Linguistic, cultural, and narrative capital: computational and human readings of transfer admissions essays

AJ Alvero1 · Jasmine Pal2 · Katelyn M. Moussavian2

Received: 12 May 2022 / Accepted: 7 September 2022 / Published online: 30 September 2022 © The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd. 2022

Abstract
Variation in college application materials related to social stratification is a contentious topic in social science and national discourse in the United States. This line of research has also started to use computational methods to consider qualitative materials, such as personal statements and letters of recommendation. Despite the prominence of this topic, fewer studies have considered a fairly common academic pathway: transferring. Approximately 40% of all college students in the US transfer schools at least once. One quirk of the system is that students from community colleges are applying for the same spots for students already enrolled in four year schools and trying to transfer. How might different aspects the transfer application itself correlate with institutional stratification and make students more or less distinguishable? We use a dataset of 20,532 transfer admissions essays submitted to the University of California system to describe how transfer applicants vary linguistically, culturally, and narratively with respect to academic pathways and essay prompts. Using a variety of methods for computational text analysis and qualitative coding, we find that essays written by community college students tend to be distinct from those written by university students. However, the strength and character of these results changed with the writing prompt provided to applicants. These results show how some forms of stratification, such as the type of school students attend, inform educational processes intended to equalize opportunity and how combining computational and human reading might illuminate these patterns.

Keywords Transfer admissions essays · Computational text analysis · Qualitative coding · Classification · Institutional stratification · sociocultural capital · Accumulated capital · Evaluative contexts

Jasmine Pal and Katelyn M. Moussavian have contributed equally to this work. Author order was determined randomly.

* AJ Alvero
aalvero@ufl.edu
https://scholar.google.com/citations?user=Kufgl3QAAAAJhl=en

Extended author information available on the last page of the article
Introduction

US higher education is distinctive in its emphasis on student choice, autonomy, and mobility despite frequently failing to meet such a promise through the many examples of educational inequality and stratification [1]. For example, the course selection decisions made by first year college students are stratified by social class [2]. One of the most frequently discussed sites of these ideas and arguments is college admissions, particularly through the well-understood correlations between household income and test scores [3]. This line of research, along with the Covid-19 pandemic, have coalesced to push many universities’ decisions to drop standardized testing from first year admissions.

A new wave of research in and around college admissions has leveraged computational methods to analyze how other elements in college applications are similarly correlated with other information about the applicant. These include examples of how admissions essays are strong predictors of gender, income, and SAT scores [4, 5]. Other studies have considered letters of recommendation [6, 7] and which high schools are targeted by universities for recruitment [8]. However, this growing literature along with the now classic approaches to analyzing variation and stratification in application materials have rarely considered an important feature of US higher education: transfer admissions.

In the US, transferring is the process where students from community colleges or bachelors degree granting institutions apply to college or university to complete their undergraduate education. These institutions are stratified by the types of degrees they are able to award as well as through their selectivity (community colleges have open admissions, four year schools do not) and economic opportunities (bachelors degrees often lead to higher paying jobs and are a prerequisite into high paying professional degree programs). Recent research has found that educational outcomes (i.e., graduation rates after transferring) are stratified by the direction of the transfer [9]. The “direction” refers to the academic pathway of the applicant: “lateral” for students transferring from a four year institution to another four year institution; “vertical” for students transferring from community college to a four year university; and “reverse” for students transferring from four year institutions to community college (the aforementioned study compared lateral and reverse transfers). Studies have examined how coursework, gender, remedial classes, and other student characteristics lead to successful vertical transfer [10]. But fewer studies have examined how, like first year application materials, dimensions of stratification can be connected to variation in transfer application materials.

Despite receiving less attention in academia and popular discourse, the transfer system is quite popular (nearly 40% of all undergraduates transfer schools at least once) and emblematic of the ideals in USA higher education of giving students flexibility and “second chances” [11]. Enrolling in community colleges to eventually transfer is also popular in Latinx communities. Latinx students are the largest ethnic

1 Gender neutral term referring to all people who descend from people across Latin America.

Springer
group enrolled in California community colleges: 46% of Latinx students graduating from the highest performing high schools in the state still end up going to community college first [12]. But because community colleges also have to serve a wider variety of students along with their mission of helping students transfer out, their experiences are much different from students already enrolled in four year schools. Analyzing application data, such as admissions essays, could provide unique snapshots into these differences as articulated by the students themselves and how they vary by academic pathway. Doing so with computational methods would also expand the burgeoning literature of computation and systems of evaluation. It could also push current notions of fairness and bias in data science, which often focuses on race and gender [13], to consider how institutional experiences and pathways are also fateful in data generation processes as well as analysis and practice.

In this paper, we apply computational and human reading of text to measure variation in transfer admissions essays based on the institutional stratification of the applicants. Generally, students at community colleges have access to very different educational resources and experiences than students enrolled in four year schools, and these differences likely go beyond traditional outcome metrics. Our work therefore highlights another way stratification can shape educational mechanisms but also how these differences can be measured in text and how the transfer application itself could help reduce these differences. Our analysis focuses on three elements that are stratified in many social contexts: linguistic, cultural, and narrative capital. These constructs have been used in the theoretical and empirical development of the sociology of evaluation [14] and culture. We consider both the stratified academic pathway of the applicants (lateral and vertical) and the essay prompt students responded to. We find that the essays are fairly distinctive based on the applicant’s educational pathway, though these differences are also sensitive to the language of the essay prompt they were given. Outside the research context, systems of evaluation might consider these results before or after adopting technology related to evaluation. See Fig. 1 for a conceptual diagram of the paper.
Admissions essays and beyond: computational methods and systems of evaluation

Previous work studying admissions essays has been mostly descriptive and correlational. For example, studies using both computational and qualitative methods have found that applicant gender, income, college GPA, and thematic choices have all been found to be correlated with admissions essays [4, 5, 15, 16]. Another study found that word vectors trained on essays written by applicants in the highest income quartile performed better on the Google Analogy Test Set and various word similarity tests than essays written by the two lowest income quartiles [17]. The computational studies have been consistent in their results: applicants characteristics and information are predictable from essays.

Other studies have analyzed essays using other methods. A study of admissions essays submitted to selective universities in the UK found narrative and orthographic patterning based on the type of school attended by the applicants, such as students from elite backgrounds name-dropping their elite networks and connections [18]. Findings from a study of diversity statements submitted to the University of Michigan found that students from similar class backgrounds would write about similar topics, regardless of race [19], and the information dashboards used in holistic review were found to improve equity in admissions outcomes for students from lower socioeconomic backgrounds [20]. These studies match the results of the computational analyses to a degree but not entirely due to the unique affordances of qualitatively oriented methods. We took these as cues to complement our computational analyses with qualitative coding.

Data science and institutional stratification

Some of the guiding principles of the new wave of data science for education research are the abilities to ask “old questions with new methods and data” but also “new questions with old data” [21–23]. Computational sociology has adopted similar positions by seizing opportunities to analyze new sources of data to study social interactions and processes [24]. However, most of the “new types” of data are often unstructured, non-numerical data (e.g., text, visual, audio) that have long been analyzed using qualitative methods. Sociologist Laura Nelson has taken this insight and presented frameworks to combine computational and human reading of text to both deepen analysis towards a framework of “computational grounded theory” [25] and to generate points of reference and comparison [26]. Given the relative paucity of information about transfer application materials and how they might be stratified, we follow Nelson’s frameworks towards methodological plurality to generate deeper, multiplex insights into a key social process.

Though research into transfer application materials is relatively sparse, social scientists in the US have generated insights into the transfer process from other
perspectives. For example, a national study found that students were less likely to transfer laterally if they were more socially integrated into their school [27]. Sociologists have also observed stratification in the community college system and constituent transfer processes [28]. In an even broader sense, other scholars have pointed out that there is stark stratification across all institutions of higher education, including graduate school [29, 30]. Our study applies the “new ways” and “new data” frameworks described above to examine how applicants are stratified by the type of institutions they attend at the moment of application to reframe the transfer application itself as an outcome of different types of acculturation and socialization.

Prior research on transfer and institutional stratification in higher education has shown in various ways that students combine their own sociocultural capital [31] with those acquired and accumulated at their respective institutions. If we accept the premise that lateral applicants have a distinguishable forms of sociocultural capital as at least partly evidenced by their ability to enroll in a four year school from high school, we might also expect that these differences not only change but also become more pronounced given their experiences in colleges relative to the experiences of students in community college. Likewise, students attending community colleges are also experiencing an accumulation of different types of sociocultural capital. In our study, we look specifically at how this manifests through differences in essay content and style, the culture of how closely students follow prompts, and narrative patterns. Understanding stratification in what many might assume is an “equalizing” force is important not just to scholarship but also practice.

Framework

Measuring institutional stratification and variation in transfer application materials could be done in many ways, but here we apply concepts from Bourdieusian sociocultural capital [31]. Specifically, we consider linguistic, cultural, and narrative capital with respect to academic pathway (lateral or vertical). Other computational analyses of text data have adapted similar theoretical constructs and measured macro level factors [32, 33]. We also consider how specific essay prompts might activate different types of capital that either exacerbate or reduce variation [34]. We hypothesize that since transfer applicants are stratified by the types of schools they attend, the content of their essays likewise vary with respect to their academic pathway and in response to specific essay prompts. There are many different perspectives from which to measure this variation, but we adapt sociocultural frameworks and theories in our analysis.

Linguistic capital

Bourdieu described linguistic capital as the knowledge and capacity to use linguistic forms and styles in a given social context [35]. Though Basil Bernstein used different terminology, he also noted how linguistic features varied and that schools tended
to prefer the language practices of students from higher social classes [36]. While their focus was on spoken language, we use computational methods to test these theories with textual data. In our study, we measure the strength of the relationship between academic pathway and transfer essays based on word choice patterns and writing style to observe how institutional stratification relates to linguistic variation. Given our use of the same methods, we can also compare these results with the aforementioned research on first year college admissions essays which found strong correlation between essay content and style, income, and SAT scores.

**Cultural capital**

Cultural capital, perhaps the most well known of Bourdieu’s theories of capital, is defined as the cultural practices, knowledge, and performances that are enacted in social situations [31]. We pair this concept with Durkheim’s theories of rules and education, specifically that schools teach and reinforce rules and approaches to life that students carry in and out of school [37], and examine cultural practices around how closely transfer applicants follow the wording of the essay prompts. Put differently, do lateral and vertical applicants follow the “rules” of the essay prompt in the same way? Computers can be used to identify markers of cultural capital, such as references to specific named entities in text, something we take advantage of here.

**Narrative capital**

Although narrative capital is similar in spirit to other forms of Bourdieusian capital, [38] describes it as the raw narrative materials that go into storytelling and self presentation. It therefore relies much on mastery of both language and culture [38]. Computational methods could be good at identifying and modeling narrative capital, but there also is no reported description of the narrative tropes and themes which comprise the genre of “transfer admissions essay” comparable to something well-known like Hollywood movie narrative arcs [39]. To capture narrative capital, we therefore qualitatively code and analyze the essays. Doing so could inform future work on personal statements beyond numerical measures.

**Data**

The data for the study come from transfer applications submitted by every student who identified as Latinx to the University of California in the 2015–2016 (n = 10,347) and 2016–2017 (n = 10,305) academic years. In this study, we focus on variation in the essays based on their transfer pathway: “lateral” for students applying from four year universities and “vertical” for students applying from community college. Our data include the College Entrance Examination Board (CEEB) code from the schools the students applied from; these codes were used to label students as coming from either bachelors granting institutions (lateral) or community college (vertical). See Table 1 for an overview of the data. A link to
the data and code needed to replicate the study is available in the supplementary materials. Note that the raw essays are not available due to privacy issues.

The essay prompts students wrote for changed over the years from asking students about naming specific experiences like “internships” and “interest in the field” to a more open-ended question about what students have done to prepare for their intended major (see Table 2). The format also changed from 1000 words for two essays in 2015–2016 to 1400 words for four essays in 2016–2017. To account for the change in the length of the essays, we use percentages of certain features of the text rather than raw counts. Some of our analyses take advantage of the change in essay prompts and measure how the essays changed across the two years. Our paper therefore focuses on the interplay between transfer admissions essays, academic pathways (lateral and vertical), and essay prompts.

As a first pass, we analyzed the essays using commonly used, relatively simple metrics. These included Simpson’s diversity index, a statistic sometimes used to measure lexical diversity in text [40]; the Flesch-Kincaid grade level readability metric [41]; and average word counts and standard deviations per document. All of the metrics were similar for the lateral and vertical applicants. The lexical diversity index was the same for both groups up to the fourth decimal (0.012). The average readability scores were 12.01 for lateral applicants and 11.88 for vertical applicants (not a statistically significant difference based on t-testing). Vertical applicants wrote slightly longer essays than lateral applicants on average for both years (respectively: 479 vs. 468.84 words in 2015–2016; 312.29 vs. 308.35 words in 2016–2017). The standard deviations for the word counts were also quite similar for vertical and lateral applicants (respectively: 149.73 vs. 148.32 words in 2015–2016; 85.25 vs. 87.69 words in 2016–2017). The superficial similarity of the documents combined with strong results reported elsewhere using computational approaches helped motivate our selection of methods.

Table 1 Transfer applications by academic pathway (lateral or vertical) and application year (2015–2016 and 2016–2017) and aggregate counts

| Group    | 2015–2016 | 2016–2017 | Total   |
|----------|-----------|-----------|---------|
| Aggregated | 10,347    | 10,305    | 20,652  |
| Lateral  | 850       | 789       | 1639    |
| Vertical | 9497      | 9516      | 19,013  |

Table 2 Transfer prompts in 2015–2016 and 2016–2017

University of California transfer prompts, 2015–2017

2015–2016 What is your intended major? Discuss how your interest in the field developed and describe any experience you have had in the field—such as volunteer work, internships and employment, participation in student organizations and activities—and what you have gained from your involvement.

2016–2017 Please describe how you have prepared for your intended major, including your readiness to succeed in your upper-division courses once you enroll at the university.
Methods

We use outputs from the computational models as independent variables to analyze using various statistical techniques. We then complement these analyses with inductive qualitative coding to compare the human and machine reading and provide deeper insight into the specific tropes and themes transfer applicants deploy in their essays. Each of the methods correspond to specific theories of sociocultural capital and their stratification among transfer applicants by their academic background. The results of the paper are also organized this way and are presented as study one, two, and three. We hope the three sets of findings can become reference points for future studies of transfer admissions in the US.

Study one: essay content and style

First, to capture linguistic stratification we use correlated topic modeling (CTM) [42] and the LIWC software package [43] to model variation in essay content and style, following [4]. We use CTM to model thematic patterns with the STM package in R; when the primary STM function of this specific package is used without covariates, the code defaults to a fast implementation of CTM [44]. Prior to generating the model, we removed stopwords (using the SMART list of English stopwords [45]), stemmed the words (using the Snowball stemmer [46]), removed numbers and punctuation, and lowercased all letters. After removing these, we only included the 10,000 most frequent terms to account for rare and infrequent words. These are common textual preprocessing techniques that can potentially improve the quality of the topic model and downstream analysis [47], though this is not guaranteed to always be the case [48]. These procedures were also used in the aforementioned study of essay content and style [4]. We use the ldatuning package in R [49] to determine the number of topics (see Figures SM1 through SM3 in the supplementary materials). To model writing style and other patterns, we use LIWC (specifically LIWC-15). We exclude the LIWC feature of “Word Count” due to the change in essay formats between the years. Combined, these methods can show how linguistic capital varies among transfer applicants based on their educational histories.

Differences in the prevalence of topics and LIWC features were analyzed by educational pathway using a traditional comparative approach with t tests along with a supervised learning framework to predict whether applicants are applying from community college or bachelor’s granting institutions based on their essays. To account for the data imbalances, we use Welch’s t test and two different but complementary methods within the supervised learning framework. The first is the Inter-Model Vigorish (IMV), a new technique to analyze the outcomes of predictive models (specifically logistic regression) that works by comparing the predictive accuracy of a given model with another model that only predicts the mode outcome [50]. The results from this method could be compared to the common vigorish (percentage of winnings deducted from players) of 1% in blackjack which, on average, becomes 0.0099 cents per dollar for the casino.
Along with the IMV, we use undersampling and report prediction accuracies over several thousand trials. The first step in undersampling is to make the two groups (lateral and vertical) the same size. We follow standard procedures by generating a random sample of observations from the larger group that is the same size as the smaller group. Then, we use 10-fold cross validation to predict if essays were written by lateral applicants or vertical applicants. This procedure was repeated thousands of times. Repeated undersampling allows for a robust analysis where (1) classification accuracy can be compared with random guessing (expected accuracy of 50%) and (2) create points of comparison for how well text classifiers can predict academic pathways of applicants solely based on their essays and their respective prompts. We use logistic regression as our classifier based on how well it worked in a similar study [5]. These different approaches would show how specific thematic and stylistic features of writing vary for individual variables (eg. specific topics) but also in the aggregate among transfer applicants.

**Study two: named entity recognition**

Next, we perform an observational experiment to measure how the essays changed (or did not change) with the new prompt as a way to describe cultural stratification in transfer admissions essays. Specifically, we use named entity recognition (NER) to measure the effect that the new essay prompt had on students naming specific organizations, people, businesses, places, and other entities. The older prompt solicited students to describe the specific examples of how they participated in various organizations related to their field of interest, but the newer prompt was less specific. This study will examine if transfer applicants responded to the prompt by bringing up named entities even though the new prompt did not explicitly solicit them as before. The specific “culture” being measured here is not the types of named entities (eg. businesses vs. non-profits) but rather the culture of essay prompt interpretation. Since the new prompt did not specifically solicit named entities like it did the year prior, are students making fewer references to them?

The emerging causal inference with NLP literature has so far focused on word embedding [51] and topic modeling [52] methods. We extend this rapidly growing literature in our use of a finer grained method, NER, and in our more sociological context (as opposed to political science or traditional computer science). The first step was to use NER on the essays and tag every instance of a named entity being mentioned. To do this, we used the SpaCy package [53] in Python. SpaCy’s NER tags many different types of entities, including times and numbers, but we limited the tagging to people, organizations, and concepts. To verify the tags, we took random samples of essays and checked that all of the tags were referencing named entities and reflected entities the students were associating with their intended majors in various ways.

After tagging the essays with NER, we counted the number of words and number of characters that the applicants used in the named entities for a given essay. These were used to create the outcome of interest: the percentages of the essays (based on words and characters) that were used on named entities. Our percentage approach
was chosen to reflect the different word limits applicants had on their essays from year to year (478 words on average vs. 312 words on average). Focusing on word and character based percentages helps serve as robustness check (e.g., if some NER tags have very few words or many characters). Once the NER percentages per essay were calculated, we use propensity score weighting to make statistical adjustments to the data and control for any selection biases. See Table 3 for a full list of the variables used for weighting. The propensity score weighting was calculated with the average treatment effect (ATE) as the estimand [54] with the 2016–2017 essay prompt as the treatment.

While it is possible that some students in our data applied in 2015–2016 and 2016–2017, it is an unlikely scenario because transfer applicants to the UC must have already completed a significant amount of coursework (equivalent to an associates degree) and would be unable to complete additional units. It is also not likely that applicants in 2015–2016 were even aware that a new essay prompt was going to be used in 2016–2017. Combined, we assume the stable unit treatment value assumption (SUTVA) was not violated [55], giving us additional confidence in the data and findings. Note that we do not include reported household incomes in our analysis. We did this because approximately 20% of all transfer applicants report a household income either below $10,000 or no income at all. Among students reported an income of at least $10,000, there was also a weak correlation with the essays ($R^2 = \sim 0.05$). This is likely due to many community college students working at least part time and reporting that income rather than the income of their entire household.

**Study three: inductive qualitative coding**

Finally, we model narrative stratification with qualitative inductive coding. This approach complements the previous analyses because it provides a perspective on the essays that would be difficult to capture with purely computational methods [26]. For this analysis, we coded a stratified random sample of essays submitted by transfer applicants intending to major in computer science and sociology for each year. Computer science and sociology were selected because they are very different

| Table 3 Variables used for propensity score weighting |
|------------------------------------------------------|
| Variable                                             | Counts                        |
| Gender (M/F)                                         | $F = 10,434$                  |
|                                                     | $M = 10,068$                  |
| Latinx ethnicity                                     | Mexican = 14,792              |
|                                                     | Cuban = 325                   |
|                                                     | Puerto Rican = 567            |
|                                                     | Other = 2680                  |
| HSI status                                           | 17,293                        |
| First-gen college student                            | 16,388                        |
| Research school                                      | 921                           |
| Private or public                                    | 442                           |
disciplines and students presumably have different experiences leading up to declaring each as a major; they are also both popular majors and attract a wide variety of students. We ended up coding a total 198 essays divided roughly equally into fourths by intended major and application year.

First, we developed a qualitative codebook through an iterative process to identify the narrative themes and arcs described by the students in their essays. Once the codebook was finalized, we hand coded each of the essays to census. More information about the codebook is available in Table SM5 in the supplementary materials. We then report the codes that lateral and vertical applicants used the most differently over the two years. These were the codes that one group used more or less and the other group did the opposite. For example, we would highlight a code that lateral applicants used more in the second year but vertical applicants used less. We further filter the results by focusing on codes where the differences in code frequency for one group is at least three times the difference for another group. Using the previous example, this means we would present a code used by lateral applicants more in the second year only if the increase was at least three times larger than the decrease among the vertical applicants. These are the codes that get at the strongest narrative responses to the essay prompts for each group and intended major.

Results

Study one: variation in essay content and style

Three CTM models were generated: one for the 2015–2016 transfer essays; one for the 2016–2017 transfer essays; and one for all of the transfer essays. The outputs from the ldatuning package indicated that 30 topics for the individual year essays and 50 for the combined essays would be appropriate (Figures SM1 through SM3). The first analysis describes variation in the individual topics and dictionary features for vertical and lateral applicants among transfer applicants from the 2015–2016 academic year. We use t testing (specifically Welch’s t test to account for data imbalances) to compare the magnitudes of the individual variables for this first analysis. The results for the 2016–2017 and entire corpus are available in the supplementary materials. After, we use these same variables as inputs for text classification (described in more detail below).

The results comparing the individual topic and dictionary features from the 2015-2016 essays are described in Table 4. The topic patterns were generally similar across the years aside from two notable exceptions: the number of topics strongly associated with either lateral or vertical applicants was smaller in 2016–2017 (19 in 2015–2016 vs. 14 in 2016–2017); and a topic referencing various California-based research experiences that was associated with lateral applicants in the model for both years (“California research experiences”). These were noteworthy because they were unique but also because they foreshadowed subsequent results presented in the paper.

Of the 30 topics generated by the 2015–2016 model, 19 were strongly associated with either lateral (9) or vertical (10) applicants. Vertical applicants wrote more
about helping others, interpersonal relationships (specifically family and friends), and popular majors (eg. psychology). Lateral applicants wrote more about self-realization and specialized majors, often in STEM. These results are similar to topic modeling results of first year admissions essays that found essays using thematically comparable topics as the vertical applicants tended to come from students with lower income and SAT scores [4]. Conversely, lateral applicants tended to write more like first year applicants with higher reported income and SAT scores. The different majors students tended to write are also reflective of lower paying and higher paying career pathways (psychology and art history for vertical applicants, STEM and medicine for lateral applicants). The “Film and music” is somewhat of an outlier with respect to expected career earnings and might be an artifact of lateral students applying to the UCLA School of Theater, Film, and Television.

The dictionary results were also reflective of similar SES and SAT patterns previously described in studies of first year essays. Of the 89 (“Dash” was omitted, dropping the feature count from 90 to 89) features generated by LIWC, 37 were strongly associated with lateral (16) or vertical (21) applicants. Lateral applicants tended to

### Table 4  Summary of t test results comparing topics and dictionary features by academic pathway

Comparing essay features from vertical and lateral transfer applicants, 2015 ($p < 0.05$)

| Used more by vertical applicants | Used more by lateral applicants |
|----------------------------------|--------------------------------|
| **Topics**                       | **Dictionary**                 |
| Psychology                       | Clout                          |
| Family reflection                | Educational goals             |
| Helping others                   | Pronoun                        |
| Humanities/literature            | Ppron                           |
| Educational/social activities    | We                             |
| Social identity                  | They                           |
| Intellectual goals               | Adj                            |
| Language and culture             | Compare                        |
| Math reflections                 | Interrog                       |
| Humanities: art history          | Negemo                         |
|                                  | Sad                             |
|                                  | Social                          |
|                                  | Friend                          |
|                                  | Insight                         |
|                                  | Cause                           |
|                                  | Drives                          |
|                                  | Affiliation                      |
|                                  | Power                           |
|                                  | Risk                            |
|                                  | Focuspast                       |
|                                  | Home                            |
|                                  |                                 |
| **Topics**                       | **Dictionary**                 |
|                                  | Film and music                  |
|                                  | Educational goals              |
|                                  | Realization                    |
|                                  | Animals                        |
|                                  | STEM: CS                       |
|                                  | STEM: mechanical engineering   |
|                                  | Environment                    |
|                                  | STEM: biology                  |
|                                  | STEM: medicine                 |
|                                  | AllPunc                        |
|                                  | Comma                          |
|                                  | Colon                          |
|                                  | Quote                          |
|                                  | Apostro                        |
|                                  | OtherP                         |

The dictionary results were also reflective of similar SES and SAT patterns previously described in studies of first year essays. Of the 89 (“Dash” was omitted, dropping the feature count from 90 to 89) features generated by LIWC, 37 were strongly associated with lateral (16) or vertical (21) applicants. Lateral applicants tended to
use more punctuation (e.g., commas, colons, and quotation marks) and articles (e.g., “the”) whereas vertical applicants tended to use more pronouns and words not in the LIWC internal dictionary (“Dic”). The results for the 2016–2017 and combined essays were largely similar (see Tables SM1 through SM4). These follow similar patterns for features associated with income and SAT scores from first year applicants (negatively correlated for the vertical features and positively correlated for the lateral features) as well as final college GPA among accepted students in an analysis of admissions essays from accepted students to the University of Texas at Austin [16]. These results suggest that, aside from writing about specific intended majors, community college students tend to write about topics and in similar styles as first year applicants with lower income and SAT scores; the inverse was true for lateral applicants.

We then use the topics and dictionary features as inputs for text classification. The primary sets of results, one set for the full analysis using IMV and one set for the undersampled analysis reporting classification accuracies, are presented in Table 5. The classification models used to generate these results were trained independently of one another. Figure 2 is a visualization of the results described in Table 5. Both methodological approaches yielded similar results: the educational background (lateral or vertical) of applicants was more discernible in the essays written in 2015–2016. While most (nearly 92%) of applicants are vertical applicants, including information about their essay content improves our ability to predict their academic background. For the 2015–2016 essays, the IMV between a baseline prevalence model and one that uses essays was approximately 0.03. This is comparable to the IMV for predicting death within the next two years for those approximately 90 years old given additional information about their age, sex, and education level [50]. Though the IMV metrics were comparable, the prevalence was less skewed than the transfer essays (70 vs. 90% for the transfer essays), suggesting a stronger relationship with essay content and style among transfer applicants than fundamental health correlates and age of death among the most elderly. In the models for the 2016–2017 essays and entire corpus, the IMV drops to approximately one third of the IMV score for the 2015–2016 essays.

The results for the undersampled classification were similar: the essays written in 2015–2016 were more correlated with academic background than the

| Data       | Proportions | IMV   | 95% confidence interval | Accuracy |
|------------|-------------|-------|-------------------------|----------|
| 2015–2016  | 0.918       | 0.025 | [0.687, 0.753]          | 0.721    |
| 2016–2017  | 0.923       | 0.009 | [0.562, 0.636]          | 0.598    |
| Both       | 0.921       | 0.008 | [0.576, 0.630]          | 0.604    |

Results are reported two different metrics: IMV (left) and classification accuracy after repeated undersampling (right).
2016–2017 essays or the entire corpus. The final reported accuracy for the 2015–2016 analysis was 72%, more than 10 percentage points higher in classification accuracy than the 2016–2017 essays and the full corpus. Figure 2 presents a visualization of these differences and also shows that even the lowest accuracies from the undersampling analyses for the 2015–2016 essays were still higher than the highest accuracies for the 2016–2017 essays. These results are comparable to classification accuracies of gender (high of 80% accuracy) and household income (above or below median income; high of 69% accuracy) in a study of first year admissions essays [5]. However, that study was analyzing more data (283,676 essays), a larger number of features (counts of all unigrams in the corpus), and used more flexible models (deep learning classifier).

Combined, the results for study one suggest that transfer applicants are stratified not just in their experiences in community college vs. four year colleges and universities but that these differences also emerge in their writing styles and topical choices. However, differences among transfer applicants diminished once the
new essay prompt and format was released in the 2016–2017 school year, pointing to a simple but perhaps powerful way that essay prompts relate to variation in linguistic capital in transfer admissions.

**Study two: transfer essay prompts and named entity usage**

All of the essays were tagged using SpaCy’s NER functionality to count the number of named entities the students included in their essays. However, the following analysis focuses not on the number of tags per essay but rather the percentage of words and characters the tags comprise per essay. Calculating proportions based on words and characters allowed us to account for possible discrepancies in the character count for different tags (e.g., working at a family owned business or firm compared to single word companies like “Google”) and to reflect the writing situation: students had a limited number of words they could use in their essays and therefore had to be strategic in what they referenced. Put differently, rather than focus on the volume of named entities included in essays this analysis focuses on the proportions of the documents comprised of named entities and whether or not the new prompt changed applicant behavior.

Before analyzing the essays, applicants that did not report a gender (“Male” and “Female” were the only options at the time) and/or submitted an essay shorter than 100 characters were excluded ($n = 174$), bringing the total number of essays analyzed to 20,478. After running SpaCy’s NER, the author team checked approximately 100 essays and found that, although the tags covered a broad range of named entities (e.g., school districts, university labs and working groups, non-profit organizations, careers and fields of study), they were all instances of applicants linking these named entities with their intended majors as part of their admissions pitch. The link to the full list of NER tags is available in supplementary materials.

In a basic sense, there were notable differences in the essays written for the new prompt compared to the old prompt. For the rest of study two, we refer to the 2016–2017 prompt as the “treatment” and the 2015–2016 prompt as the “control”. The average combined word and character length of named entities included in the essays fell from approximately 14 words per essay to 9 and from approximately 94 characters to 59. Though lateral applicants tended to include more named entities, generally there was not strong variation in the average word counts of named entities for lateral (15.05 words in 2015–2016; 9.80 words in 2016–2017) and vertical (13.84 words in 2015–2016; 8.78 in 2016–2017) essays. The proportion of essays that did not include any named entities nearly tripled (5–14% of essays) from 2015–2016 to 2016–2017. The distributions of percentages of essays comprised of named entities are presented in Fig. 3. The average percentages for words (control = 0.0316; treatment = 0.0305) and characters (control = 0.0363; treatment = 0.0346) were smaller in 2016-2017 but only by approximately 0.002. The seeming larger change in amount of words and characters of named entities decreasing but the overall percentage of essay dedicated to named entities not changing as much might be partly
due to the essay length requirement also shortening; this would also help justify the decision to compare percentages rather than numbers of tags or words.

Once all of the named entities were tagged and tabulated, the first step to measure the direct effect of the new essay prompt was to perform propensity score weighting. This was done using the WeightIt package in R [56]. See Figure SM4 for a visualization of this step. The largest adjustments were made for whether or not the student reported at least one college educated parent (“First Gen. Status”) and whether or not they attended a Hispanic Serving Institution (HSI). Whether or not students identified as Cuban or “Other” (meaning they had to write in how they identify) had relatively large adjustments that were comparable to gender, academic pathway, and whether or not the applicant applied from a private school. The Love plots generated
The percentages of words and characters were then measured using the weights and treatment. The full results for the analysis are in Table 6, including the disaggregated results just examining essays written by vertical and lateral applicants (respectively). Each numerical value in the table represents a percentage, with full numbers indicating 1% and above. Overall, there was a small but significant drop in the usage of named entities in the transfer essays after the new transfer essay prompt was introduced, regardless if the percentages were tabulated by words or characters. These results held even after excluding essays that did not use any named entities in their essays from the analysis.

However, there was also an important caveat to these results: the results were mixed after disaggregating by academic pathway. Vertical applicants included fewer named entities in their essays, but for lateral applicants the effect of the new prompt was null (p value for the estimate of the treatment effect was above 0.05 and the 95% confidence interval for the estimate included zero). This suggests that vertical applicants are following the text of the essay prompt more closely than lateral applicants by including fewer named entities. Combined, these results show a relationship between essay prompts and the various types of named entities they include in their essays. But some students, in this case those who were attending four year colleges and universities, seem to still see the essay as an opportunity to highlight experiences with different organizations even if the prompt does not specifically ask for it. Though we do control for different essay lengths by focusing on named entities as percentages of essays,

| Model   | Dependent variable | Term        | Estimate | Std. Error | 95% confidence interval |
|---------|--------------------|-------------|----------|------------|-------------------------|
| All     | % Words            | Intercept   | 3.18***  | 0.03       | [3.12, 3.23]            |
|         | Treatment          | − 0.14***   | 0.04     | [− 0.02, − 0.06] |
| All     | % Characters       | Intercept   | 3.64***  | 0.03       | [3.59, 3.71]            |
|         | Treatment          | − 0.18***   | 0.05     | [− 0.27, − 0.09] |
| Vertical| % Words            | Intercept   | 3.17***  | 0.03       | [3.10, 3.23]            |
|         | Treatment          | − 0.14***   | 0.04     | [− 0.23, − 0.06] |
| Vertical| % Characters       | Intercept   | 3.64***  | 0.03       | [3.57, 3.71]            |
|         | Treatment          | − 0.19***   | 0.05     | [− 0.29, − 0.10] |
| Lateral | % Words            | Intercept   | 3.59***  | 0.10       | [3.34, 3.80]            |
|         | Treatment          | − 0.05      | 0.16     | [− 0.36, 0.27] |
| Lateral | % Characters       | Intercept   | 4.09***  | 0.13       | [3.83, 4.35]            |
|         | Treatment          | − 0.10      | 0.19     | [− 0.47, 0.27] |

*p < 0.05, **p < 0.01, ***p < 0.001

for the lateral essays are presented in Figure SM5 and for the vertical essays in Figure SM6.
future studies might take a deeper examination into the specific effects of the shorter format than we do here.

**Study three: shared narrative materials in transfer essays**

Given the lack of literature on narrative strategies used in admissions essays broadly but transfer essays specifically, we took an inductive approach to qualitative coding. The authors spent several weeks reading random essays, describing narrative themes and tropes used by the applicants, and iteratively developed a final codebook to use on our stratified random sample of sociology and computer science applicants. Full descriptions of the codes are available in Table SM5. Given the relatively small sample size and short length of the documents, we opted for consensus in coding rather than inter-rater reliability [57]. We identified 21 codes that represented the variety of narrative themes, moves, arcs, and tropes that transfer students drew upon in writing their transfer essays. These codes also represent justifications and explanations for why the applicants were prepared to complete a degree in their intended major and why the campuses they selected were appropriate.

Students wrote about their past experiences from before enrolling, such as describing the humble origins of their interests (“Humble Beginnings”) and how they have been long interested in their intended major, sometimes since childhood (“Playing the Long Game”). Many students also drew upon their experiences in college, such as specific coursework, participation in student clubs and organizations, personal enrichment projects relevant to the major outside of class, and examples of academic excellence. There were also many examples of students describing interpersonal relationships as inspiration, connections indicative of strong social capital, and personal reflections on things like time served in the military, personal identity, experiences of trauma, and deep connections to the major (e.g., students describing how their ADHD diagnosis inspired their interest in psychology). Interestingly, there was one case of a student describing internships, but upon closer inspection they were describing internships they did not receive. We present the results in the aggregate in Table SM6. There were multiple instances of lateral applicants mentioning working with famous professors and labs affiliated with universities (e.g., Jet Propulsion Laboratory) and no comparable references from vertical applicants.

On average, the essays were tagged for more codes in 2015–2016. Lateral sociology students used 2.68 codes per document, whereas vertical applications used 3.76 codes. Computer science applicants used 3.6 codes (lateral) and 4 codes (vertical) per document. In 2016–2017, the average numbers of codes mostly dropped and the disparities between the code frequency also dropped, meaning the number of codes per essay started to look more similar. The sociology applicants used 2.75 (lateral) and 2.68 (vertical) codes per document with the new prompt. And for computer science students, the average numbers of codes fell to 2.48 for lateral applicants and 2.60 for vertical applicants. In this way, the new essay prompt was also associated with the applicant essays becoming more similar for the two groups.

Among the sociology and computer science students in our sample, there was notable variation in usage for some codes. Sociology applicants were at least three
times more likely to use the codes for Trauma, Personal Identity, Military Experience, and Volunteering. Computer science applicants were at least three times more likely to use the codes for Playing the Long Game and Personal Enrichment. The other 15 codes were used in more similar proportions. At first glance, this suggests that among applicants there are narrative materials that are broadly shared, regardless of intended major. In the aggregate, these results are similar in a qualitative study of first year essays which also found some parity in narrative choices [15]. The full table of codes is available in Table S7. In the next section, we look deeper into the qualitative results and disaggregate the codes by major, application year, and academic pathway (see Fig. 4).

Several notable patterns emerged after disaggregation. Here, we apply the same heuristic of comparing differences in counts that are at least three times larger for one group or another as above but also consider the directions of the changes (positive or negative) after the new essay prompt was introduced. There were codes that shifted in frequency three times as much for one group or another and differences that were directionally different for the groups, but here we only discuss the codes that fit both criteria. Of the codes presented in Fig. 4, all of the shifts except for Academic Excellence were associated with one group using the code less (e.g., Media was used 4 times less by vertical applicants than the increase in frequency among lateral applicants). Interestingly, these results connect back to study two because the most effected codes with lateral applicants theoretically connect back to named entities (Volunteering, Student Club or Organization, Internships, and Research or Teaching Experiences). Although lateral applicants did not use significantly fewer named entities in their essays, these results suggest a narrative shift observable through human reading.
Among sociology applicants, vertical applicants used fewer codes for media, volunteering, trauma, Eureka moments while these codes either increased or stayed the same for lateral applicants; vertical applicants also used academic excellence more than lateral students with the new prompt. Among lateral sociology applicants, student club and organizations codes were used less often but became more frequent among vertical applicants. Vertical applicants for sociology were using less self-description and analysis in their essays relative to lateral applicants, such as the trauma and Eureka moment codes. This shift of vertical applicants using less narratives about themselves and lateral applicants including more might help explain the drop in classification accuracy described in study one based on the topics and dictionary features.

Among computer science applicants, the only code used differently while fitting the criteria among vertical applicants was Eureka moment. For lateral applicants, the research and teaching experiences, volunteering, and internship codes were used less. The results for the lateral computer science applicants were somewhat surprising since they suggest that they were writing less about experiences involving named entities despite the null effect reported in study two. It might also be the case that they describe these types of experiences more in other essays (in 2016–2017, all University of California applicants had to write four essays). Vertical applicants in computer science and sociology both described fewer Eureka moments as well, possibly pointing to similar patterns of less self-description and analysis from them with the new prompt. This, along with the codes used less by lateral applicants and more by vertical applicants for computer science, would further help explain the results of study one.

**Discussion**

Social science has been calling on the potential for computational methods to ask new questions about new forms of data. Here, we take these calls and extend the literature on stratification of college application materials while focusing on an often underexamined group, transfer applicants. The transfer system is a popular option in the US which gives students second chances and complete degree programs. But, as we show in this paper, unique forms of stratification based on the academic pathways of the applicants are linked to variation in transfer application materials. We also show how a specific element used in many transfer application protocols, the essay, seems to vary along dimensions relating to sociocultural capital (linguistic, cultural, and narrative) but that the language of the essay prompt might mediate the strength of this variation.

It is unlikely that lateral applicants were developing less linguistic capital through their writing classes, professional development, and interactions with other students at their four year schools than the previous year. What is more likely is that the new essay prompt altered the choices that all transfer applicants made about what language to use in their essays, who to reference, and narrative structures. This effect was noticeable among the vertical applicants with respect to their lower usage rates of named entities not observed with the lateral applicants. An important aspect of
cultural capital is knowing when to break the rules, and here we showed how the vertical applicants followed the rules of the new prompt more closely than the lateral applicants. Regardless of the rules of the essay prompts, students still had to decide which narrative arcs and vignettes to provide. Vertical applicants are unlikely to have access to the same types of characters, stories, and opportunities as lateral applicants. To compensate this difference in narrative capital, vertical applicants used more codes related to themselves and personal experiences (e.g., Media and Trauma) in 2015–2016. This shifted towards personal experiences and successes more grounded in school (Academic Excellence). Our theoretical frames and computational methods allowed us to analyze patterns in the data not detectable at the surface level using popular textual metrics.

In this way, how we did the work was also important. There is a rapidly growing body of research (and in some cases practice) that applies computational methods (e.g., NLP) to data used in high stakes decision-making. Other examples of this include parole transcripts [58] and occupational mobility [59]. This paper is similar but also departs somewhat by following [25] and including human (qualitative) coding as part of the analysis. This not only provided additional insight into the narratives that transfer applicants composed as part of their applications but also provided another data point beyond their correlation with other information about the applicant.

For example, there was a stark drop in the predictive accuracy of essay content and style after the new prompt was released (72 vs. 59% classification accuracy). In the aggregate, there was also a significant drop in the amount of words and characters comprised of named entities in the essays after the new prompt was revealed. But qualitatively we also see that the applicants shifted their narrative strategies in ways that were detectable from the results of human reading. The combination of computational and qualitative results that all tell different versions of a similar story adds an extra degree of certainty and depth beyond word frequencies and co-occurrence patterns. Combined, they also show how institutional stratification of transfer applicants can be tied back to more or less measurable variation in the essays based on the essay prompt. Since so much variation in these forms of capital are connected with different types of social stratification, we hypothesize that this drop is desirable for educational equity because students appear more similar than different. Since we do not have outcomes data, we can only speculate, but additional research might extend these results to consider how admissions officers read admissions essays and how our results converge or diverge.

Our approach to research also highlights the importance of people in these evaluative processes and how we can rate and evaluate text in ways that are greater than the sums of their words. Given the strength of the computational results, there is both opportunity but also risk [60] in the downstream application of these methods to actual systems. Such risks include the potential for misinterpretation of computational results by human evaluators; the use of humans to simply “correct” the computation when errors arise rather than trust their judgement; and the possible distraction of computation in addressing the question of whether or not admissions essays should be used at all. Determining not just how but if computational toolkits
should be integrated into evaluative processes is a key question for researchers and practitioners to solve now and in future.

Before transfer essays even reach admissions offices, however, this study showed how the essays are varied based on the stratification of an applicant’s academic background and the essay prompt used to solicit said essays. Lateral and vertical applicants mentioned different types of majors they described in their essays that are generally associated with higher and lower income careers, respectively. Lateral students were also more likely to describe experiences related to internships and student research as well as include more named entities in their essays. Beyond our study, some of these patterns were also similar to results from studies of first year applicant essays. The combination of these factors shows how the essays provide evidence of an accumulated capital: students who have the capital (financial as well as sociocultural) to enroll in college (lateral applicants) add more during their time in college (and vice versa for vertical applicants). From this perspective, the drop in predictive accuracy after the new essay prompt was released could be interpreted as a drop in observable accumulated capital. Though this study cannot definitively connect this drop to educational outcomes, numerous studies of stratification and sociocultural capital hint at the possibility of these results indicating more equitable reading of these materials. Future studies could support or reject this premise.

Study two using NER provides further evidence of this from a unique perspective based on the culture of how applicants respond to essay prompts. Although lateral applicants referenced fewer named entities in the essays written for the new prompt on average, the new prompt did not significantly change the amount of named entities they included in their essays. The opposite was true for vertical applicants. This suggests that for transfer applicants, the cultures of how closely to follow the essay prompts tend to vary based on academic pathway. With respect to named entities, vertical applicants "played by the rules" of the prompt whereas lateral applicants had looser interpretations. Although we focus on the results of the propensity score weighting, it is still important to reiterate that the number of essays that did not include any named entities nearly tripled after the new prompt was introduced. Future studies of admissions systems should examine other types of essay prompts and their effects on students. Methodologically, study two also showed promise for NER and causal inference methods as tools for social science.

These results were somewhat in contrast to the results of study three, though not entirely. For example, lateral computer science applicants mentioned fewer internships, research and teaching experiences, and volunteering with the new prompt while these types of references either stayed the same or increased for vertical applicants. But, these and other results also highlight how the language of the new prompt helped essays from lateral and vertical applicants become more similar, a key finding for the other two studies.

The three studies described in this paper showed different forms of stratification in transfer essay prompts, academic pathways, and sociocultural capital using computational and human reading. For many good reasons, fairness, ethics, and bias in computational analyses of education data typically focuses on protected attributes [61], but future work might consider how academic pathways
and institutional experiences might also lead to biased results and stratified outputs due to stratified inputs. Given the similarity of transfer prompts and graduate school personal statement prompts, as well as the well known correlations of academic pathways to graduate school [62], our results seem to invite analyses based on graduate school personal statements.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42001-022-00185-5.

Acknowledgements We thank the Stanford Institute for Social Science Research and their Community College Research Experience program for helping the team come together. We also thank Ben Domingue and Klint Kanopka for their helpful, consistent feedback for the IMV section. We thank Anthony Lising Antonio and the rest of the Student Narratives Lab for their support and feedback. We thank Melissa Mesinas and Rosalía C. Zárate for contributing to a previous version of this work that was presented at the Association for the Study of Higher Education conference in 2020. Finally, we thank the editors and reviewers for great feedback that helped improve the paper.

Data availability A link to the data and code needed to replicate the study is available in the supplementary materials.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Bowles, S., & Gintis, H. (2002). Schooling in capitalist America revisited. Sociology of Education, 75, 1–18.
2. Harrison, M. H., Hernandez, P. A., & Stevens, M. L. (2022). Should I start at math 101? Content repetition as an academic strategy in elective curriculums. Sociology of Education, 95(2), 133–152.
3. Dixon-Román, E. J., Everson, H. T., & McArdle, J. J. (2013). Race, poverty and sat scores: Modeling the influences of family income on black and white high school students’ sat performance. Teachers College Record, 115(4), 1–33.
4. Alvero, A., Giebel, S., Gebre-Medhin, B., Antonio, A. L., Stevens, M. L., & Domingue, B. W. (2021). Essay content and style are strongly related to household income and sat scores: Evidence from 60,000 undergraduate applications. Science Advances, 7(42), 9031.
5. Alvero, A., Arthurs, N., Antonio, A.L., Domingue, B.W., Gebre-Medhin, B., Giebel, S., & Stevens, M.L. (2020). Ai and holistic review: informing human reading in college admissions. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 200–206.
6. Kim, B.H. (2022). Applying data science techniques to promote equity and mobility in education and public policy. PhD thesis.
7. Rothstein, J. (2022). Qualitative information in undergraduate admissions: A pilot study of letters of recommendation. Economics of Education Review, 89, 102285.
8. Salazar, K. G., Jaquette, O., & Han, C. (2021). Coming soon to a neighborhood near you? Off-campus recruiting by public research universities. American Educational Research Journal, 58(6), 1270–1314.
9. Spencer, G. (2021). Off the beaten path: Can statewide articulation support students transferring in nonlinear directions? American Educational Research Journal, 58(5), 1070–1102.
10. Crisp, G., & Delgado, C. (2014). The impact of developmental education on community college persistence and vertical transfer. Community College Review, 42(2), 99–117.
11. Schudde, L., & Goldrick-Rab, S. (2015). On second chances and stratification: How sociologists think about community colleges. Community College Review, 43(1), 27–45.
12. Malcom-Piqueux, L., Bensimon, E. M., Suro, R., Fischer, A., Bartle, A., Loudenback, J., & Rivas, J. (2013). Addressing Latino outcomes at California’s Hispanic-serving institutions. University of Southern California Center for Urban Education.

13. Quintana, R. (2021). What race and gender stand for: Using Markov blankets to identify constitutive and mediating relationships. Journal of Computational Social Science. https://doi.org/10.1007/s42001-021-00152-6

14. Lamont, M. (2012). Toward a comparative sociology of valuation and evaluation. Annual Review of Sociology, 38, 201–221.

15. Gebre-Medhin, B., Giebel, S., Alvero, A. J., Domingue, B. W., Stevens, M. L., & Antonio, A. L. (2022). Application essays and the ritual production of merit in us selective admissions. Poetics. https://doi.org/10.1016/j.poetic.2022.101706

16. Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. PloS One, 9(12), 115844.

17. Arthurs, N., & Alvero, A. J. (2020). Whose truth is the “ground truth”? college admissions essays and bias in word vector evaluation methods. In: A. N. Rafferty, J. Whitehill, V. Cavalli-Sforza, C. Romero (Eds.), Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020) (pp 342–349)

18. Jones, S. (2013). “Ensure that you stand out from the crowd”: A corpus-based analysis of personal statements according to applicants’ school type. Comparative Education Review, 57(3), 397–423.

19. Stevens, M. L. (2009). Creating a Class. Harvard University Press.

20. Bastedo, M. N., Bell, D., Howell, J. S., Hsu, J., Hurwitz, M., Perfetto, G., & Welch, M. (2021). Admitting students in context: Field experiments on information dashboards in college admissions. The Journal of Higher Education. https://doi.org/10.1080/00221546.2021.1971488

21. McFarland, D. A., Khanna, S., Domingue, B. W., & Pardos, Z. A. (2021). Education data science: Past, present, future. AERA Open, 7, 23328584211052056.

22. Fisscher, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. Review of Research in Education, 44(1), 130–160.

23. Singer, J. D. (2019). Reshaping the arc of quantitative educational research: It’s time to broaden our paradigm. Journal of Research on Educational Effectiveness, 12(4), 570–593.

24. Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational social science and sociology. Annual Review of Sociology, 46, 61–81.

25. Nelson, L. K. (2020). Computational grounded theory: A methodological framework. Sociological Methods & Research, 49(1), 3–42.

26. Nelson, L. K., Burk, D., Knudsen, M., & McCall, L. (2021). The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. Sociological Methods & Research, 50(1), 202–237.

27. Ishitani, T. T., & Flood, L. D. (2018). Student transfer-out behavior at four-year institutions. Research in Higher Education, 59(7), 825–846.

28. Dowd, A. C., & Melguizo, T. (2008). Socioeconomic stratification of community college transfer access in the 1980s and 1990s: Evidence from HS &B and NELS. The Review of Higher Education, 31(4), 377–400.

29. Gerber, T. P., & Cheung, S. Y. (2008). Horizontal stratification in postsecondary education: Forms, explanations, and implications. Annual Review of Sociology, 34, 299–318.

30. Possett, J. R., & Grodsky, E. (2017). Graduate education and social stratification. Annual Review of Sociology, 43, 353–378.

31. Bourdieu, P. (1987). Distinction: A Social Critique of the Judgement of Taste. Harvard University Press.

32. Stoltz, D. S., & Taylor, M. A. (2019). Concept mower’s distance: Measuring concept engagement via word embeddings in texts. Journal of Computational Social Science, 2(2), 293–313.

33. Kim, J. Y. (2021). Integrating human and machine coding to measure political issues in ethnic newspaper articles. Journal of Computational Social Science, 4(2), 585–612.

34. Lareau, A., & Horvat, E. M. (1999). Moments of social inclusion and exclusion race, class, and cultural capital in family-school relationships. Sociology of Education, 72, 37–53.

35. Bourdieu, P. (1991). Language and Symbolic Power. Harvard University Press.

36. Bernstein, B. (1964). Elaborated and restricted codes: Their social origins and some consequences. American Anthropologist, 66(6), 55–69.

37. Durkheim, E. (2012). Moral Education. Courier Corporation.
38. Takacs, C. G. (2020). Becoming interesting: Narrative capital development at elite colleges. *Qualitative Sociology*, 43(2), 255–270.
39. Toubia, O., Berger, J., & Eliashberg, J. (2021). How quantifying the shape of stories predicts their success. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.2011695118
40. Simpson, E. H. (1949). Measurement of diversity. *Nature*, 163(4148), 688–688.
41. Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, FOG count and Flesch reading ease formula) for Navy enlisted personnel. Naval technical training command millington TN research branch: Technical report.
42. Blei, D., & Lafferty, J. (2006). Correlated topic models. *Advances in Neural Information Processing systems*, 18, 147.
43. Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin.
44. Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). STM: An R package for structural topic models. *Journal of Statistical Software*, 91, 1–40.
45. Salton, G. (1971). *The SMART retrieval system-experiments in automatic document processing*. Prentice-Hall Inc.
46. Porter, M.F. (2001). Snowball: A language for stemming algorithms. Published online. Accessed 3 Nov 2008.
47. Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.
48. Schofield, A., Magnusson, M., & Mimno, D. (2017). Pulling out the stops: Rethinking stopword removal for topic models. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 432–436.
49. Nikita, M. (2016). Idatuning: Tuning of the latent Dirichlet allocation models parameters. R package version 0.2-0. https://CRAN.R-project.org/package=ldatuning
50. Domingue, B., Rahal, C., Faul, J., Freese, J., Kanopka, K., Rigos, A., Stenhaug, B., & Tripathi, A. (2021). *InterModel Vigorish*. A novel approach for quantifying predictive accuracy with binary outcomes: IMV.
51. Pryzant, R., Card, D., Jurafsky, D., Veitch, V., & Sridhar, D. (2021). Causal effects of linguistic properties. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4095–4109.
52. Egami, N., Fong, C.J., Grimmer, J., Roberts, M.E., Stewart, B.M. (2018). How to make causal inferences using texts. arXiv preprint arXiv:1802.02163
53. Hennibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2017). *spaCy: industrial-strength natural language processing in Python*. https://doi.org/10.5281/zenodo.1212303.
54. Angrist, J., & Imbens, G. (1995). Identification and estimation of local average treatment effects. National Bureau of Economic Research Cambridge: Mass.
55. Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444–455.
56. Greifer, N. (2020). WeightIt: Weighting for covariate balance in observational studies. R package version 0.9. 0.
57. Cascio, M. A., Lee, E., Vaudrin, N., & Freedman, D. A. (2019). A team-based approach to open coding: Considerations for creating intercoder consensus. *Field Methods*, 31(2), 116–130.
58. Bell, K., Hong, J., McKeown, N., & Voss, C. (2021). The Recon Approach: A new direction for machine learning in criminal law. *Berkeley Technology Law Journal*, 37.
59. Jayaratne, M., & Jayatilleke, B. (2021). Predicting job-hopping motive of candidates using answers to open-ended interview questions. *Journal of Computational Social Science*. https://doi.org/10.1007/s42001-021-00138-4
60. Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW), 1–24.
61. Yu, R., Lee, H., & Kizilcec, R.F. (2021). Should college dropout prediction models include protected attributes? In: *Proceedings of the Eighth ACM Conference on Learning@ Scale*, pp. 91–100.
62. Posselt, J. R. (2016). *Inside graduate admissions*. Harvard University Press.
Publisher's Note. Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

AJ Alvero1 · Jasmine Pal2 · Katelyn M. Moussavian2

Jasmine Pal
jasminep1@g.ucla.edu

Katelyn M. Moussavian
moussavian@g.ucla.edu

1 University of Florida, Gainesville, FL, USA
2 University of California, Los Angeles, Los Angeles, CA, USA