Influencing Identities: The Relationship between Creators and Character Representation in Children's Books*

Anjali Adukia, University of Chicago and NBER
Emileigh Harrison, University of Chicago
Jake Nicoll, University of Chicago
Xiling Celia Zhu, University of Maryland

Preliminary, active working **DRAFT**. Please do not cite or distribute.

February 19, 2025

Abstract

Authors and illustrators of children's books transmit values and instill beliefs in children through the messages their books convey and the characters they depict. We study the creators whose voices shape the narratives, stories, and ideas that we present to our children, and we ask how their identities influence the messages they choose to send and the content they produce. Using modern techniques from computer vision and natural language processing, we characterize the relationship between the race and gender identities of these creators and the identities of the characters they represent in influential, award-winning children's books: Mainstream awards that highlight books for their general literary value and Coretta Scott King (CSK) awards that highlight books that center the African American experience. We find that, conditional on the award won, books by Black creators depict darker-skinned characters than books by White creators. Moreover, Black creators recognized by the CSK award depict characters with darker skin than Black creators recognized by Mainstream awards. Additionally, we find that, independent of the award won, females are increasingly present in female-authored books over time, whereas their relative lack of presence in male-authored books remains somewhat stable across decades. In terms of portrayal, we detect strong stereotypical associations, with females being associated with familial roles and appearance, whereas males are associated with occupations in the labor force and positions of power. Surprisingly, these stereotypical portrayals are similarly strong in both male-authored and female-authored books, and these patterns remain consistent across time. Detailed purchase data reveal that, compared to White purchasers, Black purchasers exhibit stronger preferences for books that depict darker-skinned characters. Purchasers with at least one daughter exhibit stronger preferences for books by Female authors and books that have a higher percentage of female words. We find suggestive evidence that the purchases of books written by White authors correlate with more conservative views on racism and immigration.

*For helpful feedback, we thank seminar participants at Society for Research on Educational Effectiveness, the Symposium on Communication & Intelligence, the University of Chicago Computational Social Science Workshop, and the University of Chicago Data Science Institute. For financial support, we thank the Becker Friedman Institute at UChicago. The research reported here was also supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A200478 to the University of Chicago. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

I Introduction

For authors, historians, economists, innovators, educators, and reporters alike, their perspectives, values, and experiences may influence the content they create (Francis et al., 2022; Koning et al., 2021; Einiö et al., 2019; Akerlof, 2000). From books to newspapers and from educational curricula to blog posts, the content they produce encodes messages about the world and our position in it (Kearney and Levine, 2020; Cantoni et al., 2017). Their identities may therefore shape these messages and influence whom their knowledge and narratives are constructed for (Louie and Wilkes, 2018). Characterizing the relationship between the identities of creators and the content they create may therefore illuminate how narratives and knowledge are produced as well as demonstrate the importance of underrepresented voices in this production process.

In this paper, we investigate the role of identities of authors and illustrators (referred to as creators) in shaping the content of a collection of influential children's books. In particular, we describe the relationship between creator race and gender identities and the race and gender identities they represent in their stories. Through the use of tools from computer vision and natural language processing, we measure representation both in terms of presence, which is a quantitative accounting of how frequently different identities appear, and in terms of portrayal, which more qualitatively characterizes how these identities are depicted.

We apply these tools to measure representation in a century of influential children's books. We first examine books that received recognition from the Newbery or Caldecott awards. These books are four to five times more likely to be purchased and checked out from libraries (Adukia et al., 2023), and the receipt of these awards increases the likelihood of a book's usage in classrooms (Koss et al., 2018). We call these the "Mainstream" awards, and neither award explicitly considers identity, either of the book's creators or its characters. Second, we examine books that received recognition from the Coretta Scott King (CSK) Book Award, which highlights books that center the African American experience. Due to this award's selection criteria, we expect that the CSK collection provides an upper bound on representation of Black characters.

We use computational tools to characterize the relationship between creator identity and the identities they depict in the images and text of the books. We examine the presence of identities in images by classifying features of pictured characters (e.g. skin color, putative race, gender) and by measuring putative race and gender through textual mentions of characters, pronouns, and other related word counts. In order to study portrayal in text, we use word embeddings. This tool learns relationships between words that enables it to

effectively identify synonyms (Mikolov, Sutskever, Chen, Corrado, and Dean 2013; Mikolov, Chen, Corrado, and Dean 2013). This allows us to analyze which words emerge empirically as being implicitly synonymous with different identity groups. In ongoing work, we are also examining the presence and portrayal of topics using large language models.

We find that Black creators depict darker-skinned characters, even after conditioning on collection. Additionally, Black creators in the CSK collection depict darker-skinned characters than Black authors in the Mainstream collection. These results indicate that the founding of awards that explicitly highlight the experiences of underrepresented identities as well as the recognition of creators with those identities by existing awards both play simultaneous and distinct roles in influencing the content to which children are exposed.

Additionally, we find that female creators depict more female characters than male creators. In the text, they use relatively more female-centered words than male creators, and they depict more females in images. These patterns are consistent across our two collections. Moreover, we see that females have used female-centered words in increasing proportion over time, whereas books by male creators consistently use proportionally more male-centered words. We also examine the pattern of books featuring female characters in images more than in the text, which was identified in (Adukia et al., 2023). Despite the fact that female creators increase the presence of female characters in images and in the text relative to male creators, we find that, regardless of creator gender and collection, female characters are shown more than heard.

We also find patterns consistent with predictions by scholars of intersectionality when we consider race and gender simultaneously. In particular, we see that Black male creators increase the presence of Black characters in their books through the inclusion of more images of Black male characters. Similarly, White female creators increase the presence of female characters by including more images of White females. These patterns produce a unique site of exclusion of Black female characters and illustrate the importance of considering different dimensions of identity simultaneously (Crenshaw, 2013).

Moving from presence to portrayal, our analysis of word embeddings reveals clear stereotypical associations attached to males and females. We find that males are generally associated with occupations and positions of power, whereas females are typically associated with familial roles and appearance. These associations remain surprisingly stable across author gender. To further investigate this phenomenon, we examine how these associations change over time. Again, across author gender and collection, the same stereotypical associations emerge consistently in different time periods.

To complement this content analysis, we also study consumers' preferences for the different features we extract. Using detailed purchase-level data, we estimate a conditional logit model to assess how consumers value the presence of different race and gender identities, the identity of the author, and the award the book won. We also include parameters to assess how these preferences vary based on consumer identity. We find that purchasers with a daughter (compared to those without a daughter) place a significant premium on books written by females and on the presence of females in the text. We also find that, compared to White purchasers, Black purchasers place a significantly higher value on books with darker-skinned characters. These findings illustrate that the features that are emphasized by creators of different identities are also valued by individuals who share those identities.

Finally, we link our purchase data to the Cooperative Election Study (CCES), which contains ZIP code-level information about local attitudes and beliefs. This enables us to study how local-level beliefs correlate with the creators' identity. In particular, we find that, in ZIP codes in which residents hold relatively conservative positions on policy issues related to race and immigration, consumers are significantly more likely to purchase books by White authors.

We offer three important contributions. First, we characterize the relationship between creator identity and the identities they represent in a set of influential children's books. Previous studies that examine this relationship do so manually (Koss et al., 2017, 2018). While manual investigations can uncover specific nuances, they are relatively limited in scope, both in terms of the number of books they can evaluate and the number of features of each book they can study. Second, we apply recent innovations in the use of word embeddings to produce automated assessments of qualitative features of representation (Charlesworth et al., 2022). Moreover, we combine this approach with a lexicon of human-labeled valence, arousal, and dominance scores that is larger and more reliable than many existing lexicons (Mohammad, 2018). This allows us to supplement our discussion of the stereotypes that we detect by quantifying the differences between words associated with different identities. Third, we illustrate how children's book creators produce content for people who share their identities. In particular, we see that female creators use a higher percentage of female words than male creators, and purchasers with daughters (relative to those without daughters) prefer books with a higher percentage of female words. Similarly, Black authors depict darker-skinned characters than White authors, and Black purchasers (relative to White purchasers) prefer books with darker-skinned characters. These findings demonstrate the importance of representation - both in the content of books and in the collection of authors who are recognized by these awards - in ensuring that children have access to the content they can see themselves

portrayed in.

The rest of the article proceeds as follows. In Section II, we examine the role of creators in shaping messages and how this matters for the content of children's books. In Section III, we describe the books in our sample, our data on purchasers, and the other datasets we use throughout the article. In Section IV, we explain the computer vision and natural language processing techniques we deploy as well as the models we use to study consumer purchasing patterns. In Section V, we discuss our results. In Section VI, we summarize and conclude.

II Background

II.A The Role of Creators

In several settings, scholars have documented a relationship between the identities of innovators and the identities of people their innovations are meant to benefit. For example, medical students have been found to underdiagnose skin conditions for Black individuals (Fenton et al., 2020). This is plausibly linked to the fact that Black individuals are underrepresented in research articles and textbooks (Louie and Wilkes, 2018). Moreover, Einiö et al. (2019) find a correlation between gender, age and socioeconomic status of inventors and the people their innovations benefit. Additionally, Koning et al. (2021) find that female researchers are more likely to research diseases that impact females more than males. In economics, Francis et al. (2022) trace themes of Black economists' research, and they identify a focus on economic mobility and educational opportunity for Black Americans. Just as researchers demonstrably guide knowledge production and innovation in directions that benefit people who match their identities, children's book creators may similarly produce content that is tailored for children who match their own racial or gender identities.

These patterns could be explained by individuals' preferences to write about or illustrate content for someone who shares their identity (Akerlof, 2000). In addition, authors and illustrators may bring their own biased perceptions of identity groups into their portrayal of characters. This is analogous to what Dee (2005) finds in a classroom setting, where racial, ethnic, and gender dynamics between students and teachers have large effects on teachers' perceptions of students. This could lead authors and illustrators to depict more characters of their own identity group (and portray them in a more favorable light) than characters from groups they do not identify with.

These patterns are particularly important given the role that children's books and the messages they carry can play in children's lives. Children (and adults) absorb messages implicit in passively consumed media, and these messages can influence attitudes and preferences (Kearney and Levine, 2020; Jensen and Oster, 2009). Moreover, children's literature

used in middle school language arts and social science curricula has been shown to shape beliefs about one's self, their community, and civic action (Levstik and Tyson, 2010). Additionally, a number of psychosocial interventions demonstrate the effectiveness of video interventions that show the explicit success of individuals of a certain identity on the outcomes of individuals who share that identity (Bossuroy et al., 2022; Riley, 2024). The latter study demonstrated that girls who saw a film featuring a female role model performed better on subsequent exams and even became more likely to remain in school. The swapping of one book that depicts the experiences of a character whose identity matches that of the reader with one that does not is a light-touch intervention that parents may be unwittingly carrying out repeatedly at scale. The fact that one-time exposures to the stories of characters who share a kid's identity could have significant long-term impacts illustrates the necessity to investigate the messages constantly being sent to kids through the books they consume.

Moreover, children pick up on these influences very early on in life. Bian et al. (2017) show that, as early as the age of six, girls are less likely than boys to believe that members of their gender are "really, really smart," and they avoid activities believed to be for children who are "really, really smart." These early beliefs may result in the early formation of young girls' preferences for certain fields of study or societal domains, potentially resulting in important long-term consequences (Bian et al., 2017). These early beliefs can create a self-fulfilling prophecy. Eble and Hu (2020) illustrate that children who believe that they have low academic ability may choose to invest relatively little effort in their education. This results in low educational attainment, confirming these children's beliefs and resulting in low effort being exerted in subsequent grade levels. In this way, small discrepancies in children's beliefs about themselves at a young age can have massive long-term consequences. The fact that children can acquire these beliefs at such a young age and from implicit messages particularly motivates our investigation into representation in children's literature.

II.B Complementing and Extending Manual Content Analysis

Content analysis involves the systematic study of the content of books, including the representations of different identity groups in those books. Traditional content analysis studies have involved teams of researchers manually sifting through pages of content to identify the characters being depicted as well as to look for themes embedded in narratives and images (Neuendorf, 2017; Krippendorff, 2018). Weitzman et al. (1972) is a landmark study of this type. The author finds that girls are greatly underrepresented in lead roles and illustrations in prize-winning picture books. Additionally, when women are present in these books, they occupy traditional gender roles. They are confined to their roles as wives and mothers, while the men in these stories engage in a variety of occupations. In a later

study, Williams Jr et al. (1987) find that, even when female characters are depicted with greater frequency, they typically remain confined to traditional gender roles. In addition to illustrating traditional gender roles, Koss (2015) finds that picture books also tend to overrepresent White characters.

Some manual content analyses have studied the influence of creator identity on representation in children's books. Koss et al. (2018) find that, in books winning or honored with the Caldecott medal, the authors, illustrators, and main characters are all predominantly White, and this pattern is persistent in every decade from the inception of the award to 2017. When the racially diverse main characters do appear in a book, the authenticity of their voice might be absent, as the majority of those characters are created by individuals who do not come from the same racial groups as the characters. Similarly, Koss et al. (2017) show that most of the Latinx authors and illustrators did not create Latinx main characters, while many of the Latinx main characters were created by authors and illustrators from other cultures. Focusing on the literary experience of Black males, Davis et al. (2021) find that books with Black males as the primary character(s) account for only 15.7% of the Caldecott-recognized books in their study. When Black males do appear as the primary characters, their experiences are limited to the challenges of urban settings or stories about Black Americans during times of slavery and the civil rights movement. Among the books with Black males as the primary character(s), only 27.8% had Black male authors. They also find that Black men tended to collaborate with one another to create books that featured Black male protagonists, which may be driven by a shared concern of the under-representation of Black males in children's literature.

The strength of traditional approaches to content analysis is that researchers can identify complex themes and messages being communicated in images and the narratives of the text. This helps researchers measure the portrayal dimension of representation in addition to the presence dimension. However, the obvious drawback of these traditional approaches is that they are extremely intensive in terms of time and labor. This severely limits the potential scope of these studies and expands the potential for human errors in the analysis.

Computational content analysis, which involves leveraging AI to quantify representation in books, overcomes this limitation. With our framework, we can easily, cheaply, and rapidly process content in large collections of books. Additionally, our methods translate flexibly to different datasets. With advances in techniques used to process text, we also have access to tools which help measure representation along the portrayal dimension, giving our analyses a richness that can compete with that of traditional content analysis. Our word embedding models use the entirety of the text to determine how characters are depicted rather than

selectively procure subsections of the text that speak to different aspects of their portrayal. Our tools therefore enable us to complement and extend previous manual content analyses by expanding the studies from Caldecott winners to Newbery and CSK Book Award winners as well. Additionally, we are able to quantify a variety of features in these books and leverage the substantial amount of textual data in the latter two collections.

III Data

Corpus We focus on children's books that are awarded or honored with the Caldecott Medal, the John Newbery Medal, or the CSK Book Award. The Newbery Medal, first awarded in 1922, is given to authors of books that are considered to be the "most distinguished contribution to American literature for children." The Caldecott Medal, first awarded in 1938, is given to illustrators of "the most distinguished American picture books for children." The CSK Book Award, established in 1969, recognizes books that are created by and center the experiences of African Americans. We chose books recognized by these awards for their importance and prominence in children's lives. First, books recognized with Newbery or Caldecott awards are of greater social influence compared to non-recognized books. Adukia et al. (2023) shows that the receipt of these awards increases both purchasing and library checkouts of the books by four to five times. Moreover, according to Koss et al. (2018), the receipt of these awards increases the likelihood of a book's usage in classrooms. Therefore, books recognized by these awards form a critical part of children's learning experience, highlighting the importance of understanding the role creator identity plays in shaping the content of this particular corpus.

Identity of authors and illustrators We manually labeled the race and gender of authors and illustrators of the books in our sample. For books with multiple authors (illustrators), if the authors (illustrators) are of the same identity group, we label the aggregated author (illustrator) as having this shared identity. If the authors (illustrators) are of different identity groups, we label the aggregated author (illustrator) identity as "Mixed." In practice, there were relatively few books with multiple authors (illustrators) who have different identities, so we exclude these observations from our sample. Table 1 displays counts of the books in our sample, subdivided by collection, creator race, and creator gender. In performing our analyses, we will assess variation in our metrics of representation along these three dimensions (or some subset of them).

We recognize several limitations with this manual labeling process. First, we acknowledge that assigning a putative race to authors and illustrators based on phenotypic presentation does not necessarily capture their actual racial identities. Second, we note that racial categories are socially constructed (Roberts, 2011; Logan, 2022). As such, our own assignment of

racial labels to these creators reflects our own socialization rather than a latent attribute of the creators. Nevertheless, we expect that our manually assigned racial identity labels do not frequently differ from the racial identities that the authors and illustrators would report. We therefore view this limitation as a form of measurement error that does not fundamentally undermine our results.

Famous Figures Mentioned in the Text As discussed in Section IV.A, we apply Named Entity Recognition to extract mentions of identifiable characters in the text. To determine which of these characters refer to real and prominent historical figures, we consider characters with (at least) a first and last name, and we match them to the Pantheon 2.0 dataset. These data contain information from over 70,000 Wikipedia biographies (Yu et al., 2016). They include entries for characters with biographies in at least 15 language editions of Wikipedia, and we designate a figure as "famous" if they are included in this dataset. These data also contain information on each individual, including their gender but not including their race. We manually code the race for individuals appearing in our corpus, noting the same caveats already discussed.

Sentiment Analysis: Measures of Valence, Arousal, and Dominance As detailed in Section IV.A we construct word embeddings in order to measure the relative strengths of implicit associations assigned to different identities. When measuring these associations, we search over all words in the National Research Council Canada valence, arousal, and dominance (NRC VAD) Lexicon. This lexicon features reliable human ratings of valence, arousal, and dominance scores for more than 20,000 words (Mohammad, 2018). Valence captures the positivity and negativity of the word, arousal captures its activity or passivity, and dominance captures its dominance or submissiveness. Words are assigned scores between 0 and 1, with 1 being the most positive, active, and dominant scores of valence, arousal, and dominance respectively. These three traits emerge from several prominent factor analyses (Osgood, 1957; Russell, 1980), which indicates that they encode three important dimensions of a word's meaning (Bakker et al., 2014). Therefore, they provide us with metrics to quantitatively assess discrepancies in the implicit associations assigned to different identities. Moreover, the large size of the NRC VAD Lexicon allows us to examine implicit associations in this way without meaningfully truncating the number of associations we can test.

Book Purchase Data We also draw data from the Numerator OmniPanel, a large dataset which includes information from over 1 billion shopping trips from over 44,000 retailers between 2017 and 2020. Each purchase is matched to detailed information on consumers, including the purchaser's demographics and number and gender of their children. We subset these data to only include purchases of children's books. We manually link these books

recorded in the Numerator OmniPanel to books in our corpus based on the title and, if available, publisher and copyright year.

CCES Data To further understand consumption patterns, we link the Numerator data to survey data from the Cooperative Election Study (CCES). This survey of a nationally representative, stratified sample contains ZIP code-level information about local attitudes and beliefs. We use the ZIP code to link the two datasets. The CCES data allow us to measure how purchaser preferences for books with different features and written by authors of different identities covary with local political leanings.

IV Methods

IV.A Feature Construction

Our raw data consist of a scan of every page of every book in our corpus. In order to make use of these data, we begin by parsing these scans for images and text. We follow the methodology of Adukia et al. (2023). We include a brief overview of the data pipelines and detail the features we extract. In Section V, we present a descriptive analysis of how these features vary based on the collection as well as creator race and gender.

Image as data To study images in our books, we first detect faces in images using a fine-tuned version of Google's AutoML Vision model. The fine tuning procedure helps our model perform better at detecting faces in illustrations since the original model was trained mainly on photographs. After isolating faces in the books, we perform two tasks to extract two sets of features.

1. *Perceptual Tint:* First, we construct an aggregate measure of the representative skin color of the detected face. The first step of this procedure is to isolate the skin within the detected face. To do this, we apply a Fully Connected Convolutional Neural Network Continuous Conditional Random Field (FC-CNN CRF). This model detects important landmarks, such as the edges of the face, the nose, the eyes, and the mouth. It then uses these landmarks to extract the parts of the fact that are more likely to be skin. After isolating the skin on the face, we apply k-means clustering to group the pixels into distinct clusters in RGB space. We choose $k = 5$, and we drop the two smallest clusters. (These tend to represent shadows or non-skin features.) We then take the centroids of these three clusters, and we map these to L*a*b* space. Finally, we construct a weighted average of these three clusters, with weights given by the proportion of pixels coming from the cluster corresponding to each centroid. This weighted average provides a representative skin color for the face. We define the perceptual tint of the character to be the value of L*, which provides a 0-100 scale for

lightness in which 0 is the darkest and 100 is the lightest.

2. *Predicted Race and Gender:* In addition to constructing this measure of perceptual tint, we also pass the detected faces through a classifier to predict the putative race, gender, and age of the depicted character. This model estimates probabilities that a given depicted character is of each possible putative race and gender category. It classifies characters by choosing the labels that are most probable. We note that models which attempt to classify individuals by race based on images may suffer from a high error rate (Fu et al., 2014; Krishnan et al., 2020). Additionally, as discussed previously, we acknowledge the conceptual challenge of identifying racial identities from images. Nonetheless, we do not expect these limitations to systematically affect our results.

Text as data We begin by performing optical character recognition to convert our raw scans into text data. This procedure finds text in the images and converts it into ASCII. After performing this step, we construct several features from our text.

1. *Token Counts:* First, we identify all words with a gender association (*gendered words*). These words include gendered tokens such as titles, pronouns, or specific terms such as queen and husband. We also include names with a gender association (*gendered names*), which we automatically predict using characters' first names. After obtaining token counts for each of these gendered words, we compute the book-level percentage of gendered terms associated with females. This measure provides a naive quantification of the relative presence of males and females in our books.

2. *Famous Characters:* We also measure the presence of different identities among famous characters mentioned in the books. To find mentions of famous individuals, we perform Named Entity Recognition (NER) on the text. We then extract each entity that the model tags as a person. We match the entities to the Pantheon 2.0 dataset, and, as discussed in Section III, we designate a character as famous if and only if they appear in that dataset. With a combination of pre-labeled information from these data and additional manual annotation, we obtain the race and gender identities of the famous individuals. Finally, we construct the percentages of these famous individuals from each race, gender, or race-gender identity.

3. *Word Embeddings:* Word embeddings represent words as numerical vectors in a latent space while ensuring that words that are more "similar" to each other are "closer" in this space. This closeness is measured by their cosine similarity, which is the cosine of the angle between their corresponding vectors. Word embeddings reflect the context in which words appear, and their meaningfulness and interpretability rely on the

distributional hypothesis by Firth (1957) and Harris (1954). The distributional hypothesis asserts that the context surrounding a word provides a good representation of the meaning of that word.

Mikolov, Sutskever, Chen, Corrado, and Dean (2013) show that embedding vectors learned through word2vec capture a large number of precise syntactic and semantic word relationships. Mikolov, Sutskever, Chen, Corrado, and Dean (2013) evaluate word2vec's command of the syntactic and semantic relationships by analogy tests. For example, the semantic analogy test "Germany" : "Berlin" :: "France" : ? is solved by finding a vector corresponding to a word x such that $vec(x)$ is closest to $vec(\text{"Berlin"}) - vec(\text{"Germany"}) + vec(\text{"France"})$. This test is considered to have been answered correctly if x is "Paris". Mikolov, Sutskever, Chen, Corrado, and Dean (2013) shows that the word embedding trained by word2vec has a total accuracy of no less than 60%.[1]

(a) *Construction of Word Embeddings.* To find vector representations that capture the similarity between words in our corpus, we use word2vec, developed by Mikolov, Sutskever, Chen, Corrado, and Dean (2013) and Mikolov, Chen, Corrado, and Dean (2013). word2vec is a family of two algorithms: continuous bag-of-words (CBOW) and continuous skip-gram. We briefly sketch the idea of the skip-gram approach, which we apply in this paper.

Given a word in a text, the aim of skip-gram is to predict words surrounding the given word within a window of a certain radius r. That is, given a center word w_t and its context, which are the words that surround w_t, we index the words by their position in the context. We let the probability of those words appearing in the context be given by $\Pi_{-r\leq j\leq r, j\neq 0} p(w_{t+j}|w_t;\theta)$, where θ is a vector containing the parameters in the word embedding model. Then, given a sequence of center words w_t for $t = 1, 2, ..., T$, we minimize the objective function

$$J(\theta) = -\sum_{t=1}^{T} \sum_{-r\leq j\leq r, j\neq 0} \log\, p(w_{t+j}|w_t;\theta). \tag{1}$$

[1]Total accuracy is the average of the semantic analogy test accuracy rate and the syntactic analogy test accuracy rate. The syntactic analogy test "quick" : "quickly":: "slow" : ? is solved in the same way as the semantic analogy test, and it is considered to have been answered correctly if x is "slowly."

The basic Skip-gram formulation of the probability $p(w_{t'}|w_t)$ is softmax

$$p(w_{t'}|w_t;\theta) = \frac{exp(v_{w_{t',O}} \cdot v_{w_{t,I}})}{\sum_{s=1}^{W} exp(v_{w_{s,O}} \cdot v_{w_{t,I}})}, \tag{2}$$

where $v_{w_{s,O}}$ and $v_{w_{s,I}}$ represent the "output" and "input" vectors associated with the word w_s, and W is the number of words in the vocabulary. These vectors constitute the parameters θ to be learned.

(b) *Application of word embeddings.* We use word embeddings to uncover implicit associations attached to different identities and assess how these implicit associations vary with creator identity. We first combine texts into different subsets based on collection and author identity, and we train our word embeddings separately on each subset. We then specify group words associated with different identities. For example, we use "she," "her," and "hers" to represent females and "he," "him," and "his" to represent males. With the embedding vectors, we measure the association between sets of group words and comparison words.

In the literature, there are two approaches to specifying comparison words in order to uncover implicit associations. One commonly used approach requires a researcher to pre-specify a set of target words, a priori, to be a set of words that are representative of some domain (Garg et al., 2017; Caliskan et al., 2017; Jones et al., 2019). For example, "Chemist", "Engineer", and "Mathematician" are some words that represent the domain of "STEM career". The choices of domain words are at the discretion of the researcher, and the association between a domain and a group might be sensitive to such choices. In an alternative approach taken more recently, target domains are not pre-specified by the researcher (Charlesworth et al., 2022). We follow this approach to capture the associations that emerge empirically between identity groups and all possible words appearing in the NRC VAD Lexicon.

In particular, given a collection of group-word embedding vectors $\mathcal{G}$ and another embedding vector w, we define the Mean Average Cosine Similarity (MAC) as

$$\text{MAC}(\mathcal{G}, w) = \frac{1}{|\mathcal{G}|} \sum_{v \in \mathcal{G}} \cos(v, w),$$

where

$$\cos(v, w) = \frac{v \cdot w}{\|v\|\|w\|}.$$

Next, given two collections of group-word vectors $\mathcal{G}_1$ and $\mathcal{G}_2$, and given another vector w, the *centeredness of* w *on* $\mathcal{G}_1$ is given by

$$\text{MAC}(\mathcal{G}_1, w) - \text{MAC}(\mathcal{G}_2, w).$$

The MAC provides a measure of how associated a given word is with an identity group. Our measure of centeredness then quantifies a word's simultaneous association with one group and disassociation with another. When we consider male and female group words in Section V, we will compute the words most centered on males as well as the words most centered on females.

IV.B Measuring Consumer Preferences and Beliefs

We investigate consumers' expressed preferences for different features of the books as well as for books written by authors of different identities. To analyze these preferences, we apply a conditional logit model. Let u_{ij} denote the utility of consumer i obtained through the purchase of book j. We model u_{ij} as

$$u_{ij} = \beta_{0j} + X_{ij}\beta + \epsilon_{ij},$$

where ϵ_{ij} follows the type I extreme value distribution. We suppose that individual i chooses alternative j if and only if $u_{ij} > u_{ij'}$ for all $j' \neq j$. Then, we have that

$$\mathbb{P}(u_{ij} > u_{ij'} \ \forall \ j' \neq j) = \frac{\exp(\beta_{0j} + X_{ij}\beta)}{\sum_k \exp(\beta_{0k} + X_{ij}\beta)}.$$

In our model, we include the book-level percentage of female words, the book-level mean perceptual tint of detected faces, an indicator for whether the author is Black, an indicator for whether the author is female, an indicator for whether the book is from the CSK collection, and the following interactions: interactions of an indicator for whether the purchaser is female with the book-level percentage of female words and with an indicator for whether the book has a female author, interactions of an indicator for whether the purchaser has a daughter with the book-level percentage of female words and with an indicator for whether the book has a female author, and interactions of an indicator for whether the purchaser is Black with the book-level mean perceptual tint, with an indicator for whether the book has a Black author, and with an indicator for whether the book is in the CSK collection.

We filter the Numerator data for purchases of books in our corpus. We also filter for observations where the gender of the purchaser is either male or female, and their race is

either White or Black. We exclude shopping trips where the purchased books contained 0 gendered words. For books with no images, we imputed the mean perceptual tint to be the average of the mean perceptual tint of all other books in our corpus. To maintain the assumption of mutually exclusive options of a discrete choice model, we removed trips where multiple books were purchased. This leaves us with 1497 trips.

The choice set is defined as all books in our corpus that are published in or before the year a trip was made. For a trip made in 2017, there are 617 books in the choice set. For a trip made in 2018, 2019, or 2020, there are 619 books in the choice set. However, for a trip to a brick-and-mortar store, it is unrealistic to assume the consumer has access to all books in our corpus. As a robustness check, we remove in-store shopping trips and run the model with 1381 shopping trips made online.

We also run regressions to understand the relationship between consumer's political beliefs and the identities of the authors whose books they purchase. To do this, we link our Numerator data to our CCES data at the ZIP code level. We then analyze the relationship between ZIP-code-level purchase data and ZIP-code-level consumer beliefs. In particular, we run regressions of the form

$$Y_i = X_i\beta + \varepsilon_i, \tag{3}$$

where Y_i is the proportion of individuals in ZIP code i who agree with a given statement, and X_i is a feature of the ZIP-code-level purchase data, such as the percentage of purchased books written by White authors.

V Results

V.A Influencing Presence

We begin by analyzing how the identities of book creators influence the presence of different identities in the text. We only display trends based on author identities. Interestingly, the influences of illustrator identities are qualitatively the same, even when we consider data derived from images. This seems to reflect the fact that a book's illustrator either is the author or shares an identity with the author.

As discussed in Section III, we display variation in our metrics of representation based on collection, author race, and author gender. This will allow us to separate the influence of the author's identity on the characters they depict from the selection process that leads to the book appearing in our sample. In several instances, we only show results disaggregated by author race or by author gender when there is little variation to be seen along the other

dimension of identity.

V.A.1 Presence in Images

We begin by analyzing the distributions of perceptual tints depicted in books, disaggregated by author race and collection. The full distributions are depicted in Figure 1. We note that further disaggregating by gender reveals that, conditional on author race and collection, the distributions are similar across author genders. Using the Kolmogorov-Smirnov test, we see that all three distributions are pairwise statistically significantly different at the 0.05 level. This remains true after using the Benjamini-Hochberg procedure to control the false discovery rate. We see that depictions in books by White Mainstream authors are heavily skewed towards lighter perceptual tints, whereas depictions in books by Black CSK authors are conversely skewed towards darker perceptual tints. However, we see that Black Mainstream authors depict characters with perceptual tints between those depicted by the authors in the other subcollections. This indicates that, even conditional on the selection process associated with receiving a Mainstream award, Black authors depict darker characters than White authors. However, we also see that conditional on the race of the author, the depictions of characters across the two collections are significantly different. These results indicate that both the founding of awards that explicitly celebrate underrepresented identities and the presence of creators from those identities among winners of existing awards simultaneously play significant roles in shaping the content that children are exposed to.

Next, we investigate how these distributions have changed over time. To do this, we group skin tints into terciles of the full distribution and examine the number of characters in different terciles that the authors depict. In Figure 2, we plot the percentage of illustrated characters in each tercile by decade, again disaggregated by collection and the race of the authors. We note that the proportion of characters in each tercile among books by White Mainstream authors has remained somewhat stable over time, although a slightly higher proportion of characters in the darkest tercile have begun to appear. On the other hand, we see that Black CSK authors depict many more characters in the darkest tercile, and that this proportion has been increasing over time. In contrast, Black Mainstream authors have consistently depicted characters in the middle tercile, illustrating that the pattern discovered in Figure 1 has remained relatively stable over time. Given the increase in calls for representation and the founding of new awards, this lack of movement may be surprising.

In Figure 3, we display the shares of characters predicted to be in each race-gender gender category. We disaggregate these shares by collection, author race, and author gender. We see that Black Mainstream authors are about three times more likely to depict Black characters than White Mainstream authors, and Black CSK authors are about two to three times as

likely to depict Black characters as Black Mainstream authors.

We also observe that our results corroborate classic patterns predicted by scholars of intersectionality. We note that, conditional on depicting a Black character, Black male CSK authors are about three times more likely to depict a male than a female, whereas Black female CSK authors are only 1.5 times as likely to do so. A similar pattern holds for Black Mainstream authors. On the other hand, while White female Mainstream authors depict females in about 51.4% of images compared to White male Mainstream authors, who depict them in about 41.3% of images, most of this discrepancy is driven by the fact that White female Mainstream authors depict many more White female characters than White male Mainstream authors. These results suggest that Black male authors increase the representation of Black characters by including more Black male characters and, correspondingly, White female authors increase the representation of female characters by including more White female characters. These patterns leave Black female characters systematically excluded.

Another striking feature of Figure 3 is that, among all groups of authors, over 50% of images are of White characters. This is especially surprising given that we would expect Black authors to be more likely to center Black characters in their books. Moreover, since the CSK selection committee explicitly values portrayals of the Black experience, we expect these authors to further center Black characters. Given the difficulty of the task of predicting putative race and gender from a face, we do expect there to be some measurement error. Nonetheless, even if our model overstates the presence of White characters in images, we do not expect this direction of mislabeling to vary across author identity groups. Therefore, the conclusions of our previous discussion, in which we conditioned on a predicted race-gender identity and compared across author identities, are still credible.

V.A.2 Presence in Text

In order to further assess the depiction of different races, we turn to an analysis of their presence in the text. We begin with an investigation into the presence of famous individuals. Although these famous individuals only constitute a share of the characters in our sample, tracing their mentions allows us to confidently predict when the text is actually referencing characters of different identity groups. It also allows us to circumvent the challenge of programatically predicting putative races based on images.

In Figure 4, we show the proportion of famous characters of different predicted races mentioned in the text by authors, fully disaggregated by collection, author race, and author gender. Additionally, we display each race's relative share of the U.S. population according to

the Census. We see that almost all famous characters mentioned are either White or Black. White Mainstream authors mention famous White characters disproportionately frequently compared to their share of the U.S. population. Moreover, this level of depiction has remained relatively consistent since 1970, even as the share of the U.S. population that is White has decreased over this time period. Black authors reference far more Black famous individuals. Interestingly, we see that from roughly 1960-1980, Black Mainstream authors referenced a similar proportion of White famous individuals as White Mainstream authors. However, that proportion quickly declined in subsequent decades to match the proportion featured in books by Black CSK authors, which has remained relatively constant since the establishment of the award. We also note that, conditional on author race and collection, authors of different genders mention famous characters of different races in similar proportions.[2]

Although our comparison of the Mainstream and CSK collections makes questions about race particularly salient, we also investigate how creators' gender correlates with their depictions of different gender identities. We began some of this analysis with our presentation of Figure 3. In this Section, we will more explicitly investigate the depiction of gender in our sample of books.

In Figure 5, we display the book-level percentages of female words appearing in the text, broken down by the gender of the authors and their collections. We group texts by author gender, collection, and decade, and we display the average of book-level percentages of female words in these groups. We see that female authors consistently use a greater percentage of female words than male authors across all decades. Nonetheless, their usage of female words was still below the proportion of females in the U.S. census in each decade. Moreover, we note that, conditional on gender, the proportion of female words used by authors is roughly consistent across collections.

In Figure 6, we display the densities of proportions of female words across all books, broken down by author gender and collection. We plot separate densities for different time periods. We see that male-authored books frequently use a relatively low percentage of female words, whereas the distributions of proportions by female authors are more symmetric about 50%. We also note that, conditional on gender, there is again little variation across collections.

Interestingly, we see that the time dynamics of these distributions vary substantially by

[2]We note that these proportions of detected identities do not logically contradict the findings displayed in Figure 3. It may be that the distribution of identities depicted in images differs from the distribution of identities discussed in the text. Additionally, the distribution of identities of famous characters could differ from the distribution of identities among all characters. In particular, some authors may center Black famous characters and yet continue to depict many surrounding and supporting characters as White.

author gender. The skew towards male words in male-authored books has remained relatively consistent over the decades. In contrast, the shapes of the distributions for female-authored books have changed substantially. Over time, female-authored books have shifted from a distribution skewed towards male words to a distribution that reflects approximately equal use of female and male words. This may reflect an increasing intention of female authors to create books for female readers. It could also reflect changes in award selection criteria that provide female authors with more flexibility to depict the characters they have in mind.

Next, we combine our text and image data to quantify presence in both of these domains. In Figure 7, we plot the proportion of female faces depicted in books against the proportion of female words among all gendered words. Each point represents average percentages for books in a given decade. We see that female-authored books have higher proportions of female faces and higher proportions of female words than male-authored books. We do not see substantive differences by collection after conditioning on the gender of the author. However, consistent with the patterns found in Adukia et al. (2023), we see that proportions of female faces are consistently higher than proportions of female words. Moreover, female-authored books still follow this same pattern, even though they depict a higher proportion of female faces and use a higher proportion of female words. This means that, across all groups of authors, female characters are more frequently visualized than heard.

V.B Influencing Portrayal

We now shift our attention from whether different identities are represented to how they are depicted. As discussed in Section IV, we apply word embeddings to detect implicit associations that emerge empirically as being attached to different identity groups. In particular, we find the words that are most centered on males and the words that are most centered on females. We disaggregate our books by author gender and collection to detect variation in these portrayals along these dimensions. As noted in Section IV when we described our centeredness metric, the list of female group words is "she," "her," and "hers," and the list of male group words we use is "he," "him," and "his."

Table 2 displays our results. The top panel shows the top five words most centered on males for each subcollection together with the average valence, arousal, and dominance scores for these words.[3] The bottom panel shows the analogous results for female-centered words. The results reflect obvious stereotypes featured in these books. We observe that the words centered on males feature occupations and power, and several exhibit a connection to warfare. In contrast, the female-centered words typically refer to familial roles and appearance. The

[3]The conclusions in this section are essentially the same when we consider the top 10 or 20 words.

VAD scores corroborate these qualitative findings, as female-centered words typically have higher valence, lower arousal, and lower dominance than male-centered words. We also observe that the word most centered on females in both male-authored subcollections is "husband." At the same time, "husband" is the fifth most-centered word on female in female-authored CSK books, and it ranks in the top 20 for female-authored Mainstream books, so it is a highly female-centered word in all subcollections.

This observation points to a more general (and surprising) feature of Table 2, which is the striking similarity between the word lists across author gender. For both male-centered and female-centered words, the word lists remain qualitatively consistent across collection and author gender. The VAD scores again corroborate the intuition gleaned from the word lists; there is relatively little variation in VAD scores across subcollections for either the male-centered or female-centered words.

In light of the results displayed in Figure 6, which showed that the female authors used greater proportions of female words in books over time, we replicated our word embeddings analysis for the same three time periods used in Figure 6. In Table 3, we display word embeddings for our female-authored subcollections, and we further disaggregate into books that won awards between 1920-1970, books that won between 1970-1990, and books that won after 1990. Table 4 shows the analogous results for our male-authored subcollections. We observe that the lists of the top five words most centered on females and most centered on males, along with the stereotypes they convey, remain consistent throughout each time period. This result holds across author gender. This result implies that, even while female authors may be including more references to female characters over time, the implicit associations attached to these characters have remained consistent.

V.C Relationship between creator's and consumer's identity

We use our Numerator data to estimate consumers' preferences for different features of books as well as for books by authors of different identities. As discussed in Section IV, we only consider observations for which the purchaser is known to be either male or female, the purchaser is known to be either Black or White, exactly one book was purchased, and the book has some gendered words. In our model, we include the book-level percentage of female words, the book-level mean perceptual tint of detected faces, an indicator for whether the author is Black, an indicator for whether the author is female, an indicator for whether the book is from the CSK collection, and the following interactions: interactions of an indicator for whether the purchaser is female with the book-level percentage of female words and with an indicator for whether the book has a female author, interactions of an indicator for whether the purchaser has a daughter with the book-level percentage of female words

and with an indicator for whether the book has a female author, and interactions of an indicator for whether the purchaser is Black with the book-level mean perceptual tint, with an indicator for whether the book has a Black author, and with an indicator for whether the book is in the CSK collection.

The estimated parameters are displayed in Table 5. We observe a positive coefficient on the book-level mean perceptual tint. The indicator for books in the CSK collection and the indicator for books having a female author have negative coefficients. This suggests that, overall, consumers prefer books with lighter-skinned characters, Mainstream books to CSK books, and books authored by males.

We note that the two interaction terms involving the indicator for a purchaser having a daughter are both significant at the 0.01 level. Consumers with daughters (relative to those without daughters) prefer books with a higher percentage of female words and books authored by a female. In contrast, we do not observe that female purchasers (relative to male purchasers) prefer female-authored books or books with a higher percentage of female words.

We also note that the interaction term between the average perceptual tint and the indicator for a Black purchaser is significant at the 0.05 level. Black purchasers (relative to White purchasers) exhibit preferences for books with darker-skinned characters. However, Black purchasers place only an insignificant albeit positive premium on Black-authored books and books in the CSK collection.

The choice set has more than 600 books. As noted in Section IV, it is more plausible to assume consumers have access to the entirety of the choice set when the purchase is made online than in-store. As a robustness check, we re-estimate our model with 1381 online trips and display the results in Table 6. Limiting the shopping trips to those made online, we see that the results remain qualitatively similar.

V.D Relationship between creator's identity and consumer's beliefs

In this section, we investigate the relationship between authors' identities and the political beliefs of the purchasers of their books. In order to do this, we link CCES data with our Numerator data at the ZIP code level. We then run regressions of the form given in Equation 3. We regress political beliefs separately against the percentage of books purchased in a given ZIP code that are female-authored, male-authored, or White-authored. The results are displayed in Tables 7-10.

We find that, while political beliefs do not correlate with the gender of the authors of

books being frequently purchased, we do find that higher proportions of purchases of White-authored books are associated with significantly more conservative views. In particular, people living in ZIP codes in which higher proportions of White-authored books are purchased are (significantly) more likely to disagree with the statement that White people in the U.S. have certain advantages because of the color of their skin, less likely to say they are angry that racism exists, and more likely to express support for withholding federal funds from localities that do not follow federal immigration laws.

VI Conclusion

The authors and illustrators of children's books play profound roles in shaping the content to which children are exposed. We find that these creators are more likely to depict characters who share their identities, both in the text and in images. However, the vast majority of authors and illustrators in the Mainstream collection, which consists of the most influential children's books, are White. This overrepresentation of White authors translates to an overrepresentation of White characters in the books. Given children's ability to detect messages about identities, including their presence or absence in certain spaces, this imbalance may have profound consequences for the messages they absorb.

In this study, we make three important contributions. First, we apply a suite of tools from computer vision and natural language processing to systematically characterize the relationship between the identities of authors and illustrators and the identities of the characters they represent. We find that Black authors depict darker-skinned characters and more Black characters than White authors, and we find that female authors depict more female characters than male authors. Nonetheless, Black authors in the Mainstream collection depict lighter characters than Black authors in the CSK collection, indicating that award selection criteria also play a role in filtering the content that reaches children. Second, we apply innovative applications of word embeddings to uncover the implicit associations that attach themselves to gender identities. This unsupervised approach reveals that females are associated with stereotypical familial roles whereas males are associated with occupations of influence. Moreover, these stereotypes emerge in all books, regardless of the author's gender, and they remain relatively consistent over time. Finally, we connect the book awards, author identities, and features we extract with consumers' preferences. We find that Black purchasers reveal stronger preferences than White purchasers for books that feature darker-skinned characters, and we find that purchasers with daughters (relative to purchasers without daughters) prefer female-authored books and books with a higher percentage of female words.

Our study has several limitations. First, we only have 24 Black-authored books in our

Mainstream collection. The vast majority of the winners and honorees in our Mainstream collection are written by White authors. This is an important finding to document in its own right. Despite this small sample size, we were able to detect distinguishable differences in our measures of representation between White-authored Mainstream books and Black-authored Mainstream books. Second, while we uncover implicit associations in the text, we cannot detect the sentiment attached to those stereotypes. For example, it may be the case that female authors associate women with familial roles because they are truthfully depicting historical realities that their characters would have faced. It is possible that these female authors attach relatively negative sentiment to these stereotypes, and future work could investigate this possibility. Third, we note that many messages reach children beyond those in the books we study. Television, movies, and social media may produce greater influences than these books. Fourth, we note some flaws in our measurement of representation. In particular, we restricted our attention to male and female gender identities, and we acknowledge that race is a socially constructed attribute, which creates a conceptual dilemma for any research attempting to assign or predict it.

Our application of word embeddings following the methodology proposed by (Charlesworth et al., 2022) allows us to detect the implicit associations that arise empirically without having to pre-specify a set of stereotypes to search for. This circumvents a challenge faced by many other studies that use word embeddings (Caliskan et al., 2017; Garg et al., 2017; Jones et al., 2019). In these settings, researchers must specify both an identity group and a domain whose association with that group they want to test. Our unsupervised approach allows us to avoid this manual process. Moreover, our approach still yields obvious and interpretable stereotypical associations, which demonstrates its power to detect these associations when they are present.

Our study also demonstrates the pattern of creators producing content for people who share their identities. Female authors use a higher percentage of female words than male authors, and Black authors depict darker-skinned characters than White authors. Consistent with the notion that these patterns reflect the production of books by authors for kids who share their identities, we see that purchasers with daughters prefer books with a higher percentage of female words, and Black purchasers prefer books with darker-skinned characters. These findings demonstrate the importance of representation among authors for meaningfully shaping the content that reaches children.

References

Adukia, A., A. Eble, E. Harrison, H. B. Runesha, and T. Szasz (2023). What We Teach About Race and Gender: Representation in Images and Text of Children's Books. The Quarterly Journal of Economics.

Akerlof, G. A. (2000). Economics and identity* george a. akerlof and rachel e. kranton. Technology 715.

Bakker, I., T. Van Der Voordt, P. Vink, and J. De Boon (2014). Pleasure, arousal, dominance: Mehrabian and russell revisited. Current Psychology 33, 405–421.

Bian, L., S.-J. Leslie, and A. Cimpian (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. Science 355(6323), 389–391.

Bossuroy, T., M. Goldstein, B. Karimou, D. Karlan, H. Kazianga, W. Parienté, P. Premand, C. C. Thomas, C. Udry, J. Vaillant, et al. (2022). Tackling psychosocial and capital constraints to alleviate poverty. Nature 605(7909), 291–297.

Caliskan, A., J. J. Bryson, and A. Narayanan (2017, 4). Semantics derived automatically from language corpora contain human-like biases. Science 356(6334), 183–186.

Cantoni, D., Y. Chen, D. Y. Yang, N. Yuchtman, and Y. J. Zhang (2017). Curriculum and ideology. Journal of political economy 125(2), 338–392.

Charlesworth, T. E., A. Caliskan, and M. R. Banaji (2022, 7). Historical representations of social groups across 200 years of word embeddings from Google Books. Proceedings of the National Academy of Sciences of the United States of America 119(28), e2121798119.

Crenshaw, K. W. (2013). Mapping the margins: Intersectionality, identity politics, and violence against women of color. In The public nature of private violence, pp. 93–118. Routledge.

Davis, J. M., N. Pearce, and M. Mullins (2021). Missing Boys: The Limited Representation of Black Males in Caldecott Books. Journal of Children's Literature 47(1), 10–20.

Dee, T. S. (2005, 5). A Teacher Like Me: Does Race, Ethnicity, or Gender Matter? American Economic Review 95(2), 158–165.

Eble, A. and F. Hu (2020). Child beliefs, societal beliefs, and teacher-student identity match. Economics of Education Review 77, 101994.

Einiö, E., J. Feng, and X. Jaravel (2019). Social push and the direction of innovation. Available at SSRN 3383703.

Fenton, A., E. Elliott, A. Shahbandi, E. Ezenwa, C. Morris, J. McLawhorn, J. G. Jackson, P. Allen, and A. Murina (2020). Medical students' ability to diagnose common dermatologic conditions in skin of color. Journal of the American Academy of Dermatology 83(3), 957–958.

Firth, J. R. (1957). A Synopsis of Linguistic Theory, 1930-1955.

Francis, D. V., B. L. Hardy, and D. Jones (2022). Black economists on race and policy: Contributions to education, poverty and mobility, and public finance. Journal of Economic Literature 60(2), 454–493.

Fu, S., H. He, and Z.-G. Hou (2014). Learning race from face: A survey. IEEE transactions on pattern analysis and machine intelligence 36(12), 2483–2509.

Garg, N., L. Schiebinger, D. Jurafsky, and J. Zou (2017, 11). Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. Proceedings of the National Academy of Sciences of the United States of America 115(16), E3635–E3644.

Harris, Z. S. (1954). Distributional Structure. WORD 10(3), 146–162.

Hlavac, M. (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables. Bratislava, Slovakia: Social Policy Institute. R package version 5.2.3.

Jensen, R. and E. Oster (2009). The power of tv: Cable television and women's status in india. The Quarterly Journal of Economics 124(3), 1057–1094.

Jones, J. J., M. R. Amin, J. Kim, and S. Skiena (2019). Stereotypical Gender Associations in Language Have Decreased Over Time. Sociological Science 7, 1–35.

Kearney, M. S. and P. B. Levine (2020). Role models, mentors, and media influences. The Future of Children 30(1), 83–106.

Koning, R., S. Samila, and J.-P. Ferguson (2021). Who do we invent for? patents by women focus more on women's health, but few women get to invent. Science 372(6548), 1345–1348.

Koss, M., M. G. Martinez, and N. J. Johnson (2017, 8). Where are The Latinxs? Diversity in Caldecott Winner and Honor Books. Bilingual Review/Revista Bilingüe 33(5), 50.

Koss, M. D. (2015). Diversity in contemporary picturebooks: A content analysis. Journal of Children's Literature 41(1), 32–42.

Koss, M. D., N. J. Johnson, and M. Martinez (2018). Mapping the Diversity in Caldecott Books From 1938 to 2017: The Changing Topography. Journal of Children's Literature 44(1), 4–20.

Krippendorff, K. (2018). Content analysis: An introduction to its methodology. Sage publications.

Krishnan, A., A. Almadan, and A. Rattani (2020). Understanding fairness of gender classification algorithms across gender-race groups. In 2020 19th IEEE international conference on machine learning and applications (ICMLA), pp. 1028–1035. IEEE.

Levstik, L. S. and C. A. Tyson (2010). Handbook of research in social studies education. Routledge.

Logan, T. D. (2022). American enslavement and the recovery of black economic history. Journal of Economic Perspectives 36(2), 81–98.

Louie, P. and R. Wilkes (2018). Representations of race and skin tone in medical textbook imagery. Social Science & Medicine 202, 38–42.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013, 1). Efficient Estimation of Word Representations in Vector Space. 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean (2013, 10). Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers), pp. 174–184.

Neuendorf, K. A. (2017). The content analysis guidebook. sage.

Osgood, C. E. (1957). The measurement of meaning. Urbana: University of Illinois Press.

Riley, E. (2024). Role models in movies: the impact of queen of katwe on students' educational attainment. Review of Economics and Statistics 106(2), 334–351.

Roberts, D. (2011). Fatal invention: How science, politics, and big business re-create race in the twenty-first century. New Press/ORIM.

Russell, J. A. (1980). A circumplex model of affect. Journal of personality and social psychology 39(6), 1161.

Weitzman, L. J., D. Eifler, E. Hokada, and C. Ross (1972). Sex-role socialization in picture books for preschool children. American journal of Sociology 77(6), 1125–1150.

Williams Jr, J. A., J. Vernon, M. C. Williams, and K. Malecha (1987). Sex role socialization in picture books: An update. Sociology Department, Faculty Publications, 8.

Yu, A. Z., S. Ronen, K. Hu, T. Lu, and C. A. Hidalgo (2016). Pantheon 1.0, a manually verified dataset of globally famous biographies. Scientific data 3(1), 1–16.

Figures and Tables

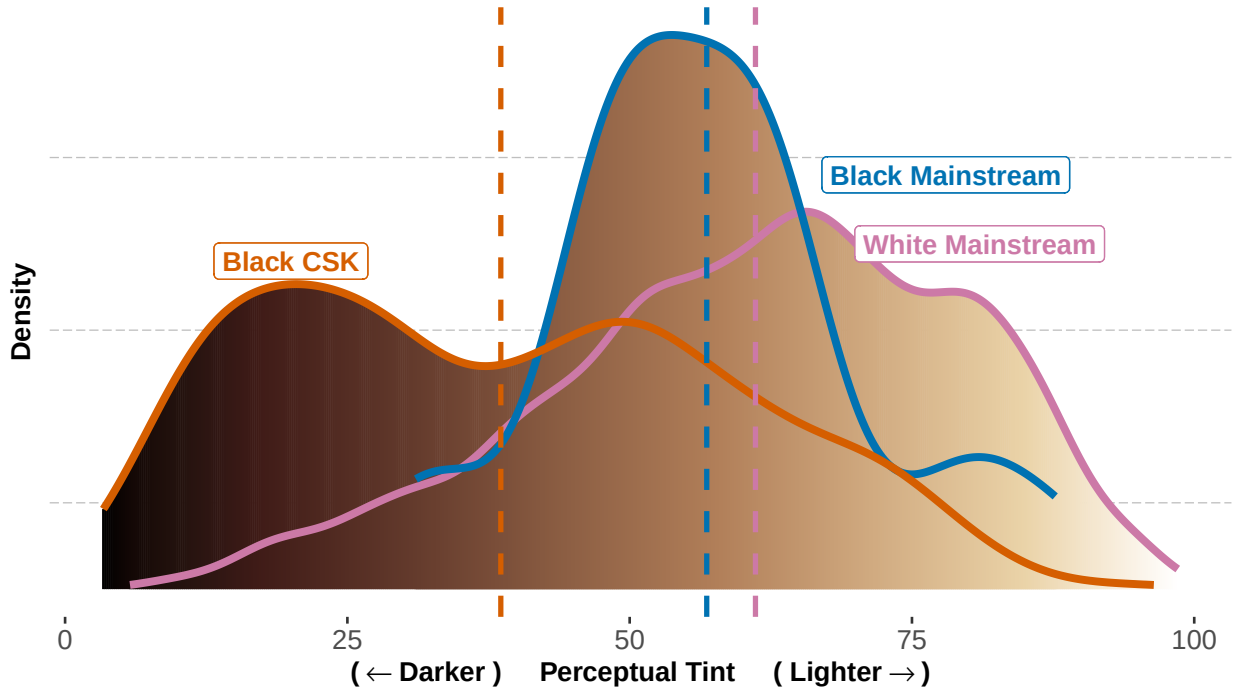
Table 1. Sample Composition by Author Race and Gender and Book Award

	Mainstream Award		***Coretta Scott King Award***	
Panel A: Authors	**Black**	**White**	**Black**	**White**
Female	13	272	86	0
Male	11	170	31	0

Panel B: Illustrators	**Black**	**White**	**Black**	**White**
Female	2	117	9	0
Male	14	200	33	0

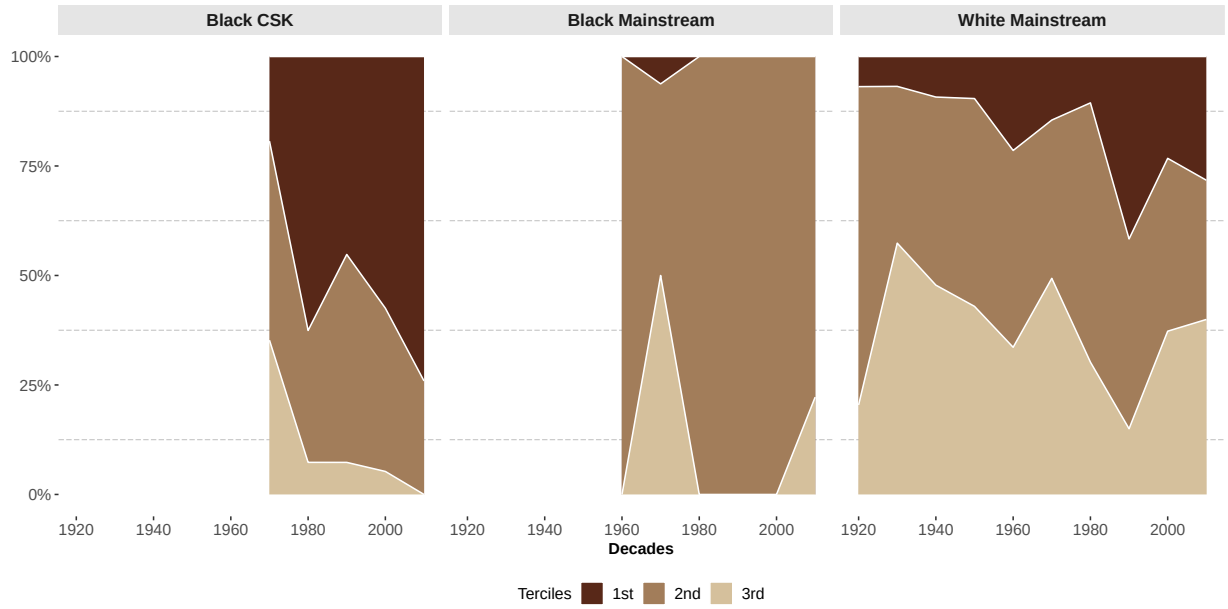
Notes: In Panel A, we display the number of authors of each race-gender identity among books in our Mainstream and CSK collections. Panel B displays the same figures for illustrators. We note that books may be included in the Mainstream and CSK collections simultaneously. We also note that not every book has an illustrator.

Figure 1. Perceptual Tints of Illustrated Characters by Author Race and Award



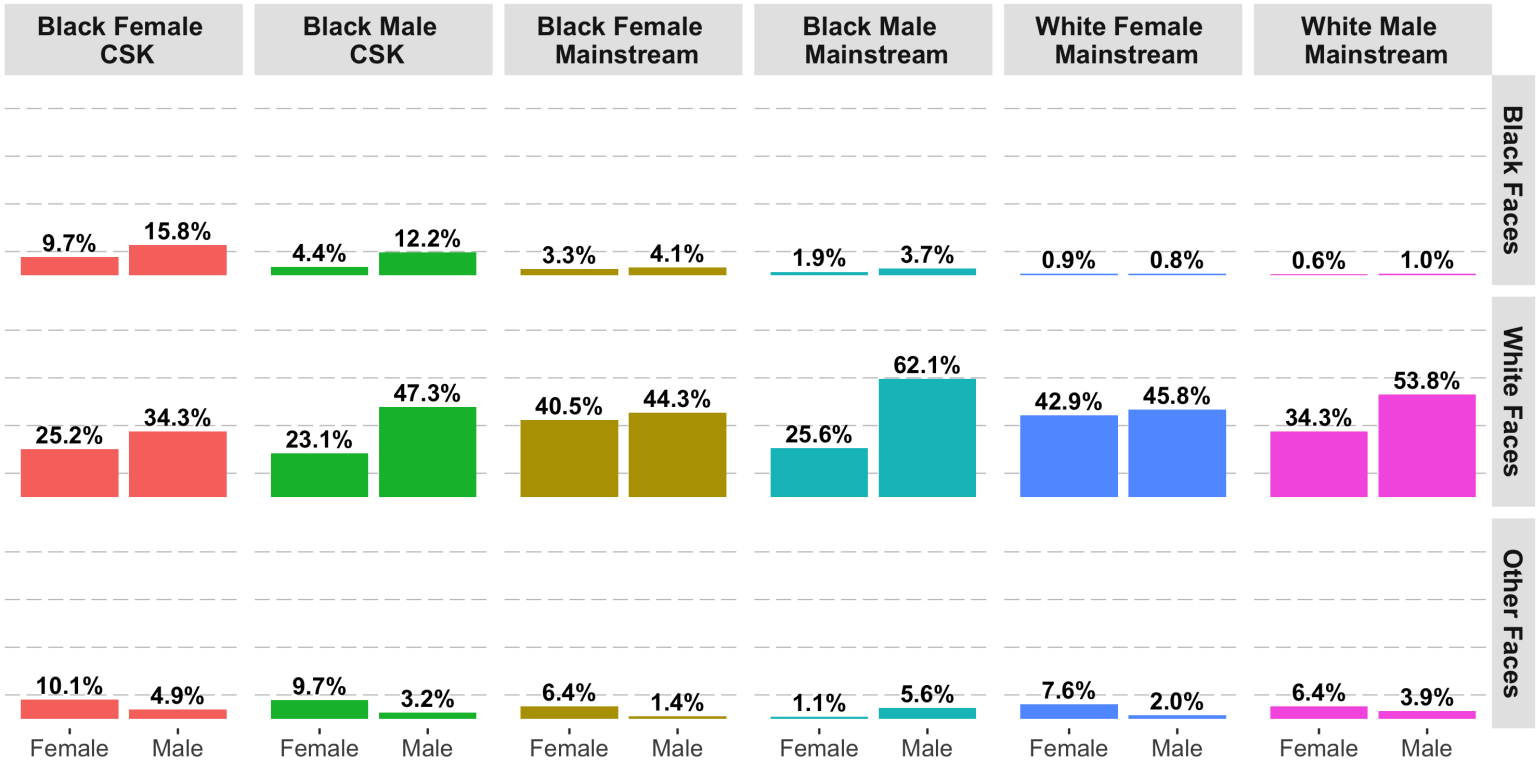
Notes: We display the densities of perceptual tints in faces detected three different subcollections: Black-authored Mainstream books, White-authored Mainstream books, and Black-authored CSK books.

Figure 2. Terciles of the Perceptual Tint Distribution Over Time, by Author Race and Award



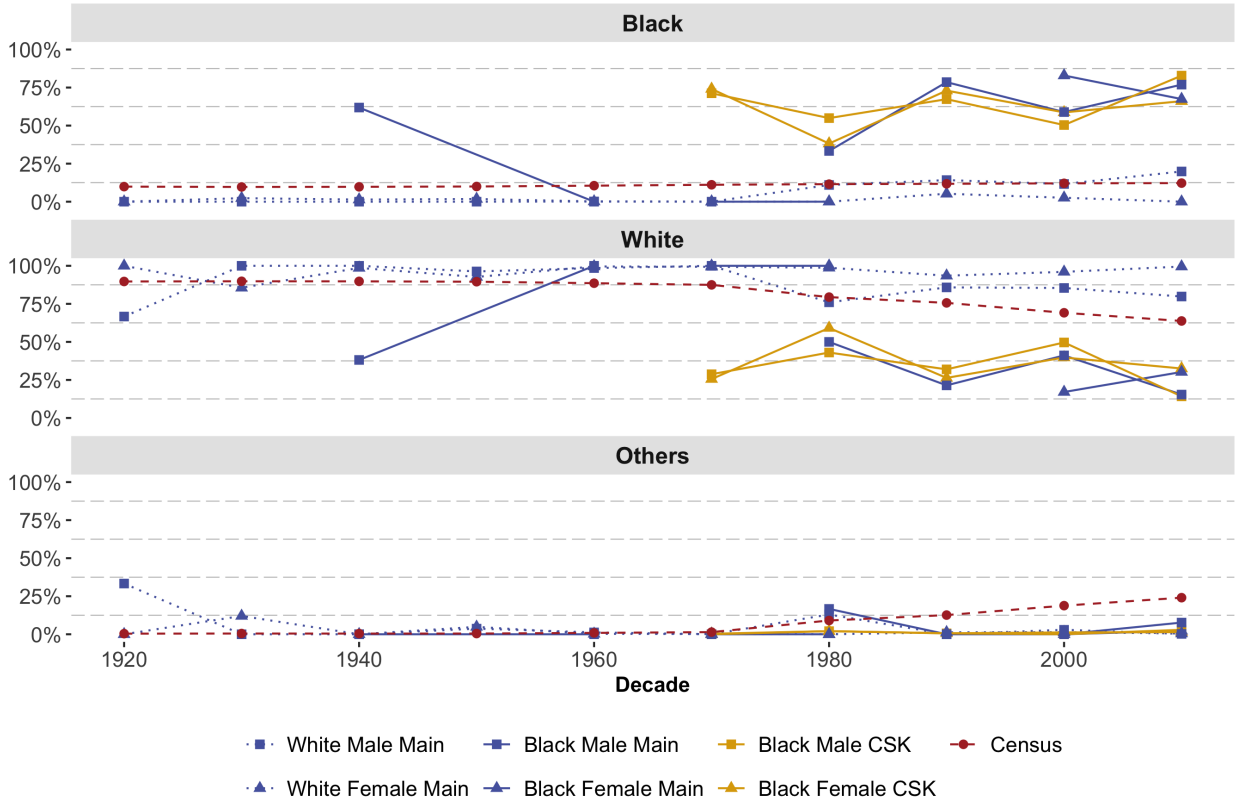
Notes: We construct terciles of the distribution of perceptual tints in detected faces in our entire collection, and we display the percentage of detected faces in each tercile in the same three subcollections considered in Figure 1. Decades are displayed on the horizontal access to examine how these depictions change over time.

Figure 3. Proportion of Pictured Characters of each Race-Gender Identity by Author Race, Author Gender, and Award



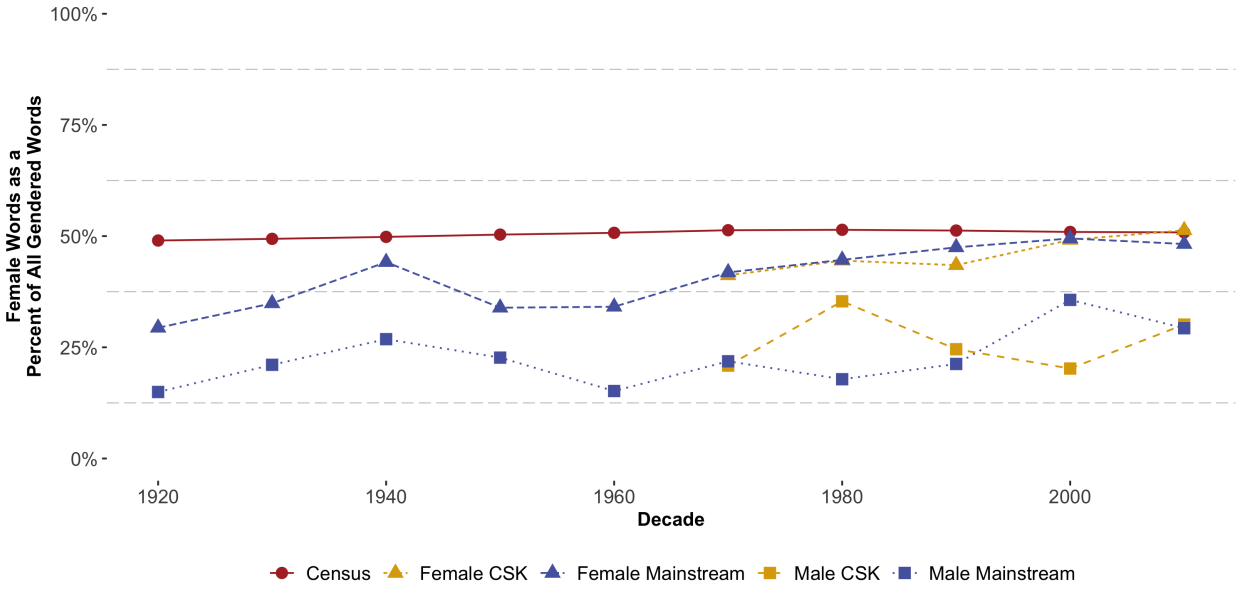
Notes: We display the percentages of detected faces predicted to be in different race-gender categories for each subcollection. Subcollections are defined for each award, author race, and author gender category.

Figure 4. Presence of Famous Individuals of Different Races Over Time, by Author Race, Author Gender, and Award



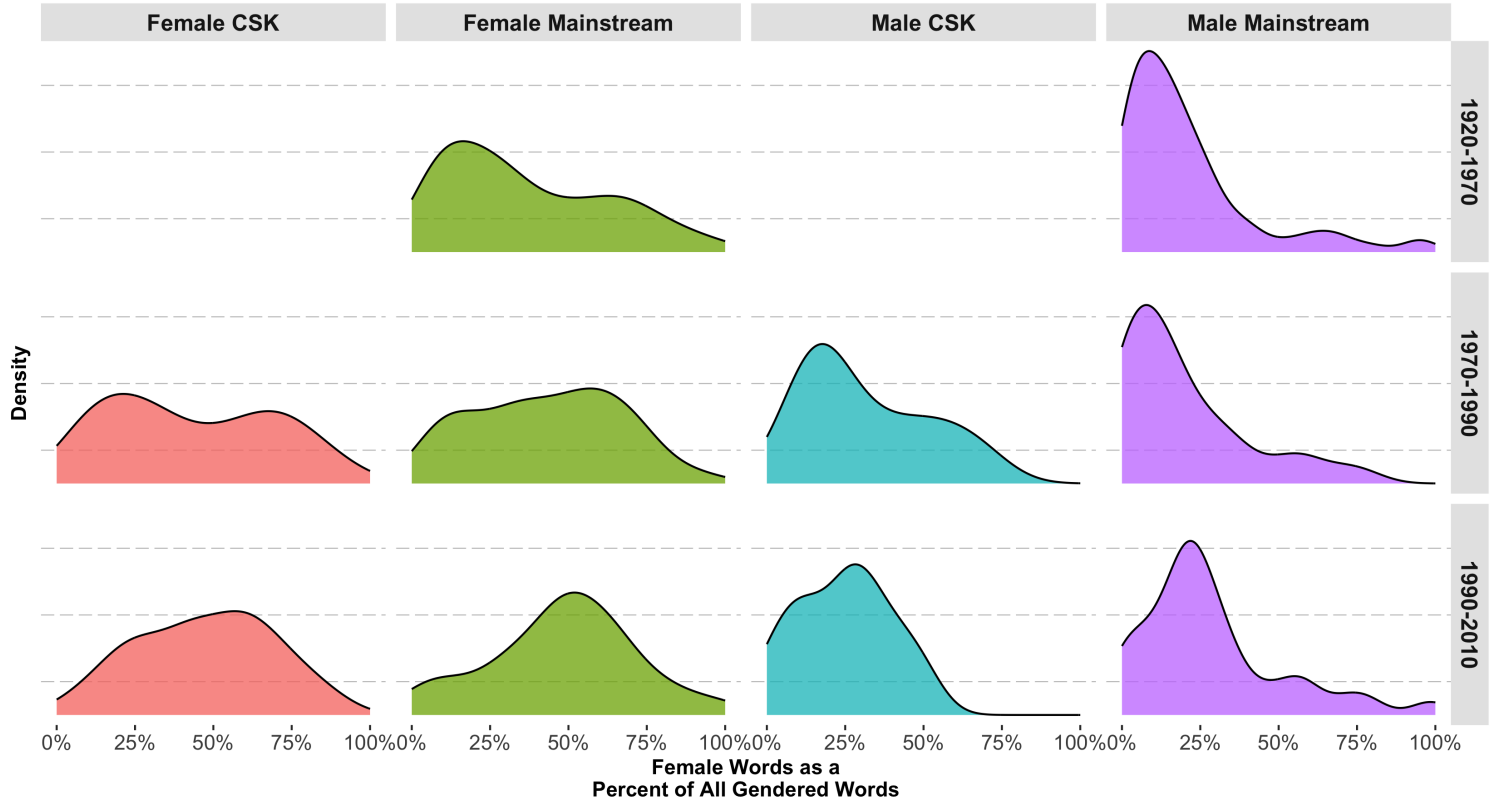
Notes: We display the percentages of famous individuals of each race-gender identity for each subcollection. Subcollections are defined for each each award, author race, and author gender category. Famous individuals are identified in the text using NER and matching the detected names to the Pantheon 2.0 dataset. Decades are displayed on the horizontal axis to show changes over time. For comparison, the percentage of the U.S. population in each racial group is plotted in red.

Figure 5. Percentages of Female Words over Time, by Author Gender and Award



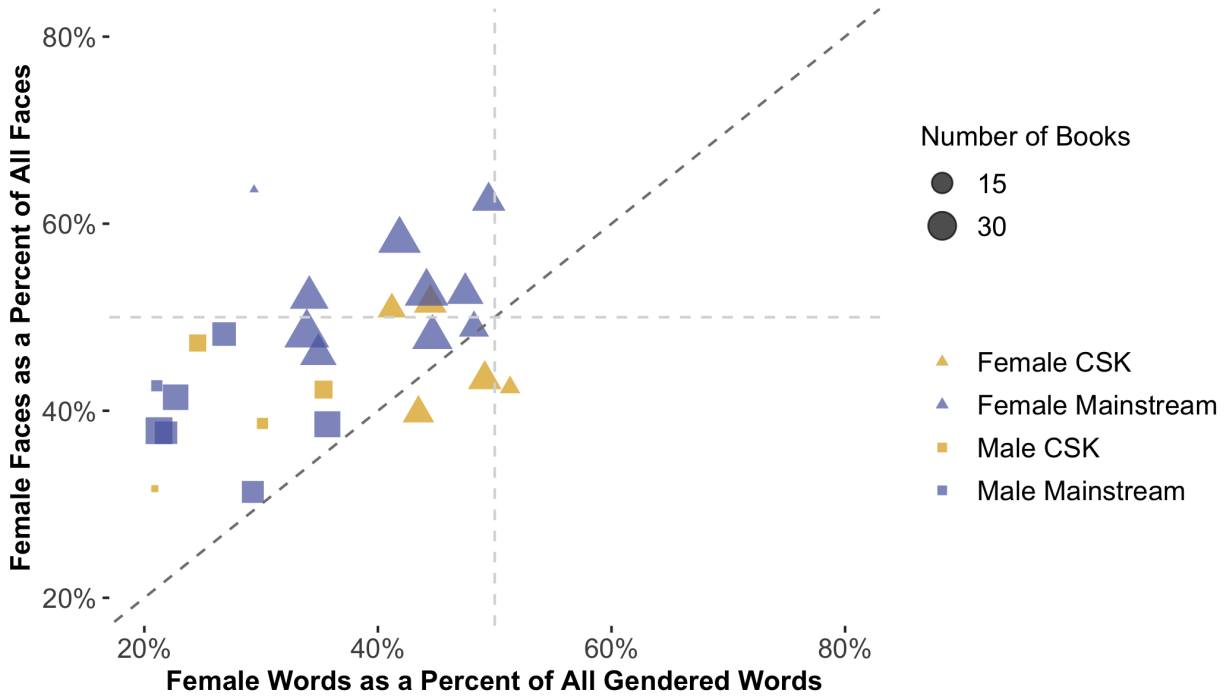
Notes: We display the percentages of female words (out of all gendered words) in each subcollection. Subcollections are defined for each each award and author gender category. Gendered words include titles, pronouns, specific terms such as queen and husband, and names with a gender association, which we automatically predict using characters' first names. Decades are displayed on the horizontal axis to show changes over time. For comparison, the percentage of the U.S. population of each gender is plotted in red.

Figure 6. Percentages of Female Words in books with Male and Female Authors, by Author Gender and Award



Notes: We display the distributions of percentages of female words (out of all gendered words) across all books in each subcollection. Subcollections are defined for each each award and author gender category. Gendered words include titles, pronouns, specific terms such as queen and husband, and names with a gender association, which we automatically predict using characters' first names. We construct these distributions for three time periods, indicated on the vertical axis.

Figure 7. Depictions of Female Characters versus Percentage of Female Words, by Author Gender and Award



Notes: We display the distributions of percentages of female words (out of all gendered words) on the horizontal axis and the percentages of faces predicted to be female on the vertical axis in each decade for books in each subcollection. Subcollections are defined for each each award and author gender category. Gendered words include titles, pronouns, specific terms such as queen and husband, and names with a gender association, which we automatically predict using characters' first names. The 45° line is also shown as a benchmark for when female characters are depicted in images proportionally as much as they are discussed in the text.

Table 2. Top 5 Words Most Centered on Male and Female, by Author Gender and Award

		Mean		
Top Words Centered on Male (vs. Female) Pronouns		V	A	D
Female CSK	colonel, deputy, papa, farmer, gun	0.474	0.563	0.676
Female Mainstream	attorney, fellow, rifle, tavern, gun	0.467	0.658	0.646
Male CSK	track, press, bank, legs, hit	0.455	0.524	0.520
Male Mainstream	rifle, bat, team, pilot, gun	0.439	0.674	0.667
Top Words Centered on Female (vs. Male) Pronouns		V	A	D
Female CSK	mom, skirt, apartment, cute, husband	0.788	0.444	0.536
Female Mainstream	skirt, blouse, dress, pink, skirts	0.674	0.352	0.317
Male CSK	husband, beauty, baby, leaving, kiss	0.716	0.564	0.551
Male Mainstream	husband, dress, grandma, baby, sewing	0.758	0.342	0.460

Notes: In the top panel, we display the top five words most centered on male for each subcollection. In the bottom panel, we display the top five words most centered on female. The average valence, arousal, and dominance scores of these top five words are also displayed. Subcollections are defined for each each award and author gender category.

Table 3. Implicit Associations in Word Embeddings Over Time, by Author Gender and Award

		Mean		
Female CSK				
Top Words Centered on Male (vs. Female) Pronouns		V	A	D
All Periods	colonel, deputy, papa, farmer, gun	0.474	0.563	0.676
1970-1990	master, horse, owner, leg, hairy	0.631	0.440	0.591
1990-2010	coach, farmer, team, horse, father	0.693	0.512	0.648
Top Words Centered on Female (vs. Male) Pronouns		V	A	D
All Periods	mom, skirt, apartment, cute, husband	0.788	0.444	0.536
1970-1990	telephone, angel, husband, dress, felt	0.755	0.379	0.539
1990-2010	doll, cooking, stomach, skirt, plate	0.595	0.388	0.359
Female Mainstream				
Top Words Centered on Male (vs. Female) Pronouns		V	A	D
All Periods	attorney, fellow, rifle, tavern, gun	0.467	0.658	0.646
1920-1970	murder, agent, errand, president, demand	0.379	0.639	0.665
1970-1990	gun, walker, wife, shot, rider	0.473	0.684	0.593
1990-2010	millstone, guard, guitar, truck, boat	0.599	0.455	0.557
Top Words Centered on Female (vs. Male) Pronouns		V	A	D
All Periods	skirt, blouse, dress, pink, skirts	0.674	0.352	0.317
1920-1970	skirt, dress, quilt, apron, lovely	0.701	0.310	0.357
1970-1990	husband, pretend, fed, teacher, neat	0.700	0.407	0.611
1990-2010	lovely, daughter, dress, grandmother, skirt	0.796	0.383	0.439

Notes: We display the top five words most centered male and the top five words most centered on female, as well as their average valence, arousal, and dominance, for each subcollection of female-authored books. In the top panel, we consider female-authored books that won the CSK award, and we divide this set into books that won in different time periods in order to track changes in implicit associations over time. In the bottom panel, we perform the same exercise for female-authored books that won the Mainstream award.

Table 4. Implicit Associations in Word Embeddings Over Time, by Author Gender and Award

		Mean		
Male CSK				
Top Words Centered on Male (vs. Female) Pronouns		V	A	D
All Periods	track, press, bank, legs, hit	0.455	0.524	0.520
1970-1990	office, grandfather, plenty, press, teach	0.695	0.445	0.652
1990-2010	stick, prison, river, suddenly, key	0.496	0.444	0.491
Top Words Centered on Female (vs. Male) Pronouns		V	A	D
All Periods	husband, beauty, baby, leaving, kiss	0.716	0.564	0.551
1970-1990	husband, soul, patience, aunt, lovely	0.779	0.318	0.543
1990-2010	baby, grandma, pretend, sands, sister	0.737	0.346	0.485
Male Mainstream				
Top Words Centered on Male (vs. Female) Pronouns		V	A	D
All Periods	rifle, bat, team, pilot, gun	0.439	0.674	0.667
1920-1970	rifle, office, boots, chair, shirt	0.524	0.345	0.417
1970-1990	shot, ball, team, second, arrow	0.504	0.448	0.456
1990-2010	weight, split, force, plutonium, base	0.489	0.544	0.635
Top Words Centered on Female (vs. Male) Pronouns		V	A	D
All Periods	husband, dress, grandma, baby, sewing	0.758	0.342	0.460
1920-1970	husband, baby, sow, witch, rabbi	0.605	0.473	0.550
1970-1990	dress, aunt, auntie, grandma, sassy	0.627	0.356	0.405
1990-2010	dress, grandma, aunt, pie, baby	0.788	0.313	0.416

Notes: We display the top five words most centered male and the top five words most centered on female, as well as their average valence, arousal, and dominance, for each subcollection of male-authored books. In the top panel, we consider male-authored books that won the CSK award, and we divide this set into books that won in different time periods in order to track changes in implicit associations over time. In the bottom panel, we perform the same exercise for male-authored books that won the Mainstream award.

Table 5. Conditional Logit Estimation of Consumer Preference Parameters: All Shopping Trips

	Purchased
% Female Words	0.509
	(0.319)
Average Perceptual Tint	0.010***
	(0.002)
Black Author	−0.116
	(0.136)
CSK Author	−0.412**
	(0.172)
Female Author	−3.028***
	(0.299)
% Female Words:Female Purchaser	−0.103
	(0.318)
% Female Words:Purchaser Has Daughter	0.782***
	(0.206)
Average Perceptual Tint:Black Purchaser	−0.019**
	(0.008)
Black Author:Black Purchaser	0.421
	(0.484)
CSK Author:Black Purchaser	0.214
	(0.173)
Female Author:Female Purchaser	−0.189*
	(0.112)
Female Author:Purchaser Has Daughter	1.727***
	(0.647)
Number of trips	1497
Number of choices in 2017	617
Number of choices in 2018, 2019, 2020	619
Observations	925,925

Notes: We display the structural parameters for consumer preferences estimated from purchaser data. We only consider observations for which the purchaser is known to be either male or female, the purchaser is known to be either Black or White, exactly one book was purchased, and the book has some gendered words (Hlavac, 2022). *p<0.1; **p<0.05; ***p<0.01

Table 6. Conditional Logit Estimation of Consumer Preference Parameters: Online Shopping Trips

	Purchased
% Female Words	0.747**
	(0.324)
Average Perceptual Tint	0.007***
	(0.002)
Black Author	−0.975***
	(0.211)
CSK Author	−0.408**
	(0.178)
Female Author	−2.235***
	(0.347)
% Female Words:Female Purchaser	−0.083
	(0.322)
% Female Words:Purchaser Has Daughter	0.666***
	(0.210)
Average Perceptual Tint:Black Purchaser	−0.020***
	(0.008)
Black Author:Black Purchaser	0.836
	(0.628)
CSK Author:Black Purchaser	0.309*
	(0.178)
Female Author:Female Purchaser	−0.250**
	(0.117)
Female Author:Purchaser Has Daughter	1.382*
	(0.765)
Number of trips	1381
Number of choices in 2017	617
Number of choices in 2018, 2019, 2020	619
Observations	854,153

Notes: We display the structural parameters for consumer preferences estimated from purchaser data. We only consider observations for which the purchaser is known to be either male or female, the purchaser is known to be either Black or White, exactly one book was purchased, and the book has some gendered words (Hlavac, 2022). *p<0.1; **p<0.05; ***p<0.01

Table 7. ZIP Code-Level Beliefs about Advantages of White Americans and Purchasing Patterns

	% of Respondents that somewhat or strongly agree		
	White people in the U.S. have certain advantages because of the color of their skin		
% of all female authored books	0.010		
	(0.014)		
% of all male authored books		−0.010	
		(0.014)	
% of all white authored books			−0.077***
			(0.021)
Constant	55.063***	56.025***	61.948***
	(0.918)	(0.843)	(1.849)
Observations	6,023	6,023	6,023
R^2	0.0001	0.0001	0.002
Adjusted R^2	−0.0001	−0.0001	0.002
Residual Std. Error (df = 6021)	40.520	40.520	40.479
F Statistic (df = 1; 6021)	0.443	0.481	12.940***

Notes: We consider the ZIP code-level proportion of individuals who agree with the statement, "White people in the U.S. have certain advantages because of the color of their skin," and we regressed this variable separately against the ZIP code-level percentage of female-authored books purchased, percentage of male-authored books published, and percentage of White-authored books published. *p<0.1; **p<0.05; ***p<0.01

Table 8. ZIP Code-Level Beliefs about Racism and Purchasing Patterns

	% of Respondents that somewhat or strongly agree		
	I am angry that racism exists		
% of all female authored books	0.001		
	(0.011)		
% of all male authored books		−0.0003	
		(0.011)	
% of all white authored books			−0.033**
			(0.017)
Constant	80.109***	80.199***	82.945***
	(0.728)	(0.669)	(1.468)
Observations	6,023	6,023	6,023
R^2	0.00000	0.00000	0.001
Adjusted R^2	−0.0002	−0.0002	0.0005
Residual Std. Error (df = 6021)	32.144	32.144	32.133
F Statistic (df = 1; 6021)	0.016	0.001	3.844**

Notes: We consider the ZIP code-level proportion of individuals who agree with the statement, "I am angry that racism exists," and we regressed this variable separately against the ZIP code-level percentage of female-authored books purchased, percentage of male-authored books published, and percentage of White-authored books published. *p<0.1; **p<0.05; ***p<0.01

Table 9. ZIP Code-Level Beliefs about Identifying Undocumented Immigrants and Purchasing Patterns

	% of Respondents who think the U.S. government should		
	Identify and deport undocumented immigrants		
% of all female authored books	0.013		
	(0.014)		
% of all male authored books		−0.009	
		(0.014)	
% of all white authored books			−0.003
			(0.021)
Constant	36.137***	37.250***	37.024***
	(0.890)	(0.817)	(1.794)
Observations	6,023	6,023	6,023
R^2	0.0001	0.0001	0.00000
Adjusted R^2	−0.00002	−0.0001	−0.0002
Residual Std. Error (df = 6021)	39.272	39.273	39.274
F Statistic (df = 1; 6021)	0.850	0.466	0.015

Notes: We consider the ZIP code-level proportion of individuals who think that the U.S. government should identify and deport undocumented immigrants, and we regressed this variable separately against the ZIP code-level percentage of female-authored books purchased, percentage of male-authored books published, and percentage of White-authored books published. *p<0.1; **p<0.05; ***p<0.01

Table 10. ZIP Code-Level Beliefs about Funding Localities that do not Follow Federal Immigration Laws and Purchasing Patterns

	% of Respondents who think the U.S. government should		
	Withhold federal funds from localities that do not follow federal immigration laws		
% of all female authored books	0.014		
	(0.014)		
% of all male authored books		−0.013	
		(0.014)	
% of all white authored books			0.057***
			(0.021)
Constant	55.529***	56.869***	51.565***
	(0.909)	(0.835)	(1.833)
Observations	6,023	6,023	6,023
R^2	0.0002	0.0001	0.001
Adjusted R^2	0.00001	−0.00004	0.001
Residual Std. Error (df = 6021)	40.142	40.143	40.122
F Statistic (df = 1; 6021)	1.036	0.780	7.219***

Notes: We consider the ZIP code-level proportion of individuals who think that the U.S. government should withhold federal funds from localities that do not follow federal immigration laws, and we regressed this variable separately against the ZIP code-level percentage of female-authored books purchased, percentage of male-authored books published, and percentage of White-authored books published. *p<0.1; **p<0.05; ***p<0.01