

Skill Measurement, Returns and Mismatch: Evidence from Narrative Evaluations*

Marco A. Badilla Maroto[†]

October 26, 2026

Abstract

Academic transcripts typically compress student performance into grades and GPA, obscuring other dimensions observed by instructors. This paper develops measures of cognitive and fine-grained soft skills from narrative evaluations: paragraph-length written assessments provided by instructors for students in each course. I study the University of California, Santa Cruz, where narrative evaluations historically coexisted with letter grades and were later discontinued, and link academic records to California administrative wage data from 2015 to 2023. Narrative evaluations contain predictive information beyond conventional GPA: a one-standard-deviation increase in a student's narrative evaluation relative to classmates predicts approximately 1.9 percent higher wages after controlling for GPA. I then use narrative text to construct student-level measures of cognitive skills and thirteen soft skills. Cognitive skills carry an approximately 4 percent wage premium. Among soft skills, self-control ranks highly for both labor force participation and wages, while leadership and teamwork rank more highly for participation and integrity and perceptiveness for wages. These patterns suggest that the skills related to obtaining and maintaining employment are not necessarily the same as those rewarded through higher wages. Finally, mismatch between a student's revealed skill profile and the typical skill bundle of the student's major predicts lower wages: a 10-point increase in mismatch is associated with approximately 1.3 percent lower wages, while its relationship with labor force participation is close to zero. Together, the results show that narrative evaluations contain information about economically relevant skills and their alignment with educational choices that is not captured by conventional grades alone.

*I am especially grateful to Sydnee Caldwell, Jesse Rothstein, and Christopher Walters for their invaluable guidance and support. I also thank Benjamin Faber, Patrick Kline, and participants in the UC Berkeley Economics Seminars for their helpful comments and feedback. I am grateful to Zachary Bleemer and UC Santa Cruz for providing access to the data and for their support. This research was supported by the American Educational Research Association through its AERA-NSF Grants Program, funded by the National Science Foundation under award NSF-DRL #2427569. Opinions reflect those of the author and do not necessarily reflect those of AERA, UC Santa Cruz or the NSF. All results have been reviewed to ensure that no confidential information is disclosed. All remaining errors are my own.

[†]UC Berkeley. Email: mbadilla@berkeley.edu

1. Introduction

The labor market rewards more than grades—it rewards skills. Yet most academic records compress student performance into a GPA, offering little visibility into the specific cognitive and soft skills students demonstrate in the classroom. This limitation is increasingly important given the growing labor market returns to soft skills (Deming (2017), Edin et al. (2022)) and their complementarity with cognitive abilities (Weinberger (2014), Deming and Kahn (2018)). Despite this growing interest, most empirical work treats soft skills as a single, undifferentiated construct—often broadly labeled “social skills”—rather than unpacking their distinct dimensions and relative economic value. Moreover, existing measures are typically drawn from the demand side, such as job postings, which primarily capture employer demand; from self-reported platforms such as LinkedIn, which may reflect self-presentation rather than directly demonstrated skills; or from letters of recommendation, which are typically available only for a selected subset of students and therefore provide an incomplete picture.

In this paper, I develop a novel approach to measuring cognitive skills and fine-grained soft skills—spanning both social and non-social dimensions—by extracting them from narrative evaluations: detailed written assessments provided by professors on student performance in each class at the University of California, Santa Cruz (UCSC). This setting is uniquely suited for such an analysis, as narrative evaluations are written systematically for each student, generated class-by-class by instructors, and provide instructor assessments of demonstrated performance at the individual level. Narrative evaluations historically served as UCSC’s primary assessment system; letter grades were expanded to most courses in 1997, became mandatory for students admitted in 2001 and later, and narratives were effectively discontinued in 2010. I leverage this institutional setting to address three central questions. First, do narrative evaluations contain predictive information about labor market outcomes beyond what is summarized by conventional GPA, and do the two grading systems provide complementary, substitutable, or approximately independent signals? Second, which specific soft skills are most strongly associated with labor market outcomes? Third, how are labor market outcomes related to misalignment between students’ revealed skill profiles and the dominant skill orientation of their chosen major? More broadly, I also examine the returns to cognitive and soft skills, their joint predictive patterns and heterogeneity across gender and underrepresented minority (URM) groups. This approach allows me to recover a richer, instructor-generated measure of demonstrated skills than is typically available in administrative or survey data.

To answer these questions, I combine comprehensive academic records for all UCSC graduates since its foundation, with California administrative employment and wage microdata from the Employment Development Department (EDD) spanning 2015 to 2023. These data allow me to link detailed classroom-level performance and narrative-based skill measures to long-run labor market outcomes, including employment, wages, and industry of work.

I begin by examining whether narrative evaluations are related to students’ later labor market outcomes in two complementary ways. First, I exploit UCSC’s transition from a narrative-based grading system to one in which narratives coexisted with letter grades, and eventually disappeared altogether. Comparing UCSC cohorts to the broader University of California system, I show that relative wages rise during the mixed grading regime and decline sharply as exposure to narratives falls around their discontinuation. Relative wages increase by roughly 3 percent for cohorts exposed to both narratives and letter grades, but subsequently decline as exposure to narratives falls, eventually reaching approximately 5 percent below the benchmark

1994 cohort. While this exercise is descriptive rather than causal, the timing is consistent with several possible channels, including improved signaling and job matching when richer evaluation information is available, as well as the loss of academically relevant information once narratives disappear. Changes in student composition may also play an important role, so these cohort patterns are best viewed as motivating evidence rather than estimates of the causal effect of narrative evaluations.

Second, I directly test whether narratives contain predictive information beyond what is summarized by conventional GPA. Using a hand-coded dictionary based on UCSC's historical evaluation guidelines, I construct a narrative sentiment score and demean it within each course-instructor-term environment to remove differences in average evaluation levels across classroom settings. I then take a credit-weighted average of these relative narrative scores for each student and standardize the resulting student-level measure. Using student-labor-market-year observations, I find that a one-standard-deviation increase in a student's average relative narrative evaluation is associated with approximately 1.9 percent higher wages, conditional on conventional credit-weighted GPA and rich academic and socio-demographic controls. I then use the interaction between GPA and the narrative measure to distinguish whether the two grading systems behave as complementary, substitutable, or approximately independent signals of later wages. If narratives are complementary to grades, their incremental predictive content should be stronger at higher GPA; if they are substitutes, it should be weaker; and if they provide approximately independent information, the incremental narrative signal should not vary systematically with GPA. The estimated interaction is close to zero, supporting the third interpretation: narratives add predictive information beyond GPA, but that additional signal is approximately additive rather than concentrated among students with especially high or low grades. The narrative signal also does not vary systematically with URM or STEM major status. I do, however, find heterogeneity by gender: the narrative-wage association is smaller for women than for men. These results suggest that narrative evaluations contain information about later labor-market outcomes that is not fully summarized by conventional GPA and that, along the GPA dimension, the two grading systems provide distinct rather than systematically reinforcing or offsetting predictive signals.

I then use the narrative text to construct student-level measures of cognitive skills and thirteen fine-grained soft skills spanning both social and non-social dimensions, using a new dictionary-based framework grounded in O*NET. I evaluate several alternative skill-assignment rules and adopt the most stringent specification, which is also the one that aligns most closely with classifications produced by a context-aware language model validator. In the main specification, I assign skills using a conservative binary rule that classifies a student as possessing a given skill when any of its associated keywords appear in at least 50 percent of the student's narrative evaluations, thereby capturing persistent rather than incidental signals. This measurement approach reveals that soft skills are pervasive: under the strictest classification rule, 85 percent of students exhibit at least one soft skill and 78 percent exhibit at least one social skill, with teamwork and attention to detail emerging as the most frequently detected traits. Skill prevalence is systematically higher for women across nearly all categories, while URM and non-URM students display broadly similar patterns and students above the median GPA exhibit slightly higher skill prevalence.

I further show that detection of these skills varies over the college career, with cognitive and social-skill detection increasing after the second year, around the period of major declaration and academic specialization, while non-social soft-skill detection changes little. I also document

that early skill profiles predict subsequent field choice: students with early social and non-social soft-skill signals are less likely to sort into STEM majors, while early cognitive-skill signals are positively associated with STEM selection.

Turning to labor market outcomes, I find little evidence of meaningful returns to the three broad narrative-derived skill categories on labor force participation. Wage outcomes display a clearer pattern. Cognitive skills carry a positive wage return in the baseline specification and retain a positive premium after controlling for major, industry, and firm location. By contrast, the negative baseline returns to social and non-social soft skills attenuate to approximately zero once these academic and labor-market controls are included, suggesting that those raw relationships largely reflect sorting across fields and employment environments. Adding skill measures constructed from first-year narratives as controls attenuates the cognitive premium from approximately 4.0 to 2.7 percent, but it remains positive and statistically significant. This pattern suggests that part of the cognitive premium reflects skills already present early in college, while cognitive-skill signals emerging later in students' college careers retain additional predictive content for subsequent wages.

I also examine whether the wage associations of the broad skill dimensions depend on the combinations in which they appear. The pooled interaction specification provides little evidence of complementarity. Among men, however, the three-way interaction among cognitive, social, and non-social soft skills is positive, providing suggestive evidence of complementarity across skill dimensions. No comparable pattern appears among women, URM students, or non-URM students.

I then examine the relative associations of individual soft skills within the population of students who exhibit at least one soft-skill signal. Most individual soft-skill coefficients are not statistically distinguishable from the omitted category, but their rankings differ across labor-market margins. Self-control ranks highly for both labor force participation and wages, while leadership and teamwork rank more highly for participation and integrity and perceptiveness for wages. These patterns suggest that the skills related to obtaining and maintaining employment are not necessarily the same as those rewarded through higher wages.

Finally, I examine whether labor market outcomes are systematically related to misalignment between students' revealed skill profiles and the dominant skill bundle of their chosen major. To do so, I construct a student-major skill similarity index based on cosine similarity between each student's narrative-derived skill vector and the historical average skill profile of their field of study. I find that mismatch predicts lower wages: a 10-point increase in mismatch is associated with approximately a 1.3 percent reduction in wages, while the relationship with labor force participation is close to zero. In subgroup specifications, the wage penalty is statistically significant among women and non-URM students, but not among men or URM students. I also provide descriptive cohort evidence that greater exposure to narrative evaluations coincides with higher student-major alignment. Together, these findings suggest that long-run labor market outcomes are related not only to measured skill levels, but also to how those skills align with educational pathways.

Taken together, these findings contribute to three strands of literature. First, this paper contributes to the large and growing literature quantifying the labor market returns to cognitive and non-cognitive skills (Lindqvist and Vestman (2011), Weinberger (2014), Deming (2017), Deming and Kahn (2018), Edin et al. (2022)). While prior work typically aggregates non-cognitive skills into a single dimension, I disaggregate soft skills into six social and seven non-social categories based on the O*NET classification. Although some recent studies ex-

amine specific soft skills—particularly teamwork (Deming and Weidmann (2021), Adhvaryu et al. (2023), Aghion et al. (2024))—they rarely compare multiple skill dimensions jointly to assess their relative economic value. Moreover, unlike Deming and Kahn (2018) and Dorn et al. (2025), who infer skill profiles from job advertisements or self-reported online platforms, this paper constructs individual-level skill profiles from detailed, instructor-generated narrative evaluations written across multiple courses, providing a more direct, instructor-generated measure of demonstrated skills.

Second, this paper contributes to the literature on skill mismatch, which documents wage penalties and other labor market consequences when workers’ skills do not align with the demands of their occupational environments (Fredriksson et al. (2018), Guvenen et al. (n.d.), Lise and Postel-Vinay (2020)). I advance this literature in two ways. First, I measure mismatch using a richer and more granular vector of cognitive and soft skills than has typically been available in administrative data. Second, I extend the mismatch framework to the educational setting by comparing students’ revealed skill profiles to the dominant skill bundle of their declared majors, allowing the analysis of skill misallocation before labor market entry.

Third, this paper contributes to the literature on the signaling value of educational metrics for academic and labor market outcomes. Prior work has examined the predictive content of high school grades relative to standardized test scores (UC-Academic-Council (2020), Chetty et al. (2023)), the signaling value of diplomas and degree completion (John H. and Willett (2000), Clark and Martorell (2014), Arcidiacono et al. (2010), Aryal et al. (2022)), and the role of high-stakes test scores within school settings (Diamond and Persson (2016), Schwerdt and Woessmann (2017)). I contribute to this literature by showing that qualitative narrative evaluations written for each college class carry a statistically meaningful predictive signal for long-run labor market outcomes even conditional on conventional GPA. I further show that the narrative–GPA interaction is close to zero, consistent with the two grading systems providing approximately independent/additive predictive information rather than systematically complementary or substitutable signals. Together, these results suggest that the standard transcript summary does not fully capture economically relevant dimensions of student performance.

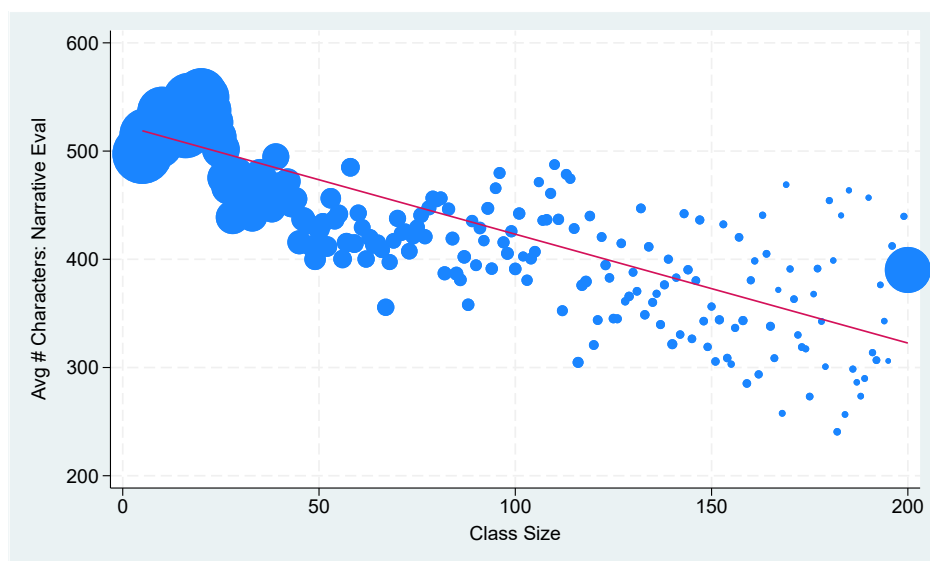
The remainder of the paper is organized as follows. Section 2 introduces the institutional setting and UCSC’s grading-policy transitions. Section 3 outlines the administrative academic and labor market datasets. Section 4 examines whether narrative evaluations contain predictive information beyond conventional GPA. Section 5 develops the skill-extraction framework and documents patterns of skill prevalence, skill dynamics, and academic sorting. Section 6 estimates the labor market returns to broad skill categories and to specific soft skills. Section 7 studies whether mismatch between students’ revealed skill profiles and the dominant skill bundle of their major predicts later labor market outcomes. Section 8 concludes.

2. Setting

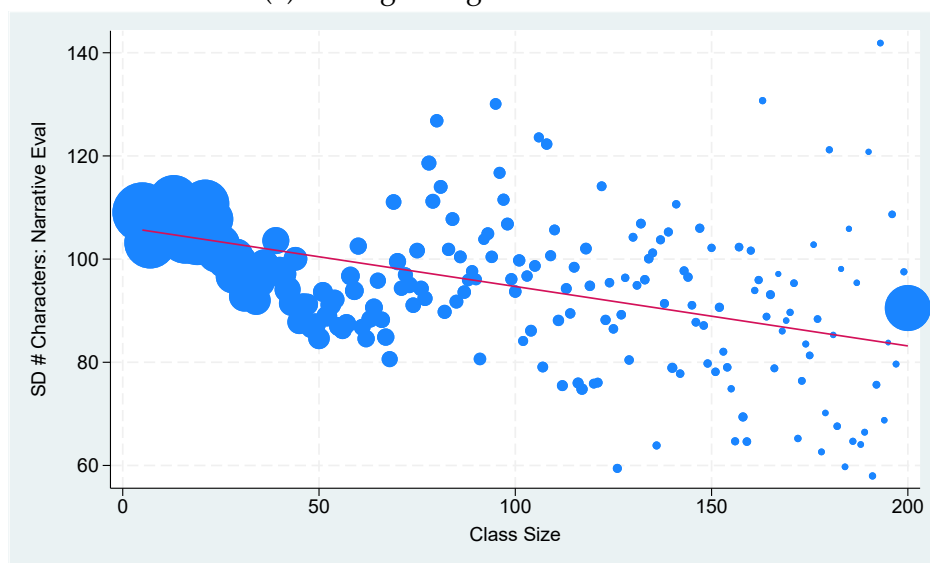
UC Santa Cruz (UCSC), located along Monterey Bay, is one of ten campuses in the University of California system. The campus enrolls over 19,000 students across a wide range of disciplines, including the arts, social sciences, environmental studies, engineering, economics, and the natural sciences. Founded in 1965, UCSC adopted a distinctive educational philosophy built around progressive pedagogy and student-centered learning. From its inception, the university rejected the notion that academic achievement should be reduced to a single number—such as a GPA—and instead adopted narrative evaluations as its primary assess-

ment system. Instructors wrote paragraph-length qualitative assessments for every completed course, describing a student's strengths, weaknesses, contributions, and patterns of growth. A Pass/No Pass mark was entered into the Registrar's system for transcript reporting alongside these written evaluations.

Narrative evaluations were designed to provide students with nuanced feedback and insight into their academic progress—something a single grade could not capture. Narrative evaluations were generally richer in smaller classes, where faculty interacted more closely with students and could provide individualized assessments. In larger lecture settings, where such familiarity was less feasible, narratives tended to be shorter and more standardized. As shown in Figure 1, both the average length of narrative evaluations and the within-class variation in their length decline as class size increases. Because the narrative text contains less information in large lecture courses, I restrict the analysis to classes with no more than 40 enrolled students—corresponding to the median section size at UCSC. Classes with fewer than five students are also excluded, as they typically correspond to independent study courses without formal instruction.



(a) Average Length: # Characters



(b) Variation in Length: SD # Characters

Figure 1: Summary Statistics for Narrative Evaluations by Class Size

Note: This figure presents the class average number of characters written in narrative evaluations and the within-class variation in character count, aggregated by class size. Circle size is proportional to the number of university classes within each class size category. Classes with more than 200 students are grouped together.

Beginning in 1971, students were allowed to request letter grades in a limited number of courses, although GPAs were not calculated. Pressure to align with other universities and concerns that UCSC graduates were disadvantaged in graduate admissions and labor markets eventually led the campus to adopt traditional letter grading. In 1997, this option was expanded to most courses, and for students admitted in 2001 and later, both letter grades and GPA calculations became mandatory. During this transition, a GPA was computed only if at least two-thirds of a student's total accumulated credits were letter-graded. Narrative evaluations continued to coexist with letter grades until 2010. After that year, narrative evaluations became formally optional under campus policy and were, in practice, abandoned almost im-

mediately. Table 1 summarizes the key changes in grading policy.

Table 1: UCSC General Timeline on Grading Policies

Year	Policy change
1965	Foundation of UCSC with Pass/Fail grading and narrative evaluations.
1971	Letter grading option introduced for <u>some</u> courses.
1997	Letter grading option extended to <u>most</u> courses.
2001	Mandatory letter grading and GPA reporting for all admitted students.
2001	Pass/Fail credit capped at 25% of total earned units.
2001	Chromatic modifiers (+/-) added to A–D grades.
2010	Narrative evaluations become optional and effectively discontinued.

Figure 2 illustrates the rapid adoption of letter grades, beginning with cohorts already admitted and enrolled when optional letter grading was introduced and culminating in full implementation with the 2001 entering cohort. Because the main analysis requires both individual letter-grade/GPA information and narrative evaluations, I restrict the sample to cohorts where such data are consistently and jointly observed. This corresponds to the 2001–2010 admission cohorts, when letter grades were mandatory rather than selectively assigned. The broader set of admission cohorts from 1985 to 2010 is used only for a few secondary analyses that do not rely on letter grades.

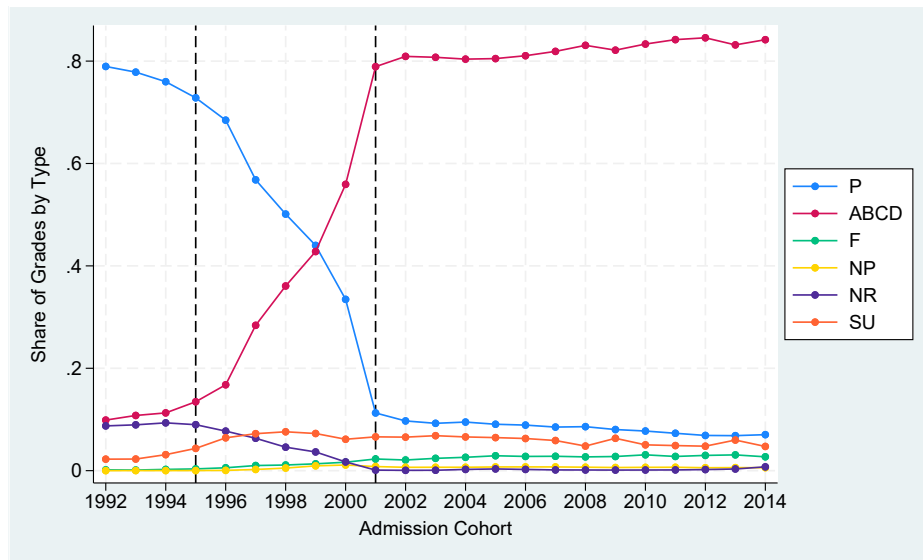


Figure 2: Grading Policies by Admission Cohort at UCSC

Note: This figure plots the share of course grades awarded by type (Pass, letter grades A–D, Fail, No Pass, Not Reported, and Satisfactory/Unsatisfactory) for UCSC students by admission cohort. The vertical dotted lines indicate the policy transition: the introduction of optional letter grading (1997) and the shift to mandatory letter grading (2001). Because the 1997 policy applied to coursework taken during the 1997 academic year, cohorts admitted prior to 1997 began earning letter grades before graduating, which explains the increase beginning with the 1995 entering cohort.

3. Data

This project draws on the universe of historical administrative and academic records of UC Santa Cruz students, spanning from the university's founding in 1965 to 2020¹. Data access was approved by the UC Santa Cruz Office of the Registrar and is hosted on a secure server maintained by Prof. Zachary Bleemer.² Anonymized student identifiers are used to merge datasets at the individual level.

UCSC Academic Transcripts

This dataset contains course-level records for every student who enrolled in a class at UC Santa Cruz. Each observation includes the course name, course code, academic department, credit value, instructor, section, term, and year, along with the student's recorded grade (letter, narrative, or both). The dataset includes approximately 5 million student-course transcript records.

UCSC Graduation Records

This dataset includes graduation information for all students who completed a degree at UCSC. Variables include degree level, major(s), minor(s), graduation year, graduate's birth year, birth ZIP code and country, college application ZIP code and country, admission year and term, ethnicity, gender, and transfer status. A proxy for parental income is constructed by merging a student's application ZIP code with publicly available IRS Adjusted Gross Income statistics. The dataset covers roughly 166,000 UCSC graduates.

UCSC Graduates' Labor Outcomes

The California Employment Development Department (EDD) provided annual wage records from 2015 to 2023 for all UC Santa Cruz graduates.³ Annual wages are measured as the sum of quarterly wages, converted to 2018 dollars using the Consumer Price Index (CPI), and winsorized at the top and bottom 1%. Labor force participation is proxied using positive wage records. For employed graduates, the dataset also reports the ZIP code of the employer's headquarters and the six-digit NAICS industry.

The EDD administers California's Unemployment Insurance (UI), Disability Insurance (DI), and Paid Family Leave (PFL) programs. Employers paying more than \$100 in wages in any calendar quarter are required to register with EDD, ensuring coverage for employees working in California. Wages are not reported for individuals outside UI coverage—such as the self-employed, federal workers, or those employed out of state. Approximately 76% of UCSC graduates have positive covered wages in the dataset.

¹For details on digitization efforts and procedures for older records, see Bleemer (2018). Additional information about the academic records is available in Bleemer (2019).

²Professor Bleemer is an Assistant Professor of Economics at Princeton University in the Industrial Relations Section and a faculty research fellow at the National Bureau of Economic Research. He is also a research associate at Opportunity Insights and the Center for Studies in Higher Education. Website: <https://zacharybleemer.com/>

³Micro-level wage records are only available for UCSC. For the remaining UC campuses, the EDD provided admission-cohort average wages aggregated across the UC system.

4. Do Narrative Evaluations Matter?

A natural first question is whether narrative evaluations contain information that is related to students' subsequent outcomes. If narratives contain little information beyond conventional academic measures, there should be little systematic relationship between exposure to narratives, the content of those narratives, and subsequent outcomes. By contrast, if narratives capture economically relevant information, two types of evidence are informative: (i) systematic shifts in outcomes across cohorts as narratives coexist with letter grades, and eventually cease to be written, and (ii) whether students with stronger relative narrative evaluations have different subsequent outcomes even conditional on conventional GPA. The remainder of this section examines both dimensions—first through the lens of UCSC's grading policy transitions, and then through direct tests of whether narrative text contains incremental information beyond conventional GPA.

Evidence from Grading Policy Transitions

UCSC's staggered transition from a narrative-only system to the coexistence of narratives and letter grades, and eventually to their elimination, offers a useful setting to test whether narratives have bite. To explore this question, I construct a simple event-style comparison that benchmarks UCSC graduates against students admitted to the full University of California campuses over the same period. For the comparison campuses, I observe only average wages by admission cohort—rather than micro-level administrative records—which precludes estimation of standard errors and statistical inference. The exercise therefore compares average log annual wages for each UCSC admission cohort to the corresponding UC-system cohort average and tracks how this gap evolves as UCSC students become increasingly exposed to letter grades and later, lose access to narrative evaluations. I normalize the UCSC–UC system wage difference to zero for the 1994 entering cohort—one year before the first cohort with an uptick in exposure to letter grades—so that subsequent movements reflect how relative outcomes shift as grading regimes change.

Two broad patterns emerge from Figure 3. First, relative wages for UCSC students rise beginning with the 1995–1997 entering cohorts, with the UCSC–UC wage gap stabilizing around three percent higher for cohorts admitted during the mixed regime when both narratives and letter grades were recorded. Second, relative wages decline sharply for cohorts with reduced exposure to narratives around their discontinuation, eventually eliminating the earlier advantage and falling roughly five percent below the wages of comparable UC graduates in the benchmark 1994 cohort.

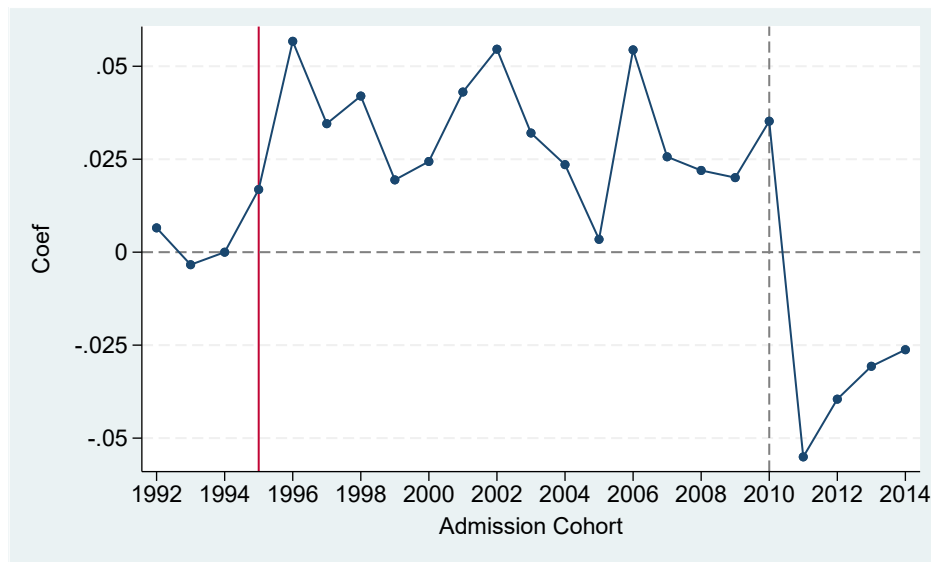


Figure 3: Relative Wages of UCSC Graduates vs. UC System, by Admission Cohort

Note: This figure plots the difference in average log annual wages between UCSC graduates and the UC systemwide cohort average by admission year. The series is normalized to zero for the 1994 entering cohort. The solid red vertical line (1995) marks the first admission cohort with meaningful exposure to letter grades after their 1997 expansion, and the dotted vertical line at the 2010 admission cohort coincides with the calendar-year discontinuation of narrative evaluations.

These patterns should be viewed as descriptive rather than causal. In particular, contemporaneous admission-policy changes affecting all UC campuses may contribute to movements in the relative wages gap if their effects are heterogeneous across campuses. Even so, the timing is consistent with plausible mechanisms. The rise in relative wages after the introduction of letter grades may reflect improved signaling to employers, better job matching into higher-paying firms, or compositional changes if the mixed evaluation system attracted or retained students with stronger academic preparation. These compositional channels may be less relevant for the earliest affected cohorts, whose admissions decisions predated the policy shift, although the aggregate comparison does not allow these mechanisms to be separated cleanly. Conversely, the post-2010 decline aligns with the removal of narrative evaluations, which may reflect either compositional changes in who enrolls once narratives are no longer offered, or the loss of information to students that could have helped guide academic development, shape course choices, or translate classroom performance into labor market outcomes.

Do Narrative Evaluations Provide Information Beyond Grades?

A complementary way to assess whether narrative evaluations matter is to ask whether they contain predictive information about students' later labor market outcomes beyond what is summarized by conventional GPA, and whether the two grading systems provide complementary, substitutable, or approximately independent signals. To evaluate this question, I first convert each narrative evaluation into a numeric sentiment score using the UCSC 2010 Guidelines on Writing Narrative Evaluations, which implicitly map evaluative language to a letter-grade scale (A = 4.0 "Excellent," B = 3.0 "Very Good," C = 2.0 "Good or Fine," D = 1.0 "Bad," and F = 0.0 "Very Bad"). I construct a dictionary of approximately ten representative synonyms for each category and assign each word root the corresponding numeric value; the full set of

proxy terms is reported in the Appendix. I also apply light text cleaning to account for common spelling deviations. A narrative’s score is computed as the simple average of the values associated with all matched roots appearing in the evaluation.

Because instructors and classes may differ systematically in the level of narrative evaluations they assign, I demean the narrative score within each course–instructor–term environment by subtracting the corresponding class mean. I then collapse these deviations to the student level, taking a credit-weighted average across the student’s eligible narrative evaluations. The resulting measure captures whether a student tended to receive more positive or more negative narrative evaluations than other students facing the same instructor in the same class environment. Finally, I standardize this student-level measure across students, so that a one-unit change corresponds to a one-standard-deviation increase in the student’s average relative narrative evaluation.

The benchmark grade measure is the student’s conventional credit-weighted GPA, calculated over all letter-graded coursework. I deliberately retain GPA in its usual transcript-wide form rather than restricting it to the courses used to construct the narrative measure or converting grades into within-class rankings. This makes the comparison directly answer the question of whether narrative evaluations provide information beyond the standard GPA signal contained in an academic record. After constructing these student-level academic measures, I reshape the labor-market data to the student–year level, so that each annual wage observation enters the analysis once for each student rather than being repeated across the classes the student took.

I estimate the predictive value of narrative evaluations using the following specification:

$$\log(Wage_{it}) = \alpha + \theta_1 GPA_i + \theta_2 Narr_i + \delta X_{it} + \epsilon_{it}, \quad (1)$$

where $Wage_{it}$ denotes annual wages for student i in labor-market year t , GPA_i is the student’s conventional credit-weighted GPA across all letter-graded coursework, and $Narr_i$ is the standardized credit-weighted average of the student’s within-class demeaned narrative scores. The vector X_{it} includes family income and fixed effects for major, ethnicity, sex, admission cohort, transfer status, and labor-market year. Standard errors are clustered at the student level. The coefficient θ_2 therefore asks whether students who tended to receive systematically stronger narrative evaluations relative to their classmates subsequently earn more, conditional on overall conventional GPA and similar observed characteristics. These estimates are predictive associations rather than causal effects.

Table 2 reports the results. Column (1) presents the baseline specification. GPA is strongly predictive of later wages: a one-point increase in overall GPA is associated with approximately 8.7 percent higher wages. More importantly for the question of interest, narrative evaluations retain statistically significant predictive content conditional on GPA. A one-standard-deviation increase in a student’s average relative narrative evaluation is associated with approximately 1.9 percent higher annual wages. Thus, narrative evaluations contain information about later labor-market outcomes that is not fully summarized by conventional GPA.

Table 2: Predictive Value of Narrative Evaluations Conditional on GPA

	(1)	(2)	(3)	(4)	(5)
	Baseline	Signal Interaction	Female	URM	STEM
	$\beta/(SE)$	$\beta/(SE)$	$\beta/(SE)$	$\beta/(SE)$	$\beta/(SE)$
GPA	0.087*** (0.019)	0.087*** (0.019)	0.088*** (0.019)	0.088*** (0.019)	0.088*** (0.019)
Narrative Evaluation (SD)	0.019** (0.008)	0.027 (0.045)	0.034*** (0.010)	0.021** (0.009)	0.020** (0.008)
GPA $\times$ Narrative		-0.003 (0.014)			
Narrative $\times$ Female			-0.028** (0.013)		
Narrative $\times$ URM				-0.008 (0.014)	
Narrative $\times$ STEM					-0.005 (0.015)
Base Controls	YES	YES	YES	YES	YES
Adj-R ²	0.135	0.135	0.135	0.135	0.135
Obs (K)	158.6	158.6	158.6	158.6	158.6
# Stud (K)	17.0	17.0	17.0	17.0	17.0
Avg Wage	100,419	100,419	100,419	100,419	100,419

Note: This table estimates model 1 using student–labor-market-year observations. Narrative Evaluation (SD) is constructed by demeaning the narrative sentiment score within each course–instructor–term environment, taking a credit-weighted average of these deviations across each student’s eligible narrative evaluations, and standardizing the resulting student-level measure to have a standard deviation of one. GPA is the student’s conventional credit-weighted GPA calculated over all letter-graded coursework. Base controls include ethnicity, sex, transfer status, admission-year fixed effects, and family income; all specifications also include major and labor-market-year fixed effects. Standard errors are clustered at the student level. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Column (2) uses the GPA–narrative interaction to characterize how the two grading systems relate as predictive signals. I interpret a positive interaction as evidence of complementarity: narrative evaluations would be more informative about later wages among students with higher GPA. A negative interaction would instead indicate substitutability, with the incremental narrative signal becoming weaker as GPA rises. An interaction close to zero corresponds to approximately independent, additive signals: narratives may predict wages beyond GPA, but their incremental predictive content does not systematically depend on the level of GPA. The estimated interaction is close to zero and statistically insignificant. Combined with the positive baseline narrative coefficient in Column (1), this pattern supports the independent-signal interpretation rather than complementarity or substitutability: narratives contain additional predictive information, but that information does not appear to reinforce or offset the GPA signal systematically.

Columns (3)–(5) examine heterogeneity across student groups. The gender interaction is negative and statistically significant: the association between narrative evaluations and wages is 2.8 percentage points smaller for women than for men. By contrast, the interactions for underrepresented minority (URM) students⁴ and STEM majors are small and statistically insignificant. Thus, the results provide evidence of heterogeneity by gender, but not by URM or STEM status. These subgroup patterns remain descriptive and should not be interpreted causally.

What, then, might narrative evaluations be capturing beyond conventional GPA? Because

⁴URM includes all racial and ethnic groups except White and Asian.

narratives are free-text and unconstrained by grading rubrics, they can encode dimensions of student performance that do not map cleanly into a single transcript average. These include work habits such as effort, improvement, and resilience; behavioral traits such as leadership, reliability, and collaboration; and academic behaviors such as engagement, creativity, or contribution to class discussions. Narratives can also reference task-specific strengths such as writing quality, quantitative reasoning, persistence, or teamwork that are not directly summarized by GPA.

The predictive results therefore suggest that instructors systematically communicate—or detect—student attributes that are not fully captured by conventional GPA and that are associated with later labor-market outcomes. This motivates the next step of the analysis: identifying, categorizing, and quantifying the specific cognitive, social, and non-social soft skills embedded in narrative evaluations, and assessing how they differ across students, disciplines, and time, and how they are associated with later labor-market outcomes.

5. Measuring Skills from Narrative Evaluations

Motivated by the evidence that narrative evaluations contain predictive information beyond conventional GPA, I now turn to identifying the specific competencies embedded in narrative text. This section develops a systematic method for extracting and classifying cognitive, social, and non-social soft skills from the universe of narrative evaluations.

Skill Taxonomy and Dictionaries

I construct a comprehensive skill framework grounded in the U.S. Department of Labor’s O*NET occupational taxonomy. Following this classification, I focus on fourteen broad skills: one cognitive skill and thirteen soft skills, which I group into six “social” soft skills and seven “non-social” soft skills.⁵

- **Cognitive:** General cognitive / analytical reasoning
- **Social Soft Skills:** Persuasion, Teamwork, Teaching, Leadership, Interpersonal Relationships, Perceptiveness
- **Non-Social Soft Skills:** Integrity, Attention to Detail, Self-Control, Adaptability, Initiative, Persistence, Dependability

Each skill is represented by a dictionary of approximately fourteen proxy terms or synonyms. For example, the dictionary for *Teamwork* includes: *team, collaborate, communication, coordinate, compromise, unity, cooperate, contribute, helpful, partner, participate, member, group, project*. These dictionaries are stemmed so that root forms capture variation in tense or phrasing. The full set of proxy terms for all skills is reported in Appendix.

⁵Unlike soft skills, which encompass a wide and heterogeneous set of abilities that can be meaningfully divided into social and non-social categories, cognitive skills are typically defined in O*NET as a single unified construct. This reflects their narrower conceptual scope, and I therefore retain cognitive ability as one category rather than splitting it further.

Skill Assignment

Narrative evaluations are processed using a string-search algorithm that detects the presence of the proxy term roots associated with each skill. Skill classification is conducted under multiple alternative rules:

1. **Broad Detection Rule:** A student is coded as possessing a skill if any of its proxy term roots is referenced at least once in any of their narrative evaluations. An alternative broad version flags a skill when a single instructor mentions proxy terms at least twice within the same narrative, under the intuition that repeated emphasis reflects intentional evaluation rather than incidental phrasing. Even then, a single instructor’s impression may not reflect a stable student attribute, so the strict rule requires repeated identification across a student’s college career.
2. **Persistent-Signal Rule (Strict):** A student is coded as possessing a skill only if it appears (through any of its proxy terms) persistently across their evaluations. This requires appearance in a minimum share of the student’s narratives (e.g., at least 25 percent), making repeated detection across the student’s record more demanding than a one-off mention. A 50 percent threshold yields the most stringent feasible variant, as skill detection becomes rare beyond that point given how few students receive highly repeated descriptive signals. This rule is intended to capture stable student characteristics rather than course- or instructor-specific remarks.

In the analyses that follow, I use the strict 50 percent persistence threshold to construct binary indicators of whether a student exhibits each skill. While a continuous measure based on the share of narratives mentioning a skill is theoretically possible, small differences in mention frequency are difficult to interpret as meaningful variation in underlying ability rather than noise or writing style. The conservative binary rule therefore prioritizes reliability over granularity and ensures that the skill assignments reflect traits that appear repeatedly across a student’s college career.

While the detection rules track positive skill terms, misclassification from negative signals is unlikely to contaminate the classifications. Empirically, negative evaluative language is extremely rare in UCSC narratives: a full corpus scan using sentiment-scoring algorithms identified virtually no instances of dictionary stems associated with “Bad” or “Very Bad” descriptors (e.g., *poor teamwork*, *lacks initiative*), and the distribution of sentiment scores shows almost no mass in the negative range. This pattern aligns with historical accounts of narrative practices at UCSC, where instructors generally avoided writing explicitly critical assessments.

To validate and complement the dictionary-based approach, I also benchmark skill assignments against a context-aware classifier. Using a secure installation of the meta-llama-3.1-8b-instruct model, I draw a random subsample of narrative evaluations and prompt the model to identify whether each excerpt exhibits the target skills. Unlike the string-matching rules, which rely solely on the presence of proxy terms, the LLM interprets sentences in context—capturing cases where skills are expressed implicitly or where proxy words carry different meanings depending on usage. The machine-based classifications are compared against the rule-based indicators, and agreement is highest under the strictest variant of the dictionary method—particularly the 50 percent persistence threshold—suggesting that the repeated cross-evaluation criterion best captures persistent skill signals. Taken together, the dictionary framework and the LLM cross-check provide complementary evidence that

the resulting measures are more likely to reflect persistent skill signals embedded in the narratives rather than incidental or course-specific phrasing. This measurement pipeline lays the foundation for the descriptive and outcome-based analyses that follow.

Skill Prevalence and Group Heterogeneity

Having established a method for identifying skills from narrative text, I next document how widely these skills appear across the student population. Table 3 reports the share of students coded as possessing each skill under the four detection rules described above: (i) any single keyword mention, (ii) at least two mentions by a single instructor, (iii) appearance in at least 25 percent of a student’s narratives, and (iv) appearance in at least 50 percent of narratives (the strictest feasible measure). These statistics summarize not only how common each skill is, but also how persistence filters shift the classification of students as skill-holders.

Several patterns stand out. First, under the broadest measure (one keyword anywhere), virtually every student has evidence of at least one soft skill, and 97 percent display cognitive-skill terms. Teamwork and attention to detail remain among the most frequently referenced traits even under the two-keyword requirement. Second, when requiring persistent mention across a student’s coursework, frequencies decline sharply, reflecting the fact that repeated reinforcement of the same quality across narratives is rarer than one-off descriptive phrases. Under the strictest 50 percent persistence rule, 85 percent of students exhibit at least one soft skill, 78 percent display at least one social soft skill, and 49 percent are coded as demonstrating cognitive ability. Thus, even the conservative persistence rule identifies a sizeable set of repeatedly recognized competencies while substantially reducing the prevalence of individual skill classifications.

Table 4 turns to heterogeneity, reporting skill prevalence under the 50 percent strict rule across gender, underrepresented minority (URM) status, above- and below-median GPA, and STEM-major status. Several patterns emerge. Women are more likely than men to be coded as possessing most skills, with particularly visible differences in teaching, teamwork, attention to detail, and cognitive skills. URM and non-URM students exhibit broadly similar prevalence patterns. Students with above-median GPA generally exhibit higher rates of persistent skill signals, with particularly visible differences in teamwork, attention to detail, and the aggregate social- and soft-skill measures. Finally, non-STEM students are substantially more likely than STEM students to exhibit social and other soft-skill signals. Cognitive-skill prevalence is also higher among non-STEM students, although the gap is smaller than for the aggregate soft-skill measures.

Skill Dynamics and Academic Sorting

This section examines whether the skills identified in narrative evaluations are stable or vary over the course of students’ college careers, and how they relate to academic specialization. I conduct this analysis among students, for whom progression through a standard four-year college sequence provides a natural measure of academic time.

To study skill dynamics, I estimate within-student regressions of skill incidence on indicators for year in university (corresponding to first through fourth academic year after enrollment). Skills are measured at the student-narrative level: for each student–class observation, a skill indicator equals one if the narrative contains at least one proxy term associated with that

Table 3: Skill Frequency (% of students) across different detection measures

Skill	1Keyword	+2Keyword	+25%Narr	+50%Narr
Adaptability	24	1	2	0
Integrity	53	7	10	1
Self-control	57	8	12	1
Relationships	76	23	34	7
Initiative	79	30	39	7
Persuasion	80	31	40	8
Persistence	84	35	47	10
Leadership	86	38	48	12
Perceptiveness	89	53	64	22
Dependability	93	49	75	29
Teaching	95	53	79	36
Detail	96	76	86	49
Teamwork	98	86	93	70
Cognitive	97	80	87	49
Non-social	99	86	94	62
Social	99	93	97	78
Soft	100	96	99	85

Note: This table reports the share of students coded as exhibiting each skill under progressively stricter detection rules, ranging from a single keyword match to appearance in at least half of a student's narrative evaluations. Social, Non-Social, and Soft skill categories aggregate the listed component skills, and a student is coded as having a category if any of its constituent skills is detected. Percentages are computed relative to all students with at least one narrative evaluation.

Table 4: Skill Frequency (% of students) by Student Characteristics

Skill	Male	Female	URM	nURM	High GPA	Low GPA	STEM	nSTEM
Adaptability	0	0	0	0	0	0	0	0
Integrity	1	2	1	1	2	1	1	2
Self-control	1	1	1	1	2	2	1	1
Relationships	5	8	7	7	10	6	5	8
Initiative	6	8	7	7	9	6	8	7
Persuasion	8	9	8	9	10	7	3	10
Persistence	8	11	10	10	11	9	8	11
Leadership	11	13	12	12	16	13	11	12
Perceptiveness	17	25	21	22	22	16	14	24
Dependability	23	33	30	28	34	25	22	31
Teaching	29	42	40	35	39	36	31	38
Detail	42	55	49	49	44	34	38	53
Teamwork	62	76	73	69	80	69	51	76
Cognitive	42	54	50	48	49	44	40	52
Other Soft	54	68	62	62	60	49	48	66
Social	71	84	81	77	85	77	60	84
Soft	79	90	87	85	89	81	69	91

Note: This table reports the share of students possessing each skill under the strict 50% persistence rule, disaggregated by gender, URM status, above- versus below-median GPA, and STEM-major status. High and Low GPA correspond to students at or above and below the sample median GPA of 3.0, respectively. Because conventional GPA is available for the post-2000 admission cohorts, the GPA comparison is based on that subset of students. URM includes all racial and ethnic groups except White and Asian. Social, Other Soft, and Soft categories aggregate the listed component skills, and a student is coded as having a category if any of its constituent skills is detected.

skill. I then estimate:

$$\text{Skill}_{ic} = \sum_{k \neq 2} \beta_k \mathbf{1}\{yr_uni_i = k\} + \alpha_i + \gamma_{j(c)} + \varepsilon_{ic} \quad (2)$$

where Skill_{ic} is an indicator for whether a given skill is detected in a narrative evaluation of student i in class c , α_i denotes student fixed effects, and $\gamma_{j(c)}$ are instructor fixed effects corresponding to the instructor teaching class c . Errors are clustered at the student level. Including instructor fixed effects ensures that the estimates are not driven by systematic differences in grading styles or narrative conventions across instructors that may be correlated with college year. The omitted category corresponds to the second academic year after enrollment.

Figure 4 reports the estimated coefficients. Relative to the second academic year, cognitive-skill detection is lower in the first year and progressively higher in the third and fourth years. Social-skill detection is similar in the first and second years but rises in the third and fourth years. By contrast, non-social soft-skill detection is relatively stable after the second year, with only a small positive difference in the first year. The increase in social-skill detection after the second year coincides with the period when students typically declare their major. One possible interpretation is that students begin to select into fields and classes that better match their interests, which may in turn be associated with greater expression or detection of particular skills. Overall, the results show that the narrative-based measures of cognitive and social skills vary systematically over the course of college, while non-social soft-skill detection is comparatively stable.

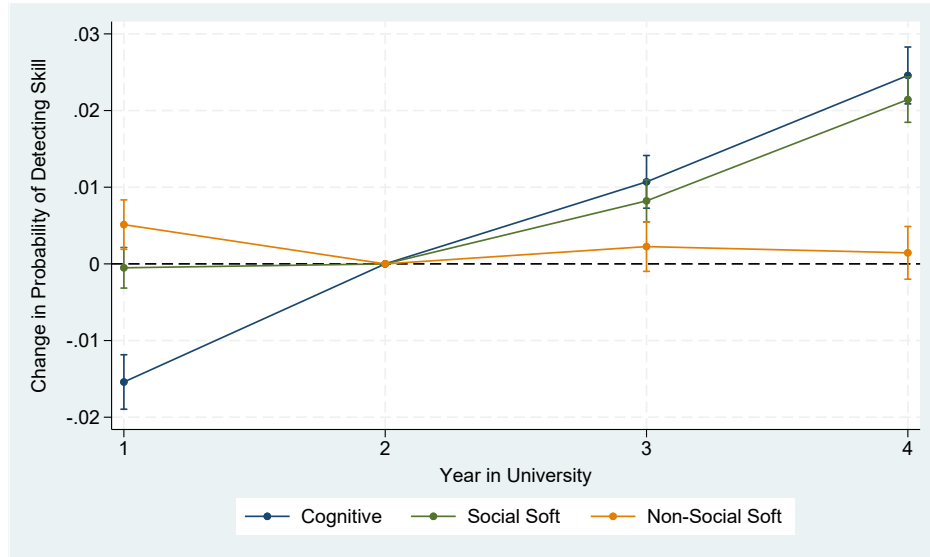


Figure 4: Skill Detection over years in College after Enrollment

Note: This figure estimates model 2. It plots within-student estimates of changes in the probability of detecting cognitive, social, and non-social soft skills across years in university, relative to the second academic year after enrollment. All specifications include student and instructor fixed effects, and standard errors are clustered at the student level. Error bars represent 95% confidence intervals.

I next examine whether early skill profiles are associated with subsequent academic specialization. This provides a complementary perspective on whether skill measures capture dimensions of performance that are relevant for students' academic choices. To do so, I construct

student-level skill indicators using narrative evaluations from the first two academic years after enrollment, generally preceding or overlapping the period in which students declare a major. A skill is assigned if it is detected in at least 50% of a student's early narratives. I then estimate linear probability models of STEM major choice:

$$\text{STEM}_i = \alpha + \beta_1 \text{Cognitive}_i + \beta_2 \text{Social}_i + \beta_3 \text{NonSocial}_i + \delta X_i + \varepsilon_i \quad (3)$$

where X_i includes credit-weighted GPA measured over the first two academic years, family income, and fixed effects for ethnicity, sex, admission cohort, and transfer status. Standard errors are heteroskedasticity-robust.

Students with early cognitive-skill signals are 1.3 percentage points more likely to subsequently select into STEM fields ($p = 0.040$). In contrast, early social-skill signals are associated with a 17.0 percentage point lower probability of selecting into STEM ($p < 0.001$), while non-social soft-skill signals are associated with a 4.3 percentage point lower probability ($p < 0.001$). These patterns show that early skill profiles are strongly associated with subsequent academic specialization, even conditional on academic performance during the first two years and other observed student characteristics.

6. Returns to Skills in the Labor Market

This section evaluates the labor-market returns to the skills surfaced in narrative evaluations. With students classified according to the competencies repeatedly described in their evaluations, I study whether cognitive, social, and non-social soft skills predict labor force participation and wages after graduation.

The baseline specification takes the form:

$$Y_{it} = \alpha + \theta_1 \text{Cognitive}_i + \theta_2 \text{Social}_i + \theta_3 \text{NonSocial}_i + \delta X_{it} + \epsilon_{it} \quad (4)$$

where Y_{it} denotes either labor force participation or log wages for student i at time t . The key regressors are indicators equal to one if the student is classified as possessing each skill category under the strict 50-percent persistence rule. The control vector X_{it} includes ethnicity, gender, transfer status, admission-year and labor outcome year fixed effects, family income, and graduation GPA; in preferred specifications, I also include additional controls to examine specific mechanisms. Errors are clustered at the student level. Despite the rich set of controls, these coefficients should be interpreted as predictive associations rather than causal estimates.

Table 5 reports the results for labor force participation. The overall pattern is largely null: the three broad skill measures do not systematically predict labor force participation once the controls are included. This conclusion is similar when first-year skill measures are added as controls and when the interaction specification is estimated separately by gender and URM status. When the three skill dimensions are allowed to interact in Column (4), however, there is some evidence of a positive return to cognitive skills and suggestive evidence of complementarity across all three broad skill dimensions.

Table 6 turns to wage returns conditional on having observed wages. In the baseline specification, cognitive skills carry a positive wage premium, whereas social and non-social soft skills carry negative estimated premiums. Once major, industry, and firm-location fixed effects are included, both soft-skill premiums attenuate to approximately zero, while the positive

cognitive-skill premium remains: students with a persistent cognitive-skill signal earn approximately 4 percent higher wages. This attenuation suggests that the negative baseline soft-skill returns largely reflect sorting across academic fields and labor-market environments rather than a stable conditional wage penalty. Adding first-year skill measures as controls, which serve as proxies for skills students bring into college, helps distinguish whether the returns to the full-college skill measures primarily reflect pre-existing skills or also capture skill signals that emerge over the course of college. The cognitive-skill premium declines from approximately 4.0 to 2.7 percent but remains positive and statistically significant. This attenuation suggests that part of the cognitive premium reflects skills already present early in college, while the remaining premium indicates that cognitive-skill signals emerging later in students' college careers contain additional information about subsequent wages. The social and non-social soft-skill coefficients remain close to zero.

Columns (4)–(8) examine whether the broad skill dimensions exhibit complementarity in their wage returns. In the pooled sample, the interaction estimates provide little evidence of complementarity. The subgroup estimates, however, provide some evidence of complementarity among men. The positive interaction across all three skill dimensions indicates an additional wage premium for men who display cognitive, social, and non-social soft skills together. No comparable pattern appears among women, URM students, or non-URM students.

Table 5: Skill Premiums on Labor Participation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Base	+ Major FE	+ First-Year	Interactions	Male	Female	URM	Non-URM
Cognitive	0.011*	0.009	0.007	0.039*	0.038	0.040	0.027	0.043
	(0.006)	(0.007)	(0.007)	(0.023)	(0.032)	(0.034)	(0.046)	(0.027)
Social	-0.002	0.005	0.005	0.004	0.017	-0.013	-0.008	0.009
	(0.009)	(0.009)	(0.010)	(0.013)	(0.017)	(0.019)	(0.024)	(0.015)
Non-social	0.002	0.007	0.007	0.001	0.007	-0.007	-0.044	0.019
	(0.007)	(0.007)	(0.007)	(0.022)	(0.030)	(0.032)	(0.042)	(0.025)
Social $\times$ Cognitive				-0.035	-0.041	-0.029	-0.034	-0.035
				(0.026)	(0.036)	(0.037)	(0.051)	(0.030)
Non-social $\times$ Cognitive				-0.070*	-0.101*	-0.045	-0.053	-0.073
				(0.041)	(0.057)	(0.059)	(0.076)	(0.049)
Social $\times$ Non-social				0.007	0.003	0.014	0.066	-0.018
				(0.024)	(0.033)	(0.035)	(0.045)	(0.028)
All Skills				0.076*	0.096	0.055	0.062	0.079
				(0.044)	(0.061)	(0.062)	(0.081)	(0.052)
Base Controls	YES	YES	YES	YES	YES	YES	YES	YES
Major FE	NO	YES	YES	YES	YES	YES	YES	YES
First-Year Skills	NO	NO	YES	NO	NO	NO	NO	NO
Interactions	NO	NO	NO	YES	YES	YES	YES	YES
Adj-R ²	0.0255	0.0309	0.0309	0.0311	0.0262	0.0385	0.0342	0.0310
Observations	200,772	200,772	200,772	200,772	88,944	111,828	57,384	143,388
Students	16,731	16,731	16,731	16,731	7,412	9,319	4,782	11,949
Avg Y	0.632	0.632	0.632	0.632	0.640	0.626	0.653	0.624

Note: This table estimates model 4 using labor force participation as the dependent variable. Base controls include ethnicity, gender, transfer status, admission-year and labor outcome year fixed effects, family income, and graduation GPA. Column (2) additionally includes major fixed effects. Column (3) controls for first-year indicators of the same three skill dimensions. Columns (4)–(8) include all pairwise interactions among the three full-college skill indicators and their three-way interaction; Columns (5)–(8) estimate this interaction specification separately by gender and URM status. Standard errors are clustered at the student level. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Skill Premiums on Wages

	(1) Base	(2) + Labor-Market FE	(3) + First-Year	(4) Interactions	(5) Male	(6) Female	(7) URM	(8) Non-URM
Cognitive	0.047*** (0.016)	0.040*** (0.013)	0.027* (0.014)	0.001 (0.044)	0.023 (0.055)	-0.012 (0.075)	-0.038 (0.117)	0.008 (0.048)
Social	-0.201*** (0.021)	0.007 (0.018)	0.008 (0.019)	-0.006 (0.024)	-0.015 (0.033)	0.001 (0.037)	-0.014 (0.052)	0.005 (0.028)
Non-social	-0.059*** (0.016)	-0.003 (0.013)	-0.002 (0.013)	0.016 (0.039)	-0.040 (0.059)	0.082 (0.055)	0.022 (0.088)	0.015 (0.043)
Social $\times$ Cognitive				0.049 (0.049)	-0.006 (0.065)	0.093 (0.080)	0.087 (0.123)	0.041 (0.054)
Non-social $\times$ Cognitive				-0.074 (0.074)	-0.198* (0.110)	0.005 (0.105)	-0.105 (0.161)	-0.058 (0.085)
Social $\times$ Non-social				-0.019 (0.043)	0.024 (0.066)	-0.067 (0.060)	-0.017 (0.096)	-0.020 (0.048)
All Skills				0.071 (0.079)	0.239** (0.119)	-0.040 (0.112)	0.080 (0.170)	0.064 (0.091)
Base Controls	YES	YES	YES	YES	YES	YES	YES	YES
Major FE	NO	YES	YES	YES	YES	YES	YES	YES
Industry / Location FE	NO	YES	YES	YES	YES	YES	YES	YES
First-Year Skills	NO	NO	YES	NO	NO	NO	NO	NO
Interactions	NO	NO	NO	YES	YES	YES	YES	YES
Adj-R ²	0.0972	0.3351	0.3352	0.3352	0.3854	0.3254	0.3473	0.3560
Observations	127,909	127,379	127,379	127,379	56,946	70,000	37,254	89,714
Students	13,779	13,755	13,755	13,755	6,108	7,626	3,976	9,756
Avg Y	99,710	99,710	99,710	99,710	113,142	88,761	93,084	102,486

Note: This table estimates model 4 using log annual wages as the dependent variable. Base controls include ethnicity, gender, transfer status, admission-year and labor outcome year fixed effects, family income, and graduation GPA. Column (2) additionally includes major, industry (NAICS), and firm-location (ZIP code) fixed effects. Column (3) controls for first-year indicators of the same three skill dimensions. Columns (4)–(8) include all pairwise interactions among the three full-college skill indicators and their three-way interaction; Columns (5)–(8) estimate this interaction specification separately by gender and URM status. Standard errors are clustered at the student level. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

To assess the relative importance of individual soft skills, I restrict attention to students who display at least one soft skill and compare outcomes across specific traits. The goal of this exercise is to determine whether certain soft skills are rewarded more than others once we condition on having any soft-skill signal at all.

I estimate the following model:

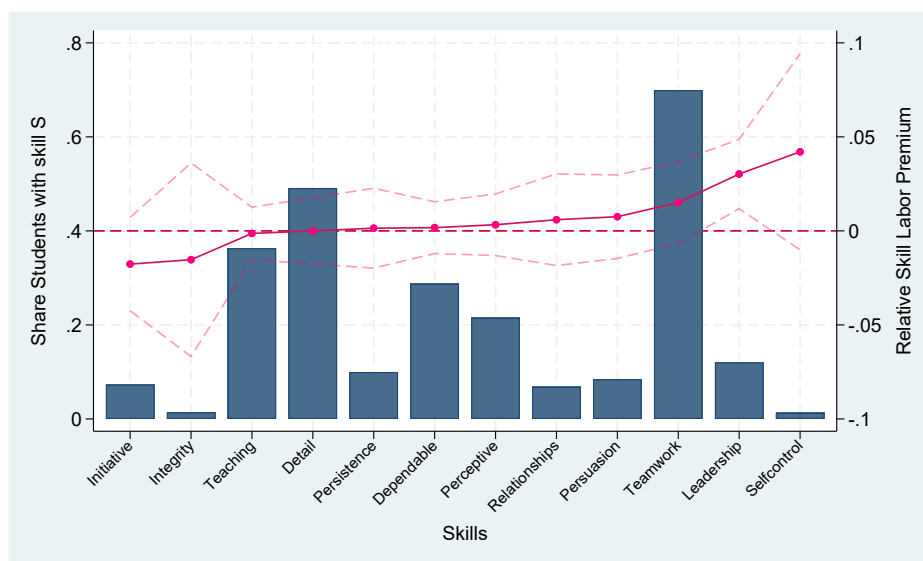
$$Y_{it} = \alpha + \sum_{SoftSkill \neq Detail} \theta_s SoftSkill_i + \theta Cognitive_i + \delta X_{it} + \epsilon_{it} \quad (5)$$

In this specification, each θ_s captures the relative return to soft skill s , with *Attention to Detail* as the omitted soft-skill category. Because the sample is restricted to students who display at least one soft skill, the coefficients compare students within the soft-skill group rather than with students who display no soft-skill signal. The control vector includes cognitive skill, graduation GPA, family income, ethnicity, gender, transfer status, major, admission-year, and labor outcome year fixed effects. The wage specification additionally includes industry and firm-location fixed effects. Errors are clustered at the student level.

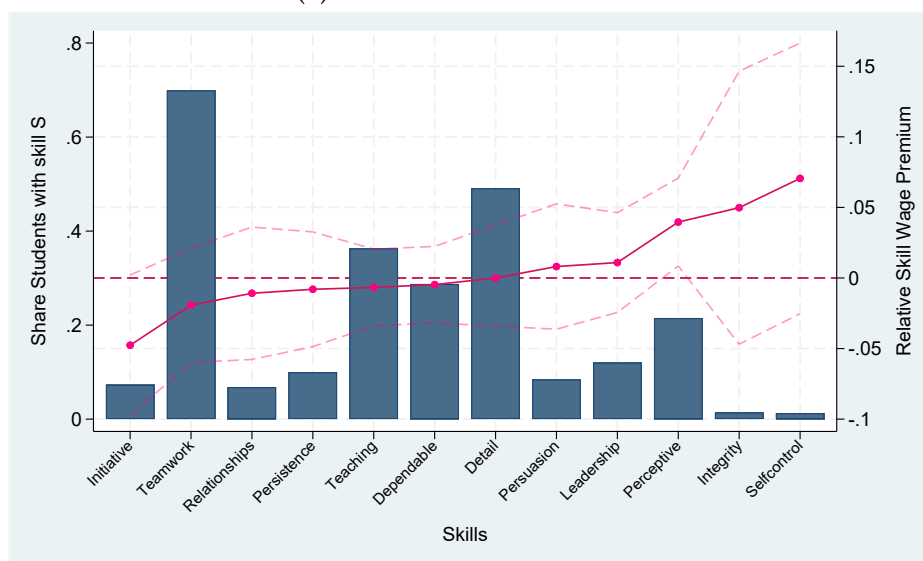
Figure 5 plots the full-sample estimates for the individual soft skills. The point estimates vary, but most are statistically indistinguishable from the omitted category. For labor force participation, self-control, leadership, and teamwork have the three largest positive estimated premiums. For wages, self-control, integrity, and perceptiveness have the three largest positive estimated returns. Self-control therefore stands out as a skill associated with both labor force

participation and wages. Beyond self-control, however, the skills most strongly associated with participation differ from those associated with wages, suggesting that the skills related to obtaining and maintaining employment are not necessarily the same as those rewarded through higher wages.

Tables 7 and 8 summarize the three largest positive point estimates in the full sample and by gender and URM status. Although the rankings vary somewhat, they are broadly similar across subgroups. Self-control features prominently for both labor force participation and wages, particularly among men and non-URM students, while perceptiveness, leadership, and integrity also frequently appear among the highest-ranked skills. URM students are the main exception to this broad similarity: integrity ranks first for both labor force participation and wages and appears more prominently for URM students than for other groups.



(a) Relative Labor Premium



(b) Relative Wage Premium

Figure 5: Relative returns to individual soft skills.

Notes: This figure reports estimates from model 5 for individual soft skills and each labor-market outcome, together with the share of students in the analysis sample who exhibit each soft skill. Bars report skill prevalence on the left axis; coefficient estimates and 95% confidence intervals are shown on the right axis. Attention to Detail is the omitted soft-skill category, so the plotted coefficients are relative to that category.

Table 7: Heterogeneity of Top 3 Soft Skills by Estimated Returns to Labor Participation

Rank	Full Sample	Male	Female	URM	Non-URM
1	Self-Control	Self-Control	Self-Control	Integrity	Self-Control
2	Leadership	Leadership	Leadership	Teamwork	Leadership
3	Teamwork	Dependability	Teamwork	Leadership	Relationships

Note: The table ranks the three largest positive estimated returns from regressions of labor force participation on the individual soft-skill indicators, estimated in the full sample and separately by gender and URM status. The sample is restricted to students exhibiting at least one soft skill. Regressions control for cognitive skill, graduation GPA, family income, ethnicity, gender, transfer status, major, admission-year, and labor outcome year fixed effects. Standard errors are clustered at the student level.

Table 8: Heterogeneity of Top 3 Soft Skills by Estimated Returns to Wages

Rank	Full Sample	Male	Female	URM	Non-URM
1	Self-Control	Self-Control	Perceptiveness	Integrity	Self-Control
2	Integrity	Integrity	Integrity	Persuasion	Perceptiveness
3	Perceptiveness	Perceptiveness	Self-Control	Persistence	Integrity

Note: The table ranks the three largest positive estimated returns from regressions of log annual wages on the individual soft-skill indicators, estimated in the full sample and separately by gender and URM status. The sample is restricted to students exhibiting at least one soft skill. Regressions control for cognitive skill, graduation GPA, family income, ethnicity, gender, transfer status, major, industry (NAICS), firm-location (ZIP code), admission-year, and labor outcome year fixed effects. Standard errors are clustered at the student level.

7. Student–Major Skill Mismatch

The results above show that narrative evaluations can be used to recover persistent skill signals and that these signals predict long-run labor outcomes. Skill signals may matter not only through their *levels* (who has which skills) but also through allocation if students fare differently when their revealed skill profiles align more closely with the typical skill bundle of their chosen major. This motivates an application that uses the narrative-derived skill measures to quantify student–major match quality and to test whether mismatch predicts labor market outcomes. I then examine descriptively whether greater exposure to narrative evaluations coincides with higher match quality, a pattern consistent with individualized feedback supporting academic sorting.

Measuring student–major skill similarity

I represent each student i with a binary vector of $K = 14$ skills extracted from narrative evaluations under the strict persistence rule in Section 5. Let $s_{ik} = 1$ if student i exhibits skill k in at least 50% of their narrative evaluations, and $s_{ik} = 0$ otherwise. I represent each major j with an average skill vector computed among students who declared that major between 1985 and 2010. For each skill k , let $m_{jk} \in [0, 1]$ denote the share of students in major j who exhibit skill k . Student–major match quality is measured using cosine similarity between the student and major vectors:

$$\text{Skill Similarity}_{ij} = \frac{\sum_{k=1}^K s_{ik} \cdot m_{jk}}{\sqrt{\sum_{k=1}^K s_{ik}^2} \cdot \sqrt{\sum_{k=1}^K m_{jk}^2}} \times 100 \quad (6)$$

which is bounded between 0 and 100 for students with at least one detected skill. Smaller values correspond to weaker alignment or mismatch between a student’s revealed skills and the typical skill bundle of a major. Figure 6 plots the distribution of Skill Similarity, illustrating substantial dispersion in match quality across students even within the same institution and major.

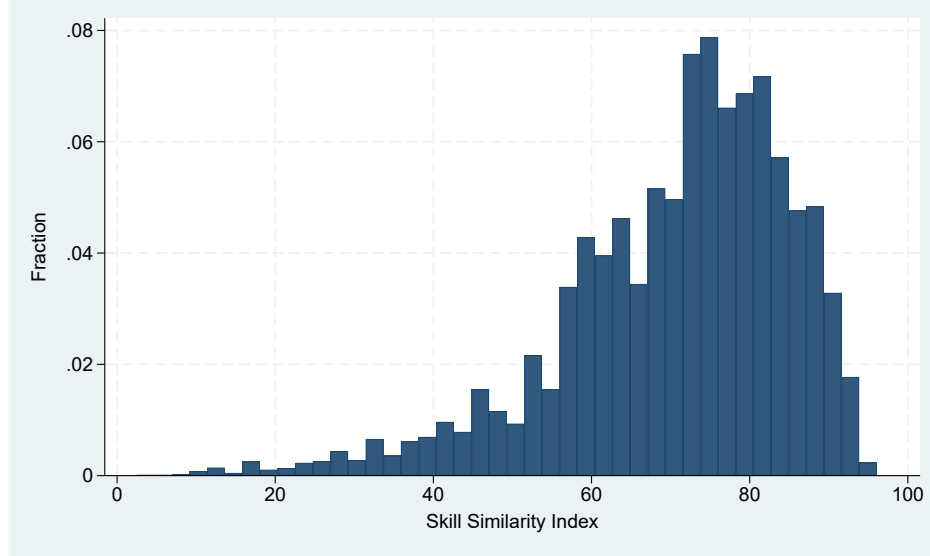


Figure 6: Distribution of Student–Major Skill Similarity

Note: This figure plots the distribution of skill similarity between each student’s narrative-derived skill vector and the average skill vector of their declared major. Skill similarity is measured using Equation 6 and is defined for students with at least one detected skill. Higher values indicate better alignment between a student’s revealed skills and the typical skill bundle of the major.

Mismatch Penalty

I estimate whether mismatch predicts labor market outcomes by regressing labor force participation and log annual wages on student–major mismatch (the negative of the Skill Similarity index), controlling for socio-demographics, academic performance, and academic environment. The baseline specification is:

$$Y_{it} = \alpha + \beta \text{Mismatch}_i + \delta X_{it} + \varepsilon_{it} \quad (7)$$

where Y_{it} denotes labor force participation or log wages for student i at time t . The control vector X_{it} includes ethnicity, gender, transfer status, family income, graduation GPA, major, admission-year and labor outcome year, so the coefficient is estimated from variation among students with similar observed characteristics, including students within the same major and admission cohort, who differ in how closely their narrative-based skills align with their field’s typical skill profile. Errors are clustered at the student level. As in prior sections, these estimates should be interpreted as predictive associations rather than causal effects.

The results indicate a meaningful mismatch penalty for wages, but little relationship with labor force participation. As reported in Table 9, a 10-point increase in mismatch is associated with approximately 1.3 percent lower wages in the full sample. In the subgroup specifications, the wage penalty is statistically significant among women and non-URM students, but not

among men or URM students. In contrast, mismatch has little systematic relationship with labor force participation. The full-sample estimate is small and statistically insignificant, with only weak evidence of a negative relationship among women. This pattern is consistent with the earlier finding that the narrative-derived skill measures are more informative about wages than about the extensive margin of labor-market participation.

Table 9: Student–Major Skill Mismatch and Labor Market Outcomes

	(1) Labor	(2) Male	(3) Female	(4) URM	(5) nURM	(6) Wage	(7) Male	(8) Female	(9) URM	(10) nURM
Mismatch	-0.00032 (0.00020)	-0.00006 (0.00031)	-0.00051* (0.00028)	-0.00050 (0.00038)	-0.00024 (0.00024)	-0.00128*** (0.00048)	-0.00045 (0.00074)	-0.00187*** (0.00062)	-0.00089 (0.00086)	-0.00143** (0.00057)
Base Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adj-R2	0.0301	0.0259	0.0358	0.0342	0.0292	0.1241	0.1325	0.1151	0.1246	0.1247
Obs (K)	218	91	127	62	156	140	59	81	41	99
# Stud (K)	18	8	11	5	13	15	6	9	4	11
Avg Y	0.64	0.65	0.63	0.66	0.63	100K	114K	89K	94K	103K

*Note: This table estimates model 7. Columns (1)–(5) use labor force participation as the dependent variable; Columns (6)–(10) use log annual wages. Mismatch is defined as the negative of the 0–100 student–major Skill Similarity index, so higher values indicate weaker alignment between a student’s narrative-derived skill profile and the typical skill profile of the student’s major. Base controls include ethnicity, gender, transfer status, family income, and graduation GPA. All specifications include major, admission-year, and labor outcome year fixed effects. Standard errors are clustered at the student level. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.*

Are narrative evaluations associated with greater student–major alignment?

A final question is whether exposure to narrative evaluations is associated with greater student–major alignment. Narrative evaluations provide individualized, multidimensional feedback that may help students better understand their skillset and how it aligns—or misaligns—with different majors. Figure 7 presents cohort-level evidence consistent with this channel by plotting student–major skill similarity by admission cohort. The figure is descriptive, but its timing is suggestive: average match quality rises through the mid-1990s, remains relatively high over the following cohorts, and declines among later cohorts with progressively less exposure to narratives before their discontinuation in 2010. This pattern is consistent with a potential channel through which qualitative evaluations could matter for long-run wages even in the presence of letter grades: narrative feedback may help students select fields that better align with their comparative strengths. The cohort evidence cannot, however, distinguish this mechanism from changes in student composition or other contemporaneous shifts.

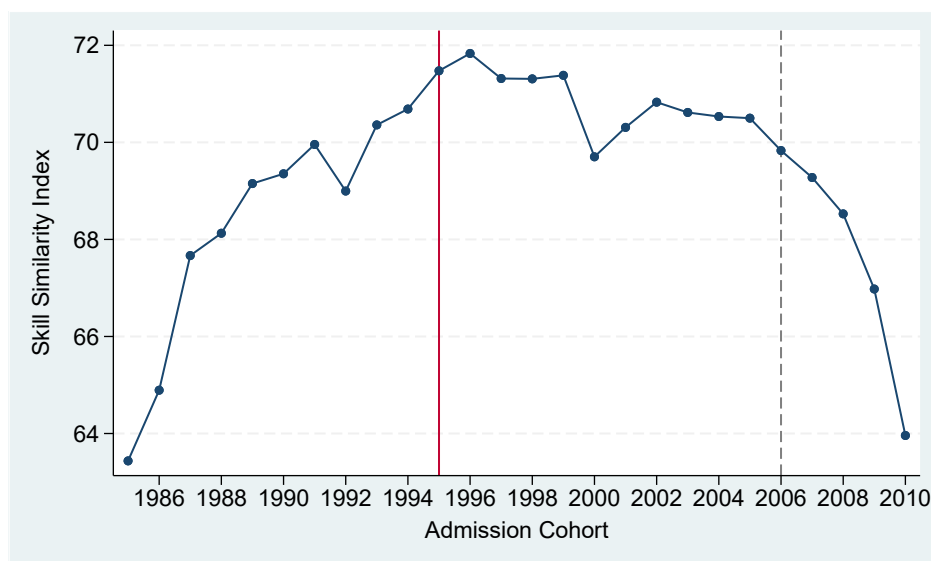


Figure 7: Student–Major Skill Similarity by Admission Cohort

Note: This figure plots average student–major skill similarity by admission cohort. Skill similarity is measured as the cosine similarity between each student’s narrative-derived skill vector and the average skill vector of their declared major (Equation 6). Higher values indicate better alignment between students’ revealed skills and the typical skill bundle of their chosen major. The solid vertical line at the 1995 admission cohort marks the point at which entering cohorts begin to receive meaningful exposure to optional letter grading after its 1997 introduction; the dashed line at the 2006 cohort marks the beginning of cohorts whose college years increasingly overlap the phase-out of narratives before their 2010 discontinuation.

8. Conclusion

This paper shows how instructor-generated narrative evaluations can be used to recover signals of student skills and competencies that conventional grades do not separately record. Using UC Santa Cruz academic records linked to California administrative wage data, I first show that narrative evaluations contain predictive information about later wages beyond conventional GPA.

I then use the narrative text to construct student-level measures of cognitive and fine-grained soft skills. These measures reveal persistent differences in the skills instructors detect across students, as well as systematic variation across the college career and across academic pathways. Their labor-market relevance differs by outcome. Broad skill measures have little systematic relationship with labor force participation, while cognitive skills retain a positive wage premium after accounting for academic field and labor-market sorting. The individual-skill results further show that the soft skills most closely associated with participation are not necessarily the same as those most closely associated with wages, and the interaction results provide some evidence that combinations of skills may matter for particular groups.

Finally, I use the narrative-derived skill profiles to study student–major mismatch. Students whose revealed skills are less aligned with the typical skill bundle of their chosen major subsequently earn lower wages, while mismatch has little relationship with labor force participation. Cohort patterns also show that greater exposure to narrative evaluations coincides with stronger student–major alignment. Although this evidence is descriptive, it is consistent with the possibility that richer feedback can help students understand how their strengths align with different academic paths.

Taken together, the findings highlight the informational limits of conventional transcript measures. Narrative evaluations provide a richer record of student performance, reveal economically relevant dimensions of skill, and make it possible to study how those skills align with educational choices. More broadly, the results suggest that both the content of student skills and their alignment with academic pathways matter for later labor-market outcomes. Because the empirical relationships in this paper are primarily predictive and associational, they do not establish that narrative evaluations or the measured skills causally improve labor-market outcomes or academic sorting. Establishing whether richer qualitative assessment systems can causally affect skill development, educational choices, or labor-market matching remains an important question for future research.

9. References

- Adhvaryu, A., Kala, N., & Nyshadham, A. (2023). "Returns to On-the-Job Soft Skills Training". *Journal of Political Economy*, 131(8) (cit. on p. 5).
- Aghion, P., Bergeaud, A., Blundell, R., & Griffith, R. (2024). "Social Skills and the Individual Wage Growth of Less Educated Workers". *Working Paper* (cit. on p. 5).
- Arcidiacono, P., Bayer, P., & Hizmo, A. (2010). "Beyond Signaling and Human Capital: Education and the Revelation of Ability". *American Economic Journal: Applied Economics*, 2(4), 76–104 (cit. on p. 5).
- Aryal, G., Bhuller, M., & Lange, F. (2022). "Signaling and Employer Learning with Instruments". *American Economic Review*, 112(5), 1669–1702 (cit. on p. 5).
- Bleemer, Z. (2018). "The UC ClioMetric History Project and Formatted Optical Character Recognition". *Center for Studies in Higher Education Research Paper Series*, 18(3) (cit. on p. 9).
- Bleemer, Z. (2019). "Gender Stereotypes in Professor-Student Interactions". *Working Paper* (cit. on p. 9).
- Chetty, R., Deming, D., & Friedman, J. (2023). Diversifying Society's Leaders? The Determinants and Causal Effects of Admission to Highly Selective Private Colleges. *NBER Working Paper* 31492. <https://doi.org/10.3386/w31492> (cit. on p. 5).
- Clark, D., & Martorell, P. (2014). The Signaling Value of a High School Diploma. *Journal of Political Economy*, 122(2), 282–318 (cit. on p. 5).
- Deming, D. (2017). "The Growing Importance of Social Skills in the Labor Market". *The Quarterly Journal of Economics*, 132(4), 1593–1640 (cit. on pp. 2, 4).
- Deming, D., & Kahn, L. (2018). "Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals". *Journal of Labor Economics*, 36(S1) (cit. on pp. 2, 4, 5).
- Deming, D., & Weidmann, B. (2021). "Team Players: How Social Skills Improve Team Performance". *Econometrica*, 89(6), 2637–2657 (cit. on p. 5).
- Diamond, R., & Persson, P. (2016). "The Long-term Consequences of Teacher Discretion in Grading of High-stakes Tests". *Working Paper* (cit. on p. 5).
- Dorn, D., Schoner, F., Seebacher, M., Simon, L., & Woessmann, L. (2025). "Multidimensional Skills on LinkedIn Profiles: Measuring Human Capital and the Gender Skill Gap". *Working Paper* (cit. on p. 5).
- Edin, P.-A., Fredriksson, P., Nybom, M., & Öckert, B. (2022). "The Rising Return to Noncognitive Skill". *American Economic Journal: Applied Economics*, 14(2), 78–100 (cit. on pp. 2, 4).
- Fredriksson, P., Hensvik, L., & Nordstrom, O. (2018). "Mismatch of Talent: Evidence on Match Quality, Entry Wages, and Job Mobility". *American Economic Review*, 108(11), 3303–3338 (cit. on p. 5).
- Guvenen, F., Kuruscu, B., Tanaka, S., & Wiczer, D. (n.d.). "Multidimensional Skill Mismatch". *American Economic Journal: Macroeconomics*, 12(1), 210–244 (cit. on p. 5).
- John H., R. J. M., & Willett, J. B. (2000). "Estimating the Labor Market Signaling Value of the GED". *The Quarterly Journal of Economics*, 115(2), 431–468 (cit. on p. 5).

- Lindqvist, E., & Vestman, R. (2011). "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment". *American Economic Journal: Applied Economics*, 3, 101–128 (cit. on p. 4).
- Lise, J., & Postel-Vinay, F. (2020). "Multidimensional Skills, Sorting, and Human Capital Accumulation". *American Economic Review*, 110(8), 2328–76 (cit. on p. 5).
- Schwerdt, G., & Woessmann, L. (2017). "The Information Value of Central School Exams". *Economics of Education Review*, 56, 65–79 (cit. on p. 5).
- UC-Academic-Council. (2020). *Report of the UC Academic Council Standardized Testing Task Force (STTF)*. University of California System. (Cit. on p. 5).
- Weinberger, C. (2014). "The Increasing Complementarity between Cognitive and Social Skills". *The Review of Economics and Statistics*, 96(5), 849–861 (cit. on pp. 2, 4).

10. Appendix

Sentiment Dictionary for Narrative Evaluations

This appendix reports the full set of proxy terms (in natural language form) used to construct the sentiment-based evaluation score. Each term is mapped to a numeric value corresponding to the UCSC narrative evaluation scale (Excellent = 4, Very Good = 3, Good/Fine = 2, Bad = 1, Very Bad = 0). In the empirical implementation, all terms are reduced to their stemmed roots (e.g., “excellent,” “excellence” → “exc”) to capture variation in tense and phrasing.

Excellent (4):	excellent, outstanding, superior, flawless, perfect, remarkable, brilliant, exemplary, master, top-notch
Very Good (3):	very good, strong, impressive, competent, commendable, admirable, effective, solid, positive, quite good, above average, well
Good / Fine (2):	good, fine, satisfactory, adequate, decent, fair, acceptable, average, okay, fairly good, successful, pass
Bad (1):	bad, insufficient, weak, limited, poor, inconsistent, needs improvement, struggling, marginal, below average, minimal
Very Bad (0):	very bad, fail, unacceptable, deficient, unsatisfactory, incompetent, lacking, dismal, unsuccessful, inferior

Skill Dictionaries

This appendix reports the full set of proxy terms (in natural language form) used to construct each skill indicator. In the empirical implementation, all terms are reduced to their stemmed roots (e.g., “collaborate,” “collaborating” → “collabora”) to capture variation in tense and phrasing. A skill is flagged when any of the corresponding stems appears in the narrative text.

Cognitive Skill

Analytical:	analysis, logic, rationality, reasoning, data, modeling, estimation, quantitative, interpretation, scientific, investigation, inquiry, memory, mathematics, problem-solving, research, cognition, statistics
--------------------	--

Social Soft Skills

Teamwork:	team, collaborate, communication, coordinate, compromise, unity, cooperate, contribute, helpful, partner, participate, member, group, project
Teaching:	teach, education, tutor, mentor, explain, train, guide, coach, present, clarity, patience, instruction, communication
Persuasion:	compromise, mediation, resolution, agreement, persuasion, strategy, negotiation, rhetoric, communication, argument, appeal, opinion, deliberation, compelling

Perceptiveness:	empathy, awareness, sensitivity, observation, intuition, listening, emotional, responsive, perceptive, thoughtful, helpful, compassion, consideration, attentive, receptive, tact
Leadership:	leadership, inspire, influence, decision-making, judgment, vision, organization, strategy, delegation, planning, direction, guidance, initiative, command management
Relationships:	personable, friendly, approachable, empathy, interpersonal, relationships, connection, social, networking, inclusive, interaction, energetic, enthusiasm

Non-Social Soft Skills

Persistence:	determination, persistence, perseverance, motivation, commitment, resilience, achievement, effort, endurance, grit, tenacity, diligence, patience, resolve, dedication
Initiative:	proactive, decisiveness, motivation, drive, enthusiasm, initiative, autonomy, confidence, resourcefulness, independence, creativity, curiosity, inquiry
Adaptability:	openness, versatility, responsiveness, resilience, agility, resourcefulness, readiness, adaptability, flexibility, adjustment, multitasking, improvisation
Self-Control:	composure, patience, discipline, restraint, professionalism, poise, resilience, stability, endurance, balance, control, perseverance, calmness, tolerance
Attention to Detail:	meticulousness, precision, carefulness, thoroughness, detail, accuracy, quality, organization, diligence, consistency, high standards
Dependability:	reliability, dependability, consistency, completion, trustworthiness, responsibility, accountability, punctuality, commitment, hard work, follow-through
Integrity:	ethics, honesty, uprightness, honor, trust, transparency, fairness, morality, credibility, principles, accountability, responsibility, integrity, values