, Figure 10 shows that math skills

[54] Figure H.5 reports the CV curve. Our results are robust to using a power kernel $(\max(0, C_{m,j})^p, \quad p \geq 1)$ with the p selected by cross-validation instead.

cluster at higher price levels than reading skills. However, the granularity of skills provided by the CCSS allows us to recover two novel results. First, in contrast to the results presented in Figure 5, we find that not all math skills dominate reading skills. The highest-price reading skill – *"compare text structures and their effects"* – ranks higher than approximately 300 lower-price math skills. Nonetheless, the highest-price math skills have much higher estimated prices than the highest-ranked reading skills; the point estimates at the top end for math are about double those for reading.

Second, consistent with the results presented in Figure 5, the highest-price reading skills often involve basic reading comprehension and text summarization. Twelve of the top 20 reading skills require students to *"summarize"* or *"identify [the] main idea"* of the text. In contrast, none of the bottom-ranked reading skills require any text summarization. Instead, they focus on *"word meaning", "word choice"* or *"word tone"*.

Finding 3: Computation may matter more than conceptual understanding. Within math, differences across the distribution are more nuanced than for reading. From a simple comparison of skills on either end of the distribution, we can nonetheless see a few key differences. First, the top math skills lean heavily toward what we might call computational or procedural fluency that requires executing multi-step procedures accurately using formal notation to produce numerical results. Often these skills emphasize verbs such as *"derive", "measure," "find"*, and they tend to focus on whole-number computation, measurement, and basic coordinate-plane reasoning.

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

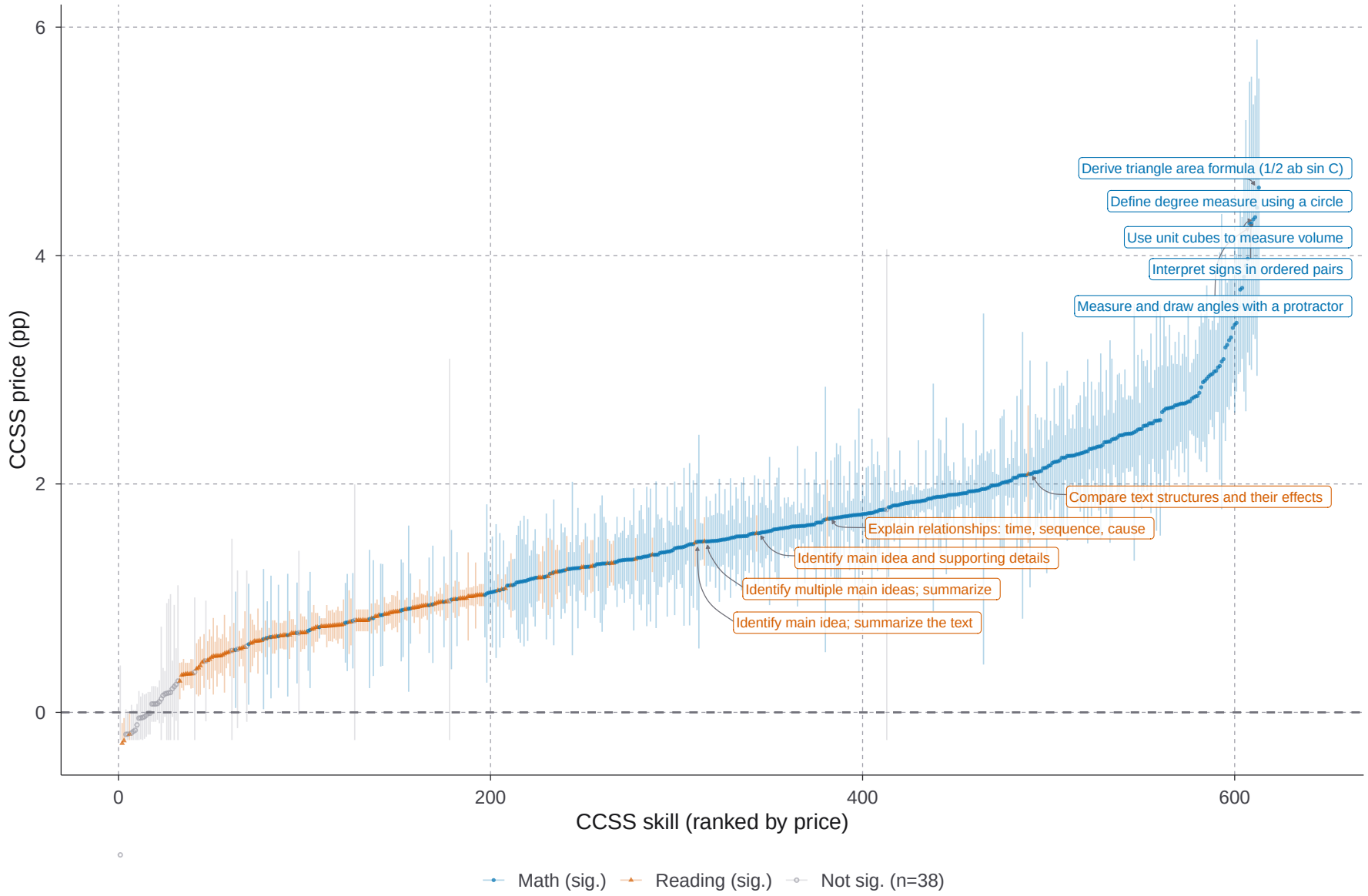
For example, while geometry-related skills are located at both ends of the price distribution, their emphasis differs. Geometry skills at the top of the price distribution favor procedural use of geometric tools (*"Derive triangle area formula"*; *"Define degree measure using a circle"*), while low-price geometry skills emphasize conceptual properties and transformations (*"Prove parallelogram theorems"*; *"Identify and draw line of symmetry"*).

Table 2: The 20 Top- and Bottom-Ranked CCSS Skills by Subject

Top 20 Skills			Bottom 20 Skills		
Rank	Grade	CCSS Short Description	Rank	Grade	CCSS Short Description
Panel A. Math					
1	HS	Derive triangle area formula (1/2 ab sin C)	517	HS	Link triangle congruence to matching sides/angles
2	4	Define degree measure using a circle	518	4	Use multiples to multiply fractions by whole numbers
3	5	Use unit cubes to measure volume	523	8	Describe congruence via rigid motions
4	6	Interpret signs in ordered pairs	528	4	Interpret a/b as a multiple of 1/b
5	4	Measure and draw angles with a protractor	532	3	Define fractions as parts of a whole
6	4	Recognize angles and angle measurement	534	HS	Justify steps when solving equations
7	6	Write order comparisons in context	536	HS	Use matrix inverses to solve systems
8	4	Use additive angle measures to find unknowns	544	HS	Use rigid motions to decide congruence
9	4	Relate angle measure to degrees	545	HS	Find inverse functions
10	5	Define volume by packing unit cubes	550	HS	Use zero/identity; link determinant to inverse
11	6	Order rational numbers; use absolute value	551	3	Understand equivalence as same size/point
12	6	Generate equivalent expressions	553	6	Report number of observations
13	HS	Interpret graphs as solution sets	567	3	Explain and compare equivalent fractions
14	5	Define a coordinate system with axes	573	4	Explain equivalent fractions using models
15	HS	Derive circle/solid formulas informally	582	HS	Solve using inverses; write inverse expressions
16	8	Identify linear vs non-linear functions	584	3	Generate and justify equivalent fractions
17	6	Evaluate expressions using values and formulas	586	4	Compare fractions using benchmarks/common units
18	6	Graph points; find distances using abs	587	4	Identify and draw lines of symmetry
19	4	Draw and identify basic geometric objects	588	3	Compare fractions with same numerator/denominator
20	HS	Fit a linear function to data	613	HS	Prove parallelogram theorems
Panel B. Reading					
125	8	Compare text structures and their effects	593	9–10	Analyze cumulative impact of word choice
233	3	Explain relationships: time, sequence, cause	594	3	Use context to self-correct while reading
271	3	Identify main idea and supporting details	595	4	Use context to self-correct while reading
299	5	Identify multiple main ideas; summarize	596	5	Use context to self-correct while reading
304	4	Identify main idea; summarize the text	597	4	Use phonics and morphology to decode words
327	7	Analyze author's organizational structure	598	5	Use phonics and morphology to decode words
336	4	Describe text structure and organization	599	3	Determine word meanings in context
348	7	Identify multiple central ideas; summarize	600	5	Determine word meanings in context
349	11–12	Analyze multiple central ideas and interactions	601	7	Analyze sound devices' effect
351	8	Analyze paragraph structure and sentence roles	602	3	Interpret words: literal vs figurative
359	8	Analyze central idea development; summarize	603	5	Interpret figurative language
360	6	Determine central idea; summarize objectively	604	8	Analyze word meaning and tone; allusions
364	9–10	Analyze theme development; summarize objectively	605	9–10	Analyze cumulative word choice and tone
365	9–10	Analyze central idea refinement; summarize	606	8	Analyze word meaning and tone; allusions
373	8	Analyze theme development in narrative; summarize	607	4	Determine word meanings in context
378	7	Analyze theme/central idea development	608	11–12	Analyze word choice effects on meaning and tone
379	11–12	Analyze multiple themes/ideas and interactions	609	6	Determine word meanings, including technical
381	6	Determine theme/central idea; summarize objectively	610	6	Analyze word choice impact on meaning/tone
383	4	Determine theme and summarize	611	7	Analyze word meanings and tone
386	9–10	Analyze structure and pacing effects	612	4	Interpret words, including mythological allusions

Notes: This table displays the 20 top- and bottom-ranked CCSS skills alongside their ranks. Panel A presents the ranked math skills, while Panel B presents the ranked reading skills. The descriptions shown are abbreviated. Standards with the same short description correspond to the same learning objective specified at different grade levels. "HS" denotes high-school standards.

Figure 10: Distribution of CCSS Skill Prices



Notes: Each point represents one of 613 CCSS standards, ranked by estimated price ($\tilde{\omega}_j$). Vertical bars show 95% confidence intervals. Prices are computed following equation (9) using a softmax kernel ($\beta = 45$) with precision weighting ($1/\text{SE}^2$). Skills with statistically insignificant (at 5%) prices are shown in gray. The five top- and bottom-ranked skills per subject are labeled.

8.6 What Predicts High-Price Skills?

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

Procedural/Routine. Given the patterns observed in Figure 10 and Table 2, we assess how CCSS skill prices vary with the extent to which a skill relies on procedural, multi-step solutions. We operationalize this analysis using two distinct but complementary measures of how "procedural" a CCSS skill is.

1. *Depth of Knowledge (DOK)* – We categorize each CCSS skill using the Depth of Knowledge (DOK) framework, which captures the cognitive complexity of academic tasks and classifies them into four ordered categories. DOK 1, Recall & Reproduction, involves basic recall of facts, definitions, or simple procedures. In math, DOK 1 tasks involve straightforward actions such as recalling multiplication facts, identifying the formula for the area of a rectangle, or performing a routine computation. DOK 2, Skills & Concepts, requires cognitive processing such as comparing, classifying, interpreting, or applying concepts in familiar situations. For example, DOK 2 tasks include comparing two fractions with unlike denominators, selecting an appropriate operation to solve a word problem, or interpreting information from a simple table or graph. DOK 3, Strategic Thinking, demands reasoning, justification, planning, and evidence-based decision-making, often involving abstract or non-routine problems. For example, DOK 3 tasks include analyzing the structure of an algebraic expression to determine equivalence, or determining the most efficient strategy for solving a multi-step percentage problem.[55]

2. *Routine vs. Non-Routine* – We adapt the task-based framework of Simon (1960), Polanyi (1966), Autor et al. (2003), Autor (2013), Spitz-Oener (2006) to assess the degree to which a CCSS skill can be performed via the rote application of a set of rules. We create an ordered scale taking five values. A value of 5 on this scale means the skill is entirely based on calculation and could in principle be executed mechanically by a calculator, spreadsheet, or computer (with pre-generative-AI capabilities). Examples of level-5 tasks include all arithmetic problems, algebraic simplification, evaluating formulas, checking spelling, solving systems of linear equations, etc. By contrast, level-1 tasks rely on skills that cannot be codified into a simple algorithm. Examples include general reading comprehension, assessing whether an analogy or comparison is reasonable, and finding mistakes in a mathematical argument.

[55] We do not include DOK 4: Extended Thinking, as there are no skills in the CCSS that match this level. This level encompasses complex, sustained work such as designing investigations, synthesizing information across sources, or applying concepts in novel contexts over time. See Webb (1997, 2007) for more details.

Spatial reasoning. We classify each skill according to whether it requires spatial reasoning. Spatial reasoning has been found to be highly predictive of academic success, particularly in STEM and high-earning fields (Uttal et al., 2013a,b).

To characterize each CCSS standard, we provide a classification protocol to an LLM to code each of these three dimensions, summarized in Table 3.[56] Each standard is presented to the model with its full text description, stripped of grade identifiers to prevent anchoring. Importantly, the classification is performed independently for each standard, and the model receives no information about the standard's estimated price. The full classification protocol, dimension distributions, and inter-dimension correlations are reported in Appendix H and Online Appendix. Particularly for math, these three measures are only modestly correlated, suggesting that they are picking up distinct skill dimensions.

Table 3: Skill Classification Dimensions

Dimension	Scale	Subjects	Definition	Literature
Depth of Knowledge	1–3	Both	Level/complexity of analysis needed: 1 = recall/reproduction; 2 = concepts/multi-step; 3 = justification/complex problems.	Webb (1997), Hess (2009)
Routine vs. Non-routine	1–5	Both	How completely can skill be mastered through rote rule application? 1 = non-routine (e.g., "analyze how characters develop"); 5 = fully routine (e.g., "compute perimeter from coordinates").	Simon (1960), Autor et al. (2003), Autor (2013), Polanyi (1966), Spitz-Oener (2006)
Spatial reasoning	0/1	Math	Does the skill require the mental manipulation of shapes, positions, or coordinate systems? 1 = spatial (e.g., "graph points on a coordinate plane"); 0 = non-spatial (e.g., "solve linear equations").	Uttal et al. (2013b) Uttal et al. (2024)

Notes: Each of 613 CCSS standards is classified independently by Claude Opus 4.6. Grade identifiers are stripped from the standard text before classification to prevent anchoring. "Both" = dimension applies to math and reading standards. Routine vs. Non-routine scale is coded so that higher values indicate more routine skills. Examples in the Definition column are illustrative CCSS standards (paraphrased).

8.6.1 Average Skill Prices by Skill Dimension

We begin by examining skill differences along each dimension presented above, separately. We also assess price differences by the grade of the CCSS skill (the grade assigned to the skill by the CCSS). Because these classifications are discrete, Figure 11 presents separate precision-weighted regressions, one for each dimension, of CCSS price on indicators for the level of the dimension, fully interacted with subject. For grade, routine intensity, and DOK, the reference cell is math at grade = HS, routine intensity = 1, and DOK = 3. The spatial reasoning panel is from a math-only regression with spatial = 0 as the reference; reading is omitted because spatial is not classified for

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

reading standards. The outcome is standardized and presented in terms of the standard deviation of CCSS prices across all standards.

Finding 1: Routine/low-DOK math skills have higher prices. For math, more routine skills and lower DOK skills have higher average prices. For routine skills in particular, the main difference appears to be between those that are completely non-routine (scale = 1) versus those that have at least some routine elements (scale > 1). Both dimensions suggest that procedural, computational skills outperform more conceptual ones.

Finding 2: Routine/low-DOK reading skills have lower prices. In contrast to math, more routine reading skills tend to have lower prices. The DOK pattern is also different for reading – low-DOK skills have the lowest average prices, while DOK level 2 is the highest. Skills that are more complex, such as summarization, have higher prices than more basic reading tasks.

Finding 3: Spatial skills have higher prices. Spatial reasoning skills have average log-earnings prices that are about 0.25 SD higher than non-spatial math skills.[57] This is in line with the literature that shows that spatial skills strongly predict STEM educational and occupational outcomes (Wai et al., 2009, Lubinski, 2010) as well as labor-market earnings (Baker and Cornelson, 2019).

Finding 4: Skill prices do not vary with CCSS-designated grade except for 8$^{\text{th}}$ grade math. For math, we do not find any statistically significant differences in average skill prices by CCSS-designated grade, except for 8$^{\text{th}}$ grade math, where the average skill prices are significantly higher than all other grades. For reading, the average skill prices are nearly constant for all grades. In every grade, the average skill prices for math are much higher than the average for reading – the within-grade gaps average around 1.25 SD.

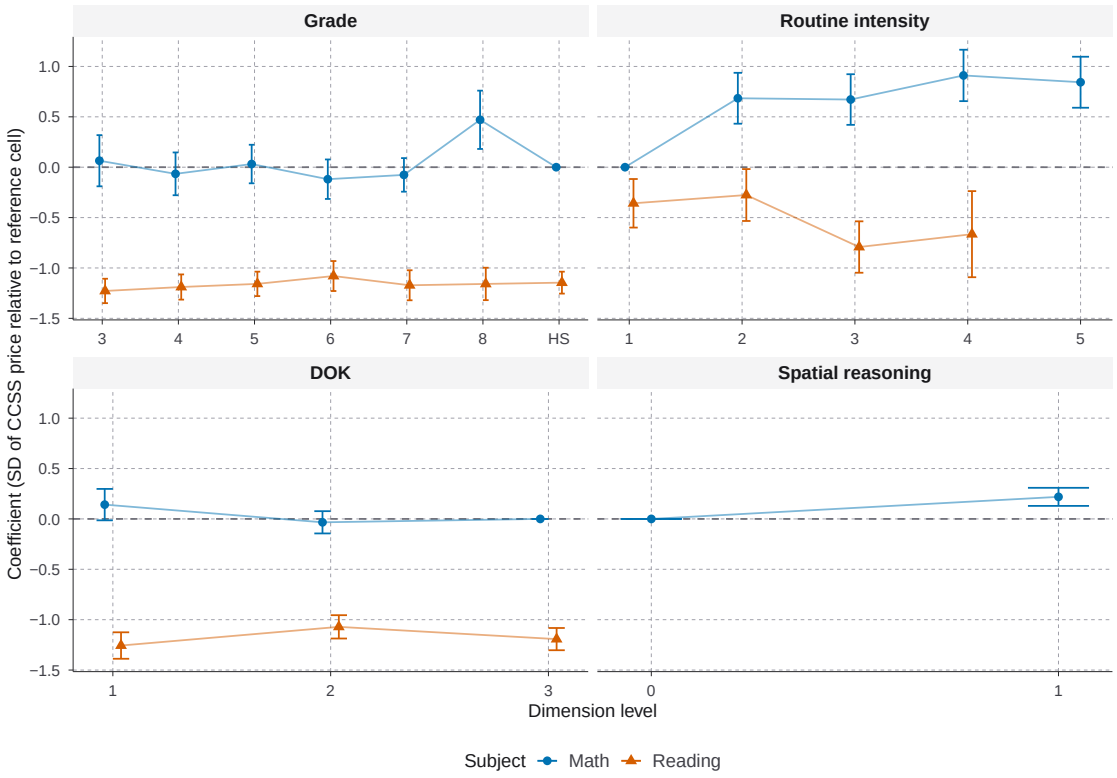
Figure H.1 in Appendix H presents analogous figures for a model that estimates skill-level coefficients for all dimensions jointly in a single regression. The joint regression estimates are qualitatively the same as the individual regression results for grade, routine intensity, and spatial reasoning. The DOK results are different, however; controlling for routine intensity yields a flat relationship between DOK and price for math and a clear negative relationship for reading.

8.6.2 Visualizing Skill Prices by Dimension

Finally, Figure 12 provides a more intuitive summary by organizing math CCSS prices along the three dimensions in the form of a heat map. Each cell shows the precision-weighted mean price together with a representative example skill (we select the standard with the highest cosine similarity to any item in the data, so that each example is well-identified). The heatmap illustrates the

[57] Recall that spatial reasoning does not apply to reading skills.

Figure 11: Skill Dimensions and CCSS Prices, Separate Specification



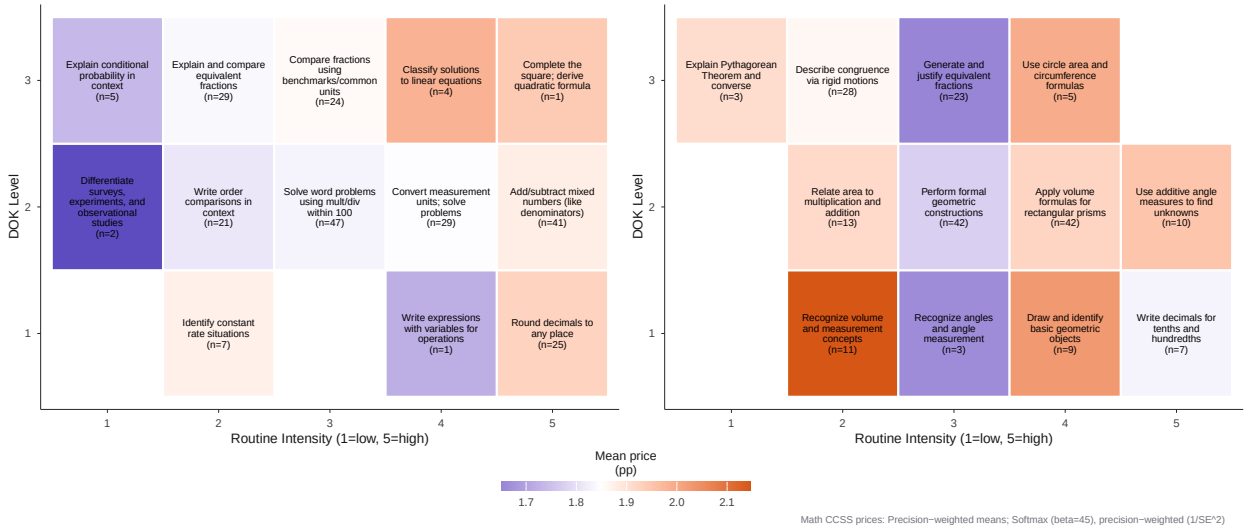
Notes: Estimates from four separate precision-weighted regressions of CCSS price on factor indicators for one dimension at a time, fully interacted with subject. For grade, routine intensity, and DOK, the reference cell is math at grade = HS, routine intensity = 1, and DOK = 3. The spatial reasoning panel is from a math-only regression with spatial = 0 as the reference; reading is omitted because spatial is not classified for reading standards. Empty subject-by-level cells—e.g., reading at routine intensity = 5—are also omitted.

results from the previous analysis: skills with lower DOK and that are more routine tend to have higher prices. The higher average prices for skills involving spatial reasoning are also apparent.

8.7 Robustness

White Male Subsample. Our main specification residualizes student demographics before estimating the item-level prices. As a robustness check, we re-estimate item prices using only white male test-takers ($\widehat{\boldsymbol{\Omega}}^{WM}$ – see Section 4), addressing the concern that demographic composition or the residualization procedure drives the results. The CCSS-level prices under the two approaches are highly correlated ($r = 0.82$, $\rho = 0.88$), and the joint skill-dimension regressions yield substantively identical patterns: the same dimensions predict higher prices in both specifications, with comparable R^2 values (math: 0.21 vs. 0.22; reading: 0.19 vs. 0.19). See Appendix H for the corresponding scatterplot and coefficient plots.

Figure 12: Math CCSS prices by DOK and Routine Intensity



Notes: Precision-weighted mean price (pp) in each DOK × routine intensity cell. One example skill per cell (highest max cosine similarity to any item). Color scale: purple = below average, orange = above average. Reading equivalent is presented in Figure H.2.

Weighting Kernel. As noted above, Appendix H documents that the CCSS prices under the softmax kernel ($\beta = 45$) are highly correlated with those from the power kernel ($p = 30$, selected by the same cross-validation procedure). The correlation between the two sets of prices is 0.98, and the skill dimension regression results are substantively unchanged.

Alternative Skill-Extraction Models. The main results use skills extracted by Qwen3-8B ("qwen3-skill"). As an alternative, we extract skills using o3 ("o3-skill"), a different model with substantially different architecture. The correlation between CCSS prices under the two extraction methods is close to one ($r \approx 0.91$), and the point estimates for the skill dimension regressions are nearly identical (see Appendix H for the corresponding coefficient and factor-level plots).

8.8 Discussion and Interpretation

The empirical results in this section—especially that more routine, lower-depth math skills have higher log-earnings prices—are quite striking and suggest a number of interesting (and not mutually exclusive) explanations. We examine these possibilities below, and we provide additional discussion of the method and data to help put them in context. This discussion is speculative and not exhaustive; we leave a detailed analysis of these and other explanations for future work.

High-price CCSS skills are directly valuable in the labor market. This is just a "straight-read" of our results. It could be that the types of math skills most rewarded in the labor market are procedural, computational skills (i.e., those that have low DOK or are easily codifiable (highly

routine)) as well as those that require spatial reasoning. That is, employers might directly value such skills and pay more for them accordingly.

High-price CCSS skills predict the future acquisition of valuable skills. Alternatively, it could be that these types of skills, assessed during childhood and young adulthood, predict the future acquisition of economically valuable skills. In this case, the math skills valued directly by the labor market need not be procedural, low-depth, or spatial in nature – all that matters is that their presence is predicted by the high-price CCSS skill categories. Thus, for example, it could be that the types of higher-level math skills used intensively in STEM careers are best predicted in earlier schooling by mastery of basic computation and spatial tasks. Indeed, prior research, including some item-level analysis (Nielsen, 2025b), provides support for this "accumulative" model of skill acquisition.[58]

High-price CCSS skills reflect soft skills. Our results indicate that math skills involving procedural, multi-step computation are associated with higher prices than more conceptual or interpretive tasks. One possible interpretation is that these patterns partially reflect returns to soft skills rather than purely academic competencies. If procedural skills require more repeated practice and reinforcement than do more conceptual skills, then they might correlate more strongly with traits such as grit, conscientiousness, or perseverance. This explanation is consistent with the discussion of $\widehat{\mathbf{\Omega}}$ in Section 2 which noted that an individual item's price might load partially on non-cognitive skills.

Our ability to speak directly to the role of soft skills in this context is limited. Even with rich item-response and text data, reconstructing meaningful measures of soft skills is difficult because these constructs differ fundamentally from the academic competencies typically assessed by reading or mathematics exams. One approximation for such traits readily available in our setting is the position of the item within the test.[59] To the extent that an item's position in the test captures variation in cognitive endurance or grit, including item position as a control absorbs that soft-skill component before the text-based prediction $\widehat{g}(\mathbf{E}_m)$ enters the regression. Figure 6 shows that even when the regression already includes item position alongside subject, difficulty, and discrimination, adding $\widehat{g}(\mathbf{E}_m)$ raises the corrected R^2 by roughly seven percentage points (from 0.27 to 0.34). The marginal contribution of the text channel is therefore the component of $\widehat{g}(\mathbf{E}_m)$ that is orthogonal to item position, and is unlikely to be driven primarily by the soft skills that position is a proxy for.

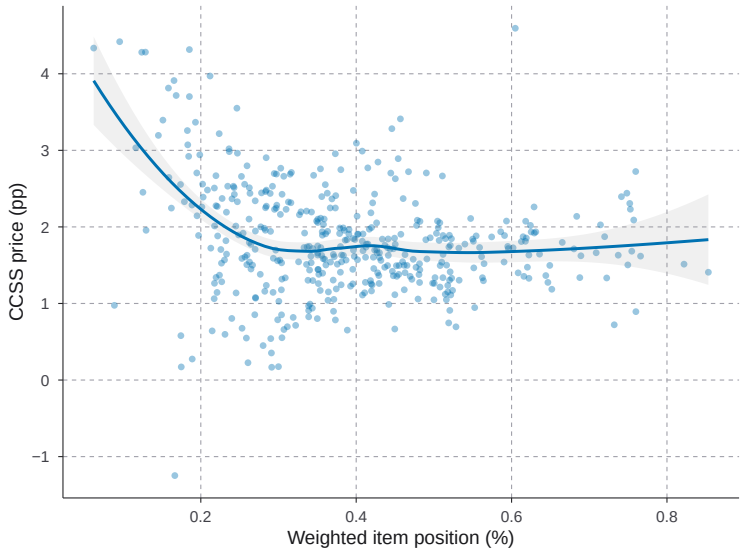
[58] See also Duncan et al. (2007), Jordan et al. (2009) for research connecting early academic skills to later academic skills. Ritchie and Bates (2013) finds that early math skills are strong predictors of later SES.

[59] Items appearing later in the test plausibly load more heavily on cognitive endurance or sustained effort, which could be interpreted as a behavioral manifestation of grit/conscientiousness (Brown et al., 2025, Brunello et al., 2021, Borgonovi et al., 2021, Debeer et al., 2014, Weirich et al., 2017, Reyes, 2023, Reyes et al., 2024). Indeed, we find evidence consistent with this interpretation. As shown in Panel (a) of Figure 5, items located toward the end of the test are associated with systematically higher estimated prices. Importantly, this finding persists even after controlling for key item characteristics such as subject, difficulty, and discrimination. Figure A.1 plots the relationship between an item's position in the test and IRT parameters.

As a final check, we examine whether high- and low-price math CCSS skills are differentially distributed across the test. Figure 13 plots each math CCSS skill's precision-weighted average item position against its estimated CCSS price. If cognitive endurance were the principal channel through which math CCSS skills predict $\widehat{\Omega}$, higher-price standards would map to items appearing later in the test on average. The data show the opposite: among math CCSS skills, those with higher estimated prices are associated, if anything, with items appearing *earlier* in the test.

Together, these results suggest that while soft skills — particularly cognitive endurance — may play some role in test performance, our empirical strategy adjusts for these factors. The observed heterogeneity in prices across CCSS skills is therefore unlikely to be driven by the kinds of non-cognitive traits that are proxied by item position. Analogous analyses for difficulty and discrimination (Figure D.3) confirm that CCSS prices are similarly unrelated to the psychometric characteristics of their constituent items. Of course, we cannot rule out explanations based on dimensions of non-cognitive skill not well captured by our available controls and item metadata.

Figure 13: Math CCSS Prices and Test Position



Notes: Each dot is a math CCSS standard. The horizontal axis is the precision-weighted average position (as a percent of the test length) of the items associated with the standard; the vertical axis is the standard's estimated CCSS price in percentage points. The line is a locally weighted regression fit with a 95% confidence band. Higher-price math CCSS skills are associated, on average, with items appearing earlier in the test — the opposite of the pattern that would be expected if cognitive endurance were the principal channel generating high CCSS skill prices.

The TAAS is an easy assessment. The typical question on the TAAS exams is quite easy—the typical IRT-estimated difficulty is well below -1, meaning that the items tend to be most effective at differentiating students with below-average achievement (a standard deviation or more below the mean). The low difficulty of the test relates to its purpose, assessing whether students are meeting broad, basic learning objectives. Consistent with the low estimated item difficulty, Table 1 shows that the TAAS items are answered correctly about 80% of the time pooling across grades

and years. These items were simply not designed to differentiate very high performing students. The items may therefore not cover rarer, higher-level skills that may have significant labor market returns. Put differently, the results are *conditional on the pool of TAAS items* and are thus most informative about the value of the skills spanned by this pool, that is, the skills covered by the Texas state curriculum.

9 Conclusion

This paper presents a framework for repurposing standardized-test items to generate achievement measures aligned with long-run economic outcomes. We also develop a method that allows us to map item-level responses to a broad set of skills that the traditional test scales might obscure. We illustrate the framework using early-adult earnings as the anchor outcome, applying it to over 3,500 digitized items from the Texas Assessment of Academic Skills administered to Texas public-school students in grades 3-8 and on the exit exam typically administered in grade 10 or 11, during 1996–2002. We link these items to approximately 1 billion student–item responses, drawn from roughly 12 million student-by-grade records, and to earnings at age 25 via state unemployment-insurance records. This combination of item content, item responses, and long-run outcomes allows us to produce the first standards-based evidence on which curricular skills—at the granularity of the more than 600 Common Core State Standards—predict adult earnings.

The results point clearly in several directions. Within mathematics, procedural, computational, and spatial skills carry the highest estimated prices, substantially more so than conceptual or interpretive tasks. Within reading, basic comprehension and text summarization dominate more fine-grained skills such as analyzing tone or determining word meanings. More broadly, the language of a test question carries economically relevant information that standard psychometric parameters do not capture: machine-learning models trained on item-text embeddings explain meaningful variation in item prices above and beyond difficulty, discrimination, and broad learning objectives. The framework also confirms that the choice of how to aggregate item responses is consequential: item-anchored scales yield racial achievement gaps roughly 45% larger than conventional scales and substantially reorder individual student rankings.

Our results have wide-ranging implications for education research and policy. The method we develop for mapping items to skill taxonomies is general: it applies to any test for which item text is available and to any taxonomy a researcher or policymaker wishes to evaluate. We apply it here to the CCSS and to earnings, but the same approach could be used with other curricular frameworks and other anchor outcomes (college completion, health, labor-force participation), each of which would, by construction, generate a different ranking of skill priorities. Which outcome should guide curricular decisions is a normative question that our framework does not resolve, but it supplies the empirical inputs that such a conversation requires.

Separately, the finding that item text explains economic variation beyond psychometric characteristics suggests that existing item metadata misses important dimensions of what a question mea-

sures. Identifying those dimensions, and understanding why procedural and spatial skills emerge as particularly predictive, are important questions for future work. For practitioners, our estimates suggest that curricular weight on procedural, computational, and spatial mathematics, as well as on basic reading comprehension, is likely under-emphasized relative to its long-run economic return.

The item and skill prices we recover are not necessarily causal, although we argue that the influence of confounders is likely modest. We therefore view our estimated skill prices as a starting template rather than a final verdict. For example, future work could assess whether targeted interventions on the high-price skills we identify increase earnings and other economic outcomes. Ultimately, we envision an iterative process in which our methods identify candidate high-price skills, those candidates are evaluated causally, promising ones are then incorporated into school curricula, and the cycle repeats. Such a process would extend and sharpen the iterative refinements already common in curriculum and test design, where standards are periodically revised and test items are continually evaluated against psychometric criteria. The expanded process would add earnings and other long-run economic outcomes, among other possible targets, as explicit inputs into curricular and assessment design.

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.

Al-Khuzaey, S., Grasso, F., Payne, T. R., and Tamma, V. (2024). Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914.

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

American Educational Research Association (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association and American Psychological Association and National Council on Measurement in Education.

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

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

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

Bach, P., Kurz, M. S., Chernozhukov, V., Spindler, M., and Klaassen, S. (2024). Doubleml-an object-oriented implementation of double machine learning in R. *Journal of Statistical Software*, (3):1–56.

Baker, M. and Cornelson, K. (2019). Title IX and the spatial content of female employment—out of the lab and into the labor market. *Labour Economics*, 58:128–144.

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*.

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

Bettinger, E. P., Evans, B. J., and Pope, D. G. (2013). Improving college performance and retention the easy way: Unpacking the act exam. *American Economic Journal: Economic Policy*, 5(2):26–52.

Betts, J. R. and Grogger, J. (2003). The impact of grading standards on student achievement, educational attainment, and entry-level earnings. *Economics of Education Review*, 22(4):343–352.

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. *Review of Economics and Statistics*, 95(5):1468–1479.

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

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

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

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

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

Bruhn, J., Gilraine, M., Ludwig, J., and Mullainathan, S. (2025). Do test scores misrepresent test results? An item-by-item analysis. NBER Working Paper w34484, National Bureau of Economic Research.

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

Cawley, J., Heckman, J., and Vytlacil, E. (1999). On policies to reward the value added by educators. *Review of Economics and Statistics*, 81(4):720–727.

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

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

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

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

Cobb, P. and Jackson, K. (2011). Assessing the quality of the Common Core State Standards for mathematics. *Educational Researcher*, 40(4):183–185.

Conrad, C., Pope, N. G., and Zuo, G. W. (2026). Skills that pay: Subject-specific test scores and long-run outcomes. Unpublished manuscript.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hemphill, F. C. and Vanneman, A. (2011). Achievement Gaps: How Hispanic and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress. Statistical Analysis Report. NCES 2011-459. *National Center for Education Statistics*.

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

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

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

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

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

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

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

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

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

Lee, S. and Schaelling, M. (2025). Content relatability and standardized testing: Evidence from Texas. Unpublished manuscript.

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

Lubinski, D. (2010). Spatial ability and stem: A sleeping giant for talent identification and development. *Personality and Individual Differences*, 49(4):344–351.

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

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

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*.

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. *European Journal of Health Economics*, 15:265–279.

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

Murnane, R. J., Willett, J. B., and Levy, F. (1995). The growing importance of cognitive skills in wage determination. *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.

NGA Center and CCSSO (2010). Common core state standards. Technical report, National Governors Association Center for Best Practices and Council of Chief State School Officers.

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 manuscript.

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

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. Unpublished manuscript.

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

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*.

Polanyi, M. (1966). *The Tacit Dimension*. Doubleday & Company, Garden City, NY.

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

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

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

Reardon, S. F. and Galindo, C. (2009). The Hispanic-White achievement gap in math and reading in the elementary grades. *American Educational Research Journal*, 46(3):853–891.

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

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

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. Statistics for Social and Behavioral Sciences. Springer, New York, NY.

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

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

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

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

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

Rose, H. and Betts, J. R. (2004). The effect of high school courses on earnings. *Review of Economics and Statistics*, 86(2):497–513.

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

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

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

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

Simon, H. A. (1960). *The new science of management decision.* Harper & Brothers.

Spitz-Oener, A. (2006). Technical change, job tasks, and rising educational demands: Looking outside the wage structure. *Journal of Labor Economics*, 24(2):235–270.

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

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

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

Train, K. E. (2009). *Discrete choice methods with simulation.* Cambridge University Press.

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.

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

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

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

Wai, J., Lubinski, D., and Benbow, C. P. (2009). Spatial ability for stem domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, 101(4):817.

Webb, N. L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. Research Monograph 6, National Institute for Science Education and Council of Chief State School Officers. ERIC No. ED414305.

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

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

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

Online Appendices

A Data Appendix

A.1 Texas Assessment of Academic Skills

A.1.1 Subjects and Objectives

The Texas ERC makes available item-level data for statewide assessments starting with the 1996 school year. This 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
TAAS 1996–2002	Reading	✓	✓	✓	✓	✓	✓	–	✓*	–
	Mathematics	✓	✓	✓	✓	✓	✓	–	✓*	–
	Writing	–	✓	–	–	–	✓	–	–	–

Notes: ✓ indicates the subject is tested at that grade level; – indicates not tested. * Exit exam. Typically administered in 10th grade, occasionally in 11th grade.

Table A.2 provides an overview of the test structure and administration details across grades for both reading and math. On average, the reading tests have fewer items than the math tests. Note that the number of items for each grade-year combination did not change throughout the period of our study (1996-2002). Students tend to respond correctly to reading items at a higher rate than math ones. Finally, the number of items increases as grade level increases.[60]

Table A.2: Item-level Statistics

	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

Notes: Item-level statistics by grade with own estimation of IRT parameters based on final sample.

The TAAS was designed to measure broad learning objectives consistently over time. We observe which learning objective each test item corresponds to. Table A.3 presents a list of these objectives for the reading and mathematics tests.

[60] 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

Notes: The objectives listed above are those recorded in the TxERC item-response data across all grade–year combinations of the TAAS sample (1996–2002).

A.1.2 Scaling

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

$$P(D_{i,m} = 1) = \frac{\exp(\theta_i - \delta_m)}{1 + \exp(\theta_i - \delta_m)}. \tag{10}$$

The Rasch model yields a difficulty estimate for each test item ($\hat{\delta}_m$) and latent ability estimate for each student ($\hat{\theta}_i$), estimated by maximum likelihood.

The TAAS scale scores are calibrated to a 70%-correct standard. That is, θ_{standard} is defined as the student ability level that on average leads to a 70% correct item response rate. θ_{standard} is also known 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.

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

$$score = \frac{\hat{\theta} - \theta_{\text{standard}}}{\hat{\sigma}_\theta} * 200 + 1500. \tag{11}$$

Under this transformation, scale scores range from 400-2400 with the passing standard set at 1500 (corresponding to 70% correct). 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 the scale scores 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 TAAS items that fell under their specific content-area knowledge or their teaching/curriculum development experience. Notably, many of these 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. Additionally, HEM assessed items for their depiction of minority, gender, and other demographic groups.

Candidate items were then submitted to the Texas Education Agency (TEA) for review. For these reviews, 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. 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.[61]

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 state. Student responses for these items were not used in their scale score calculation. Rather, these data were reviewed to determine whether new items would be included in the following testing cycle.[62]

Altogether, the TAAS items were thus subjected to at least three rounds of development and review: first by HEM, then by TEA, and finally through field testing. This extensive vetting and review process supports our preferred interpretation of the item prices as reflecting labor market returns to skills more so than the influence of confounders.

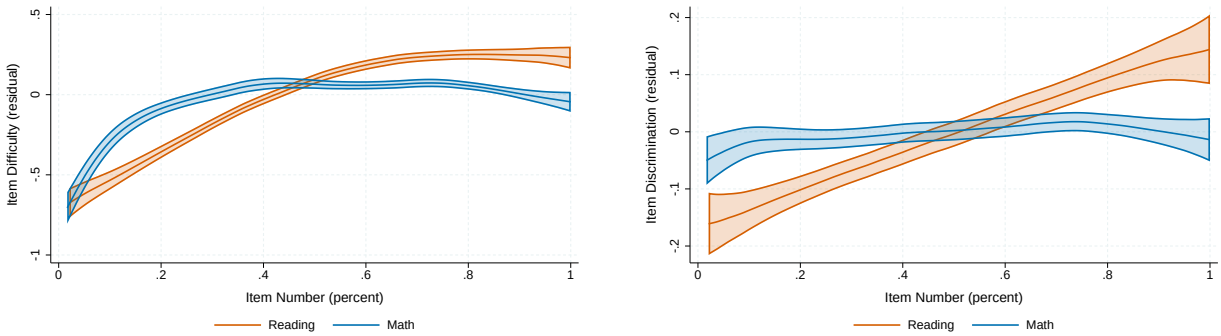
[61] The TEA review was exhaustive. For example, these committees met 75 times in 1999 alone to review all newly developed test items and all new field test data.

[62] Annual test releases to the public did not include these field-tested items.

A.1.4 TAAS Item Details

Figure A.1: Item Number and Psychometric Characteristics

(a) Difficulty (b) Discrimination

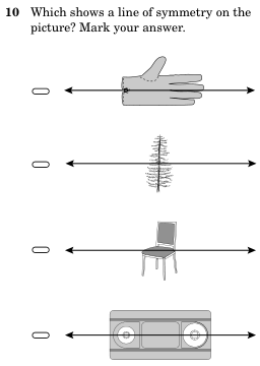


Notes: Each panel shows a local linear polynomial fit (bandwidth = 0.15) of the relevant residualized psychometric characteristic on the item's number (as a percent of the relevant exam), along with the associated 95% confidence intervals. Both difficulty and discrimination are residualized using fully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

Figure A.1 examines the relationship between an item's placement on the exam and its difficulty and discrimination. Panel (a) shows that the items that come later in an exam tend to be more difficult. However, these relationships are not linear – after a certain point items do not become more difficult as one progresses through the exam. Panel (b) shows that for math there is no relationship between an item's placement and its discrimination, while the relationship is strongly positive for reading.[63]

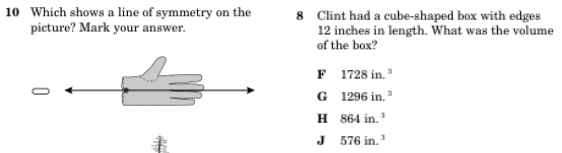
Figure A.2: Item Examples By Type

(a) No Image - Image



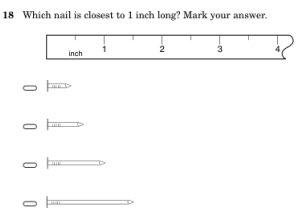
Math 2000 Grade 3 - Item 10

(b) No Image - No Image



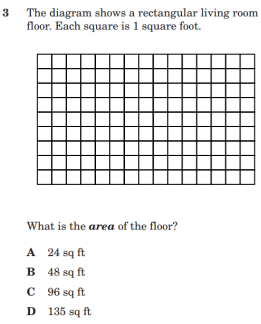
Math 2001 Grade 8 - Item 8

(c) Image - Image



Math 1999 Grade 3 - Item 18

(d) Image - No Image

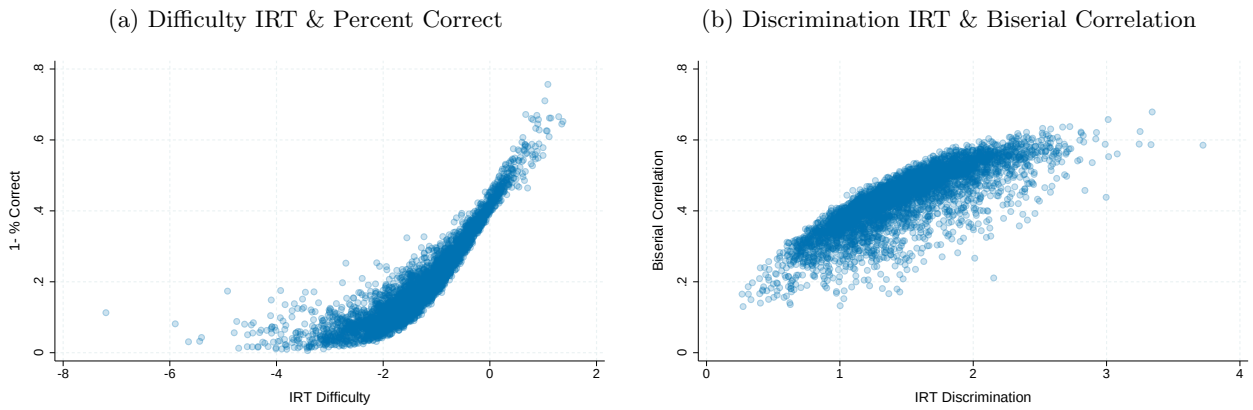


Math 2000 Grade 5 - Item 3

Notes: Examples of the four item types, varying by image presence in the question stem/response options.

[63] Even in this case, item position explains only about 20% of the variation in discrimination for reading. The patterns within each grade are quite similar to what is presented in Figure 13 and are thus omitted for brevity.

Figure A.3: Relationship Between IRT Estimates and Proxies



Notes: Each dot is one test item. Panel (a) shows that (1 − % correct) proxies for IRT difficulty. Panel (b) shows that the item–total biserial correlation proxies for IRT discrimination.

A.2 Earnings Timelines

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

Panel (b) of Figure A.4 displays the share of students for whom we observe earnings at age 25 across grades and time. On average, we observe earnings at age 25 for around 72% of test-takers, with the share very stable across all grades and years.

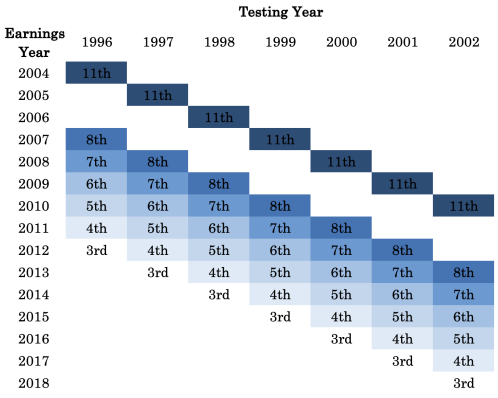
Panels (c) and (d) of Figure A.4 show analogous plots for earnings at age 30. At this age, panel (c) shows that some early grades tested at later years (e.g., third graders in 2002) could not feasibly have earnings matches as they would not have yet turned 30 by 2019. This is corroborated by panel (d), which 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 (e) and (f) show analogous plots for earnings at age 35. At this age, we are unable to match earnings for most grades and years, as shown by the low match rates in panel (f).

A.3 Booklet Data

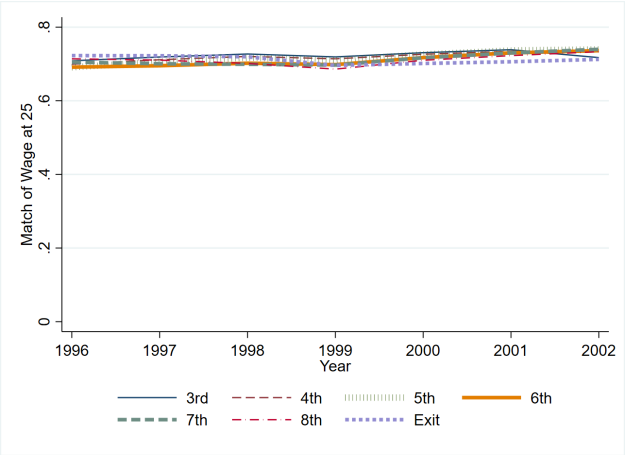
Table A.4 presents test booklet availability by grade-year. A check mark (✓) indicates that the test booklet for that grade-year was recovered and digitized; an en-dash (–) indicates not recovered. All test booklets in 1995–1996 are missing. For subsequent years, some grades are missing booklets. Overall, we were able to recover 76% of the test booklets for the years 1996-2002.

Figure A.4: Earnings Timelines for Test-takers

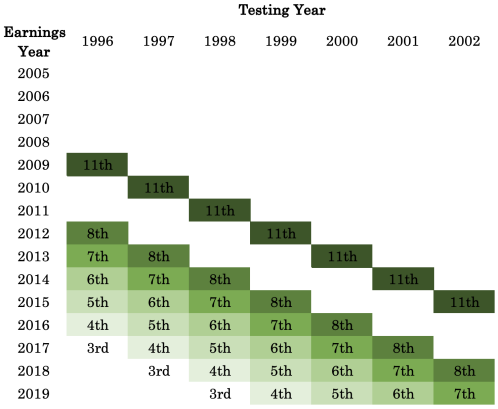
(a) Anchoring Timeline @ 25



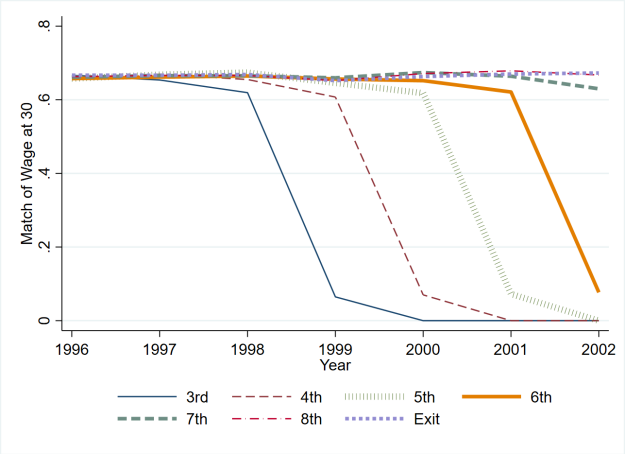
(b) Earnings Match Rate @ 25



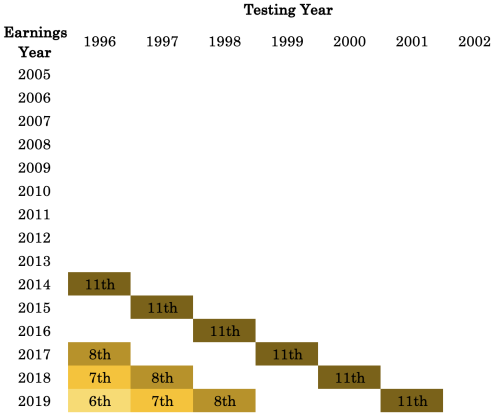
(c) Anchoring Timeline @ 30



(d) Earnings Match Rate @30



(e) Anchoring Timeline @ 35



(f) Earnings Match Rate @35

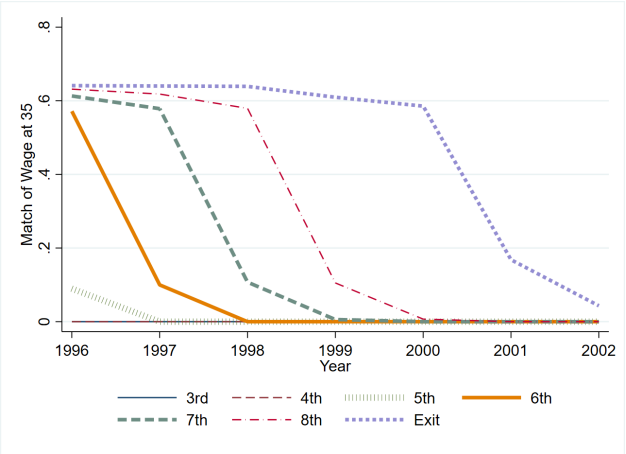


Table A.4: Booklet Availability

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

Notes: ✓ indicates the test booklet for that grade-year was recovered and digitized; – indicates not recovered. All 1995–1996 booklets are missing.

B Achievement Gaps and Attenuation Bias

To illustrate how item-anchored and given achievement gaps can differ, even when they are on the same scale, suppose that the (population) item-anchored (A_i) and given scores (Z_i) are both given by weighted sums of the items, where both sets of weights are normalized to sum to one:[64] $A_i = \sum_m \omega_m D_{i,m}$, $Z_i = \sum_m \alpha_m D_{i,m}$. 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

$$\sum_m (\omega_m - \alpha_m)(p_{m,H} - p_{m,L}). \tag{12}$$

This equation shows that, abstracting from scaling differences, item-anchored and given achievement gaps will differ 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 the mean for group-L students. However, this estimator will be biased – the irreducible error in $\hat{A}$ will tend to attenuate this naive estimator towards zero. Recall that the best we could identify is $\tilde{A}_i = A_i + \nu_i$.[65]

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

$$\mathbb{E}[S|\hat{A}_i] = R_{A,\nu}\hat{A}_i + (1 - R_{A,\nu})\bar{A}, \text{ where } R_{A,\nu} \equiv \frac{\sigma_A^2}{\sigma_A^2 + \sigma_\nu^2}. \tag{13}$$

Equation (13) 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

$$\text{plim}_{N\to\infty}(\hat{A}_H - \hat{A}_L) = R_{A,\nu}(\bar{A}_H - \bar{A}_L) < (\bar{A}_H - \bar{A}_L). \tag{14}$$

Equation (14) 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 ordinary 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,

[64] 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.

[65] If f is correctly specified, which is the assumption we maintain throughout this analysis, the total error in $\hat{A}$ will come from ν_i as well as through sampling variability in $\hat{\mathbf{\Psi}}$. This sampling error will not produce bias, but ν_i will, which is what the IV procedure is meant to correct.

via Equation 2, $\hat{\gamma}$ will be attenuated towards zero. To solve this errors-in-variables problem, we seek an instrument for S_i, some Z_i that is correlated with A_i and uncorrelated with η_i.

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

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

$$Z_i^{(1)} = N_i^{-1} \sum_{j \neq i : \hat{A}_j^{(2)} = \hat{A}_i^{(2)}} S_j, \tag{15}$$

where N_i is the number of students other than i satisfying the condition $\hat{A}_j^{(2)} = \hat{A}_i^{(2)}$. This condition ensures the relevance of the instrument, and exogeneity is guaranteed by the leave-one-out construction.

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

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

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

3. Construct $Z_i^{(1)}$ according to equation (15). 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 $\hat{\Delta}_{HL}/\hat{\gamma}_{IV}^{(1)}$.

Given the large sample sizes in our data, we modify this procedure slightly for computational speed. We randomly subdivide the observations into two groups G_1 and G_2. Then, in the construction of the instruments, we sum across j not in i's group rather than all observations not equal to i. In practice, this makes virtually no difference for either the estimated values of the instrument or the resulting estimated reliabilities.

Table B.1 shows that the even and odd items are extremely similar to each other on every psychometric and metadata dimension (besides question number, where the difference is mechanical given the size of the exams). This similarity justifies the subdivision of the TAAS exams into even and odd items.

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

	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		4,739	

Notes: Even-vs-odd item comparison: each row reports the mean of the listed variable over even-numbered items (column 1) and odd-numbered items (column 2); column 3 is the difference. Standard errors are clustered at the (grade, year, subject) level. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

C Assessing Different Anchor Models

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

C.1 Anchor Results Under Different Model Specifications

Table C.1 shows that we can rarely reject equality of $\hat{\omega}_m$ estimates between our main (preferred) specification and alternative specifications that include different observable controls in the anchor equation.

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

	(1)	(2)	(3)
Rejected	Main v. County FE	Main v. School FE	Main v. White Males
5%-level	0.13%	6.31%	1.12%
10%-level	0.19%	7.51%	1.50%

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

C.2 Monte Carlo Analysis

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

Data-Generating Process

We generate data according to the following process:

1. We fix N as the total number of students and M as the total number of items. We set $\bar{R}^2$ as the desired share of outcome variation due to the items.

2. Each student i's academic ability θ_i is drawn independently from $N(0,1)$.

3. For each item m, we set the IRT parameters and item response probabilities as follows:

   (a) Guessing: $c_m = 0.25,\ \forall m$

   (b) Discrimination: $a_m = 1,\ \forall m$

   (c) Difficulty: $b_m \sim N(0,1)$ drawn independently $\forall m$

   (d) For each (i,m), the probability of a correct response is given by the three parameter logistic IRT model:
   $$p_{i,m} = c_m + \frac{1 - c_m}{1 + e^{-a_m(\theta_i - b_m)}}.$$

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

5. For each item m, we construct the "linear" item outcome weights according to
   $$\omega_m^{(1)} = \mu + \gamma * b_m + \xi_m$$
   where $\xi_m \sim N(0, \sigma_1^2)$. In this formulation, μ gives the average outcome weight, and γ controls how strongly related are item difficulty and outcome weight.

6. Achievement under linearity is then defined by
   $$A_i^{(1)} = \sum_m \omega_m^{(1)} * D_{i,m}.$$

7. The observed outcome under linearity is
   $$S_i^{(1)} = A_i^{(1)} + \upsilon_i^{(1)}$$
   where $\upsilon_i^{(1)}$ is 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 $\psi^{(2)}_{m,m'}$ 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 $\psi^{(2)}_{m,m'} = 1$, we generate an interaction weight for (m, m') according to $\omega^{(2)}_{m,m'} \sim N(\mu_2, \sigma_2^2)$. If $\psi^{(2)}_{m,m'} = 0$, we set $\omega^{(2)}_{m,m'} = 0$.

10. The true achievement with interactions is then given by

$$A_i^{(2)} = A_i^{(1)} + \sum_m \sum_{m'>m} \psi^{(2)}_{m,m'} \omega^{(2)}_{m,m'} D_{i,m} D_{i,m'}.$$

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 $\{\omega_m^{(1)}, \omega^{(2)}_{m,m'}, \omega^{(3)}_{m,m',m''}, \psi^{(2)}_{m,m'}, \psi^{(3)}_{m,m',m''}, T^{(2)}, T^{(3)}\}$.

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 $\mathbf{D}$, 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, Breiman, 2001).

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

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

   where

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

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

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

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

Result 2: OLS and LASSO have similar bias.

Figure C.1 provides a representative picture of the OLS vs. LASSO comparisons in our Monte Carlo experiments. The left panel compares the error distributions for OLS and various lasso

models for a data generating process that features two-way item interactions but no higher-order interactions. The right panel shows analogous error distributions for a case with both two- and three-way item interactions. In both panels, it is clear that the OLS estimates have very similar mean errors as the lasso models. The LASSO models with interactions do have lower RMSEs; typically, the lasso models that are correctly specified perform best.

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

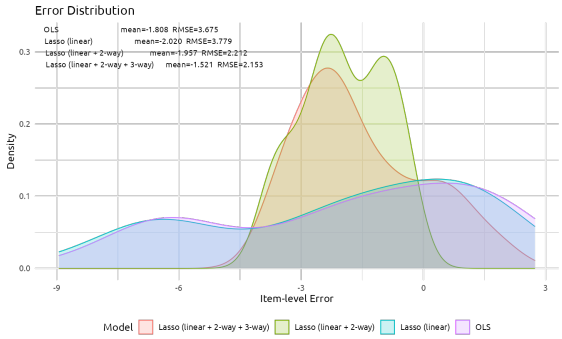


Figure C.2: Two-Way (N =20k; $M = 50$)

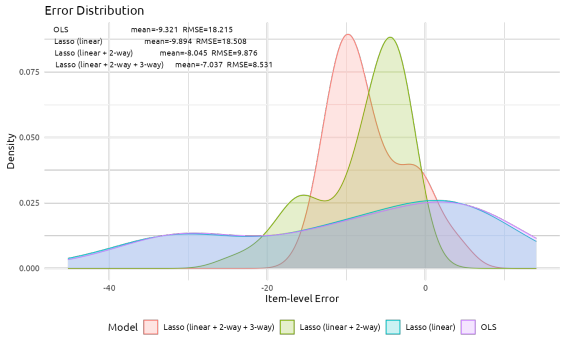


Figure C.3: Three-Way (N =20k; $M = 50$)

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

Result 3: OLS typically performs better than random forests.

Figure C.4 provides a representative picture of the OLS vs. random forest comparisons in our Monte Carlo experiments. The left panel compares the error distributions for OLS and various random forest models for a data generating process that features only two-way item interactions. The right panel shows analogous error distributions for a case with both two- and three-way item interactions. In both panels, it is clear that the OLS estimates have lower mean errors than any of the random forest estimates. Moreover, the OLS estimates also have lower RMSEs.

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

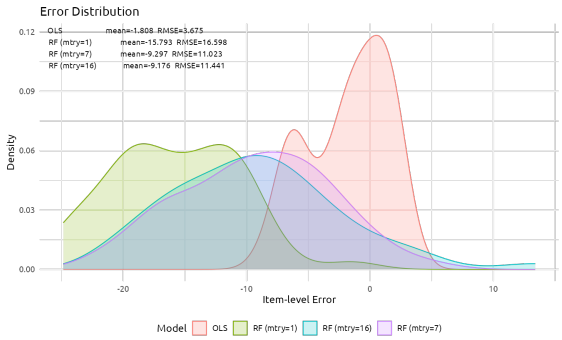


Figure C.5: Two-Way (N =20k; $M = 50$)

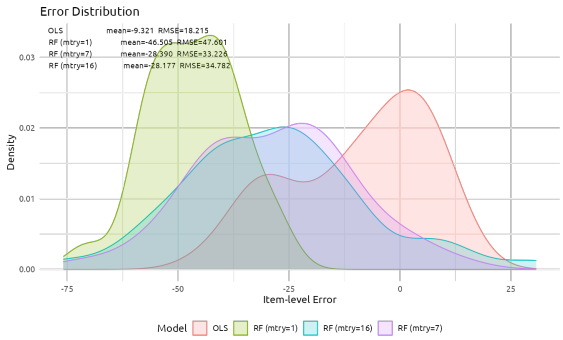


Figure C.6: Three-Way (N =20k; $M = 50$)

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

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

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

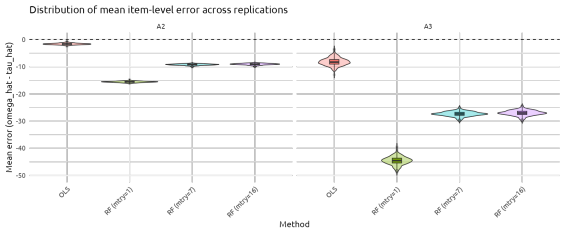


Figure C.8: Mean Errors (N =20k; $M = 50$)

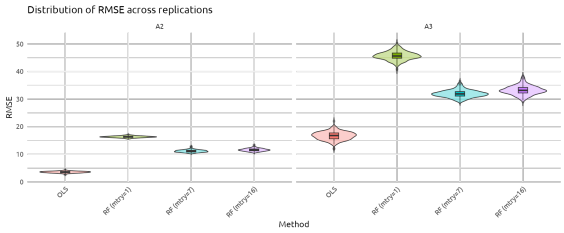


Figure C.9: RMSEs (N =20k; $M = 50$)

Notes: The left panel shows the distributions of mean errors for different models across 250 bootstrap iterations. "A2" corresponds to two-way-interaction data-generating processes, while "A3" corresponds to three-way-interaction processes. The right panel shows the corresponding RMSE distributions.

C.3 Double-Debiased Methods

The estimates for $\mathbf{\Omega}$ assessed above suffer from a number of well-known problems inherited from the ML models used to estimate f. First, because $\hat{f}$ is obtained from regularized/flexible ML models, the plug-in estimator for $\hat{\omega}$ can have non-negligible finite sample bias. Indeed, this is evident by the non-zero locations of the error distributions in Figures C.1 - C.7. Second, $\hat{\omega}_m$ averages over the marginal distribution of $\mathbf{D}_{-m}$, but the conditional distributions $\mathbf{D}_{-m}|D_m = 1$ and $\mathbf{D}_{-m}|D_m = 0$ might differ substantially. When the overlap in these distributions is limited, the plug-in estimator can be sensitive to extrapolation into regions where either $D_m = 1$ or $D_m = 0$ is rare.

Chernozhukov et al. (2018) develop a double/debiased machine learning (DML) approach that corrects for model bias and reweights to address imbalance-induced extrapolation bias.[66] The DML estimator adds to the simple plug-in estimator $\hat{\omega}$ propensity-reweighted residual corrections with cross-fitting to evaluate nuisance estimates out-of-sample.

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

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

We estimate the nuisance functions $\mu_{m,1}(x)$ and $\mu_{m,0}(x)$ via OLS, LASSO, or random forests as the case may be. The propensity scores we estimate via lasso allowing for 2-way interactions.[67] Then,

[66] 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).

[67] The DML estimates when the propensities are estimated as linear functions of the items perform slightly worse in terms of noise and bias.

the per-observation score contribution is

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

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

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

Figure C.10 shows the error distributions for the OLS and random forest models estimated using the above DML procedure. Compared to Figure C.4, the DML estimates are less biased and have lower RMSEs for every model specification. However, OLS continues to perform very well against the random forest models – OLS has the lowest bias and RMSE in the 2-way case and a still-low bias with the lowest RMSE in the 3-way case.

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

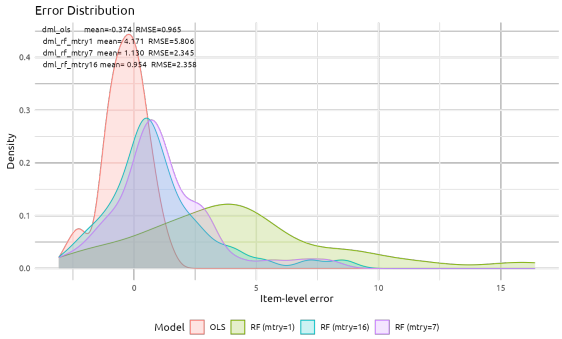


Figure C.11: Two-Way (N =20k; $M = 50$)

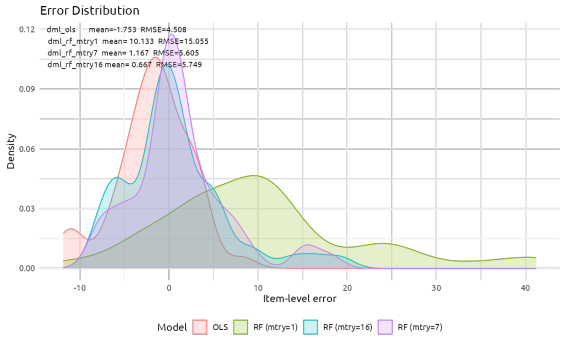


Figure C.12: Three-Way (N =20k; $M = 50$)

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

Figure C.13 likewise shows the error distributions for the OLS and LASSO models estimated using the above DML procedure. OLS continues to perform very well – the OLS estimates have

lower bias than the higher-order LASSO estimates while still achieving low RMSEs.

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

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

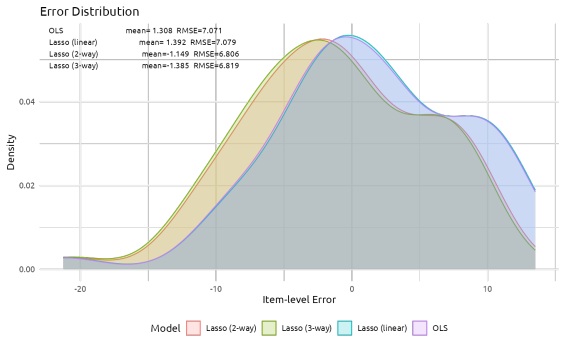


Figure C.14: Two-Way (N =20k; $M = 50$)

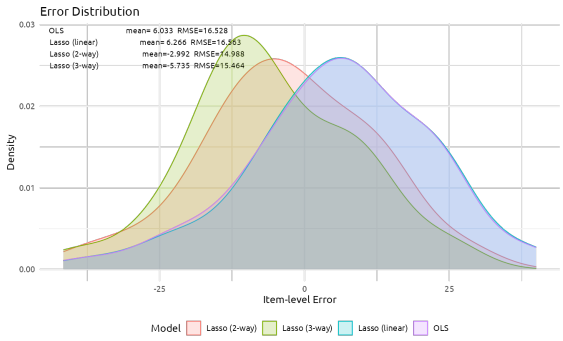


Figure C.15: Three-Way (N =20k; $M = 50$)

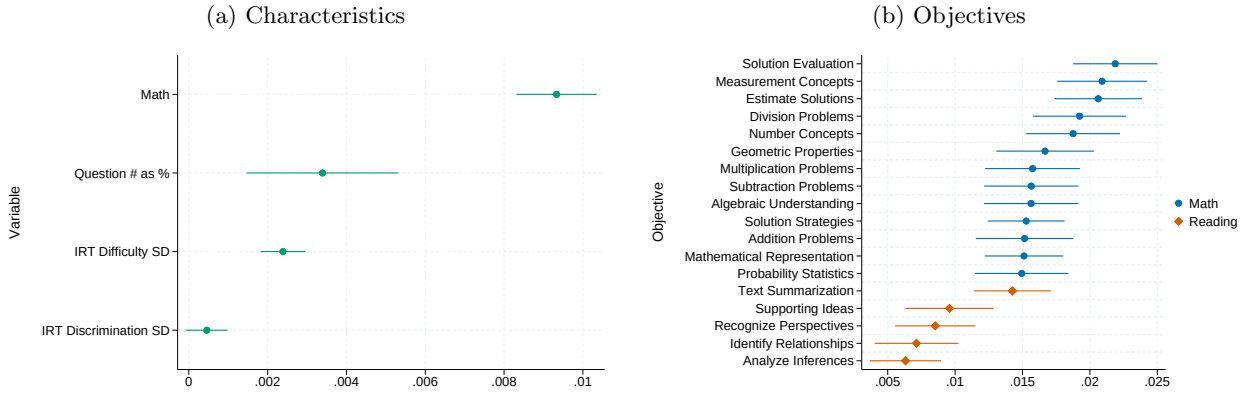
Notes: The left panel shows the error distributions for $\widehat{\boldsymbol{\Omega}}$ for the case with two-way interactions. The right panel shows the same but for data-generating processes featuring both two-way and three-way interactions.

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.

D Robustness Checks

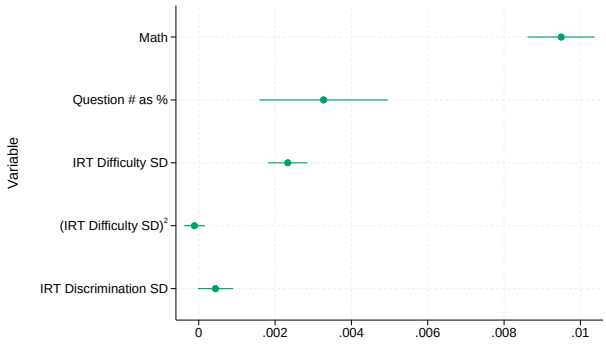
This section presents a number of pertinent robustness analyses. Figure D.1 shows that the conclusions presented in Figure 5 are unchanged when we restrict the sample to items we were able to digitize. Figure D.2 repeats the analysis from panel (a) in Figure 5 but with squared difficulty added as an additional regressor. Together, these figures show that results in Figure 5 are robust to different item inclusion criteria and different plausible ways to control for psychometric characteristics. Figure D.3 demonstrates that, for both math and reading, there is no clear relationship between the CCSS prices and the softmax-weighted psychometric characteristics (difficulty, discrimination, item position) of the items mapped to each skill.

Figure D.1: Relationship of Item Characteristics to $\hat{\boldsymbol{\Omega}}$ - Digitized Sample



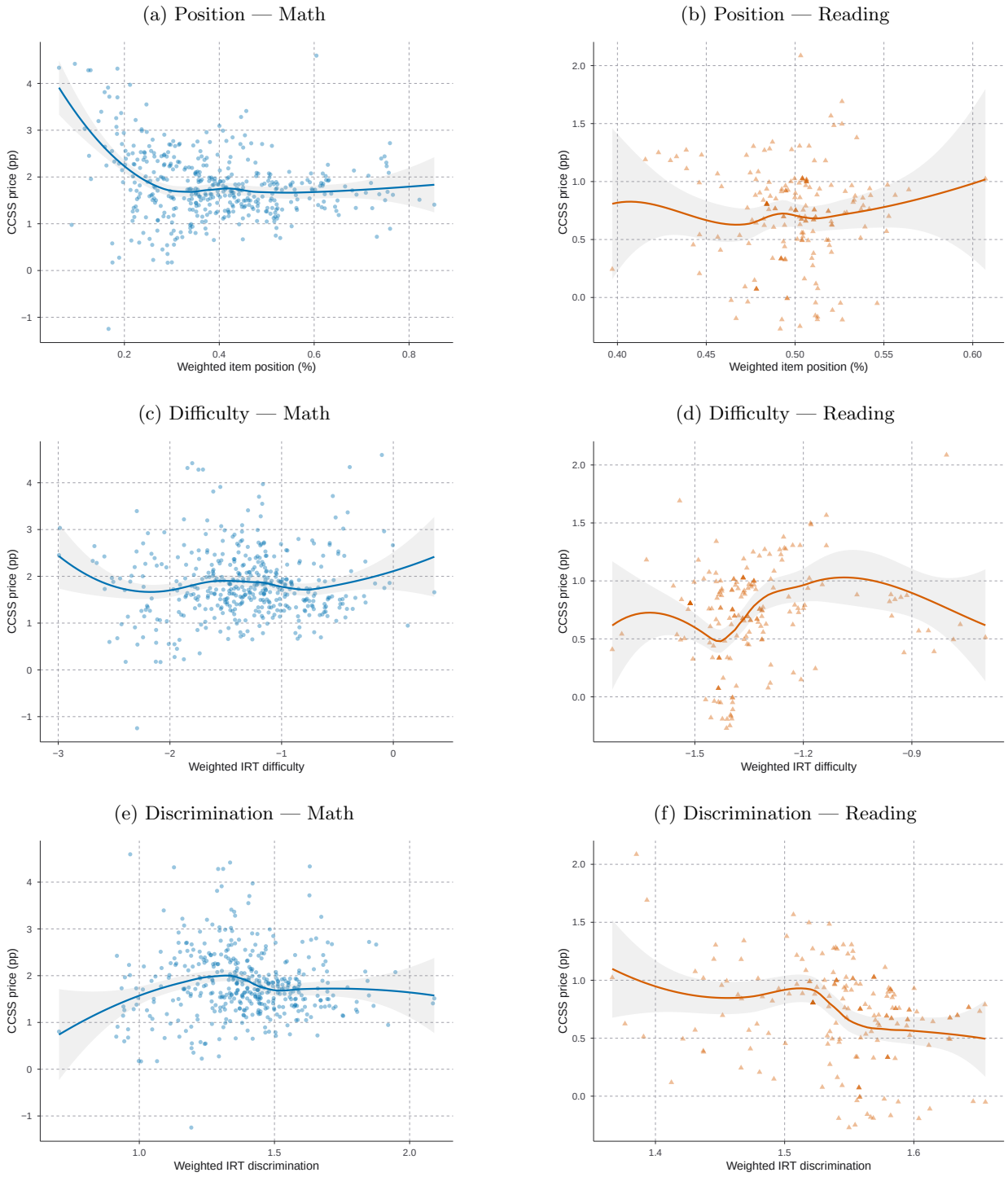
Notes: This figure plots analogous results to Figure 5 for the subset of items for which we were able to collect text data. Panel (a) of this figure shows estimates of regression coefficients of estimated item-level prices on observable item-level characteristics. All regressions have grade and year fixed effects and are weighted by the inverse of the square of the SE of the item prices 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".

Figure D.2: Relationship of Item Characteristics to $\hat{\boldsymbol{\Omega}}$ - Quadratic-in-Difficulty



Notes: This figure plots analogous results to Figure 5 while including a quadratic-in-difficulty term to control for the information content of each item. All regressions have grade and year fixed effects and are weighted by the inverse of the square of the SE of the item prices to account for estimate precision (Hedges and Olkin, 2014).

Figure D.3: CCSS Prices and Item Psychometric Characteristics



Notes: Each panel plots the CCSS estimated price ($\tilde{\omega}_j$, in percentage points) against the softmax-weighted average of the indicated item characteristic ($\beta = 45$). Rows: item position in the test, IRT difficulty, IRT discrimination. Columns: math (left) and reading (right) standards. Fitted lines are local polynomial (loess) regressions with 95% confidence bands.

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

This appendix explains how we assess the predictive performance of the machine learning models in Section 7 given that the prediction target, $\widehat{\boldsymbol{\Omega}}$, is itself estimated. Because $\widehat{\boldsymbol{\Omega}}$ contains sampling error, a naive out-of-sample R^2 will understate how well an ML model predicts $\boldsymbol{\Omega}$. Even a model that perfectly predicts $\boldsymbol{\Omega}$ would generally have an observed R^2 below one when evaluated against $\widehat{\boldsymbol{\Omega}}$. The discussion below focuses on out-of-sample predictions unless otherwise stated. Throughout, we impose the standard cross-fitting independence condition: for each item 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 $\boldsymbol{\Omega}$, we treat h as fixed with respect to the first-stage estimation error $e = \widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}$, so $\mathbb{E}[e \mid \boldsymbol{\Omega}] = 0$ and $\mathrm{Cov}(e, h \mid \boldsymbol{\Omega}) = 0$ (see Chernozhukov et al. (2018), Bach et al. (2024)).

Notationally, let b denote a given grade-year combination, and let $\widehat{\boldsymbol{\Omega}}_b$ and $\widehat{\boldsymbol{\Sigma}}_b$ denote the coefficient and variance-covariance estimates from anchor model b (for a fixed anchor outcome), both of which are consistent estimators of $\boldsymbol{\Omega}_b$ and $\boldsymbol{\Sigma}_b$, 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:

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

We use an equivalent notation for the true $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$. Define $N = \sum_b N_b = \sum_b |\widehat{\boldsymbol{\Omega}}_b|$. Also define the centering matrix $\mathbf{C} \equiv \mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}'$.[68]

The observed variance of $\widehat{\boldsymbol{\Omega}}$ will thus be

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

Consider next the expected value of $V(\widehat{\boldsymbol{\Omega}})$ conditional on the unobserved $\boldsymbol{\Omega}$:[69]

[68] 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 $\mathbf{C}/N$ for $\mathbf{P}$ such that $\mathbf{P} = \mathrm{diag}(\varphi) - \varphi\varphi'$. One such matrix would be the inverse-variance weighting matrix where $\varphi_i \propto 1/\Sigma_{ii}$, such that item prices that are more precisely estimated get more weight. Another alternative would be to use a Mahalanobis metric, penalizing errors according to the full covariance of estimation uncertainty, not just the diagonal.

[69] The second step comes from the fact that for any random vector $\mathbf{X}$ with mean m and covariance $\mathbf{S}$ and any symmetric matrix $\mathbf{A}$, $\mathbb{E}[\mathbf{X}'\mathbf{A}\mathbf{X}] = tr(\mathbf{A}\mathbf{S}) + m'\mathbf{A}m$.

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

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

Next, 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

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

Taking expectations,

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

where "$\approx$" denotes 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

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

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

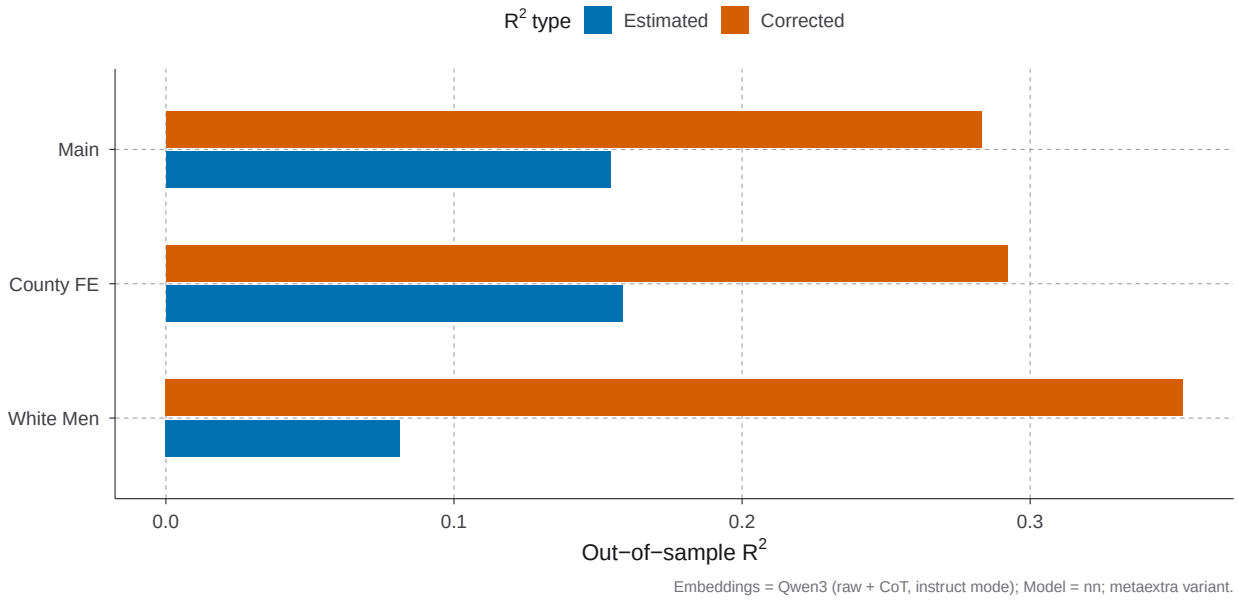
$$\widehat{V}_{\text{signal}} = \widehat{\mathbf{\Omega}}'\mathbf{C}\widehat{\mathbf{\Omega}} - tr(\mathbf{C}\widehat{\mathbf{\Sigma}}).$$

Putting it all together, our correct R^2 is given by

$$R^2_{\text{corrected}} = 1 - \frac{\widehat{\text{SSE}}_{\text{true}}}{\widehat{V}_{\text{signal}}} = 1 - \frac{(\widehat{\boldsymbol{\Omega}} - h)'(\widehat{\boldsymbol{\Omega}} - h) - tr(\widehat{\boldsymbol{\Sigma}})}{\widehat{\boldsymbol{\Omega}}'\mathbf{C}\widehat{\boldsymbol{\Omega}} - tr(\mathbf{C}\widehat{\boldsymbol{\Sigma}})}. \tag{16}$$

Figure E.1 compares the corrected and uncorrected out-of-sample R^2s for models that predict $\widehat{\boldsymbol{\Omega}}$ as a function of the item-text embeddings only. This is in contrast to the models in Section 7 which include psychometric characteristics and test metadata in addition to the item text embeddings. This figure shows the importance of correcting for the first-stage estimation error in $\widehat{\boldsymbol{\Omega}}$—the corrected R^2s for the embeddings-only models 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 subset of the full analysis sample. Interestingly, the corrected R^2s are quite similar across different anchor models, unlike for the uncorrected R^2s. Additionally, the high corrected R^2s demonstrate that the item texts alone can explain a substantial share of all item price variation in our setting.

Figure E.1: Sampling-Error Correction Applied to Item-Price R^2



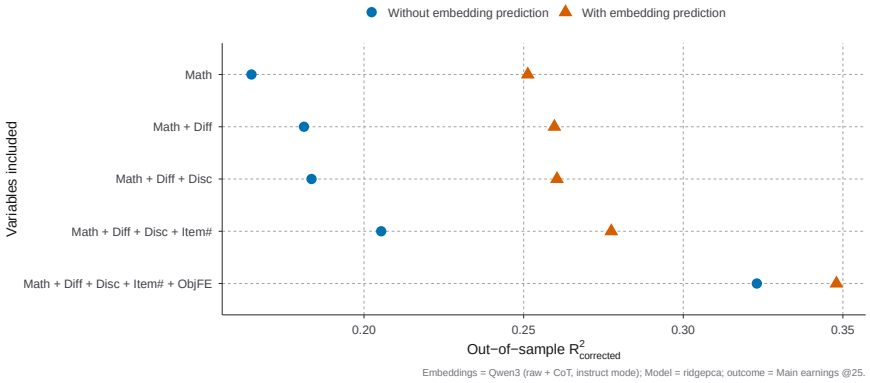
Notes: Estimated (uncorrected) and sampling-error-corrected out-of-sample R^2 using only item-text embeddings for the three $\widehat{\boldsymbol{\Omega}}$ specifications used in the paper (Main, County FE, White Men), all at age 25. The prediction model is the same neural network on Qwen3 metaextra embeddings (raw question + CoT-extracted skills) used in Figure 6. The correction (defined in Equation 16) increases the apparent fit by 50–200% across specifications, with the largest increase for the "White Men" $\widehat{\boldsymbol{\Omega}}$ which is estimated on a smaller sample.

F Alternative Model: Ridge + PCA on Qwen3 Embeddings

Figure 6 uses a neural network to flexibly estimate $g(\mathbf{R}, \mathbf{E})$. As a robustness check, we re-estimate the same five nested specifications using ridge regression on the leading 20 principal components of the Qwen3 metaextra embedding matrix, similar to Gilbert et al. (2025). The principal components are computed on the training fold, and the ridge tuning parameter is selected by 5-fold cross-validation within the training set. Both the embedding inputs and the OOS evaluation procedure are identical to the main analysis; only the prediction model differs.

Figure F.1 shows the analogous results to Figure 6. The qualitative pattern is the same: each predictor set's corrected R^2 increases substantially when the embedding prediction is added, and the magnitude of the lift is similar across specifications. While the ridge + PCA estimator has a slightly lower R^2 than the neural network (0.245 vs. 0.283 corrected), it nonetheless produces qualitatively identical conclusions about the incremental information contained in the item texts.

Figure F.1: Embeddings Recover New Elements that Explain Prices — Ridge + PCA Robustness

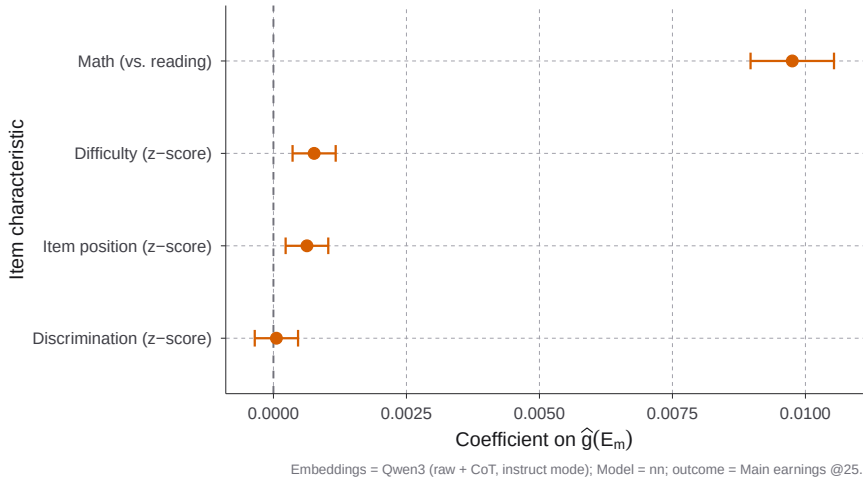


Notes: This figure is the counterpart to Figure 6 but using ridge regression on the leading 20 principal components of the Qwen3 metaextra embeddings in place of the neural network. The outcome, five nested predictor sets, and sampling-error correction are the same as in Figure 6. Blue circles report corrected out-of-sample R^2 values without the embedding-based prediction; orange triangles report the corresponding values after adding the embedding-based prediction.

F.1 What Does the Item-Text Prediction Load On?

To better understand what information the text channel encodes, Figure F.2 reports coefficients from a regression of the NN's item-text prediction $\widehat{g}(\mathbf{E}_m)$ on observable item characteristics: a math indicator, IRT difficulty, IRT discrimination, and item position within the test (each continuous regressor in z-score units). The largest loading is on the math indicator; item difficulty and item position carry smaller but statistically distinguishable coefficients, while item discrimination is indistinguishable from zero. The R^2 of this regression is approximately 0.49, indicating that more than half of the variation in $\widehat{g}(\mathbf{E}_m)$ is unexplained by these observable item-level characteristics.

Figure F.2: Text Predictions of $\widehat{\omega}_m$ and Item Characteristics



Notes: Coefficients (with 95% confidence intervals) from a regression of the NN's item-text-based prediction $\widehat{g}(\mathbf{E}_m)$ on item-level observables: a math indicator, IRT difficulty (z-score), IRT discrimination (z-score), and item position within the test (z-score of percent-of-test). Outcome is log earnings item prices from our main specification; embeddings are Qwen3 (raw question + chain-of-thought-extracted skills, instruct mode).

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

G.1 Neural Network Architecture

For the NN model we use a fully connected feedforward 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 L_2 (ridge) penalties to all dense layers. Both the dropout rates and the L_2 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.

G.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.

H CCSS Skill Analysis Details

Figure H.1 repeats the analysis in Figure 11 but where all four dimensions (grade, routine intensity, DOK, and spatial) are estimated jointly in a single regression (with each dimension interacted with subject). These results are qualitatively very similar to the individual regression results in Figure 11.

Figure H.2 presents the heat map of mean CCSS prices by routine intensity and DOK for reading. This figure makes clear that within each DOK level, mean prices are higher when routine intensity is lower. The highest mean price cell is (routine intensity = 2, DOK = 3) followed by (routine intensity = 1, DOK = 2), confirming that higher DOK and lower routine intensity skills tend to have the highest prices.

Figure H.3 shows the distribution of all three classification dimensions across math and ELA standards. Figure H.4 shows precision-weighted correlation matrices for the skill dimensions.

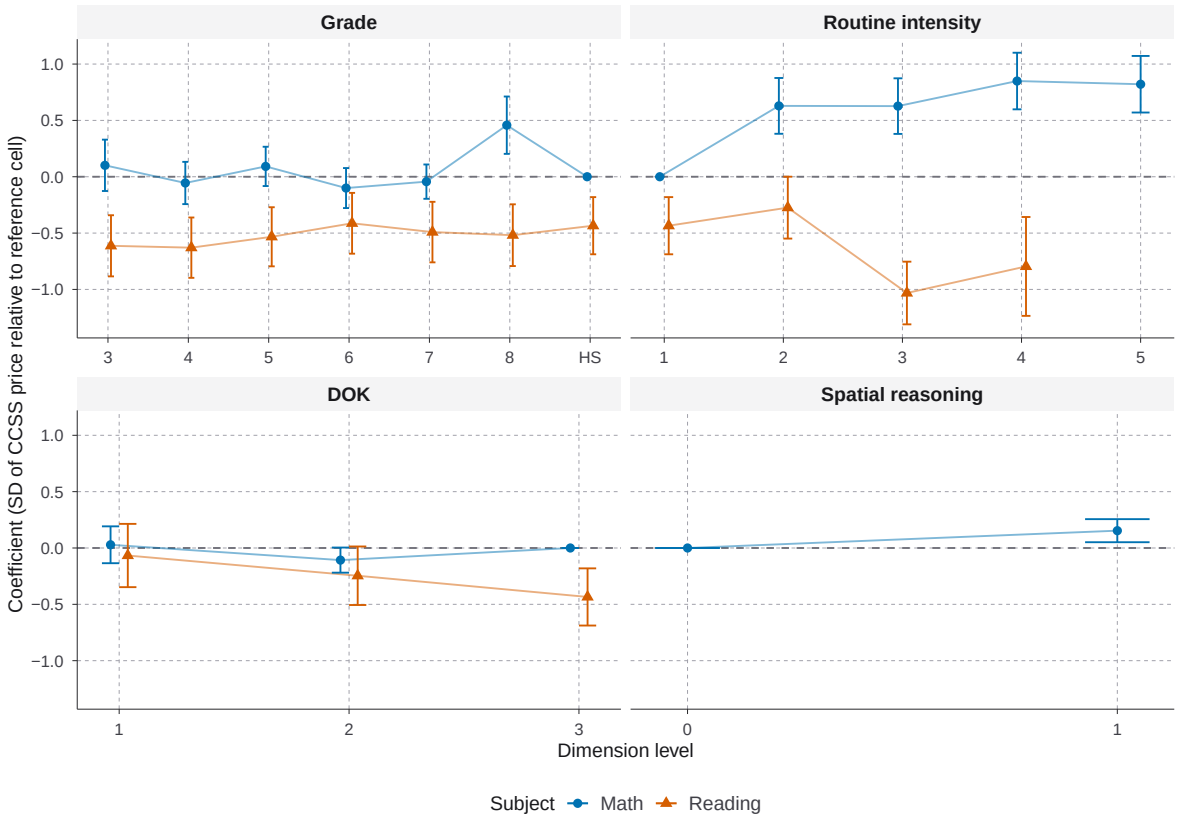
Turning to the softmax mapping procedure, Figure H.5 shows the full CV curves for both the softmax (β) and power (p) kernels. For softmax, the lowest MSE is achieved at $\beta = 45$, but we find broadly similar performance for a range of scale parameters near the optimum. For the power kernel, we find a minimum at around $p = 30$, with similar MSEs in a range around this value. Figure H.6 shows the relationship between CCSS prices estimated under the softmax kernel and the power kernel. The correlation between the CCSS skills prices across both methods is about 0.98. The resulting mappings are thus quite similar under either method.

The inclusive value $IV_j = \log \sum_m \exp(\beta \cdot C_{m,j})$ measures how well each CCSS standard is matched to the item pool. Standards with higher IV have at least some items that match well; those with low IV are effectively unidentified in the data. Figure H.7 shows the distribution. Most standards have IV values concentrated in a narrow range, suggesting that the embedding space provides reasonable matches for the majority of standards.

Figure H.8 shows the distribution of CCSS prices by subject under the softmax kernel. The higher average price for CCSS math skills is evident in the figure. However, despite the dominance of math, the figure also shows that some reading skills have higher prices than many math skills.

Figure H.9 shows that the relationships between the routine intensity, DOK, and spatial dimensions and the estimated CCSS prices are very similar for item skill prices estimated on the full sample using our preferred specification and those estimated on the white male subsample of the data. Moreover, the high correlation ($r = 0.82$, $\rho = 0.88$) in the estimated CCSS prices, presented

Figure H.1: Skill Dimensions and CCSS Prices, Joint Specification

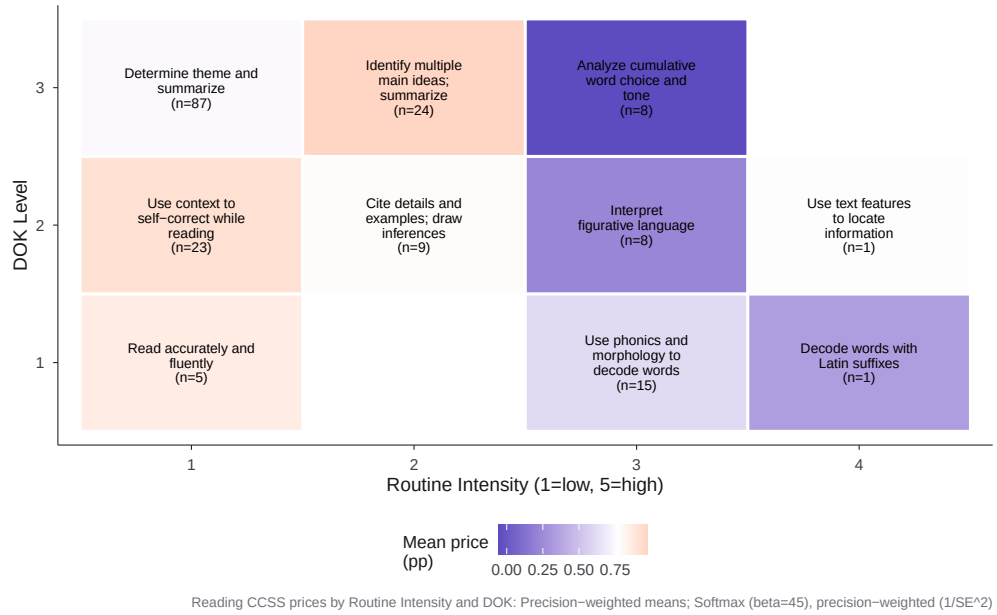


Notes: Estimates from a single precision-weighted regression of CCSS price on factor indicators for routine intensity, DOK, spatial reasoning, and CCSS grade, fully interacted with subject. The reference cell is math at grade = HS with routine intensity = 1, DOK = 3, and spatial = 0; all coefficients are plotted as deviations from that cell. The reading line in each panel sits below the math line by the (Reading − Math) gap evaluated at the reference cell, with the panel-specific level effect added. Spatial reasoning is classified for math standards only; reading is therefore omitted from the spatial panel. Empty (subject, level) cells—e.g. reading at routine intensity = 5, where no standards are classified—are also omitted. Per-dimension individual regressions are reported in Figure 11.

in Figure H.10, further confirms that the skill-price patterns are quite stable across estimation samples for the item anchor models.

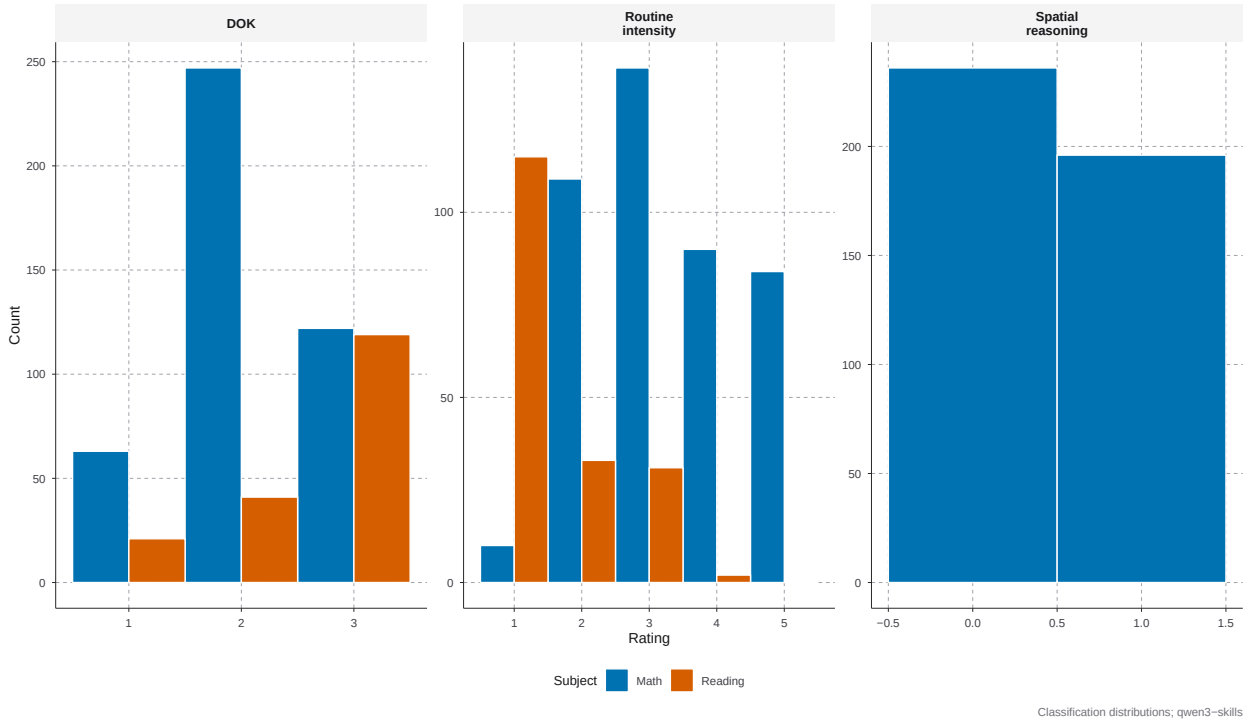
Finally, Figure H.11 shows that there is a strong correlation between using the skills extracted by Qwen3, our preferred model, or an alternative model (o3).

Figure H.2: Reading CCSS Prices by DOK and Routine Intensity



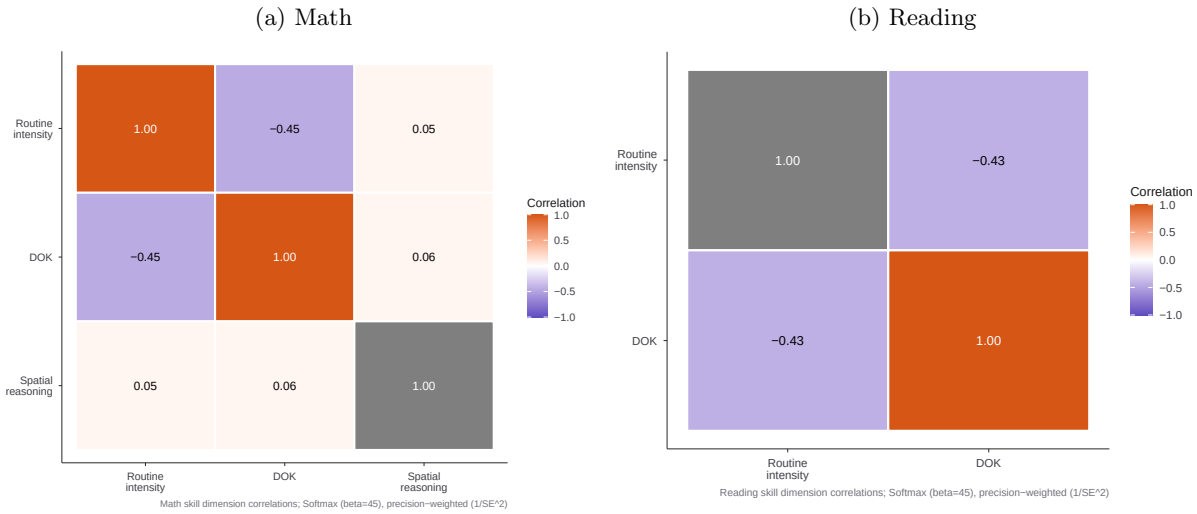
Notes: Precision-weighted mean price (pp) in each DOK × routine intensity cell. One example skill per cell (highest maximum cosine similarity to any item). Color scale: purple = below average, orange = above average.

Figure H.3: Distribution of Skill Classifications



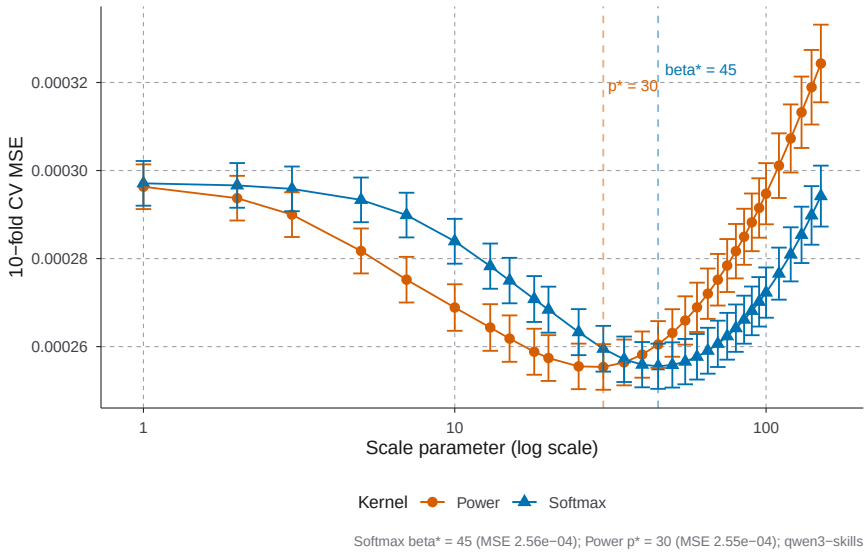
Notes: Distribution of LLM-classified skill dimensions across 613 CCSS standards. Math: blue; ELA: orange. Spatial reasoning is binary (0/1, math only). Routine intensity (1–5) and DOK (1–3).

Figure H.4: Skill Dimension Correlations



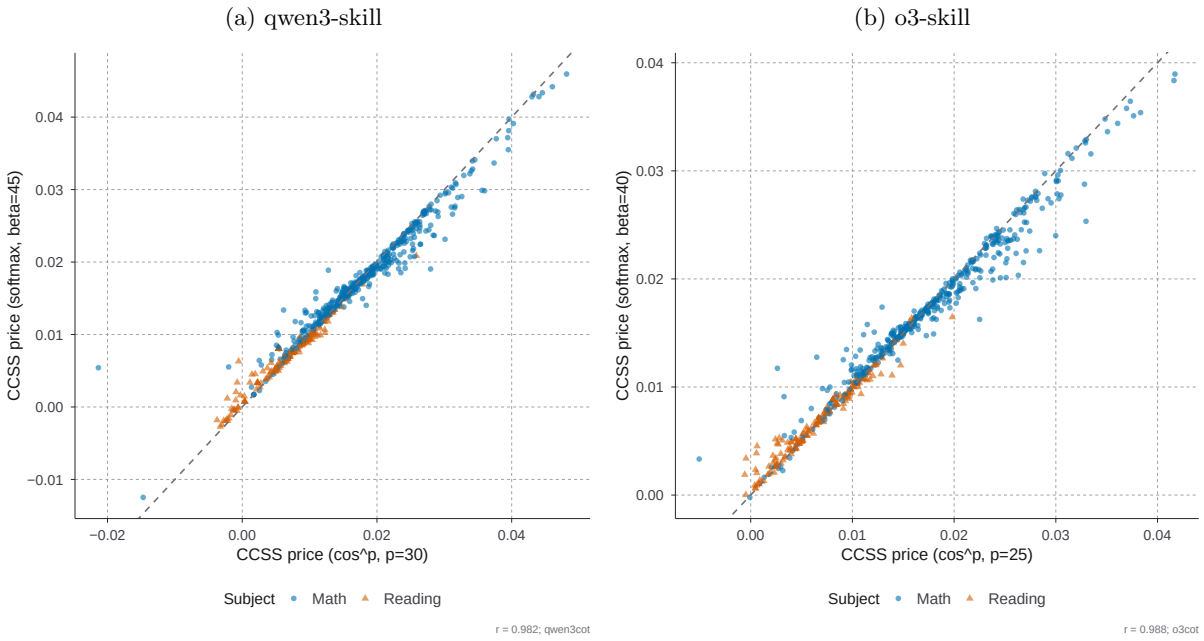
Notes: Precision-weighted pairwise correlations among the LLM-classified skill dimensions (routine intensity, DOK, and—for math—spatial reasoning) across the 613 CCSS standards, shown separately for math (left) and reading (right). Spatial reasoning is classified for math standards only and therefore does not appear in the reading panel.

Figure H.5: Cross-Validation: Softmax vs. Power Kernel



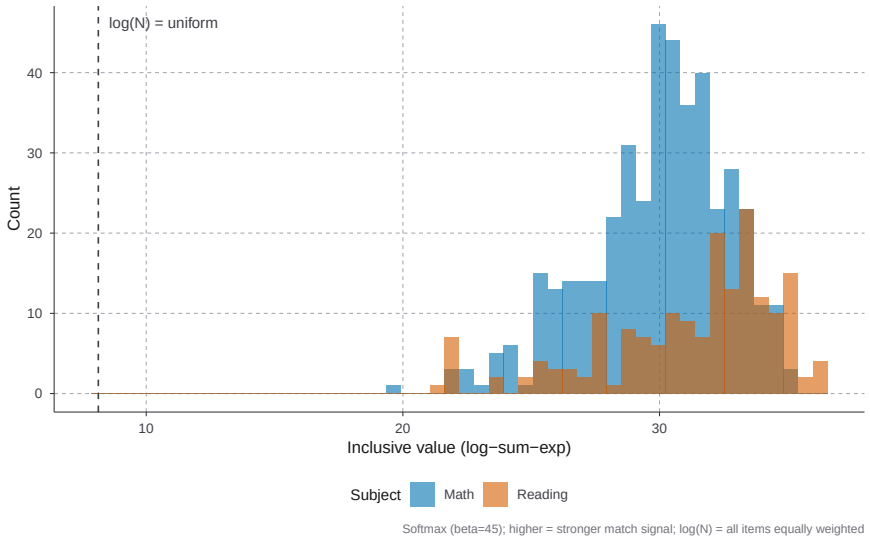
Notes: This figure presents the 10-fold cross-validation MSE as a function of β and p. Each fold holds out a stratified sample of items, computes CCSS prices from the training set, and predicts held-out item prices. Bars represent +/- 1 SD of mean MSE across folds.

Figure H.6: Softmax vs. Power Kernel CCSS Prices



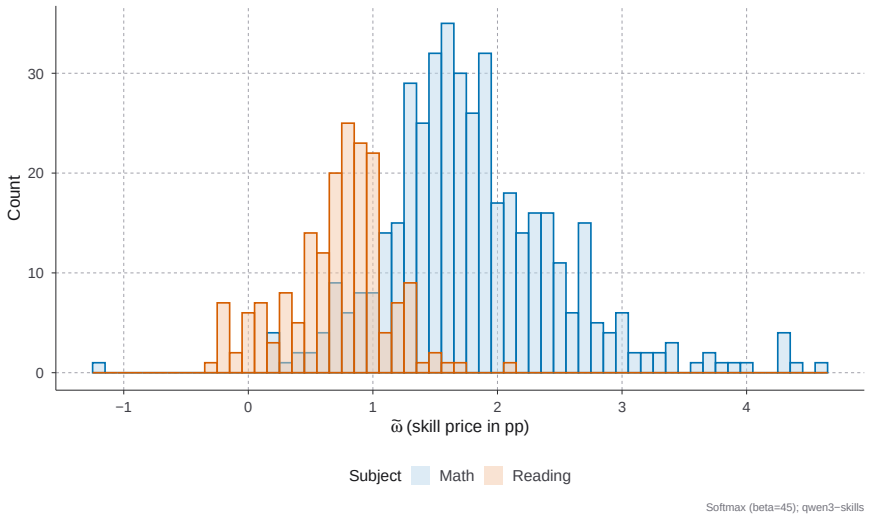
Notes: Each point is a CCSS standard; axes are its estimated price (pp) under the softmax kernel ($\beta = 45$) versus the power kernel ($p = 30$). Panel (a) uses Qwen3-extracted skills; panel (b) uses o3-extracted skills. The dashed line is the 45° line. Prices are highly correlated across kernels ($r \approx 0.99$).

Figure H.7: Inclusive Value Distributions



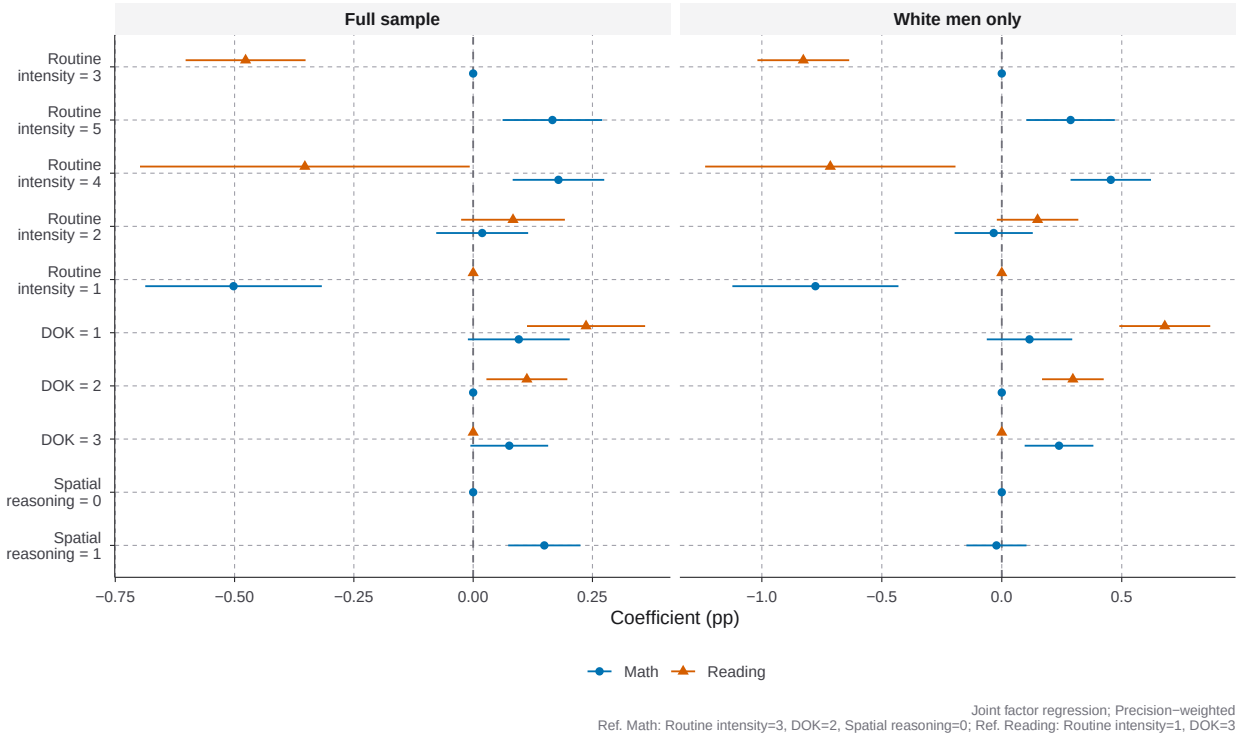
Notes: Distribution of the inclusive value $IV_j = \log \sum_m \exp(\beta \cdot C_{m,j})$ ($\beta = 45$) across CCSS standards, where $C_{m,j}$ is the cosine similarity between item m and standard j. Higher values indicate standards matched well by at least some items; low values indicate standards that are effectively unidentified in the item pool.

Figure H.8: CCSS Price Distributions by Subject (Softmax)



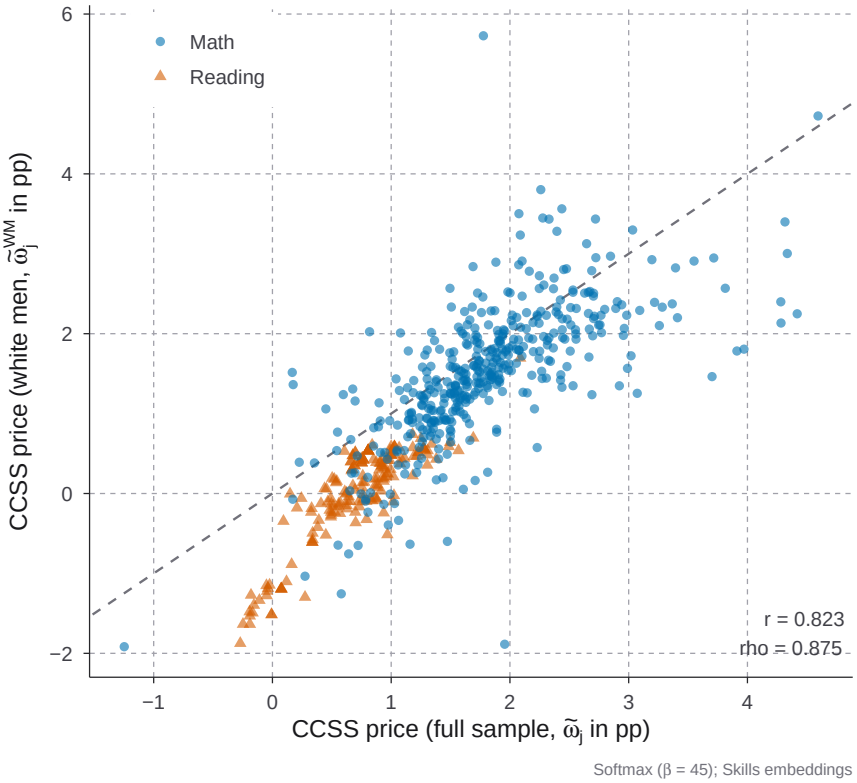
Notes: Distribution of estimated CCSS prices ($\tilde{\omega}_j$, in percentage points) under the softmax kernel ($\beta = 45$), by subject (math vs. reading). Qwen3-extracted skills.

Figure H.9: Characteristics that explain High CCSS Prices: Main Sample vs White Men Sample



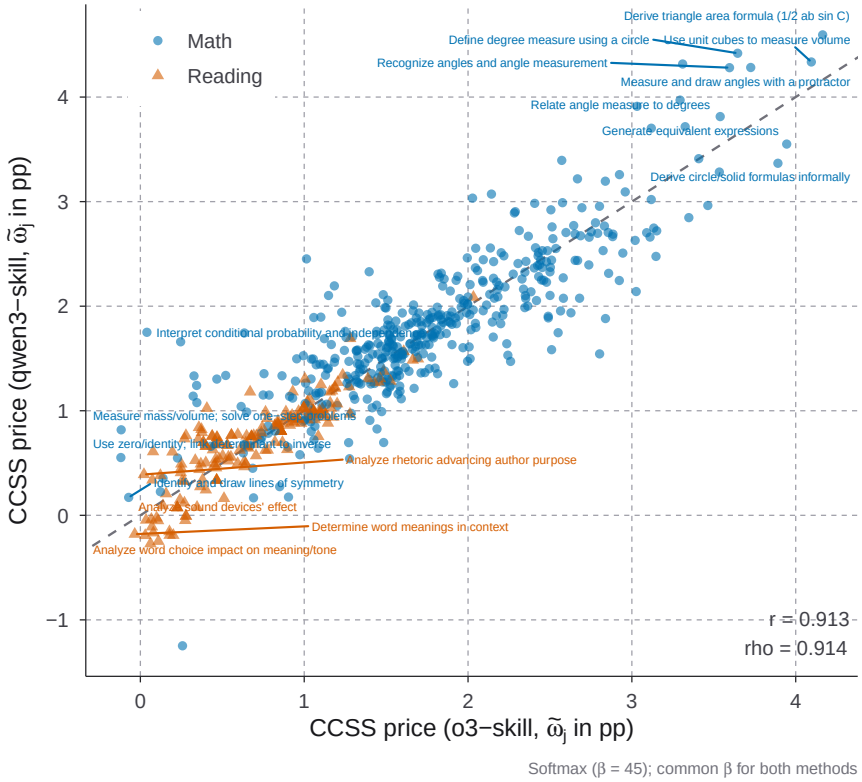
Notes: Precision-weighted joint-regression coefficients of CCSS price on factor indicators for routine intensity, DOK, and spatial reasoning, estimated separately on the full sample ($\tilde{\omega}_j$) and the white-men subsample ($\tilde{\omega}_j^{WM}$). Coefficients are plotted as deviations from the modal level of each dimension (the reference cell); spatial reasoning is classified for math standards only. Softmax kernel ($\beta = 45$); Qwen3-extracted skills.

Figure H.10: CCSS Prices: Main Sample vs White Men



Notes: Each point is a CCSS standard; axes are its estimated price (pp) under the full-sample anchor model ($\tilde{\omega}_j$) versus the white-men anchor model ($\tilde{\omega}_j^{WM}$). The dashed line is the 45° line; prices are highly correlated across samples ($r = 0.82$, $\rho = 0.88$). Softmax kernel ($\beta = 45$); Qwen3-extracted skills.

Figure H.11: CCSS Skill Prices: Qwen3 vs. o3



Notes: Estimated prices for CCSS skills across different skill extraction models (Qwen3 vs o3).