CA Journal of Cultural Analytics

Lucy, Li, et al. "Racial and Ethnic Representation in Literature Taught in US High Schools." *Journal of Cultural Analytics*, vol. 10, no. 1, Mar. 2025, https://doi.org/10.22148/001c.131682.

ARTICLE

Racial and Ethnic Representation in Literature Taught in US High Schools

Li Lucy[1], Camilla Griffiths[2], Claire Ying[1], JJ Kim-Ebio[1], Sabrina Baur[1], Sarah Levine[2], Jennifer L. Eberhardt[2], David Bamman[1], Dorottya Demszky[2]

[1] University of California, Berkeley, [2] Stanford University

Keywords: canonicity, english literature, curriculum, education, race, natural language processing

https://doi.org/10.22148/001c.131682

Journal of Cultural Analytics

Vol. 10, Issue 1, 2025

We quantify the representation, or presence, of characters of color in English Language Arts (ELA) instruction in the United States to better understand possible racial/ethnic emphases and gaps in literary curricula. We contribute two datasets: the first consists of books listed in widely-adopted Advanced Placement (AP) Literature & Composition exams, and the second is a set of books taught by teachers surveyed from schools with substantial Black and Hispanic student populations. In addition to these book lists, we provide an unprecedented collection of hand-annotated sociodemographic labels of not only literary authors, but also their characters. We use computational methods to measure all main characters' presence through three distinct and nuanced metrics: frequency, narrative perspective, and burstiness. Our annotations and measurements show that the sociodemographic composition of characters in books recommended by AP Literature has not shifted much for over twenty years. As a case study of how ELA curricula may deviate from the curricula prescribed by AP, our teacher-provided sample shows a greater emphasis on books featuring first-person, primary characters of color. We also find that only a few books in either dataset feature both White main characters and main characters of color. Arguably, these books may uphold a view of racial/ethnic segregation as a societal norm.

1. Introduction

School curricula represent kinds of historical and cultural knowledge that society prioritizes and codifies at certain historical moments. Designed to prepare younger generations for life and career (Tedesco et al. 10), its contents can shape students' values, beliefs, and identities. In the US, English Language Arts (ELA) teaches literary and writing skills to students, typically through the study of assigned texts. These texts—poetry, plays, articles, and especially novels—are focal points of curriculum, around which educators organize skills and conceptually-based lessons. These texts influence not only students' understanding of themselves, but also their sociocultural understanding of others (Dee and Penner). In particular, racial/ethnic representations in these literary texts matter for students' self-affirmation and prejudice reduction (Bishop 3; Banas et al.). In this paper, we quantify the racial/ethnic representation of authors and characters in nearly four hundred

books taught in American high schools between 2019 and 2023, to consider what kind of sociocultural knowledge ELA has promoted in the US. If book bans in the US have disproportionately impacted children's books by authors of color and books that feature characters of color, how does literature teaching in the US approach minoritized racial and ethnic groups (Goncalves et al.)?

We leverage two snapshots of ELA in the US to examine the inclusion of minoritized racial and ethnic groups in school curricula. The first includes books suggested in Advanced Placement (AP) English Literature and Composition exams administered by the College Board, a non-profit organization that develops curricula, teaching guides, and standardized tests for high school students. Colleges often use these tests as indicators of college-readiness and may grant college credit to high school students who score well. Though some educators and universities have questioned the academic value and substance of their exams, College Board potentially has "ownership of a national curriculum" (Abrams 6); AP-to-college-credit pipelines are supported by laws and policies in 37 states.[1] In 2023, 55% of US high schools offered at least one AP course ("IES 'a Majority'"), and in 2024, English Literature and Composition was the third most popular AP exam subject, with 389k student test-takers.[2] As a foil to this established curriculum, we present a second dataset of texts taught by 189 teachers from schools across the US that predominantly serve students from low-income and minoritized backgrounds. These teachers attended a multi-year professional development program at a private R1 university and shared their current or upcoming book lists as part of a program survey.

By combining these two sources, we make several contributions to the study of racial representation in ELA course curricula. First, we introduce a richly annotated dataset of 396 books and 306 authors with over 1.5k characters. In addition to insights around our data curation process, we contribute an overview of which racial/ethnic groups are present in these books, and how authors' backgrounds relate to the presence of their characters. Second, as methodological contributions, we improve BookNLP, a popular toolkit for obtaining characters from books (Bamman), and adapt Goh and Barabási's burstiness metric from network science to characterize the rhythm of characters over the course of each book. Third, through the lens of canonicity (e.g. Guillory, *Cultural Capital*; Walsh and Antoniak; González et al.; Le-Khac and Hao), we show that over the last 20 years, despite significant changes in the American sociopolitical climate, AP Literature's suggested book lists have not substantially increased in racial/ethnic diversity of characters. Finally, across both datasets, we observe a scarcity of books that

1 "Statewide AP Credit Policies", https://reports.collegeboard.org/ap-program-results/statewide-credit-policies

2 "2024 AP Program Summary Report", https://apcentral.collegeboard.org/media/pdf/program-summary-report-2024.pdf

integrate main characters from different racial/ethnic identities. Altogether, we engage with long-standing, data-centered challenges in conducting cultural analytics research on race and set a foundation for future work studying the representation of race in literature and school curriculum.

2. Culturally relevant, responsive, and sustaining pedagogy

Emphasis on culturally relevant pedagogy in curricular decisions, as well as pushback against it, has grown since the turn of the century (Ladson-Billings, "Toward a Theory"). ELA coursework that aligns with and responds to students' racial/ethnic backgrounds improves test performance, increases school attendance and engagement with materials, and positively affects identity formation and cross-racial relationships (Aronson and Laughter; Dee and Penner; Steele and Cohn-Vargas; Byrd). Thus, measurement of racial/ethnic representation in curricular texts can equip researchers and practitioners alike with an additional tool to enhance students' educational experiences and address inequities. A preponderance of literature documents racial equity gaps in education (Darling-Hammond; Reardon et al.; Gopalan and Nelson). For example, the Center for American Progress reported in 2021 on the narrowing funnel through which Black, Latinx, and Indigenous students progress through AP coursework pipelines. These students are less likely to enroll in AP courses, take AP exams, and pass these exams compared to their White and Asian peers (Chatterji et al.).

Studies have shown that exposure to and engagement with multiculturalism benefits *all* students, not just those whose backgrounds are explicitly represented in curricular materials (e.g., Bonilla et al.). ELA can serve to broaden students' sociocultural knowledge of others. The National Council of Teachers of English's position statement on the selection of ELA curricular materials states, "Materials may be included because they meet the curriculum objective of presenting articulate voices from different eras or diverse cultures" (NCTE). Underlying these guidelines is the notion that narratives can influence readers' judgements of others and provide a more informed view of society and societal issues (Strange). In communications and psychology studies, the phenomenon of contact with an outgroup through media such as books or film is called *mediated contact*, and positive mediated contact can enable empathy and reduce intergroup anxiety (Banas et al.). In particular, media depicting contact between an ingroup and outgroup, or *vicarious mediated contact*, can reduce prejudice towards the outgroup (Banas et al.; Mazziotta et al.). Thus, racial/ethnic diversity within ELA retains pedagogical significance for all student audiences.

Though educators and researchers have reinforced the importance of culturally relevant, responsive, and sustaining pedagogy for decades (Ladson-Billings, *Different Question*; Department), school curricula often remain predominantly focused on White voices and perspectives (Lucy et al.; Hadley and Toliver; Levine). For example, Kumar's 2022 study examined the racial

demographics of authors whose books were included in ELA Guidebooks 2.0, which is a high school curriculum used across the United States. This study found that White authors wrote 27 of 29 (93.1%) titles included in the curriculum. This centering of Whiteness occurs early in primary school education (Rigell et al.), and also upstream in the US book publishing industry (Agosto et al.; So and Wezerek; McGrath 772). There is some evidence of improvement in the diversity of voices featured in ELA. So's book *Redlining Culture* found an increase in Asian American, Latinx, or Native American authors between the 1979 and 2016 editions of the Norton Anthology of American Literature, which is a popular resource used in college-level coursework (145). As another example, Adukia et al. used computer vision to show that the presence of Black people in children's award-winning books has increased over the past century, though characters in more influential books have lighter skin color on average. Our work builds upon these prior studies by surfacing two more views of what the teaching of literature encompasses or could encompass in the US, and additional insight into whether ELA has evolved over time. We not only spotlight a dominant, widely-adopted curriculum in the United States, as represented by AP's suggested book lists, but also a case study centered on books taught to minoritized students, as represented by our teacher-provided lists.

3. Race, books, and cultural analytics

In cultural analytics, a key barrier to performing analyses of race is the acquisition or detection of racial/ethnic labels. Computational text analyses that center people's sociodemographic attributes tend to focus on gender rather than race (Field et al.), as gender can be operationalized through keywords (e.g. *women*) and pronouns (e.g. *she*). For race, keywords can be more semantically ambiguous (e.g. *black*). For example, Algee-Hewitt et al.'s longitudinal study of race and ethnicity in American fiction analyzes the presence and connotations of racial and ethnic terms in books, but they exclude *black* and *white* from a collocate analysis due to their prominent use as literal color words (40). To further complicate textual measurements of race, mapping from social groups to textual racial/ethnic markers is not one-to-one. For example, though *Black*, *African*, and *African-American* are related terms, they are not equivalent categories nor comprehensive of all the ways Black people may be mentioned in text. Our work deviates from this term-based approach, and instead relies on racial/ethnic labels assigned at the character-level by trained raters. Our collection of 1.5k labeled characters facilitates descriptive analyses of *who* may be represented, brought to the forefront, and canonized in US classrooms in an unprecedented manner.

Aside from methodological challenges, cultural analytics also faces a lack of access to racially diverse book data. Researchers conducting larger-scale analyses of literature production and consumption tend to rely on open datasets curated via biased social processes (Bode). As one prominent example, Project Gutenberg is a volunteer-driven dataset of full-text literature.

Though its collections prioritize "literary works and reference items of historical significance,"[3] it mostly contains out-of-copyright books by White men (Rowberry). Algee-Hewitt et al.'s analysis of race in literature used the Gale American Fiction Corpus, which spans older books published in 1789-1920 and also skews heavily in favor of White authors. Through inspection of words that collocate with racial/ethnic terms, Algee-Hewitt et al. found that their dataset consists of "a large cast of racialized stock characters" (47). Other efforts to study minoritized groups in literature have curated their own data; for example, researchers built and used the Young Readers Database of Literature to map markers of Asian identity across 5,000 contemporary books (Nomura and Dombrowski). For our study, we also curate book data from scratch, including titles listed by teachers who teach minoritized students.

Cultural analytics can shed light on canonization, a process which imbues books with legitimacy and cultural capital (Guillory, *Cultural Capital*); (Chiang). Sociocultural processes such as the distribution of literary prizes and market reception can elevate titles into canons (Manshel 6; Sáez 3). In particular, scholars often point to educational institutions as key drivers of canonization (e.g. Chiang 33; So 145; Newman; Manshel 4). One prominent definition of "the canon" focuses on repetition in syllabi; "every construction of a syllabus institutes once again the process of canon formation" (Guillory, "Canon, Syllabus, List" 52). The "canon wars" of the 1980s marked an especially intense period, during which literary and cultural scholars debated the composition of the Western cultural canon and what should be taught in higher education. Tensions arose between those who wished to conserve traditional humanistic education, which skewed White and male, with a more multicultural one. Similar debates have continued to the present and have seeped into K-12 ELA, especially around the teaching of race in US schools (Safarpour et al.).

Canonicity has featured prominently in several prior cultural analytics studies. For example, Walsh and Antoniak juxtaposed books tagged as "classic" by Goodreads users with those listed by AP exam study guides and college syllabi, and found that authors of Goodreads classics are less racially and ethnically diverse (254). Other cultural analytics studies have focused on canons scoped to specific racial/ethnic groups and have drawn out more fine-grained inequities, such as Le-Khac and Hao's study of Asian American literature and González et al.'s analysis of Hispanic studies. In our study, we define canonization as a title's relative persistence across multiple high school ELA book lists. Concretely, we examine which books recur across AP exams and surveyed teachers who teach minoritized students. Due to the

3 "Collection Development Policy", https://www.gutenberg.org/policy/collection_development.html

prominence of AP in the American education system, the repetition of a title over multiple exam years provides a particularly strong signal of canonization at a national level.

4. Dataset Creation

We began by collecting lists of books that ELA courses may teach, followed by digitization and text preprocessing. We then obtained lists of main characters from these books and manually annotated the perceived races and genders of authors and characters.

4.1. Book lists and metadata

We collected book lists from two sources: AP Literature exams and surveys of high school teachers. To create the AP dataset, we pulled book titles suggested in AP Literature exams' essay questions conducted between 1999-2021.[4] The AP Literature course, developed by a committee of six college faculty and secondary teachers (Abrams 11),[5] covers short and long-form fiction, poetry, and drama; students may take the exam without taking the course, and vice versa. The exam consists of fifty-five multiple choice questions involving literary excerpts, and three essay questions, one of which involves "literary argument," where students are "presented with a literary concept or idea and analyze how the literary concept or idea contributes to an interpretation of a literary work."[6] Though College Board's "Course Perspective" states, "There is neither a required reading list nor a required textbook for AP English Literature" (Potter), this essay question includes a suggested list of texts. The process behind the decisions about which titles this essay question suggests is opaque, though the evolution of question wording over the years provides some indication of possible intentions. The 1999 exam states, "You may use one of the novels or plays listed below or another novel or play of similar literary quality." In later exams, the phrase "similar literary quality" is supplanted by "comparable literary merit" up until 2019. In the final year of our dataset, 2021, the exam's instructions broadened to "Either from your own reading or from the list below, choose a work of fiction in which...".[7] Overall, these statements suggest that titles featured in these suggested reading lists encapsulate what AP believes are representative of literary quality or merit.

4 "AP English Literature and Composition Exam Questions." Collegeboard AP Central. https://apcentral.collegeboard.org/courses/ap-english-literature-and-composition/exam/past-exam-questions

5 AP English Literature and Composition Development Committee. https://apcentral.collegeboard.org/courses/ap-english-literature-and-composition/course/development-committee

6 "AP English Literature and Composition: About the Exam", https://apstudents.collegeboard.org/courses/ap-english-literature-and-composition/assessment

7 This phrasing of the 2021 exam has continued into 2024.

For our second dataset, we draw from surveys of 189 ELA teachers, taken during the summers of 2016-2022. The teachers were participants in the Hollyhock Fellowship, a professional learning program sponsored by Stanford University's Center to Support Excellence in Teaching. The program served teachers with 2-8 years of experience who taught at high schools where the majority of students qualified for free and reduced lunch (see Levine et al., for detailed description of this program). In general, the teachers were "committed to social justice in education and ready to take pedagogical risks" (Levine et al. 511). This program accepted between 50% - 75% of teacher applicants each year. Surveys asked teachers to list the texts they planned to use in the upcoming school year. We note that these teacher-provided book titles do not represent curricular choices across the US to a similar extent that AP does, as these teachers are not a randomized sample of ELA teachers who teach minoritized students. Though this selection bias limits the direct comparability of the two datasets, our teacher-provided sample presents a descriptive case study of what is taught in actual high school classrooms and complements the more prescriptive nature of AP.

By including books listed by both AP and these teachers, we are able to examine distinct snapshots of ELA curricula that serve contrasting student audiences. That is, these lower-income schools taught by the surveyed teachers tend to have a higher proportion of Black and Hispanic/Latinx students than AP examinees and the general American high school student population (see Figure 1).[8] The racial/ethnic composition of the schools in the teacher dataset follow national trends, where higher proportions of Black and Hispanic students attend mid-high poverty schools than White and Asian students (2023). Both datasets include interruptions which result in missing data. Due to the global pandemic, 2020 is excluded from the AP dataset. The exam was in a different format that year. 2021 is excluded from the teacher dataset for similar logistical reasons. We focus on these two datasets to approximate what books may be taught across the US; future work could consider expanding our work to more samples of ELA books, such as state-specific recommendations[9] or titles in resources written by the National Council of Teachers of English (NCTE).

We manually matched different forms of the same book title (e.g. *Bone: A Novel* or *Bone*) across sources, and deduplicated this list to produce our final list of ELA materials. We then filter this list to include only book-length fiction and narrative nonfiction, thus excluding texts such as short stories, news articles, poetry, and drama. We focus on books because we leverage a text processing pipeline, BookNLP, which operates mainly on book-length

8 Out of 100 schools included in our survey, 92 have demographic data publicly listed in NCES. Some surveyed teachers are from the same school.

9 For example, California Department of Education's "Recommended Book List": https://www.cde.ca.gov/ci/cr/rl/.

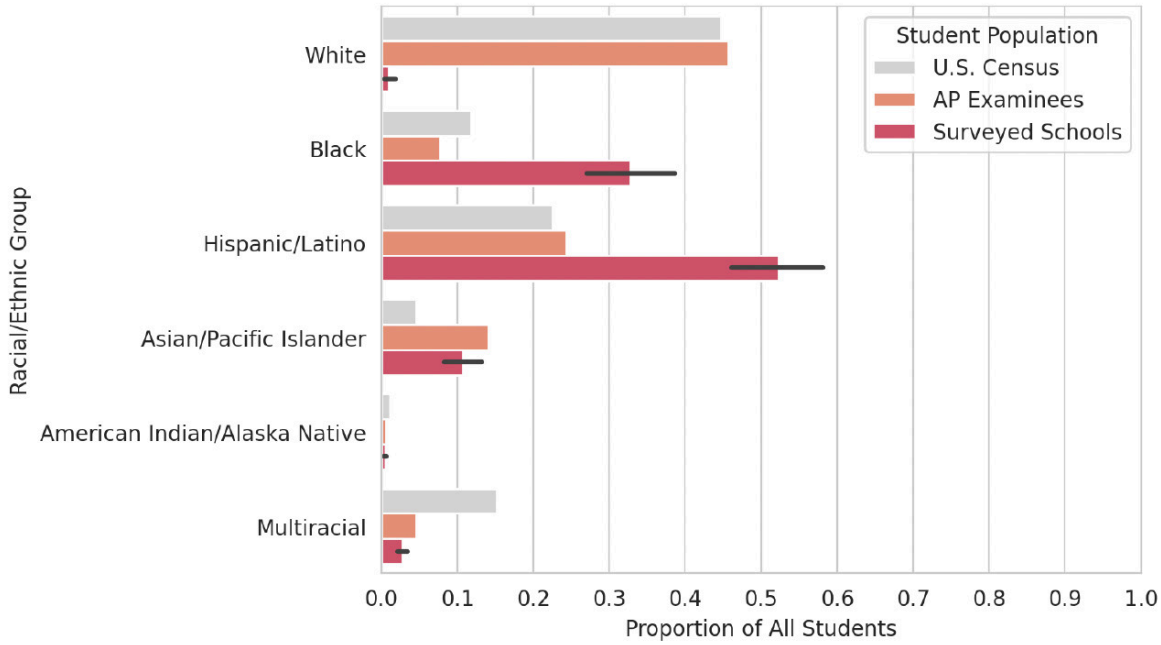

Figure 1. Student demographics of US high schools (American Community Survey), AP examinees (College Board), and surveyed teachers' schools (IES "Search Public Schools"). All data is from the year 2022, and error bars for surveyed schools are 95% confidence intervals.

prose (Bamman). We leave the construction and application of computational pipelines that can handle other media formats for future work. We attach authors to book titles, and in a few cases where multiple published books have the same title, we disambiguate based on the likelihood of a book belonging to known collections of ELA material. On average, each AP exam year listed 35.6 books (SD = 13.8) and each teacher listed an average of 3.1 books (SD = 1.8). This produced a dataset of 396 books in total; the AP dataset lists 250 books, the surveyed teachers' dataset lists 207, and 61 are in both.

4.2. Text preprocessing

We purchased and scanned physical copies of books in our dataset, except for those already available in digital form on Project Gutenberg, which mostly includes older books out of copyright. We ran optical character recognition on scanned books using ABBYY Finereader, and removed books' paratext, which includes non-narrative material such as table of contents and title pages. We also removed footers, headers, and page numbers missed by ABBYY Finereader using rule-based repetition and number heuristics. Finally, we run all books through BookNLP, an open-source pipeline that includes entity recognition, character name clustering, quote attribution, and inference of characters' referential gender.[10] Each book has 127k tokens on average (SD = 85.3k), yielding a total of 50.5 million tokens across both data sources.

10 We use BookNLP version 1.0.7 with parameters `model = big` and `spacy_model = en_core_web_trf`.

Table 1. Comparisons of our name clustering modifications with BookNLP's original implementation. We evaluate on all aliases in LitBank excerpts as well as those that appear at least 10 times in each book. All metrics range from 0 to 1, where 1 is perfect performance, and purity is an accuracy metric adapted for evaluating clustering from Manning et al. (356).

BookNLP version	Aliases	Recall of pairs	Precision of pairs	F1 of pairs	Purity of clusters
Original	all	0.637	0.708	0.671	0.871
Ours	all	0.595	0.839	0.696	0.910
Original	frequency > 10	0.686	0.698	0.692	0.852
Ours	frequency > 10	**0.675**	**0.878**	**0.763**	**0.927**

4.3. Character lists

We obtained character names using BookNLP, which combines coreference resolution and character name clustering to count all mentions and references to a character under a unique identifier. Initially, we observed that the default implementation of BookNLP yielded character lists that conflate key characters with each other or splits key characters among multiple character IDs. These errors originated from BookNLP's character name clustering module, which groups aliases into characters based on overlapping substrings, e.g. *Khalil* → *Khalil Harris* in *The Hate U Give*. These one-off named entity recognition errors can lead to main characters being incorrectly merged (e.g. Jo and Meg in *Little Women*), or split one character (e.g. Oliver in *Oliver Twist*) across multiple character IDs.[11] We implemented two modifications to this name clustering process to improve our character lists: first, we avoided grouping characters based on overlap with an alias that only appears once, and second, we forced popular proper names of people to only be attached to one character. The latter modification assumes that authors rarely give multiple main characters the same frequently mentioned name.

These changes led to observable improvements in our character lists, fixing the *Little Women* and *Oliver Twist* examples mentioned earlier. We also quantitatively evaluated our changes on human-labeled coreference annotations of publicly-available LitBank excerpts (Bamman et al.), by comparing aliases linked in these excerpts to those linked by BookNLP on the full-text of each book. Our approach improves the detection of LitBank character alias pairs, especially for more common aliases, which we prioritized in our study (Table 1). To further improve the quality of our character lists, during our demographic annotation process described in §4.4, we recorded any missing aliases not detected by BookNLP or main character IDs that should be merged. We manually merged 91 (5.2%) annotated BookNLP character IDs with another character ID, and added at least one alias to 16 (0.9%) non-narrator characters.

11 We found that the original implementation of BookNLP split 22.5% of names that appear more than 100 times in a book.

Some books feature multiple first-person narrators, who BookNLP conflates to a single narrator represented by first-person referents outside of dialogue. We used automatic approaches to disambiguate sequences of narration in cases where narrator changes are indicated by their names in chapter or section titles. Otherwise, we consulted online book summaries, which we then corroborated with book content, to manually segment books by narrator. Our dataset includes 23 books that contain multiple narrators, with three books (*Bronx Masquerade*, *As I Lay Dying*, and *There, There*) featuring more than ten narrators. Every narrator also had their name/s manually added as aliases.

We then applied a cutoff based on characters' total mention counts to determine which characters from each book to include in demographic annotation and further analyses. Who deserves to be considered a "main" character in a book can be subjective, as a character's contribution to a narrative may extend beyond the frequency of their presence in text. Books also vary in how they distributed attention towards characters; some may feature a singular central character, while others might rotate through a few of them. When selecting a cutoff, we wished to keep the main character list per book small enough to manage during demographic annotation, yet large enough to allow analysis of each book's main cast. We determined that characters who occur in at least 3% of all character mentions in a book would be considered *main characters* in our study, as this resulted in 1 to 8 main characters per book, with an average of 4.03 characters per book.[12] We analyzed 1,594 main characters in total.

4.4. Race and gender annotation

A team of six annotators assigned race and gender labels to literary authors and characters by searching through resources such as Wikipedia, online ELA study guides, authors' websites, news articles, and the books themselves for evidence of racial/ethnic identification. These annotators include the first, third, fourth, and fifth authors, and overall consisted of a doctoral student and five undergraduates who study a range of disciplines, including media studies, cognitive science, ethnic studies, and data science. The first author led the process and provided exemplars during initial rounds of annotator training.

For each author and character, we labeled gender and race/ethnicity, along with quoted evidence or rationale for the latter. We drew our base set of racial/ethnicity categories from the US Census. These categories include White, Black, Native American/Indigenous, Asian American and Pacific Islander (AAPI), Latino/x, and Middle Eastern and North African (MENA)

12 We calibrated this cutoff by comparing to books' Goodreads summaries. We found that these summaries may mention up to 5 characters per book, though the vast majority of books' summaries fall within the 0-2 characters range.

(Marks et al.). When discussing our analyses and results, we define authors or characters of color as those who are not solely White. We prioritized explicit self-identification over implied third-person observation to avoid misidentification as much as possible. We acknowledge that racialization, or the act of designating people to be part of a racial group, has been used to marginalize and discriminate, and that racial and ethnic categories are socially constructed with temporally and contextually unstable boundaries (Hanna et al.; Wimmer). We conducted this labeling process to examine racial representation in ELA, and our annotations provide the first step for future investigations around how perceptions of authors' and characters' racial backgrounds may relate to what is or is not taught in schools, and how their narratives impact students. By releasing our labels, sources, and written justifications, we hope to offer some transparency on our process and establish a necessary starting point for future studies. In addition, it is possible that our annotations can serve as a resource for future development of information retrieval methods for the humanities. That is, novel human-in-the-loop approaches may assist scholars in streamlining the process of surfacing sociodemographic signals from source collections.

For characters, we made 36.3% of annotation decisions based on explicit evidence of race, e.g. *Latino-American*. In the absence of explicit racial/ethnic markers, we drew inferences based on information around setting, time period, and other culturally relevant book content. For example, we labeled Nigerian characters who discuss tensions between speaking Igbo at home and English in public in Chimamanda Ngozi Adichie's *Purple Hibiscus* as Black. Race was particularly challenging to label for White authors and characters, given that Whiteness is often an unstated default in text (McDermott and Ferguson). Thus, in cases where we did not find clear racial/ethnic indicators, we labeled characters and authors as White. At least two annotators examined each and every character and author, and we discussed low-confidence labels. Unlike typical annotation tasks in data science, we found that annotator "disagreement" usually arose from retrieval challenges, where one annotator may be able to surface more explicit evidence of race while another may not. We changed labels only if we found more explicit or primary source evidence, and amended 4.7% of initial annotations for authors and 3.4% of them for characters.[13] Common sources of authors' sociodemographic information included authors' websites, articles featuring author interviews, and Wikipedia biographies. For characters, our explanations of our decisions often relied on book content, Wikipedia, Goodreads, and online study guides such as LitCharts and SparkNotes.

13 Many changes pertained to biracial/multiracial authors and characters. For example, an author may be labeled as belonging to a Native American tribe in the start of a source article, but later on include discussion of White ancestry.

5. Methods and Analysis Approach

Recommended book lists in education, when diversity-focused, often emphasize the identities of protagonists.[14] The concept of a protagonist assumes the existence of at least one central, forefronted character, and differentiates these characters from all others. In *The One vs. the Many*, Woloch argues that characters crowd for limited attention within a book; others, such as Vermeule, have also echoed this view. Woloch defines the *character-space* as the relationship between a character and the space and position they occupy in a narrative, while a *character-system* refers to the arrangement of character-spaces among each other. Under this theory, the appearance of each character is thus bound to "disrupt the distribution of [a reader's] attention" (Woloch; Woloch 26). We are also motivated by this notion that characters compete for highly valued attention throughout a book, and that their degree of *presence* acts as a proxy for the size of the space they occupy in the reader's attention. To draw out who is forefronted and who is less so in our dataset, we define three metrics to quantify characters' presence within a book: frequency, perspective, and burstiness.

For *frequency*, we calculate the total number of times a main character is mentioned, which includes occurrences of their aliases and coreference-resolved pronouns provided by BookNLP. Mention count is how we delineate each book's primary, or most frequent, main character from secondary ones, and we normalize this value over all character mentions. We acknowledge that not all books feature a single primary character, but this delineation in our analysis allows us to differentiate which main characters are likely appearing in more supporting roles than others.

First-person narrative *perspective* can increase readers' empathy of out-group members compared to third-person perspective, and media psychology researchers (Kim et al.) characterize this mechanism as increasing the "social presence" of a character. In a related study, first person narratives featuring characters that belong to the same in-group as readers resulted in more experience-taking, which is an immersive process in which readers identify with and adopt characters' emotions, thoughts, and traits (Kaufman and Libby). Thus, we also consider perspective as another aspect of character presence. We divide characters into two categories: those who are first person narrators at any point in the book, and those who are not. That is, if a book includes multiple first-person narrators, we count a character as appearing in first-person perspective if we disambiguated them as one of these narrators in §4.3.

14 For example, the NCTE's "Build Your Stack" blog series: https://ncte.org/blog/tag/build-your-stack/

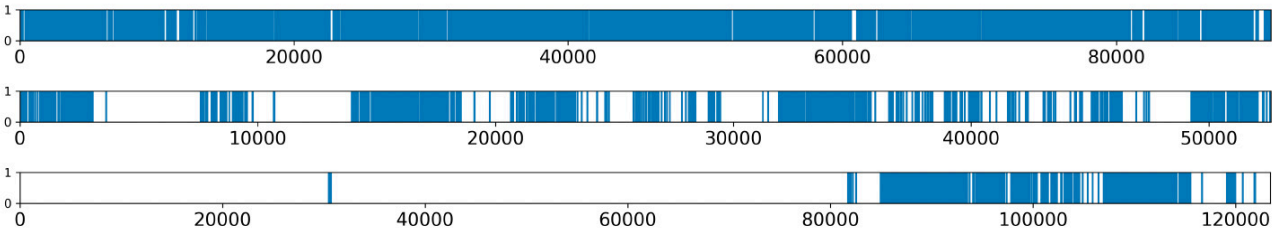

Figure 2. Three examples of characters who appear with different levels of burstiness. From top to bottom: Aristotle in *Aristotle and Dante Discover the Secrets of the Universe* (B = 0.185), Sampson in *We Beat the Street* (B = 0.639), Moushumi in *The Namesake* (B = 0.910). Each vertical line represents each mention's start token position in the book.

Finally, we calculate the *burstiness* of each main character in a book. Burstiness is a metric that characterizes patterns of event occurrence. Past researchers have used it to describe human activity patterns, such as communication in social networks (Goh and Barabási; Navarro et al.). To our knowledge, our work is the first in which it is applied to narratives, and we chose it as our third presence metric because it can capture the rhythm of attention a character is given throughout a book. We define a character c's burstiness as

$$B(c) = \frac{\sigma_c - \mu_c}{\sigma_c + \mu_c},$$

where σ_c and μ_c are the standard deviation and mean, respectively, of inter-onset intervals between c's occurrences. We measure inter-onset interval lengths by the difference between the start token index of one occurrence of a character and the start of their next occurrence. Less bursty characters are more consistently present throughout the book, while bursty ones are fleeting or isolated to a few sections (Figure 2). We theorize that burstiness may be another way in which characters of different racial/ethnic backgrounds may differ in presence within or across books. For instance, we hypothesize that on average, White characters and characters of color may occur with different levels of burstiness within a dataset.

We tie the above three metrics of character presence into analyses of book canonicity, character-author similarity, and racial integration. The remainder of this paper describes a series of results, starting with overviews of character presence and book canonicity in each of the two datasets we study (§6-7). Then, we discuss both datasets together to gain an encompassing view of trends shared across them, such as how characters' presence relates to the social identities of their authors (§8). Finally, given that content depicting intergroup contact can affect readers' attitudes towards others (Banas et al.), §2), we examine the degree to which main characters of differing racial/ethnic backgrounds appear in the same book (§9).

6. An ELA Canon: AP Literature Books

6.1. Character Presence

As discussed in §4.1, literary merit is a key motivator behind the selection and curation of AP reading lists. Literary merit is often indicated by the reception of literary prizes, which includes major awards such as the National Book Award or Pulitzer, and awards centered on media for children and young adults, such as the American Library Association's Caldecott, Coretta Scott King, and Newbery awards. Grossman et al. showed that after the year 2000, literary prize winners became increasingly less White, with Black writers winning more literary prizes than any other race in 2017. Though the institutional contexts of these prizes differs from those in education, if AP Literature were to reflect cultural shifts in the attribution of literary merit, one may expect that book suggestions by AP Literature would become less White over time.

On the contrary, we find that the number of books with main characters of color in AP Literature has wavered within consistent ranges over a twenty-plus-year span of 1999-2021 (Figure 3).[15] This consistency in composition over decades persists even when we break down proportions by specific racial/ethnic groups, and does not appear to be affected by major social justice movements such as Black Lives Matter in the 2010s. In addition to these racial/ethnic patterns, we find that higher proportion of books involving male main characters than books involving female main characters, and this binary gender gap is consistent over the years for both White main characters (Cohen's $d = 0.565$) and main characters of color (Cohen's $d = 0.813$, Appendix 13).[16] Though our sample begins in 1999, the history of College Board's AP program stretches further back, and it's possible that bigger changes in AP Literature's book lists may have occurred in the exam's earlier decades. In the 1950s, a committee of White men from preparatory schools and Ivy League universities initiated the program through lengthy discussions imbued with exclusivity and elitism, sharply differing from the program's current rhetoric around "equity and access" (Abrams 50; "Broadening Access" 3). In addition, our time frame does not illustrate the effects, if any, of the canon wars of the 1980s. Still, the sociodemographic composition of the exam's book lists in recent decades do not show a transformative shift, and as we'll discuss in §6.2, echo more historical work characterizing the Western literary canon.

15 To differentiate multiracial characters or authors with White ancestry from those who are solely White, we represent the latter with the label "White*" in figures.

16 We left out the one main character, Cal/Callie from *Middlesex*, from this binary gender analysis, as this intersex character's gender identity changes over the course of the book.

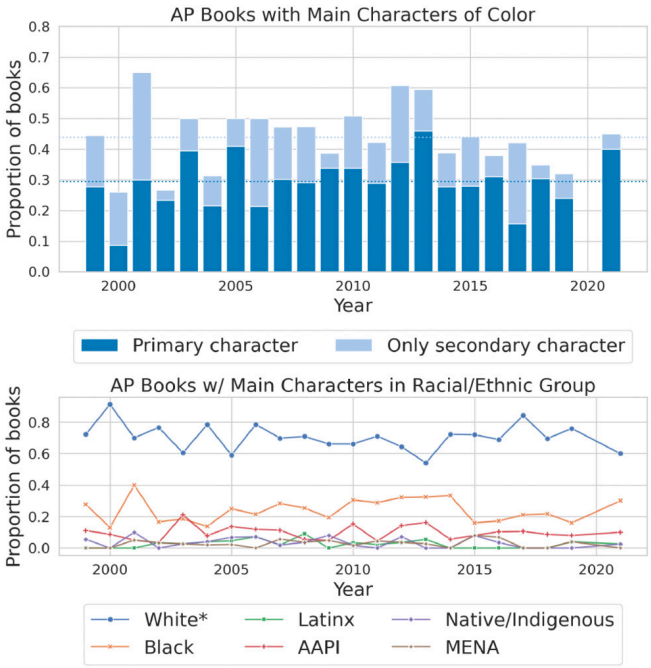

Figure 3. The top plot shows the proportion of AP books each year, out of all books listed that year, that contain at least one main character of color, who either appears as the primary, or most frequent, character, or as a secondary character. There, horizontal dotted lines are averages across all years. The bottom plot shows the proportion of books each year that include at least one main character labeled with some race/ethnicity. AAPI = Asian American & Pacific Islander, MENA = Middle Eastern & North African. *Multiracial characters are multiply counted across categories, except in White.

Figure 3 longitudinally shows books containing *any* main character in a racial/ethnic group; when we focus on AP books' primary, or most frequent, characters, we find that they are most often White (Figure 4). Together, these figures show recent levels of racial/ethnic representation in AP. Quantitative measurements like ours can inform studies around how different levels of representation impact students, and provide a comparison point for other proposed models. For example, J. Saunders Redding, a Black faculty member who taught at both historically Black colleges and Ivy League universities in the 1940s-1970s, rejected the model of supplementing existing courses with a few Black writers (Buurma and Heffernan) 113). Instead, he crafted American literature syllabi with fifty-fifty splits of works by Black and White writers, believing that these proportions best represented the country's politics and history at that time (Buurma and Heffernan 126). AP's composition of primary characters is far from a stratified sample of current US Census groups, and does not match demographic-based sampling of US students' backgrounds, either. That is, AP Literature's proportion of books containing White primary characters even exceeds the US's proportion of White high school students (Figure 1). Given the many possible pedagogical purposes of literature, it remains an open question as to how to best compose ELA book lists for high school students, especially with local contexts in mind. To contribute another view of ELA, we'll examine what teachers teach minoritized student audiences in §7.

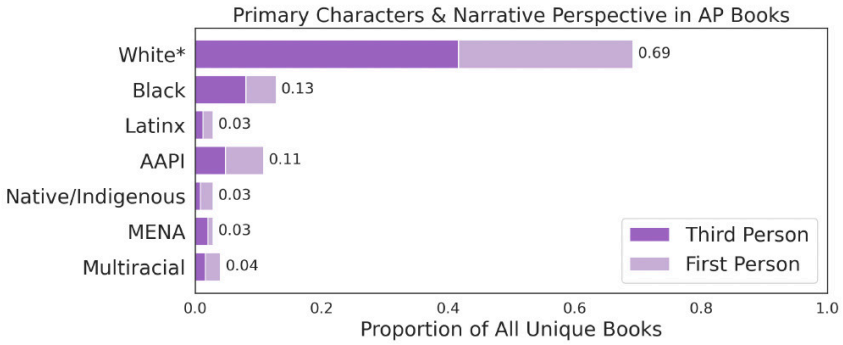

Figure 4. Primary characters, split by narrative perspective. AAPI = Asian American & Pacific Islander, MENA = Middle Eastern & North African. *Multiracial characters are multiply counted across categories, except in White.

When it comes to other metrics of presence, we did not observe a statistically significant difference in the average *burstiness* of primary characters of color and that of White primary characters, contrary to our hypothesis. In other words, primary characters across these two groups occurred with similar consistency throughout their books. In addition, when it comes to *perspective*, AP Literature includes a fairly balanced mix of first and third person books (Figure 4). We'll revisit these two presence metrics of burstiness and perspective again, as they appear in the teacher book list §7 and as they relate to author race/ethnicity §8.

6.2. Book Canonicity

Manshel's definition of the literary canon as authors and titles with "artistic timelessness" and "historical longevity" emphasize the role of time in canonization (32-33). Our longitudinal snapshot of AP Literature exams allows us to examine canonicity operationalized as temporal recurrence. Some views of Figure 3 may suggest that AP Literature is diverse "enough"; in some years, the percentage of books that include at least one main character of color dances around the 50% mark, and there is a notable set of books that feature Black main characters in some years. However, by examining these exams through the lens of canonicity, we find that many books involving primary characters of color are ephemeral, rarely securing a regular spot in yearly suggested book lists. That is, the most persistent books tend to be those that feature White primary characters (Figure 5). Some of these canonical books with White primary characters, such as *The Great Gatsby*, *Heart of Darkness*, and *Adventures of Huckleberry Finn*, do depict racial inequality or include secondary main characters of color. Still, these books' prioritization of White characters in their allocation of narrative space suggests that readers likely interact with topics of race and racism through White characters' experiences and perspectives.

When considering both race and gender, the most frequently represented subgroup of authors in our AP sample are White men (Figure 5). Nine out of 25 titles in this figure feature female primary characters (Appendix 13).

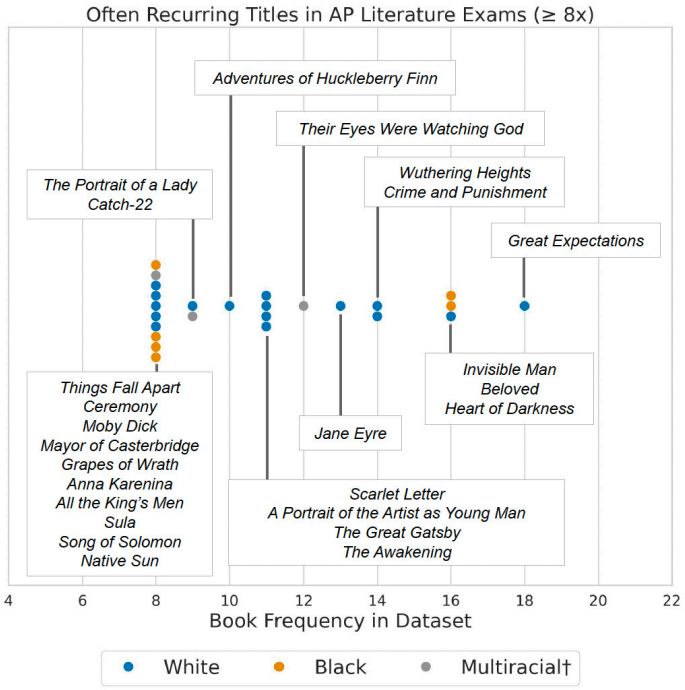

Figure 5. Titles of books that occur at least 8 times in AP Literature exams. We chose 8 as a cutoff for this visual because it encompasses the top 20 most frequent books, with ties included. Books are colored by primary character race/ethnicity. † From left to right, the three books labeled as "Multiracial" have primary characters who are Native/Indigenous & White, MENA & White, and Black & White, respectively.

Toni Morrison stands out as a frequently suggested Black woman author, having written multiple highly recurring books with Black characters (*Sula*, *Song of Solomon*, and *Beloved*). Her prominence in AP Literature echoes prior discussion by Manshel, who describes Morrison as "the novelist who best exemplifies the contentious institutionalization of minoritized writers in the late twentieth century" (22). Generally, though there may be an AP canon for White characters and perhaps a smaller one for Black characters, there are less definitive ones for other racial/ethnic groups. None of the top 15 most frequently listed books involve AAPI or Latinx primary characters. A total of 27 AAPI titles have appeared in AP exams, positioning it as the third largest racial/ethnic category (Figure 4). However, most of these titles' appearances are fleeting, and the most frequently recurring AAPI book, *Obasan* by Joy Kogawa, appears in 6 years. The absence of Latinx representation in the canon is particularly striking considering how US high schools as a whole serve substantial populations of Hispanic/Latino students (Figure 1). The most frequent Latinx book, *The House on Mango Street* by Sandra Cisneros, occurs in only 4 exam years.

"Achieving" racial representation through literary canonization is not without caveats. The canonization of a few representative characters and authors of color can burden these literary works to act as spokespeople of entire racial/ethnic groups (Sáez 83), and may have disproportionate influence on the range of accepted narrative types about a group (Manshel 31). Some educators have suggested turning away from a canon altogether and instead

advocate for curating curricula from a wide and dynamic array of texts (Boakye). Still, the sociocultural effects of the AP canon can be difficult to disregard. Greater repetition of some works over others signals whose narratives AP Literature, an influential and centralized curriculum, deems most significant and valuable. That is, canonization reworks books into cultural capital, and so decisions around what to teach impact students' status and social mobility (Guillory, *Cultural Capital*). In addition, a canon may drive rich-get-richer effects; for AP Literature, study guides often directly state that frequently appearing books in past exams are exemplars of which works students should read (McCammon; Gebauer; Albert).[17] Thus, our results show key gaps in the AP canon within the framework of culturally relevant pedagogy (Ladson-Billings, "Three Decades of Culturally Relevant, Responsive, & Sustaining Pedagogy"), especially for students of color.

Other studies also consider many of the canonized titles in our AP Literature sample as cornerstones of ELA (Kumar; Stotsky et al.; Stallworth and Gibbons; Hadley and Toliver). For example, the top three titles listed by teachers in a 2006 study of 142 Alabama teachers included *The Scarlet Letter*, *The Great Gatsby*, and *To Kill a Mockingbird* (Stallworth et al. 482), and the first two titles both recur in 11 years of AP exams, and the last title in 7 years. As another example, a national survey of 488 secondary schools in 1989 surfaced *Adventures of Huckleberry Finn* as the most frequently taught ELA book (Applebee 4). Though published over thirty years ago, Applebee's conclusion touches on enduring trends: "In all the settings which we examined, the lists of most frequently required books and authors were dominated by White males, with little change in overall balance from similar lists 25 or 80 years ago" (18). These prior snapshots of ELA in the US, combined with ours, suggest that the ELA canon may have been obdurate to societal shifts for over a century.

7. Another View of ELA: Teacher-Provided Books

7.1. Character Presence

Though the AP exam is administered widely across the US, the books that AP Literature exams suggest may not reflect what all classrooms teach. The books listed by the teachers we surveyed provide a case study for how ELA materials in schools that serve racially and ethnically diverse student audiences can differ from the lists promoted by AP Literature. We hypothesized that some teacher-provided books may have been deliberately chosen to be culturally relevant to their student audiences. Indeed, we find that the racial/ethnic composition of books listed by each teacher corresponds to the composition

17 The sources cited here are the top study guides resulting from a Google search of "*books to read for ap lit exam*".

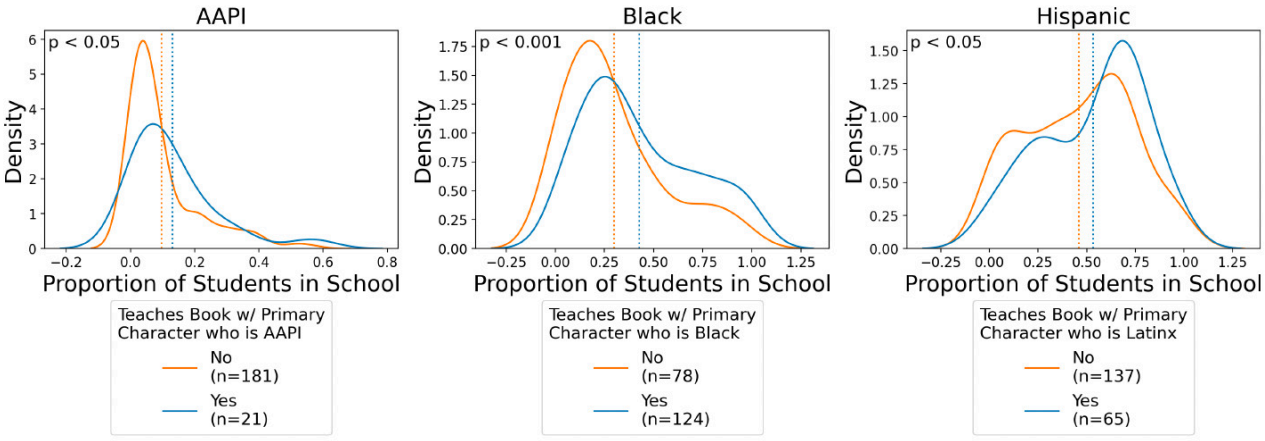

Figure 6. Students' demographics relate to the racial/ethnic composition of books in their ELA classes. These plots show kernel density estimates of demographic distributions.

of their students (Figure 6).[18] In other words, teachers in schools with more AAPI, Black, and Hispanic students tend to teach more books with primary characters of similar racial/ethnic backgrounds.

The majority of books in this dataset include main characters of color, and when present, they usually include the primary character (Figure 7). When it comes to character *perspective*, this teacher-provided dataset emphasizes first person narratives more so than third person ones, especially for primary characters of color (Figure 8). Broadly, the teacher-provided dataset includes main characters that are more evenly distributed across multiple racial/ethnic groups than AP Literature. Again, we did not find a statistically significant difference in the *burstiness* of primary characters of color and that of White primary characters. However, these teacher-provided books on average have less bursty, or more consistently present, primary characters than those in AP Literature. This finding applies to both primary characters of color (Mann-Whitney U-test, $p < 0.001$) and White primary characters (Mann-Whitney U-test, $p < 0.01$), and, when combined with this dataset's first-person emphasis, suggests that teacher-provided books heavily focus on the actions and viewpoints of primary characters. These findings illustrate how the racial/ethnic composition of characters and their degree of presence in a set of ELA books can differ from AP. That is, though AP Literature books may not sufficiently represent minoritized student populations (§6), our teacher-provided dataset serves as a demonstration of how ELA teachers can adapt their curricula to create more mirrors for their students of color.

18 Out of all 189 surveyed teachers, 14 teach AP Lit, but we did not observe a notable difference in the composition of books they teach compared to others. For example, all 14 of these teachers teach at least one book featuring a Black primary character.

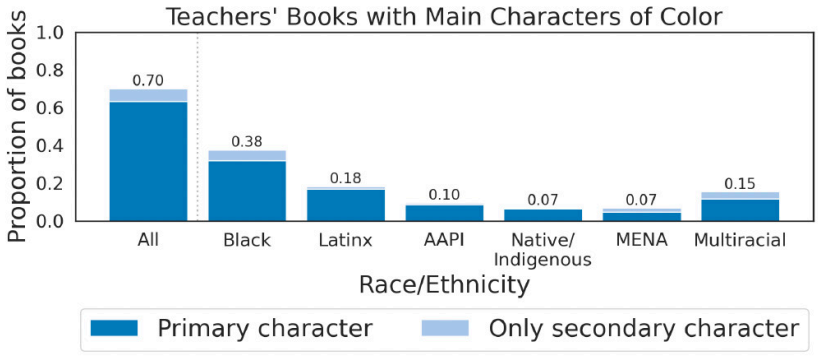

Figure 7. The racial/ethnic backgrounds of main characters in our teacher-provided dataset. AAPI = Asian American & Pacific Islander, MENA = Middle Eastern & North African. Multiracial characters are multiply counted across categories.

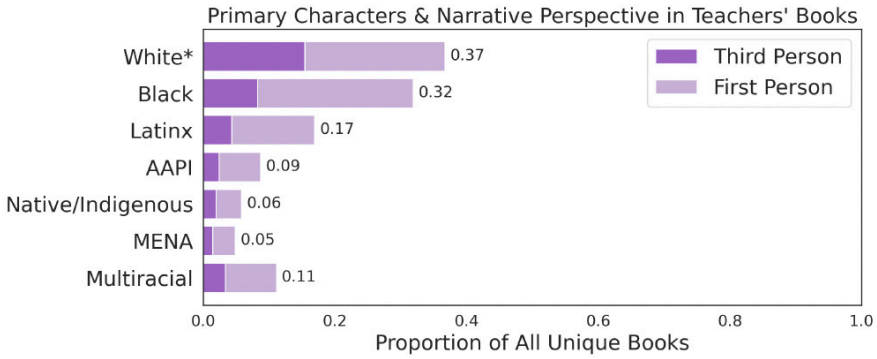

Figure 8. Primary characters, split by narrative perspective. AAPI = Asian American & Pacific Islander, MENA = Middle Eastern & North African. *Multiracial characters are multiply counted across categories, except White.

7.2. Book Canonicity

The teachers in our case study are not necessarily representative of teachers nationally, and our sample here spans a much shorter time frame than AP's. Therefore, the recurrence of titles within this dataset may be less indicative of a pervasive canon than recurrences in AP Literature. Still, some scholars, such as Manshel and Chiang, have suggested revising definitions of canonicity to go beyond the notion of a single canon unyielding to time; for instance, individuals may craft their own "personal canon" (Dalleo and Sáez 1). Titles with widespread use among the teachers in our study suggest the formation of alternative canons to the one imposed by AP (Figure 9). Popular titles include older "classics" by White authors such as *The Great Gatsby* (1925), *To Kill a Mockingbird* (1960), and *Animal Farm* (1945), but also newer books by authors of color such as *The Hate U Give* (2017) and *The Absolutely True Diary of a Part-Time Indian* (2007). These latter two titles do not appear in AP exams up to 2021 and suggest existing ways in which US teachers may deviate from AP. In fact, a majority (68.7%) of books that appear in this teacher-provided dataset but not in AP Literature feature primary characters of color. Still, the gender gap among recurring titles in Figure 9 is large, where

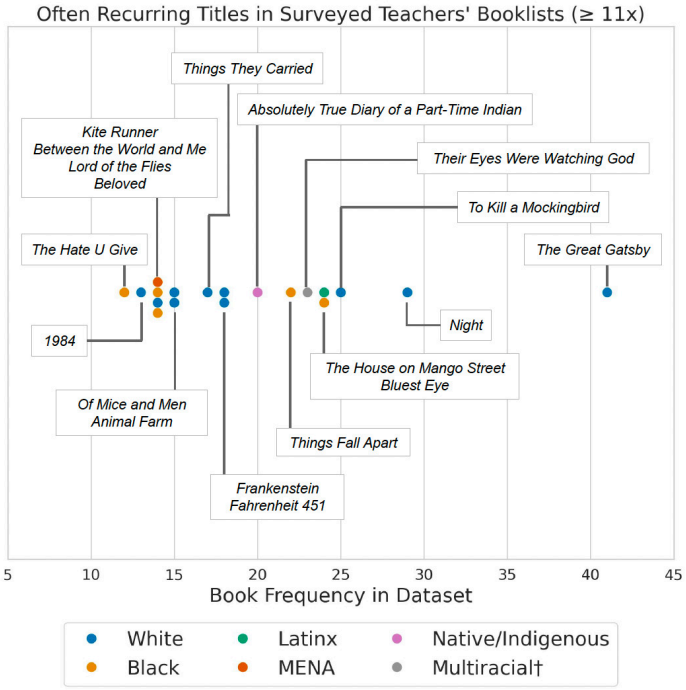

Figure 9. Titles of books that occur at least 11 times in teacher-provided lists. We chose 11 as a cutoff for this visual because it approximately includes the top 20 most frequent books. Books are colored by primary character race/ethnicity; MENA = Middle Eastern & North African. †The book labeled as "Multiracial" features a primary character who has Black and White ancestry.

only 5 of the 19 titles shown feature female primary characters (Appendix 13). Overall, the two ELA datasets in our study present distinct snapshots of how literature may be taught in the US, and together, provide concrete possibilities around what types of books and characters could be taught in classrooms.

8. Authors Mediate Characters' Presence

Book data containing a range of racial/ethnic identities among authors and characters permits us to investigate how authors' identities may relate to their characters' presence. A common characterization of diversity in literature emphasizes the social identities of authors (NCTE). In this section, we examine the frequency, narrative perspective, and burstiness of main characters who do or do not racially/ethnically mirror their authors.

Authors' personal experiences mediate their writing, with memoir and autofiction, or autobiographical fiction, as particularly striking cases of when this occurs. Thus, it may be unsurprising that we find that in both of our datasets, 28 (7.1%) books include authors writing about primary characters whose racial/ethnic identities do not overlap or match their own (Figure 10).[19] In three books, each of which only appear in our teacher-provided

19 Overlapping identities includes cases such as *Sing, Unburied, Sing*, which is written by Black author Jesmyn Ward and the main character is half-Black, half-White, but not cases where White authors write about biracial/multiracial characters.

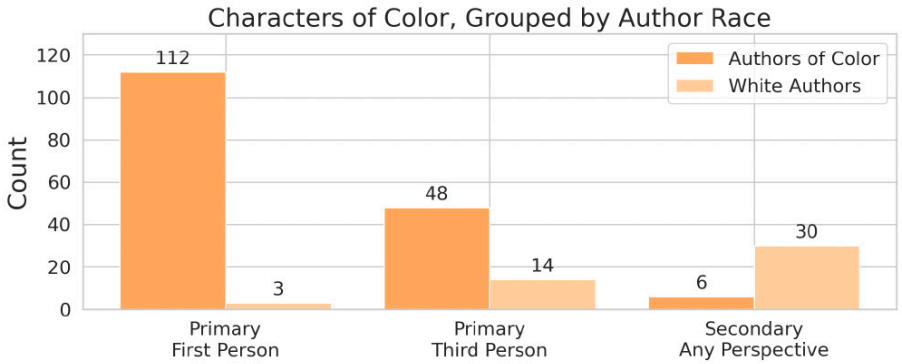

Figure 10. When it comes to frequency and perspective, authors of color tend to forefront characters of color more so than White authors.

dataset, White authors write first-person, primary characters of color: *Thus Spake Zarathustra* by Friedrich Nietzsche, *Sold* by Patricia McCormick, and *Life of Pi* by Yann Martel. The first book features a Middle Eastern primary character, while the latter two involve South Asian primary characters. When authors of color write about primary characters of color, the characters' racial identities usually match or overlap with that of their authors, with two exceptions: *White Teeth*, with a Bangladeshi primary character written by Jamaican-English author Zadie Smith, and *Coming Through Slaughter*, with a Black primary character written by Sri-Lankan-Canadian author Michael Ondaatje. Writing across genders is more common than cross-race writing in ELA; in 19.2% of the books in our datasets (76 books), authors write primary characters whose genders differ from their own.

Authors mention main characters at similar rates regardless of whether these characters' racial/ethnic identities match or overlap with theirs (Figure 11). However, the rhythm of their attention towards primary characters can vary. That is, authors tend to write about primary characters in more consistent, less bursty patterns over the course of a book if their race/ethnic backgrounds match or overlap.[20] So, even if authors write about primary characters at similar frequencies, these characters' burstiness scores are closer to those of secondary characters if they are racially/ethnically dissimilar to their authors.

Our findings about authors who write about main characters who differ from themselves generally follow our intuition that authors' lived experiences mediate the content they write and the attention they draw towards their characters. Several studies of literature have focused on measuring sociodemographic diversity of book lists by focusing on the identities of authors (e.g. Kumar), especially since retrieving information about characters can be more labor intensive than retrieving analogous information about

20 Visually, these distributions persist even if we separate out the cases where White authors write about primary characters of color from those in which authors of color write about White primary characters. Due to too low sample sizes for statistical significance testing, we grouped these two cases under an aggregated "authors and characters are dissimilar" sample.

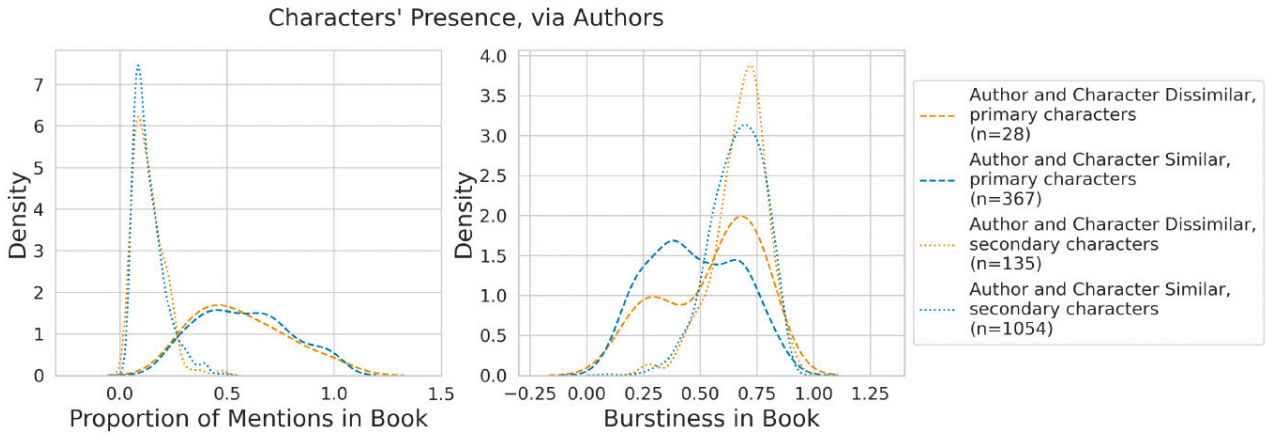

Figure 11. Despite little differences in the frequency of authors' mentions of characters who may or may not differ from themselves, the burstiness of primary characters increases if their race/ethnicity does not match their authors' (Mann–Whitney U test, $p < 0.05$).

authors. Our study provides quantitative evidence that indeed, in most cases, authors' identities match their most forefronted characters. Thus, authors' racial/ethnic identities can often act as a proxy for estimating who they emphasize within book content and informing diversity-aware book selection processes.

9. Most Books are Racially/Ethnically Homophilic

Our annotation of multiple characters per book enables us to also examine the extent to which books include characters from different racial/ethnic backgrounds. In sociology and network science, homophily is a principle where contact or association occurs at a higher rate between similar people than dissimilar ones, and interpersonal similarity may be defined by sociodemographic or behavorial characteristics (McPherson et al.). A majority of books, or 74.8% in AP Literature and 66.2% in our teacher-provided dataset, are racially/ethnically homophilic, where all main characters have matching or overlapping identities. These percentages are relatively high when compared to a null model where all race/ethnicity labels of main characters are shuffled across books in each dataset; there, only 27.5% of AP books and 22.5% of teacher-provided books are racially/ethnically homophilic.[21]

Some books that have race as their main subject matter may still emphasize a main cast of entirely White characters. One prominent example of this is To Kill a Mockingbird, which is in both datasets and is written by Harper Lee, a White author. Set in the 1930s American South, acts of racial injustice drive its plot. However, all of the main characters we annotated for this book are White: Scout, Atticus, Jem, and Dill. Major Black characters, such

21 The percentages shown for these null models are averages over fifty random shuffles, with standard deviations of 3.0 and 1.5, respectively.

as the family's cook Calpurnia or the falsely accused man Tom Robinson, do not occur often enough to meet our threshold for determining main characters. As another example, Albert Camus's *The Stranger*, which appears in both datasets, revolves around the killing of an Arab man in French-colonized Algeria, but none of the Arab characters are given names in the book. A few books stand out as cases where characters from different racial/ethnic backgrounds share similar levels of presence. Three books in the teacher-provided dataset involve multiple narrators who span distinct racial/ethnic backgrounds, and intentionally contrast their racialized experiences: *The Help* (Black and White), *Bronx Masquerade* (Black, White, Latino/a/x), and *All American Boys* (Black, White). The third book, *All American Boys*, is particularly notable in that it is also written by two authors rather than one: Brendan Kiely, who is White, and Jason Reynolds, who is Black.

Across both ELA datasets, eighty-seven books include White main characters with main characters of color, and only 8 books include main characters of color from distinct backgrounds, e.g. the MENA and Latino/a/x main characters in *Spare Parts*, a book listed by two teachers. We find that it is more common for primary characters of color to appear with White secondary characters (58.6% of interracial books, 40.2% of all books with primary characters of color) than vice versa (41.4% of interracial books, 19.8% of all books with White primary characters). This difference is analogous to patterns in the physical world: White people across the US tend to be in more racially homophilic schools and communities compared to people in other racial groups (Schaeffer; Vachuska; Goddard).

One potential reason for these interracial patterns or lack thereof is that books reflect segregated realities, especially when considering the historical and cultural contexts of these narratives. Books also typically express a singular author's perspective and lived experiences, and as we saw in §8, their identities relate directly to who is forefronted over others. Some White authors may not consider race at all, and write all characters as seemingly raceless yet presumably White.[22] Other White authors may be concerned about inadequately representing characters whose backgrounds differ substantially from their own, especially if resulting depictions may risk cultural appropriation ("'Writing with Color'"; Adams). Another potential reason for a lack of racial integration in ELA books is that the selection of books focused on characters of color counterbalances their long-standing absences in curriculum. A large proportion of older "classics" focus solely on White characters, and so it is reasonable to assume that the inclusion of books featuring characters of color need not also forefront any White characters. One consequence, however, of these choices is that racially or

22 For example, the dystopian novel *The Giver* depicts a community that values "sameness" among its members, and White actors play all characters in its film adaptation.

ethnically homophilic narratives, when read on aggregate, may suggest to students that racial/ethnic segregation is a societal norm. That is, just as the we have predominantly White neighborhoods and predominantly Black neighborhoods, we teach "White books" and "Black books." As the rate of racial integration has slowed over time in the United States despite increasing racial diversity (Lichter 373; Parisi et al. 43), it is imperative that educators consider this facet of characters' presence when considering what ELA curriculum could look like beyond the books in our two datasets.

10. Conclusion

Our work presents two contrasting datasets of ELA books hand-annotated with character and author race to provide a foundation for studies on how race is or could be taught in US schools. We draw our first set of books from exams administered by AP Literature, which prescribes a centralized curriculum for US students. Our second dataset consists of books listed by ELA teachers who primarily teach students of color, allowing us to examine how teachers may adapt their curricula for minoritized students. Our annotations of authors' and characters' racial/ethnic backgrounds, combined with our computational metrics of presence, facilitate several key findings around the racial/ethnic composition of these books. For instance, we show that AP Literature has consistently recommended a majority of books containing White primary characters for over twenty years, prioritizing these characters' experiences and perspectives in their distribution of narrative space. In addition, Latinx, AAPI, Native/Indigenous, and MENA primary characters are relatively absent in the AP canon, when canonization is defined by repeated occurrence in exams over time. We also provide a novel application of burstiness as a metric for quantifying characters' presence and show that this metric teases out differences in authors' inclusion of characters of differing races that character frequency alone cannot. In addition to surfacing concrete examples of what kinds of books may be taught in the US, we identify a notable gap in both datasets: books that integrate main characters from varying racial/ethnic backgrounds.

Measuring representation in books is complex, even when representation is scoped to *presence*—the racial/ethnic identities of an author and their characters' jointly contribute to who is present. In addition to counting how often multiple characters in each book occur, we also measure whether readers are immersed in characters' experiences through first-person perspectives, as well as characters' rhythm of occurrence throughout a book. Given the pedagogical significance of curricular materials reflecting students' identities and expanding their knowledge of others (Bishop; Banas et al.), it is critical to have transparent and rigorous ways in which educators could consider representation. For example, representation in ELA curricula can be viewed at the level of a single book or across multiple instantiations of a syllabus or course. Several titles involving primary characters of color may appear in AP Literature exams each year (Figure 3), but one could

measure representation beyond surface-level inclusion by also considering what books persist over time (Figure 5). We hope that our contributions can support evidence-backed processes that consider multiple dimensions of representation when curating ELA curricula, within both centralized institutions and individual classrooms.

One limitation of our work is that we study books isolated from their classroom contexts. That is, future work should investigate how teachers interweave these books into discussions with students, and the psychological impact characters' representation may have on students. It is impossible to make normative claims about what should be taught from book content alone, without consideration for how it is taught or socially situated. For example, there may be pedagogical value in comparing books that center racial inequality written by White authors (e.g. *The Help*, *To Kill a Mockingbird*) with those by authors of color (e.g. *Beloved*, *The Hate U Give*). As an illustration of how a classroom's local context matters, we saw some evidence in our teacher-provided dataset that curriculum may be adapted to student audiences. Our work contributes to ongoing lines of research investigating the selection and implementation of ELA curriculum in the US and its impact on students, especially during a time of increasing school book bans (Alter; Hadley and Toliver). In a landscape where efforts to diversify school curricula are actively under attack (Pollock et al.), it is important to provide tools and information that can empower educators to make purposeful and meaningful decisions around not only what to teach, but how to teach. Generally, it takes more action than simply diversifying reading lists to build a more equitable, antiracist society (Melamed).

Our work is also the first step for future studies that could develop measurements of racial/ethnic representation in literature that can reflect the substantive role a character plays in the narrative. That is, future work can build upon ours by analyzing how minoritized characters are depicted, such as the degree of essentialism in their portrayals or their positioning within a larger literary landscape. Other future work could also analyze how racial inequality and racism are illustrated through narratives, and how these depictions may vary across samples of ELA linked to different pedagogical contexts. We also encourage other researchers to critically revisit our annotations, perhaps by correcting possible misidentifications or carefully characterizing the ways in which racial/ethnic identity may be perceived by readers from textual and contextual signals. By grappling with these intricacies and practical constraints around measuring racial/ethnic representation in books at scale, we can reach a better understanding of who is forefronted and canonized in school curriculum and beyond.

Acknowledgments

We thank Jennifer Wolf, Maria Antoniak, and Richard Jean So for helpful conversations over various iterations of this work, and we thank Sebastian Orozco and Aryia Dattamajumdar for their data annotation support. Finally, we thank our paper reviewers, whose thoughtful comments helped us refine our work. The research reported in this article was supported by funding from the National Science Foundation (IIS-1942591).

Data Repository: https://doi.org/10.7910/DVN/WZQRRH
Peer reviewers: Rachel Buurma, Alexander Manshel

Submitted: July 09, 2024 EDT. Accepted: December 20, 2024 EDT. Published: March 19, 2025 EDT.

WORKS CITED

"A Majority of Grade 9-12 Public Schools Rate Themselves Favorably on Preparing Students for College." *National Center for Education Statistics*, U.S. Department of Education, 2024, https://nces.ed.gov/whatsnew/press_releases/3_19_2024.asp.

Abrams, Annie. "Shortchanged: How Advanced Placement Cheats Students." *Johns Hopkins University Press*, 2023.

Adams, Deena. *Should White Authors Write Diverse Characters?* 2021, https://deenaadams.com/should-white-authors-write-diverse-characters/.

Adukia, Anjali, et al. "What We Teach About Race and Gender: Representation in Images and Text of Children's Books." *The Quarterly Journal of Economics*, vol. 138, no. 4, Aug. 2023, pp. 2225–85, https://doi.org/10.1093/qje/qjad028.

Agosto, Denise E., et al. "The All-White World of Middle-School Genre Fiction: Surveying the Field for Multicultural Protagonists." *Children's Literature in Education*, vol. 34, 2003, pp. 257–75.

Albert. "The Ultimate AP English Literature Reading List." *Learn By Doing, Inc.*, Mar. 2022, https://www.albert.io/blog/ultimate-ap-english-literature-reading-list/.

Algee-Hewitt, Mark, et al. "Representing Race and Ethnicity in American Fiction, 1789-1920." *Journal of Cultural Analytics*, vol. 5, no. 2, Dec. 2020, https://doi.org/10.22148/001c.18509.

Alter, Alexandra. "Book Bans Continue to Surge in Public Schools." *The New York Times*, Apr. 2024, https://www.nytimes.com/2024/04/16/books/book-bans-public-schools.html.

"American Community Survey." *Explore Census Data*, United States Census Bureau, 2022, https://data.census.gov/table?q=race%20high%20school&t=-01:Education:School%20Enrollment.

Applebee, Arthur N. *A Study of Book-Length Works Taught in High School English Courses. Report Series 1.2.* 1989.

Aronson, Brittany, and Judson Laughter. "The Theory and Practice of Culturally Relevant Education: A Synthesis of Research across Content Areas." *Review of Educational Research*, vol. 86, no. 1, 2016, pp. 163–206, https://doi.org/10.3102/0034654315582066.

Bamman, David, et al. "An Annotated Dataset of Coreference in English Literature." *Proceedings of the Twelfth Language Resources and Evaluation Conference*, edited by Nicoletta Calzolari et al., European Language Resources Association, 2020, pp. 44–54, https://aclanthology.org/2020.lrec-1.6.

---. *BookNLP. A Natural Language Processing Pipeline for Books*. Github, 2021.

Banas, John A., et al. "Meta-Analysis on Mediated Contact and Prejudice." *Human Communication Research*, vol. 46, no. 2–3, May 2020, pp. 120–60, https://doi.org/10.1093/hcr/hqaa004.

Bishop, Rudine Sims. "Windows and Mirrors: Children's Books and Parallel Cultures." *California State University Reading Conference: 14th Annual Conference Proceedings*, ERIC, 1990, pp. 3–12.

Boakye, Jeffrey. "The Big Idea: Do We Need to Dismantle the Literary Canon?" *The Guardian*, June 2023, https://www.theguardian.com/books/2023/jun/12/the-big-idea-do-we-need-to-dismantle-the-literary-canon.

Board, College. "AP National and State Data Archive." *AP Central*, College Board, 2024, https://apcentral.collegeboard.org/about-ap/ap-data-research/national-state-data/archive.

Bode, Katherine. "Why You Can't Model Away Bias." *Modern Language Quarterly*, vol. 81, no. 1, Mar. 2020, pp. 95–124, https://doi.org/10.1215/00267929-7933102.

Bonilla, Sade, et al. "Ethnic Studies Increases Longer-Run Academic Engagement and Attainment." *Proceedings of the National Academy of Sciences*, vol. 118, no. 37, 2021, p. e2026386118.

"Broadening Access to Advanced Placement." *AP Central*, College Board, 2022, https://apcentral.collegeboard.org/media/pdf/broadening-access-to-ap.pdf.

Buurma, Rachel Sagner, and Laura Heffernan. *The Teaching Archive: A New History for Literary Study*. University of Chicago Press, 2021, https://doi.org/10.7208/chicago/9780226736273.001.0001.

Byrd, Christy M. "Does Culturally Relevant Teaching Work? An Examination from Student Perspectives." *Sage Open*, vol. 6, no. 3, 2016, p. 2158244016660744.

"Can White People Write POC as Main Characters?" *Writing With Color*, Tumblr, 2021, https://writingwithcolor.tumblr.com/post/656713761418346496/can-white-people-write-poc-as-main-characters.

Chatterji, Roby, et al. "'Closing Advanced Coursework Equity Gaps for All Students.'" *Center for American Progress*, American Progress, June 2021, https://www.americanprogress.org/article/closing-advanced-coursework-equity-gaps-students/.

Chiang, Mark. *Autonomy and Representation in the University*. New York University Press, 2009, https://doi.org/10.18574/nyu/9780814717004.001.0001.

Dalleo, Raphael, and Elena Machado Sáez. *The Latino/a Canon and the Emergence of Post-Sixties Literature*. Springer, 2007.

Darling-Hammond, Linda. "Inequality in Teaching and Schooling: How Opportunity Is Rationed to Students of Color in America." *The Right Thing to Do, The Smart Thing to Do Enhancing Diversity in the Health Professions*, 2001, p. 208.

Dee, Thomas, and Emily Penner. *The Causal Effects of Cultural Relevance: Evidence from an Ethnic Studies Curriculum*. Working Paper, 21865, National Bureau of Economic Research, Jan. 2016, https://doi.org/10.3386/w21865.

Department, New York State Education. *Culturally Responsive-Sustaining Education Framework*. The University of the State of New York, 2018.

Field, Anjalie, et al. "A Survey of Race, Racism, and Anti-Racism in NLP." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, edited by Chengqing Zong et al., Association for Computational Linguistics, 2021, pp. 1905–25, https://doi.org/10.18653/v1/2021.acl-long.149.

Gebauer, Dane. "AP Lit Reading List – 50 Best Books to Read." *College Transitions*, Nov. 2023, https://www.collegetransitions.com/blog/ap-lit-reading-list/.

Goddard, Isabel. "What Does Friendship Look like in America?" *Pew Research Center*, Pew Research Center, Oct. 2023, https://www.pewresearch.org/short-reads/2023/10/12/what-does-friendship-look-like-in-america/.

Goh, K. I., and A. L. Barabási. "Burstiness and Memory in Complex Systems." *Europhysics Letters*, vol. 81, no. 4, 2008, p. 48002.

Goncalves, Marcelo S. O., et al. "Book Bans in Political Context: Evidence from US Schools." *PNAS Nexus*, vol. 3, no. 6, June 2024, p. pgae197, https://doi.org/10.1093/pnasnexus/pgae197.

González, José Eduardo, et al. "Measuring Canonicity: Graduate Reading Lists in Departments of Hispanic Studies." *Journal of Cultural Analytics*, vol. 6, no. 1, Mar. 2021, https://doi.org/10.22148/001c.21599.

Gopalan, Maithreyi, and Ashlyn Aiko Nelson. "Understanding the Racial Discipline Gap in Schools." *AERA Open*, vol. 5, no. 2, 2019, p. 2332858419844613.

Grossman, Claire, et al. "Who Gets to Be a Writer?" *Public Books*, Public Books, Apr. 2021, https://www.publicbooks.org/who-gets-to-be-a-writer/.

Guillory, John. "Canon, Syllabus, List: A Note on the Pedagogic Imaginary." *Transition*, no. 52, 1991, pp. 36–54, http://www.jstor.org/stable/2935123.

---. *Cultural Capital: The Problem of Literary Canon Formation*. University of Chicago Press, 2023, https://doi.org/10.7208/chicago/9780226830605.

Hadley, Heidi Lyn, and SR Toliver. "The Monstrous Hospitality of Canonical Text Selections: The Need for a Hospitable Literacy Framework." *Journal of Literacy Research*, vol. 55, no. 4, 2023, pp. 428–49.

Hanna, Alex, et al. "Towards a Critical Race Methodology in Algorithmic Fairness." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, 2020, pp. 501–12, https://doi.org/10.1145/3351095.3372826.

Kaufman, Geoff, and Lisa Libby. "Changing Beliefs and Behavior through Experience-Taking." *Journal of Personality and Social Psychology*, vol. 103, Mar. 2012, pp. 1–19, https://doi.org/10.1037/a0027525.

Kim, Nuri, et al. "The Presence of the Protagonist: Explaining Narrative Perspective Effects through Social Presence." *Media Psychology*, vol. 23, no. 6, 2020, pp. 891–914, https://doi.org/10.1080/15213269.2019.1665548.

Kumar, Tracey. "Where Are 'Their' Voices? Authors of Color in the Secondary ELA Curriculum." *Multicultural Education*, vol. 29, 2022, pp. 15–24.

Ladson-Billings, Gloria. *Culturally Relevant Pedagogy: Asking a Different Question*. Teachers College Press, 2021.

---. "Three Decades of Culturally Relevant, Responsive, & Sustaining Pedagogy: What Lies Ahead?" *The Educational Forum*, vol. 85, no. 4, 2021, pp. 351–54, https://doi.org/10.1080/00131725.2021.1957632.

---. "Toward a Theory of Culturally Relevant Pedagogy." *American Educational Research Journal*, vol. 32, no. 3, 1995, pp. 465–91, https://doi.org/10.3102/00028312032003465.

Le-Khac, Long, and Kate Hao. "The Asian American Literature We've Constructed." *Journal of Cultural Analytics*, vol. 6, no. 2, Apr. 2021, https://doi.org/10.22148/001c.22330.

Levine, Sarah. "A Century of Change in High School English Assessments: An Analysis of 110 New York State Regents Exams, 1900–2018." *Research in the Teaching of English*, vol. 54, no. 1, 2019, pp. 31–57.

---. "How Feeling Supports Students' Interpretive Discussions about Literature." *Journal of Literacy Research*, vol. 53, no. 4, 2021, pp. 491–515.

Lichter, Daniel T. "Integration or Fragmentation? Racial Diversity and the American Future." *Demography*, vol. 50, no. 2, Feb. 2013, pp. 359–91, https://doi.org/10.1007/s13524-013-0197-1.

Lucy, Li, et al. "Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas u.s. History Textbooks." *AERA Open*, vol. 6, no. 3, 2020, p. 2332858420940312, https://doi.org/10.1177/2332858420940312.

Manning, Christopher D., et al. *Introduction to Information Retrieval*. Cambridge university press, 2008.

Manshel, Alexander. *Writing Backwards: Historical Fiction and the Reshaping of the American Canon*. Columbia University Press, 2023.

Marks, Rachel, et al. "What Updates to OMB's Race/Ethnicity Standards Mean for the Census Bureau." *US Census Bureau*, US Census Bureau, Apr. 2024, https://www.census.gov/newsroom/blogs/random-samplings/2024/04/updates-race-ethnicity-standards.html.

Mazziotta, Agostino, et al. "Vicarious Intergroup Contact Effects: Applying Social-Cognitive Theory to Intergroup Contact Research." *Group Processes & Intergroup Relations*, vol. 14, no. 2, 2011, pp. 255–74.

McCammon, Ellen. "AP Literature Reading List: 127 Great Books for Your Prep." *PrepScholar*, 2016, https://blog.prepscholar.com/ap-literature-reading-list.

McDermott, Monica, and Annie Ferguson. "Sociology of Whiteness." *Annual Review of Sociology*, vol. 48, no. Volume 48, 2022, 2022, pp. 257–76, https://doi.org/10.1146/annurev-soc-083121-054338.

McGrath, Laura B. "'Books about Race': Commercial Publishing and Racial Formation in the 21st Century." *New Literary History*, vol. 54, no. 1, 2022, pp. 771–94.

McPherson, Miller, et al. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology*, vol. 27, 2001, pp. 415–44, https://doi.org/10.1146/annurev.soc.27.1.415.

Melamed, Jodi. *Represent and Destroy: Rationalizing Violence in the New Racial Capitalism*. University of Minnesota Press, 2011, https://doi.org/10.5749/minnesota/9780816674244.001.0001.

Navarro, Henry, et al. "Temporal Patterns behind the Strength of Persistent Ties." *EPJ Data Science*, vol. 6, 2017, pp. 1–19.

NCTE. "Guidelines for Selection of Materials in English Language Arts Programs." *Position Statements*, National Council of Teachers of English, Apr. 2014, https://ncte.org/statement/material-selection-ela/.

Newman, Andrew. "These High School 'Classics' Have Been Taught for Generations – Could They Be on Their Way Out?" *The Conversation*, Sept. 2022, https://theconversation.com/these-high-school-classics-have-been-taught-for-generations-could-they-be-on-their-way-out-188197.

Nomura, Nichole Misako, and Quinn Dombrowski. "Quantifying Representations of Asian Identity in 21st-Century Anglophone Fiction for Young Readers." *Digital Humanities*, 2022, https://dh-abstracts.library.virginia.edu/works/11757.

Parisi, Domenico, et al. "Racial Segregation in a Multiracial Society: Black Exclusion and Spatial Integration in US Municipalities, 1990–2020." *Population, Space and Place*, vol. 31, no. 1, 2025, p. e2870, https://doi.org/10.1002/psp.2870.

Pollock, Mica, et al. "Keeping the Freedom to Include: Teachers Navigating" Pushback" and Marshalling" Backup" to Keep Inclusion on the Agenda." *Journal of Leadership, Equity, and Research*, vol. 8, no. 1, 2022, pp. 87–114.

Potter, Nancy. *Course Perspective: English Literature and Composition*. College Board, https://apcentral.collegeboard.org/courses/ap-english-literature-and-composition/course/course-perspective.

Reardon, Sean F., et al. "The Geography of Racial/Ethnic Test Score Gaps." *American Journal of Sociology*, vol. 124, no. 4, 2019, pp. 1164–221.

Rigell, Amanda, et al. "Overwhelming Whiteness: A Critical Analysis of Race in a Scripted Reading Curriculum." *Journal of Curriculum Studies*, vol. 54, no. 6, 2022, pp. 852–70, https://doi.org/10.1080/00220272.2022.2030803.

Rowberry, Simon. *The Early Development of Project Gutenberg c.1970–2000*. Cambridge University Press, 2023.

Sáez, Elena Machado. *Market Aesthetics: The Purchase of the Past in Caribbean Diasporic Fiction*. University of Virginia Press, 2015.

Safarpour, Alauna, et al. "Divisive or Descriptive?: How Americans Understand Critical Race Theory." *Journal of Race, Ethnicity, and Politics*, 2024, pp. 1–25.

Schaeffer, Katherine. "U.s. Public School Students Often Go to Schools Where at Least Half of Their Peers Are the Same Race or Ethnicity." *Pew Research Center*, Pew Research Center, Dec. 2021, https://www.pewresearch.org/short-reads/2021/12/15/u-s-public-school-students-often-go-to-schools-where-at-least-half-of-their-peers-are-the-same-race-or-ethnicity/.

"Search for Public Schools." *National Center for Education Statistics*, U.S. Department of Education, 2024, https://nces.ed.gov/ccd/schoolsearch/.

So, Richard Jean. *Redlining Culture: A Data History of Racial Inequality and Postwar Fiction*. Columbia University Press, 2021, http://www.jstor.org/stable/10.7312/so--19772.

So, Richard Jean, and Gus Wezerek. "Just How White Is the Book Industry?" *The New York Times*, Dec. 2020, https://www.nytimes.com/interactive/2020/12/11/opinion/culture/diversity-publishing-industry.html.

Stallworth, B. Joyce, et al. "It's Not on the List: An Exploration of Teachers' Perspectives on Using Multicultural Literature." *Journal of Adolescent & Adult Literacy*, vol. 49, no. 6, 2006, pp. 478–89, http://www.jstor.org/stable/40017605.

Stallworth, JB, and L. Gibbons. "What's on the List... Now? A Survey of Book-Length Works Taught in Secondary Schools." *English Leadership Quarterly*, vol. 34, no. 3, 2012, pp. 2–3.

Steele, Dorothy M., and Becki Cohn-Vargas. *Identity Safe Classrooms, Grades K-5: Places to Belong and Learn*. Corwin Press, 2013.

Stotsky, Sandra, et al. "Literary Study in Grades 9, 10, and 11 in Arkansas." *University of Arkansas Department of Education Reform*, 2010.

Strange, Jeffrey J. "How Fictional Tales Wag Real-World Beliefs: Models and Mechanisms of Narrative Influence." *Narrative Impact*, Psychology Press, 2003, pp. 263–86.

Tedesco, Juan Carlos, et al. "The Curriculum Debate: Why It Is Important Today." *Prospects*, vol. 44, no. 4, 2014, pp. 527–46.

Vachuska, Karl. "Racial Segregation in Everyday Mobility Patterns: Disentangling the Effect of Travel Time." *Socius*, vol. 9, 2023, p. 23780231231169261.

Vermeule, Blakey. *Why Do We Care about Literary Characters?* Johns Hopkins University Press, 2010.

Walsh, Melanie, and Maria Antoniak. "The Goodreads 'Classics': A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism." *Journal of Cultural Analytics*, vol. 6, no. 2, Apr. 2021, https://doi.org/10.22148/001c.22221.

Wimmer, Andreas. "The Making and Unmaking of Ethnic Boundaries: A Multilevel Process Theory." *American Journal of Sociology*, vol. 113, no. 4, 2008, pp. 970–1022, https://doi.org/10.1086/522803.

Woloch, Alex. *The One vs. The Many: Minor Characters and the Space of the Protagonist in the Novel*. Princeton University Press, 2009.

National Center for Education Statistics, U.S. Department of Education, May 2023, https://nces.ed.gov/programs/coe/indicator/clb/free-or-reduced-price-lunch.

Appendices

A. Additional Author Findings

Only 7.1% of books (as mentioned in §8) involve authors writing about primary characters who differ from themselves. So, trends in the racial/ethnic composition of authors are similar to those for main characters. Figure 12 shows the proportion of AP books containing authors of color over time. We also include a breakdown of authors' racial/ethnic backgrounds across all unique books in our AP Literature (Figure 13) and teacher-provided samples (Figure 14).

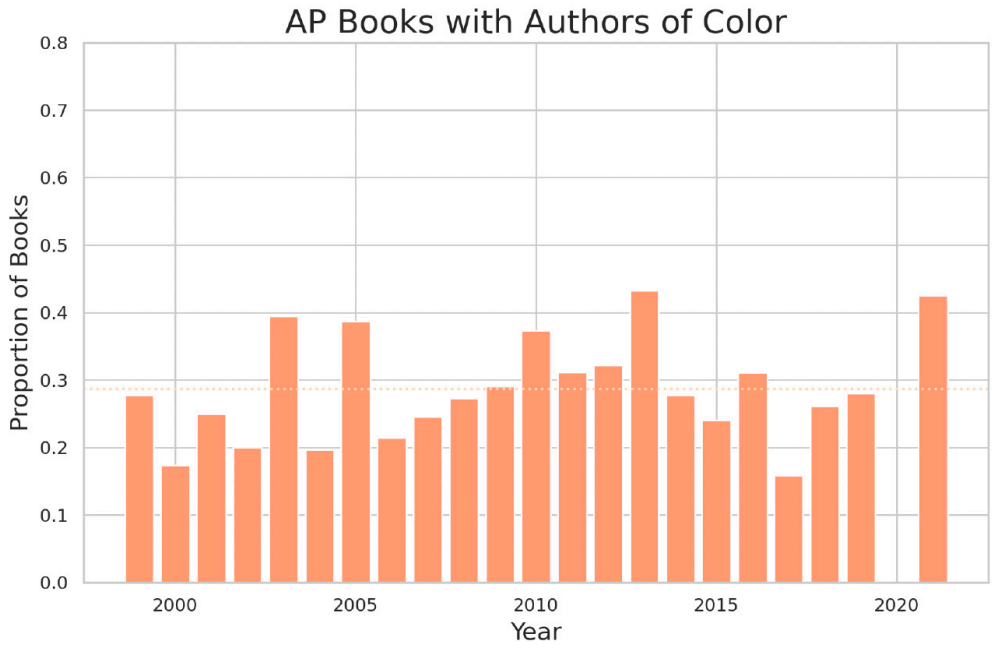

Figure 12. The top plot shows the proportion of AP books each year, out of all books listed that year, that are written by authors of color. The horizontal dotted line is the average across all years.

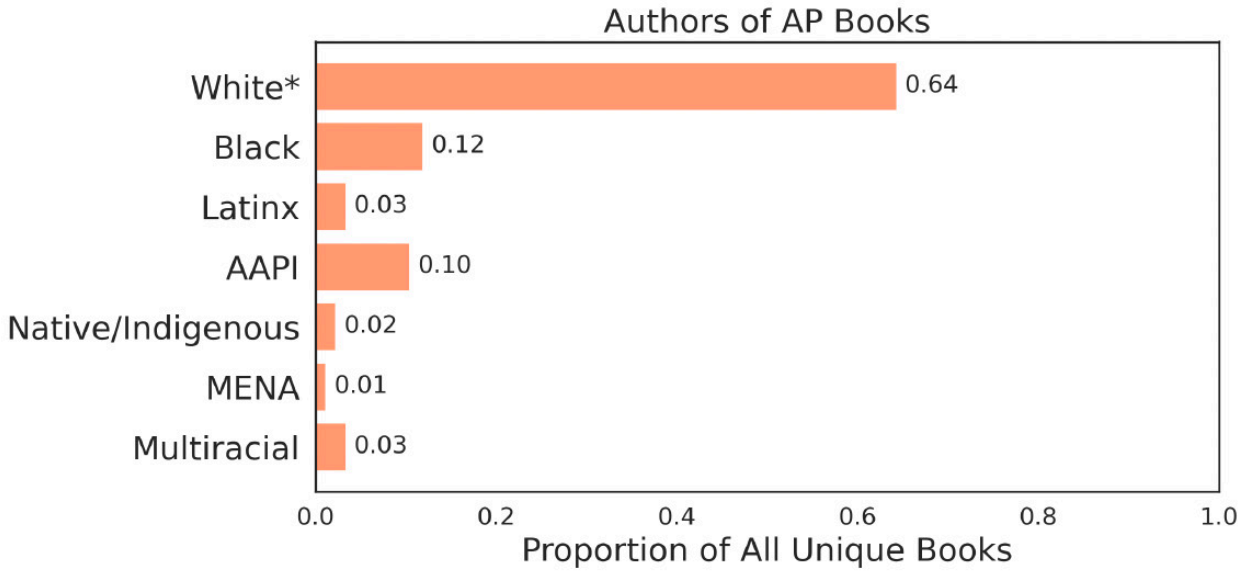

Figure 13. Authors in our AP sample. *Multiracial characters are multiply counted across categories, except White.

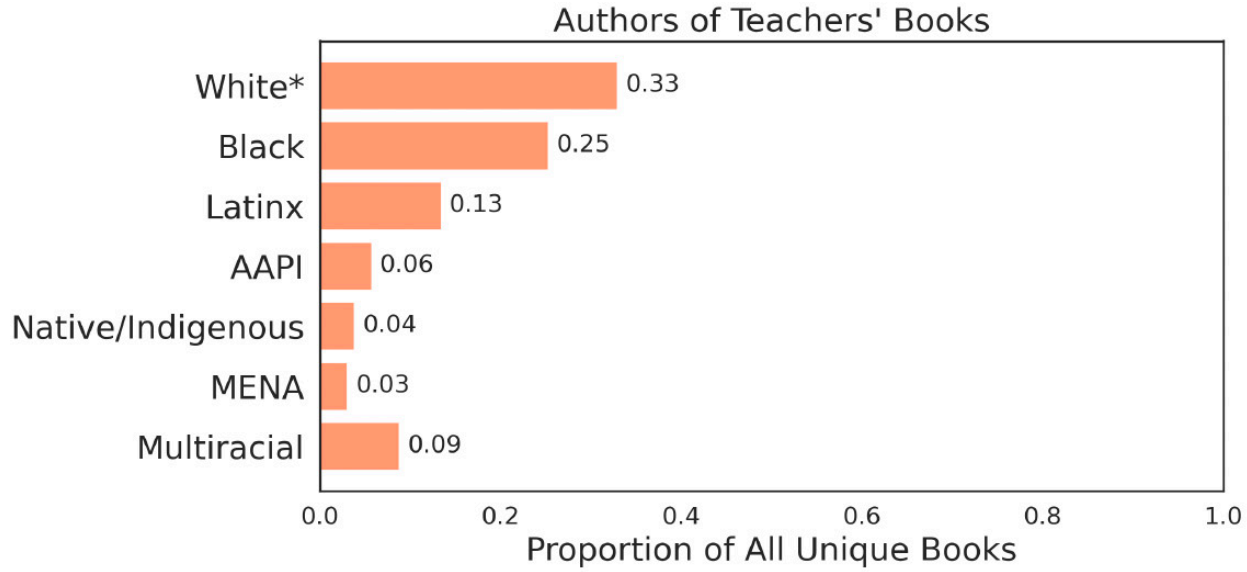

Figure 14. Authors in our teacher-provided sample. *Multiracial characters are multiply counted across categories, except White.

B. Additional Gender Findings

Our paper primarily focuses on race and ethnicity, but the main text also mentions several findings related to gender. Here, we include several figures that expand on those findings. In §6, we report a gender gap in books listed by AP Literature; Figure 15 illustrates this gap over time. Figures 16 and 17 illustrate recurring book titles in the AP Literature dataset and teacher-provided dataset, respectively, labeled with primary character gender. In §6 we also mention that among AP Literature titles that occur at least 8 times over the years, 9 out of 25 titles feature female primary characters, suggesting

a potential gender gap in the ELA canon. In the teacher-provided dataset (§7), 5 out of 19 books feature female primary characters, suggesting a gender gap there as well. For a more fine-grained breakdown of mention counts for each main character detected by our computational pipeline, readers should consult the data accompanying our paper.

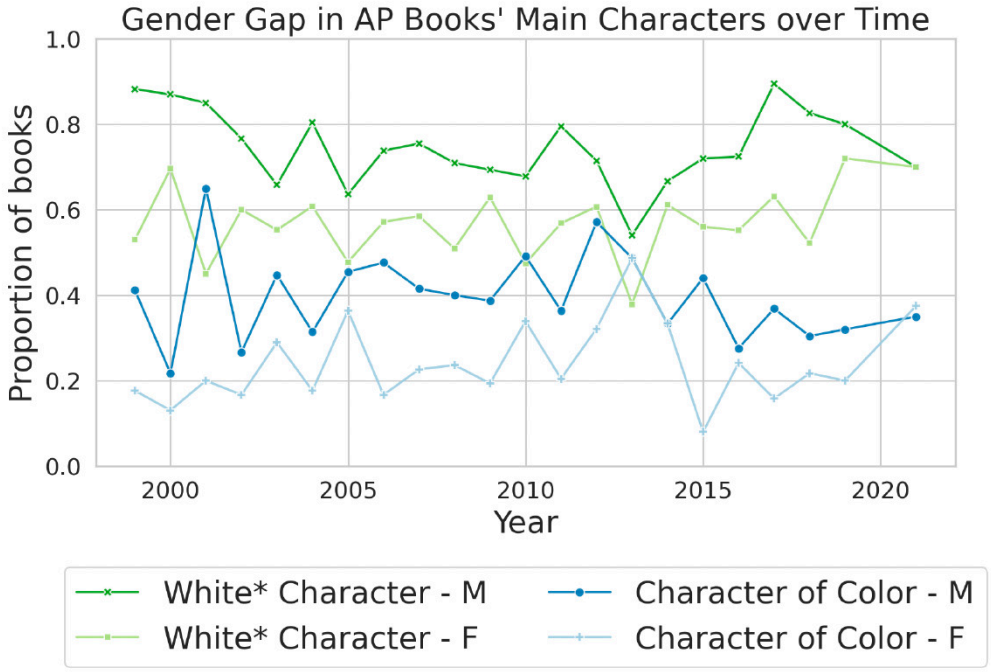

Figure 15. An intersectional look at the proportion of books containing female and male main characters in Advanced Placement (AP) Literature over twenty-two years.

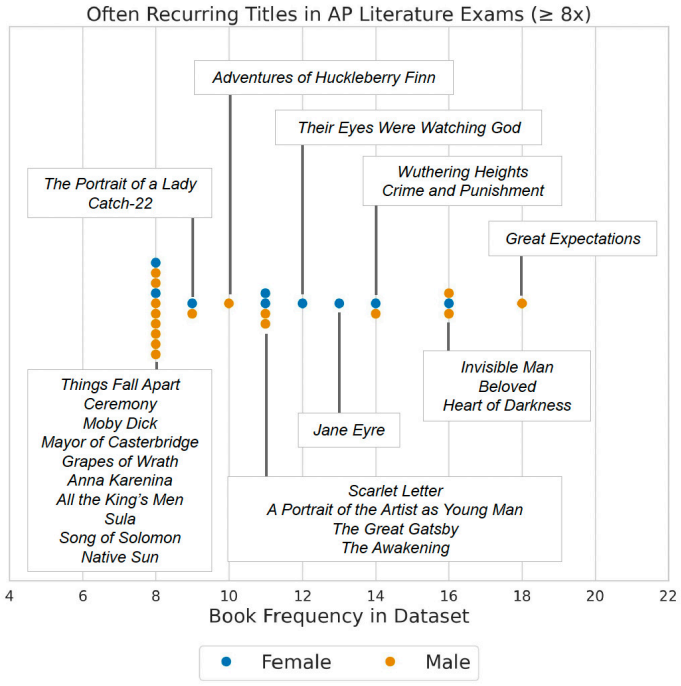

Figure 16. Titles of books that occur at least 8 times in AP Literature exams. This is a version of Figure 5 where titles are colored by primary character gender instead of their race/ethnicity.

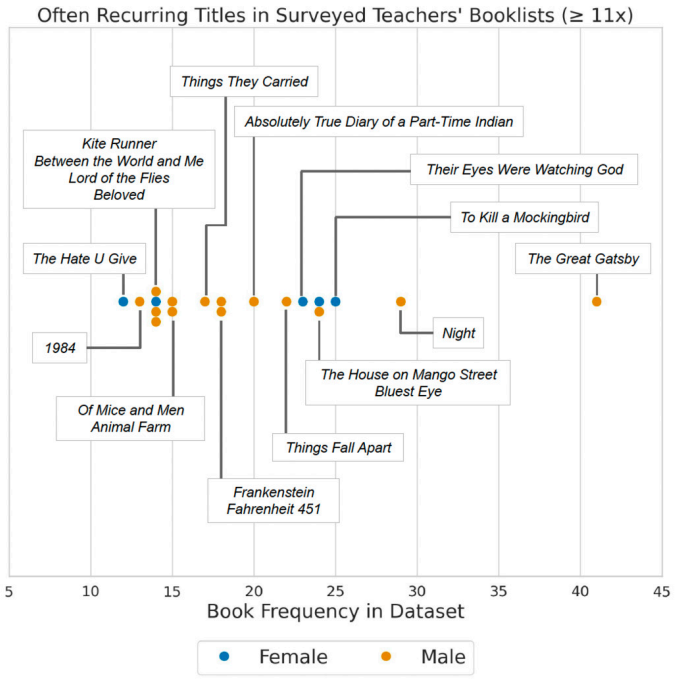

Figure 17. Titles of books that occur at least 11 times in teacher-provided lists. This is a version of Figure 9 where titles are colored by primary character gender instead of their race/ethnicity.