What We Teach About Race and Gender: Representation in Images and Text of Children’s Books

Anjali Adukia, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz

APRIL 2021
What we teach about race and gender: Representation in images and text of children’s books*

Anjali Adukia  
University of Chicago and NBER

Alex Eble  
Columbia University

Emileigh Harrison  
University of Chicago

Hakizumwami Birali Runesha  
University of Chicago

Teodora Szasz  
University of Chicago

April 15, 2021

Abstract

Books shape how children learn about society and social norms, in part through the representation of different characters. To better understand the messages children encounter in books, we introduce new machine-led methods for systematically converting images into data. We apply these image tools, along with established text analysis methods, to measure the representation of race, gender, and age in children’s books commonly found in US schools and homes over the last century. We find that books selected to highlight people of color, or females of all races, have increasingly over time depicted characters with darker skin tones; “mainstream” books, by contrast, have consistently depicted chromatically ambiguous characters with an increase in lighter skin tones in the last three decades. Children are consistently depicted with lighter skin than adults, despite no systematic differences in skin tones by age. We find that females are more represented in images than in text. There is a persistent disproportionate representation of males, particularly White males, and lighter-skinned people relative to darker-skinned people. Our data provide a view into the “black box” of education through children’s books in US schools and homes, highlighting what has changed and what has endured.

*Contact: team@miielab.com. For helpful feedback, we thank Barbara Atkin, Karen Baicker, Anna Brailovsky, Tom Brock, Steven Durlauf, Alice Eagly, Allyson Ettinger, James Evans, Adam Gamoran, Jon Guryan, Rick Hornbeck, Caroline Hoxby, Susanna Loeb, Jens Ludwig, Jonathan Meer, Martha Minow, Sendhil Mullainathan, Derek Neal, Anna Neumann, Aaron Pallas, Steve Raudenbush, Cybele Raver, Heather Sarsons, Fred Stafford, Amy Stuart Wells, Chenhao Tan, Miguel Urquiola, Alessandra Voena, and others including seminar participants at AEFP, National Academy of Education, Spencer Foundation, William T. Grant Foundation, and University of Chicago. For financial support, we thank Becker-Friedman Institute at UChicago, Center for Data and Applied Computing at UChicago, National Academy of Education, Spencer Foundation, and UChicago Career Advancement. 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. For excellent research assistance, we thank Fabiola Alba-Vivar, Celia Anderson, Ryan Atkinson, Callista Christ, Marliese Dalton, Anjali Das, Maya Escueta, Saloni Gupta, Amara Haider, Shavonna Hinton, Camilo Ibáñez, Juan Miguel Jimenez, Jahnilkh Kaur, Zipporah Khain, Jarvis Lam, Erica Lin, Ping-Chang Lin, Ping-Jung Liu, Simon Mahns, Noah McLean, Karla Monteiro, Ifeatu Oliobi, Raj Shukla, Bhargavi Thakur, Jeffrey Tharsen, Qurat ul Ain, and Charlie Wang. We also thank Ashiyyana and Kairav Adukia-Hornbeck for manual coding assistance. For access to important resources, we thank LaShanda Howard-Curry, Bridget Madden, and Kalli Mathios.
Education teaches children about the world, its people, and their place in it. Much of this happens through the curricular materials we present to children, particularly the books they read in school and at home ([Giroux](#) 1981, [Apple and Christian-Smith](#) 1991, [Jansen](#) 1997, [Van Kleeck, Stahl and Bauer](#) 2003, [Steele](#) 2010). The way that people are represented within these books is a crucial site for teaching: the identity of the people portrayed in different roles shows children what roles they and others can or cannot inhabit. Given persistent racial and gender inequality in society ([Darity and Mason](#) 1998, [O’Flaherty](#) 2015, [Blau and Kahn](#) 2017), and the potential importance of identity and representation in contributing to beliefs, aspirations, academic effort, and outcomes ([Dee](#) 2005, [Riley](#) 2017, [Gershenson et al.](#) 2018, [Porter and Serra](#) 2019), these representations offer a key means through which we either address, perpetuate, or entrench core societal inequalities.

In this paper, we use new and established tools to measure the representation of race, gender, and age in the images and text contained in influential collections of children’s books. Our work makes two core contributions. First, we develop and showcase new software tools for the systematic analysis of images, highlighting their potential use in a wide range of applications in policy, education practice, and social science research. Second, we apply these tools, alongside established text analysis methods, to measure what we have taught children about race and gender in the images and text of these books.

Our main data set is a series of books targeted to children and likely to appear in homes, classrooms, and school libraries over the past century. Specifically, we use books that have won awards from the Association for Library Service to Children, a division of the American Library Association, starting in 1922. These and other children’s books are often filled with images that transmit implicit and explicit messages to readers. Historically, content analysis to measure these messages has been done “by hand” using human coders ([Bell](#) 2001, [Neuendorf](#) 2016, [Krippendorff](#) 2018). Such analysis provides deep understanding but can generally only be done on a small set of content and necessarily reflects human behavior and biases. We apply and develop computer vision tools that use convolutional neural networks to identify and classify components of images; in our case, the tools detect characters in photos and illustrations and classify their race, gender, and age. While artificial intelligence tools also reflect bias in their training data and algorithms, they can be standardized; they also can be more replicable and applied to a much larger sample than manual content analysis permits.

Convolutional neural networks, or CNNs, are programs trained to model the way the human brain functions by being given examples and then learning to perform tasks (such as detecting faces or classifying features on these faces) by analyzing these examples without explicit instructions.
Analyzing images involves three primary components: (1) training the computer to detect faces, (2) classifying skin color, and (3) predicting the race, gender, and age of the faces. We build on existing face analysis software tools, making some key improvements. First, because most established face detection models are trained on photographs, and because the books in our sample contain a large number of illustrations, we trained our own model using illustrated faces to improve accuracy. Second, we develop a model to classify skin color of faces. This process involves isolating the skin of the detected face using convolutional neural networks, identifying the predominant colors in that segmented skin using $k$-means clustering, and then using a weighted average of those colors to classify the skin color of a character. Third, we train a new model with higher precision in its classification of race, gender, and age than those previously available.

Using these new tools to systematically analyze images, we are able to characterize representation that computers could not have previously characterized in children’s books, particularly related to the skin color, gender presentation, and predicted age of pictured characters. For the second contribution of this paper, we apply these tools to track what the educational experience of children has actually been and how this has changed over time. We divide the award-winning corpora on which we developed these tools and broadly categorize them into two primary groups: (i) “mainstream” books considered to be of high literary value but written without explicit intention to highlight an identity group (e.g., the Newbery and Caldecott Awards) and (ii) “diversity” books selected because they highlight experiences of specific identity groups (e.g., the Coretta Scott King and Amelia Bloomer Awards).

We present a series of descriptive analyses documenting patterns of representation in the images and text of these books over time. Additionally, we explore the efficacy of explicit efforts to highlight diversity and their likelihood to account for intersectional experiences. These books also represent an ideal setting for demonstrating both the challenges of analyzing heterogeneous types of images and our tools’ ability to process them into usable data.\footnote{Images in children’s books, for example, vary widely with respect to several important characteristics. These books can include illustrations or photographs. Images can be polychromatic or monochromatic (i.e., in black and white); and even when characters are polychromatic, their skin is sometimes shown in seemingly non-typical colors, such as green or blue. Characters can take human or non-human forms, and images often have shadows or highlights that add to the complexity of measurement of the representation in these images.}

We find that books in the Mainstream collection are more likely to depict lighter-skinned characters than those in the Diversity collection, even among characters of a given race. These books in the Mainstream collection are much more likely to depict characters who are racially ambiguous in terms of skin color, disproportionately using skin colors that cannot be classified either as that of light-skinned characters or as that of dark-skinned characters,
a technique to which we refer as “butterscotching.” Particularly surprising is that despite there not being systematic differences in skin tones across ages in society, children are more likely than adults to be shown with lighter skin, regardless of collection.

We also use established text analysis methods to measure gender identity, racial constructs, and age using specific words of interest, including names of characters, which serves as a complement to the image analysis results.

We compare the incidence of female appearances in images to female mentions in text, and we see that females have consistently been more likely to be visualized (seen) in images than spoken about (heard) in the text, except in the collection of books specifically selected to highlight females. This suggests there may be symbolic inclusion of females in pictures without their substantive inclusion in the actual story. Despite being half of the US population and despite substantial changes in female societal roles over time, females have persistently been more absent than males, on average, in both images and text. This is regardless of the measure used: predicted gender of the pictured character, pronoun counts, specific gendered words, famous figure gender, and character first names. Another surprising result is that, even though these books are targeted to children, adults are depicted more than children both in images and text.

The Diversity collection has broader geographic representation of famous figures born outside of the United States or Europe than the Mainstream collection. However, when either collection presents a character outside of these two regions, that character is more likely to be male. This suggests that while the Diversity collection may represent a broader range of nationalities, it is still unequal in its representation of identity at the intersection of gender and nationality. Moreover, White males comprise the majority of famous figures in all collections. Famous people from other racial groups are less likely than either White people or Black people to be represented in any collection (0 – 8 percent), but even then, males are generally more likely to be represented than females within any racial group.

This paper proceeds as follows. We present background information in Section I. Section II describes the books we use for our data. Section III explains how we convert images to data and classify skin color, gender, and age. Section IV discusses the text analysis tools. Section V synthesizes our final measures. We discuss observations from the resulting patterns of inequality and inclusion in Section VI. Section VII discusses the potential benefits and concerns to using AI models. Section VIII discusses the cost-effectiveness of machine-led approaches to analyzing content relative to traditional manual approaches. Section IX concludes.
I The importance of representation and the challenge of measurement

In this section, we briefly discuss research on the representation of race and gender and follow with a short description of the empirical challenges involved in measuring these representations.

I.A The importance of equity in representation

Our institutional practices, public policies, and cultural representations reflect the value that society assigns to specific groups. Inequality in representation, therefore, constitutes an explicit statement of inequality in value. If our records of history, culture, and society are disproportionately associated with whiteness and maleness, then the human potential of females, males of color, and non-binary individuals is devalued relative to the privileged group. In a broad range of cultural products, from news media and history books to children’s books, people who do not belong to the culturally dominant group are typically absent or portrayed through negative stereotypes (O’Kelly, 1974; Stewig and Knipfel, 1975; Dobrow and Gidney, 1998; Balter, 1999; Witt, 2000; Brooks and Hébert, 2006; Martin, 2008; Paceley and Flynn, 2012; Daniels, Layh and Porzelius, 2016).

While there exist myriad structural barriers to racial and gender equality woven throughout the organizations, laws, and customs of our society (Darity and Mason, 1998; O’Flaherty, 2015; Blau and Kahn, 2017; Muhammad, 2019; Chetty et al., 2020), (in)equality of representation is a key contributor to inequality in outcomes if it instills the belief that members of the underrepresented group are inherently deficient. Much research from different disciplines supports the conclusion that the representation gap is linked to socioeconomic inequality. For example, the experience of cultural subjugation may reduce the “capacity to aspire” (Appadurai, 2004). Indeed, the absence of positive examples of success can lead to a distorted view of the path from present action to future outcomes (Wilson, 2012). If children do not observe positive examples and role models who share their identity, they may perceive the chances of their own success as far less probable, which, in turn, could reduce their motivation (Genicot and Ray, 2017; Eble and Hu, 2020). This recursive loop from the self-image formed by the educational experience to socioeconomic success underscores the importance of closing the representation gap in educational materials. In the context of schools, this gap in representation is particularly pernicious because educational materials are explicitly intended to shape students’ image of themselves and the world around them. Importantly, the messages in these materials also shape how children view others of different identities. When children do not see others represented, their conscious or unconscious perceptions of their own potential and that of unrepresented groups is molded in detrimental ways and can erroneously shape subconscious defaults.
The good news is that the reverse is also true: improving representation may improve outcomes. Closing the representation gap by revealing previously invisible opportunities can influence beliefs, actions, and educational outcomes for females, males of color, or non-binary individuals (Dee, 2004; Stout et al., 2011; Beaman et al., 2012; Riley, 2017). While not a panacea, such “subject–object identity match” (e.g., teacher–student identity match, or content–reader identity match) can help reduce academic performance gaps among multiple marginalized groups. Moreover, including previously excluded groups could expand the subconscious defaults and assumptions held by individuals in majority-represented groups.

I.B The need for better measurement tools

Systematically addressing these issues requires a systematic method for assessing the representations contained in the content used to instruct children. Educators and curriculum developers have worked to address this representation gap by, for instance, expanding the curriculum to include individual books that elevate the presence of an identity group. These efforts, however, are inherently piecemeal. Furthermore, the incidence, levels, and impacts of such efforts are likely to vary dramatically across teachers and schools, and the sheer quantity of content that they have to review or create is too large for any individual to manually track and assess. As a result, educators, administrators, and policymakers currently lack feasible ways to systematically identify such inclusive materials.

Even those who select content with an eye towards increasing representation of particular groups are themselves often products of an education system that reflects the structural racism, sexism, and other drivers of systematic inequality woven throughout society. Thus, even deliberate efforts to address inequality in representation may inadvertently perpetuate other inequalities, thereby ignoring intersectional experiences. Different aspects of identity, such as race, gender identity, class, sexual orientation, and disability, do not exist separately from each other, but rather are inextricably linked (Crenshaw, 1989, 1990; Ghavami, Kat-siaficas and Rogers, 2016). An effort by publishers to diversify by gender, for example, is likely to overrepresent the experiences of White women relative to women of color, given the relative abundance of White women in popular media. Such narrow views of diversification neglect the notion that the meaning of each social group membership is constructed through the lens of other identities. In this paper, we explore how and whether intersectionality is addressed over time in books that are intentionally selected to highlight specific marginalized groups compared to books not selected to highlight any particular identity.

Children’s books represent a prime opportunity to “fix the institution” by increasing equity in representation, particularly in books that highlight the diverse roles that people can perform in an equal society. Identifying such books has been done through content analysis,
which historically has been conducted primarily by humans reading carefully through text, images, or other media while coding the presence of certain words, themes, or concepts by hand (Neuendorf 2016; Krippendorff 2018). Because this manual process is time-consuming and therefore costly, resource constraints have limited the scope of such work.

In this paper, we demonstrate how tools from computer vision and natural language processing can be used to systematically analyze content. We expand and develop tools for image analysis, pairing them with tools from text analysis used in recent work by Caliskan, Bryson and Narayan (2017), Garg et al. (2018), and Kozlowski, Taddy and Evans (2019), to automate core tools in content analysis. This work will help facilitate broader, cost-effective measurements of gender identity, racial constructs, and age in images and text in a larger set of content.

There are challenges to this type of numeric measurement of representation, however. Current methods measure gender identity in a binary way and neglect non-binary and gender fluid identities. Racial constructs are multi-faceted and often ill-defined. To address the latter challenge, we measure different facets of the broad concept of race in various ways: putative race (that is, assigned by society), ethnicity, birthplace, and skin color.

It is important to focus on these racial constructs, because each of these concepts has been used in systems that perpetrate oppression and inequality by asserting a system of intrinsic hierarchy. In systems of explicit and implicit racism, European facial features are privileged over non-European features, such as those seen as African, Asian, or indigenous peoples (MacMaster 2001). In colorism, lighter skin tones are similarly either more desired or more associated with desirable traits, relative to darker skin tones (Hunter 2007; Ghavami, Katsiaficas and Rogers 2016).

II Data

School libraries serve as major purveyors of sanctioned visual content for children. The books they offer are accompanied by an implicit state-sanctioned stamp-of-approval. These books are generally targeted towards specific age groups, ranging from picture books to print-only books. They are deliberately chosen and curated by librarians and school administrators, and are often selected because they transmit clear narratives about appropriate conduct, an account of important historical moments, or other, often identity-specific mes-
sages. For the purposes of our analysis, children’s books also serve as a useful test case for image analysis tool development because many contain both illustrations and photographs. By drawing from a set of materials that has a broad range of image types, we are able to develop more flexible face detection and feature classification models that can recognize a diverse set of images.

Our data come from a set of children’s books considered to be of high literary value and likely to be found in school libraries. We use books that received awards administered or featured by the Association for Library Service to Children, a division of the American Library Association (ALA). Our sample comprises 1,133 books, and each book in this sample is associated with one of 19 different awards.\(^4\)

We focus on these books because they are highly influential and common features of many children’s experiences. Winners of the Newbery and Caldecott awards are often considered part of the “canon” of children’s literature (Koss, Johnson and Martinez, 2018; Koss and Paciga, 2020). As a result, they are regularly featured in school, public, and home libraries (Smith, 2013; Koss and Paciga, 2020). Furthermore, the winners of the broader set of awards are commonly featured in numerous venues that are part of children’s learning experience, from book fairs and catalogues to school curricula and summer reading lists (Knowles, Knowles and Smith, 1997; Koss, Johnson and Martinez, 2018).

We divide these award-winning corpora up into two primary “collections”: “Mainstream” and “Diversity.” Figure 1a presents the full list of corpora in our sample and the collection(s) in which they have been categorized. Figure 1b shows the sample size of each collection by decade.\(^5\)

**Mainstream Collection.** The “Mainstream” collection comprises books that have received either Newbery Honors or Caldecott Honors, the two oldest children’s book awards in the United States. The Newbery Medal, which was 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, which was first awarded in 1938, is given to illustrators of “the most distinguished American picture books for children.” These books are explicitly chosen for their literary quality and not their popular appeal. Books receiving these awards are considered to be of general interest to all children and are quickly incorporated into mainstream outlets for children, such as school libraries (ALSC, 2007).

---

\(^4\) The 19 award corpora are comprised of 3,153 total books which either won an award or received an honorable mention; we obtained and digitized 1,133 of these books using both library and online resources. Twelve books in our sample won multiple awards.

\(^5\) Appendix Figure A1 shows the sample size of each corpus by decade.
Diversity Collection. The “Diversity” collection is a set of books featured by the ALSC that purposefully highlight the representation of typically excluded or marginalized identities. These books are also likely to be placed on “diversity lists” during events such as Black History Month or Women’s History Month. We are particularly interested in the efficacy of these books in highlighting the identity on which they focus, as well as other identities beyond their intended focus.

This collection includes books that have received the following awards: Amelia Bloomer, American Indian Youth Literature Award, Américas Award, Arab American Book Award, Asian/Pacific American Award for Literature, Carter G. Woodson Book Awards, Coretta Scott King Book Awards, Dolly Gray Award, Ezra Jack Keats Book Award, Middle East Book Award, Notable Books for a Global Society, Pura Belpré Award, Schneider Family Book Award, Skipping Stones Honor Awards, South Asia Book Awards, Stonewall Book Awards, and Tomas Rivera Mexican American Award. The first of these awards was the Coretta Scott King Awards created in 1970 specifically to highlight African American writers, partly because no such writer had received either the Newbery or Caldecott Medals as of that point. Other awards in this collection were created more recently, such as the South Asia Book Awards in 2012.

We also create smaller collections that highlight specific identity areas: people of color, African American people, females, people with disabilities, and lesbian, gay, bisexual, transgender, and queer (LGBTQ) people. Figure 1 lists the specific awards whose books comprise each of the collections that we analyze.

While different awards begin in different years, we do not limit the analysis to years in which all awards have books in the sample. The use of books persists over time, and it may be just as likely, if not more likely, for someone to select a book considered to be a “classic” (typically an older book) rather than to select a book more recently published.

The 1,133 books in the sample results in 164,482 pages of content. It would be cost-prohibitive to analyze this much content, let alone larger bodies of potential content that practitioners or policymakers might consider for inclusion in curricula, through manual content analysis techniques. As outlined in section VIII, a back-of-the-envelope calculation suggests that it would cost over $200,000 to analyze all the pages in our sample using traditional content analysis methods. The large volume of images and text these pages contain highlight the need for automated tools to analyze them. In Sections III and IV, we describe the machine-led techniques we apply and develop to analyze these books.

---

6We selected those children’s book awards featured on the ALSC website, many of which are administered by different organizations.
We present summary statistics in Table 1. This shows key information about each collection, such as the number of years the award has been in existence, as well as information about the books within, including on the length of the book (number of pages, number of words contained) and summaries of the measurements we describe in the following sections.

III Images as Data

In this section we describe our development and application of software tools to measure the content of images. The maxim “a picture is worth 1,000 words” speaks to the fact that an image often contains a wide range of messages. Despite this fact, images are not widely used as a source of data in the social sciences. This is in stark contrast to the use of text as data, which has seen substantial attention in the past decade (Gentzkow and Shapiro, 2010; Gentzkow, Shapiro and Taddy, 2019; Kozlowski, Taddy and Evans, 2019). A main contribution of this paper is to address this gap by introducing, applying, and developing tools for the computer-led analysis of the content of images.

Perhaps the first message people take from an image is that of representation: namely, who is contained in an image. The tools that we develop and apply in this paper identify faces of characters contained in images – both photographs and illustrations – and classify their skin color, race, gender, and age. In this section, we will describe the component processes: identifying characters’ faces, then identifying the color of the skin, and, separately, classifying their race, gender, and age. We depict this process in Figure 2, and refer to it as our image analysis pipeline.

III.A Face Detection

Our first step in converting images to data is to use computer vision tools to identify the faces of characters in each image. Specifically, we trained a custom transfer learning model to classify images or detect objects using Google’s AutoML Vision (Zoph and Le, 2017). Transfer learning facilitates the use of a pre-trained model as a “shortcut” to learn patterns from data on which it was not originally trained. AutoML is an artificial-intelligence-based technology for conducting “automated machine learning.” We create a model by giving it a series of labeled data sets, training the tool to recognize and identify patterns in images. The AutoML tool algorithmically optimizes its performance of classifying

---

7Computer vision involves teaching a computer to view and interpret images as a human would. Face detection is a subset of computer vision, in addition to identifying features such as colors, objects, and emotions.

8A neural network is a program designed to model the way the human brain functions, wherein the computing system is given examples and learns to perform tasks by analyzing them, usually without being given explicit instructions.
features via fine-tuning the neural networks it uses.

Classifying the identity of people represented in images is a complex problem because of the wide variance in the way people can be represented, and because of the wide variance in the characteristics of the images in children’s books which potentially further complicates these efforts. First, the images in these books consist of both illustrations and photographs. This is notable, in particular, because the existing models we considered were trained exclusively on photographs. This raised a potential concern that they might undercount illustrations. Second, these images are also likely to show both human and non-human characters. These characters could have human skin colors (different shades of beige), non-typical skin colors (such as blue or green), or monochromatic skin colors (such as greyscale). Finally, characters could be shown in different poses such as facing the viewer, shown in profile, or facing away from the viewer.

To address the potential undercounting of characters in illustrations, we trained a transfer learning face detection model (FDAI) using a manually-labeled data set of 7,000 illustrated faces drawn from two sets of books that contain a wide variety of illustrated characters. To train the model, we split the data set into training (80% of the data), validation (10% of the data, used for hyper-parameter tuning), and test (10% of the data, used for evaluating the model).

Two parameters are commonly used to evaluate the performance of this class of model: “precision” and “recall.” Take a label L for some arbitrary characteristic. Precision tells us the proportion of images to which the model assigns the label L that, in actuality, possess this characteristic. Recall, on the other hand, tells us the number of images to which the model does not assign the label L that, in actuality, possess the characteristic. Formally:

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

The higher the precision, the fewer false positives the model produces. On the other

---

9 More specifically, AutoML seeks the best network architecture and optimal hyper-parameter configuration in order to minimize the loss function of a model. Google’s AutoML Vision is based on neural architecture search and transfer learning technologies.

10 We used books in the Newbery and Caldecott corpora. A face was manually labeled if it could be easily observed. If a face was not detectable by a human, then we assumed it would not be easily detected by a machine. There were on average three detectable faces in each labeled image. We refer to our face detection model as FDAI (face detection using AutoML trained on illustrations).

11 Sometimes “recall” is also referred to as “sensitivity.”
hand, the higher the recall, the fewer false negatives the model produces. In other words, recall tells us, from all the test examples that should have had the label assigned, how many were actually assigned the label.

Our face detection model has 93.4% precision 76.8% recall. So in this case, 6.6 percent of the faces we identify may not, in truth, be faces (a false positive), while the model may neglect to identify one in 4.5 “true” faces (a false negative).

III.B Image Feature Classification: Skin Color

In this subsection, we introduce and describe our method for measuring the skin color of the faces of the characters our model identifies. Many existing artificial intelligence models have a feature which classifies the putative race of a character – meaning the race that a member of society might assign to a person based on a picture of their face. Unfortunately, these models have a high error rate, both misclassifying the race of identified people and, in “one-shot” models that identify existence of people and their race simultaneously, misclassifying people as non-human (Nagpal et al. 2019, Krishnan, Almadan and Rattani 2020, Fu, He and Hou 2014).

Thus, we measure the representation of a racial construct in images by focusing on a more well-defined parameter: the color of the skin of the characters pictured in these images. In addition to being a more clearly defined dimension that helps proxy for the broader construct of race, skin color is also, in itself, an important aspect of human categorization and is likely to be an immediate feature of an image that a viewer is likely to notice.

We develop a novel method to classify the skin color of these characters. Our skin color classification method involves a three-step process: (1) skin segmentation, (2) identifying the dominant colors in the identified skin, and (3) measuring the perceptual tints of the dominant colors and using this measure to label the skin color of each face as “Dark,” “Medium,” or “Light.” Figure 2 illustrates this process.

III.B.1 Skin Segmentation: Fully-Convolutional Conditional Random Field

We first isolate skin components of the character’s facial and non-facial skin using convolutional neural networks (CNN)\(^{12,13}\). Traditional skin segmentation methods assign

\(^{12}\)A convolutional neural network (CNN) is a multilayer, fully connected neural network, often used for machine-led image analysis.

\(^{13}\)Specifically, we use a Convolutional Neural Network (CNN) cascade which parses the skin from the detected face via a fully-convolutional continuous Conditional Random Field (CRF) neural network (Zhou, Liu and He 2017). To do so, we used the trained model proposed in Jackson, Valstar and Tzimiropoulos (2016) to automatically conduct semantic segmentation of the facial skin in which we adapt code from Lu (2018) for parsing skin and from Beyer (2018) for CRF post-processing.
a skin or non-skin label for every pixel of the cropped face image in which skin features are extracted. These labels are assigned using traditional image processing methods such as thresholding, level tracing, or watershed. These methods, however, face a number of challenges such as the need to take into account skin color (in)consistency across variations in illumination, acquisition types, ethnicity, geometric transformations, and partial occlusions (Lumini and Nanni 2020). To deal with these issues, we isolate skin from non-skin parts of the detected face using a deep learning approach called Fully-Connected Convolutional Neural Network Continuous Conditional Random Field (FC-CNN CRF).

This skin segmentation method (FC-CNN CRF) comprises three steps. First, we apply a fully-convolutional neural network (FCN), which is a type of convolutional neural network (CNN) where the last fully-connected layer is substituted with a convolutional layer that can capture locations of the predicted labels. This allows us to predict periphery landmarks such as the edges of the facial skin area, eyes, nose, and mouth. Second, we then use these predicted landmarks to extract a “mask” for the targeted facial region using the convex hull function in SciPy’s Python library. Third, we refine this mask by applying a continuous conditional random field (CRF) module, which predicts the labels of neighboring pixels (i.e., whether they are predicted to be skin or not skin) to produce a more fine-grained segmentation result. The resulting mask provides the segmented skin that we can then use to classify skin color. In Figure 2a, we show an example of this process of detecting a face and then segmenting the skin region to isolate the main facial area of interest.

### III.B.2 Skin Color Classification: k-means clustering

After segmenting the skin in a detected face, we then identify the most dominant colors showing in the identified skin using k-means clustering. k-means clustering is a traditional unsupervised machine learning algorithm whose goal is to group data containing similar features into $k$ clusters. Specifically, k-means clustering partitions all the pixels in the “segmented” skin into $k$ clusters, each pixel being assigned to the cluster with the nearest mean. For our analysis, we use 5 clusters (i.e., where $k$ takes a value of 5).

Next, we convert the $k$ colors from RGB space (which is the color space used in the k-means clustering output) into the L*a*b* color space. L*a*b* is a perceptually uniform

---

14 An equivalent term for this is Fully-Convolutional Continuous Conditional Random Field. “Fully-Convolutional” implies fully-connected CNN in this case.

15 Conditional random field (CRF) is a class of statistical modeling that uses a probabilistic graphical model.

16 Clustering entails partitioning data points into a small group of clusters.

17 We used the $k$-means clustering function implemented in the scikit-learn Python library Sculley (2010). Researchers use $k$-means clustering for different applications.
space that expresses color as three values: \( L^* \) for perceptual tint, and \( a^* \) and \( b^* \) for the four unique colors of human vision: red, green, blue, and yellow. This conversion allows us to interpret a given numerical change in the color values as a similar perceived change in color.

After this conversion, we take the weighted average of the \( k \) colors where the weights represent the relative amount that each \( k \) color contributes to the skin segmented face. By separately identifying the top \( k \) colors detected and taking their weighted average, we can more accurately classify the actual skin color depicted and minimize misclassification due to irregularities or idiosyncrasies in the image, for example, from the presence of shadows or hair that the skin segmentation process failed to remove.

### III.B.3 Skin Color Classification: Perceptual Tint

After collapsing the top \( k \) colors to their weighted average, we separate them into three categories: (1) monochromatic skin colors (e.g., black & white, sepia), (2) polychromatic human skin colors (e.g., brown, beige), and (3) polychromatic non-typical skin colors (e.g., blue, green). In the RGB color space, the closer the R, G, and B values are to each other, the less vibrant the color. For this reason, we classify a face as monochromatic if the standard deviation between the R, G, and B values associated with the weighted average of the face’s top \( k \) skin colors is less than \( T \). So a given face \( i \) is classified as monochromatic using the following equation:

\[
\text{Monochromatic}_i = \mathbb{1} \left[ \sqrt{\frac{(R_i - \mu_i)^2 + (G_i - \mu_i)^2 + (B_i - \mu_i)^2}{3}} \leq T \right]
\]

Where \( \mu \) is equal to the average of the R, G, B values.

To choose a threshold \( T \), we took the following steps. First, we manually labeled a random sample of 2,836 detected faces (stratified by collection) as either monochromatic or polychromatic. Then we found the mean squared error between our predicted labels using the equation above for every integer value of \( T \) between 0 and 100. Then we averaged these mean squared errors over 1,000 bootstrapped samples. The threshold that minimized the mean squared error on average is given by \( T = 13 \). We find this method to be 82.9% accurate on average.

After separating the monochromatic faces from the polychromatic faces, we then separate the human skin colors from the non-typical skin colors. This is important because some of the faces detected in our children’s books are non-human and may have colorful skin.

\[^{18}\text{A more common term for } L^* \text{ is “perceptual lightness,” but to decenter and de-emphasize “lightness” or “brightness” relative to “darkness,” we refer to the concept as “perceptual tint,” or “tint.”}\]
tones (for example, aliens, monsters, or characters found in Dr. Seuss books). Using the R, G, and B values associated with the weighted average of a face’s top \(k\) skin colors, conditional on the face not being monochromatic and if \(R > G > B\), we classify the skin color of the face as a human skin color \cite{Vezhnevets, Sazonov and Andreeva, 2003}. Otherwise, it is classified as a non-typical skin color. We find this method to be 82.1% accurate using the same set of 2,836 manually labeled faces that we use to test the accuracy of our monochromatic classification methods.

To classify how light or dark a monochromatic, polychromatic human, or polychromatic non-typical skin color is, we use the perceptual tint, or L* value, associated with the average of the \(k\) colors in L*a*b* space. This value ranges from 0 to 100 where 0 is the color black and 100 is the color white, and there is a range of colors in between. We split this range of possible L* values in thirds and classify skin colors as belonging to dark, medium, or light tercile. These classifications allow us to estimate the proportion of light-skinned, medium-skinned, and dark-skinned characters in each book. In the absence of data on race, we use both the continuous measure of perceptual tint along with our dark, medium, or light skin tone classifications as a proxy for race.

### III.C Image Feature Classification: Race, Gender, and Age

In this section, we discuss how we classify race, gender, and age of detected faces in images. We build a method for the analysis of race, gender, and age by training a multi-label classification model using Google’s AutoML Vision. Due to the large amount of manually labeled data necessary to train these deep learning models and due to the fact that there are no public data sets using illustrations, we use transfer learning to predict gender and age for both photographs and illustrations in our children’s books using a model trained on photographs. To train this model, we used the UTKFace public data set \cite{Zhang and Qi, 2017}, which contains over 20,000 photographs of faces with manually verified age, gender, and ethnicity labels.\(^{19}\) We trained a model using all three labels. To train the model, we split the data set into three parts: training (80% of the data), validation (10% of the data), and test (10% of the data). The resulting model has 90.6% precision and 88.98% recall. In other words, 9.4 percent of the images assigned a given race, gender, or age label will, in truth, not possess that trait (a false positive), while 11 percent of the images not assigned the label for that trait would, in truth, possess it (a false negative).

Because feature classification is subject to error, we calculate confidence intervals for

\(^{19}\) The labels in the data set include: Gender (male or female), Age (infant (0-3), child (4-11), teenager (12-19), adult (20-64), senior (65+)), Race (Asian (a combination of Asian and Indian), Black, White, and others (e.g. Latinx, Middle Eastern).
these feature estimates using the feature classification model’s estimates of precision and recall. Recollect that the model produces fewer false positives when the precision is higher and produces fewer false negatives when the recall is higher.

We can then use precision to generate lower bounds and recall to generate upper bounds. The lower bound comes from subtracting the number of false positives from the estimate, and the upper bound comes from adding the number of false negatives to the estimate. The number of false positives comes from multiplying the estimate and the imprecision \((1 – \text{precision})\) and then subtracting this product from the estimate. The number of false negatives comes from multiplying the complement of the estimate – or the estimate of those not classified as having the trait – by the lack of recall \((1 – \text{recall})\) and then adding this product to the estimate. In equation form, take a parameter \(\mu\) which is the estimated proportion of faces classified as a given gender. The lower and upper bounds, \(\mu_L\) and \(\mu_U\), would be calculated:

\[
\mu_L = \mu - \{\mu \times (1 - \text{precision})\}
\]

\[
\mu_U = \mu + \{(1 - \mu) \times (1 - \text{recall})\}
\]

The main drawback of this model is that it was trained on photographs while the majority of the faces in our children’s books are illustrations. In order to train a model to more precisely predict the race, gender, and age of faces detected in illustrations, we would ideally have a manually labeled dataset of illustrated faces to use as training data.

**Race Classification (Images).** The model assigns the probability that a detected face is of a given race category; we aggregate these into four categories: Asian, Black, Latinx + Others, White. We assign race for a given face based on which race category has the highest predicted probability.

**Gender Classification (Images).** For each face detected, we predict the probability that the face is female- (or male-) presenting. We also label a face as female if the predicted probability that the face is female presenting is greater than 50 percent, and similarly we label a face as male using the same 50 percent probability threshold.

---

20 A more comprehensive confidence interval would capture two stages of misclassification: misclassification of whether or not an image contains a face and misclassification of the labels assigned to detected faces. Our method takes into account the latter, but future work could account for the former.

21 In a random sample, 84.2% of the detected faces were illustrated not photographs.

22 The race labels in the original model are: Asian, Black, Indian, Others, where “Others” includes Latinx and Middle Eastern, and White. We combine Asian and Indian predictions into a broader Asian category.

23 We compare these predictions to a manually labeled random sample of 2,836 detected faces and present the results in the Data Appendix.
We recognize that these classifications imperfectly identify the performative aspect of gender presentation, as they are trained based on how humans classify images. We apply these methods from a reflective perspective, validating them and documenting the frequency and nature of misclassifications. As we progress, we wish to incorporate fluid and nonbinary gender identities in our analyses.

Age Classification (Images). The model assigns the probability that a detected face is of a given age category (infant, child, teenager, adult, senior). We aggregate these categories into two bins: child and adult. We sum the probabilities for infant and child for the child bin and for teenager, adult, and senior for the adult bin. We then assign age for a given face based on which bin is above 50 percent.

IV Text as Data

We describe the tools we use to measure representation in the text of books. Social scientists have analyzed the messages contained in text of printed material for centuries (Neuendorf, 2016; Krippendorff, 2018). Recent work by economists and sociologists showcases how the computational speed and power of (super)computers can be harnessed to conduct automated text analysis, greatly accelerating work traditionally done by hand (Gentzkow, Kelly and Taddy, 2019; Kozlowski, Taddy and Evans, 2019). In this paper, we draw from this work and, in particular, a series of natural language processing tools that take bodies of text – e.g., from a book – and extract various features of interest. In Figure 2b, we show our process of extracting text from digitized books and then analyzing it; we refer to this as our “Text Analysis Pipeline.”

The first step in conducting this analysis is to extract text from digital scans of books. We conduct this extraction using Google Vision Optical Character Recognition (GVOCR). We input the raw files into GVOCR, which then separately identifies images and text (e.g., illustrations and photographs) in each file. It then applies its own OCR software to the text sections of the scans, generating the text data we analyze.

24We compare these predictions to a manually labeled random sample of 2,836 detected faces and present the results in the Data Appendix.

25There are other commonly used OCR interfaces. However, over the past five years, researchers have consistently identified Google Cloud Vision OCR as the best technology for converting images to text. In one study, Tafti et al. (2016) compare the accuracy of Google Docs (now Google Vision), Tesseract, ABBYY FineReader, and Transym OCR methods for over 1,000 images and 15 image categories, and found that Google Vision generally outperformed other methods. In particular, Google Vision’s accuracy with digital images was 4 percent better than any other method. Additionally, the standard deviation of accuracy for Google Vision was quite low, suggesting that the quality of OCR does not drastically change from one image to the next. A test of OCR tools by programmers compared the performance of seven different OCR tools (Han and Hickman, 2019). This analysis also found Google Vision to be superior, specifically when extracting results from low resolution images. In another study that focused on comparing results from multiple image
We clean these raw text data to remove erroneous characters and other noise generated by the OCR process, increasing the precision of our measurement of features in the text. The cleaning process removes numerical digits and line breaks but maintains capitalization, punctuation, and special characters. It also standardizes the various permutations of famous names (for example, all variations of MLK become “Martin Luther King Junior”).

From these text data, we then extract several features. These features include: token (single word) counts, the presence of famous people, and the first names of non-identifiable characters. Next we describe how we use these features to construct measures of the representation of gender identity, racial constructs, and age in each book.

IV.A Text Analysis: Token Counts

One branch of traditional content analysis consists of enumerating words that represent a particular attribute (Krippendorff, 2018). We generated a set of tokens associated with identities related to gender, race, or age. We aggregate counts of these words by their respective identity category (such as female or male) by book, generating our “token count” measures of the representation of each identity in each book (Neuendorf, 2016).

Gender (Token Counts). To calculate gender representation in token counts, we calculate the proportion of words with a gendered meaning that refers to females. For our main analysis, we combine specific gendered words (e.g., queen, husband, daughter, and son) with gendered pronouns (e.g., him, her, she, and he).

We show how gender representation varies on three additional dimensions: one, whether the gendered identity is represented by individuals (singular) or groups (plural); two, whether the character is placed as the subject or object of a sentence; and three, by age. To analyze singular and plural representation separately, we separate gendered tokens into those referring to singular cases (e.g. daughter) and plural cases (e.g. daughters). To analyze whether the character is the subject or object of a sentence, we generate counts of formats (including .jpg, .png, and .tif). Vijayarani and Sakila (2015) found that Google surpassed all other OCR tools. We also tested OCR using ABBYY FineReader and Google Tesseract. Our comparison of their performance relative to manual coding also showed GVOCR performed the best.

Tokens are a maximal sequence of non-delimiting consecutive characters. In our context, a token is an individual word. The vocabulary used for each of these lists is available in the Data Appendix.

We use the spaCy library to generate these counts, but we see similar patterns in our findings when we use NLTK instead.

Traditional content analysis often restricts gendered words to pronoun counts. We show the sensitivity of our findings related to this construct by restricting the analysis to gendered pronouns only in Appendix Figure A3. Our results are robust to this alternate specification.

We calculate the total number of words in a book by removing all punctuation from the text and then dividing the text into a list of words using either the “nlp” package in the spaCy library or the “word_tokenize” package in the NLTK library. The length of this list provides the total number of words in a given text.
the number of gendered pronouns that are capitalized versus lowercase, under the theory that an individual who is the subject of a sentence is in a position of more active importance than the same character when used as the object and thus occupying a more passive role. To analyze representation of gender by age, we generate a list of “younger” gendered words (e.g. princess, boy) and “older” gendered words (e.g. queen, man).

*Color (Token Counts).* As another proxy for race, we calculate the proportion of all words that refer to colors (e.g. blue, black, or white).

### IV.B Text Analysis: Named Entity Recognition

We also wish to measure the representation of gender and race among named characters in these stories, be they fictional or historical. To do so, we use a tool from Natural Language Processing called Named Entity Recognition (NER) \(^{30}\). NER identifies and segments “named entities,” or proper nouns, starting with a pre-defined library of such entities and also identifying new entities through the application of neural nets. NER recognizes these entities in strings of text and tags each entity with a type, such as person, location, or date. Applying NER to our data, we identify these entities and count how many times each specific named entity is mentioned in a given book. We then associate these frequency counts with traits of the person, such as their race, gender, or place of birth. There are two types of named entities that we identify: (1) famous characters (described below) and (2) first names of characters.

#### IV.B.1 Identifiable famous people

To identify the instances of famous characters represented in books, such as Martin Luther King, Jr. or Amelia Earhart, we match all of the entities identified by NER that have at least two names (for example, a first and last name) with a pre-existing data set called Pantheon 2.0 which contains data from over 70,000 Wikipedia biographies that have a presence in more than 15 language editions of Wikipedia (Yu et al. 2016). This method generates a data set of over 6,000 famous people. We count the number of unique books each famous person is mentioned in as well as the number of times they are mentioned in each book.

*Gender and Birthplace (Famous People).* The Pantheon 2.0 data set also contains data on the gender and place of birth of these famous people, which we include in our

\(^{30}\)We run our NER analysis using the open-source software library spaCy, which employs convolutional neural networks for both text categorization and NER. Another commonly-used library for NER is NLTK, but it only recognizes single words for NER, whereas spaCy can recognize strings of words as a distinct entity. For example, “Martin Luther King” would be recognized as one entity in SpaCy but as three entities with NLTK (“Martin,” “Luther,” and “King”).
Race (Famous People). We then manually code race for each identified person. This coding is subject to the individual biases and perceptions of each human coder and is imperfect. We collapse the following identities: East Asian, Middle Eastern, and South Asian into the Asian category; North American indigenous peoples and South American indigenous peoples into the Indigenous category; and African-American and Black African into the Black category. If an individual was coded as having more than one race, they were then classified as multiracial.

IV.B.2 Character First Names

We also study the representation of gender among people who are named but not identified as “famous” using the methods described above. Using the named entities identified by the spaCy NER engine, we limit the sample to those entities categorized as a person and remove the famous characters we found by applying the process described in Section IV.B.1. We then categorize the remaining named entities and construct a data set containing the name of each unique character and the number of times that character is mentioned in a given book. Below we describe how we classify gender representation in this data set of names. We do not identify race directly using first names. This is because it is notoriously difficult to do so; even cutting-edge NLP methods have found it virtually impossible to distinguish between first name / last name classifications of Black and White people, for example (Garg et al., 2018).

Gender (Character First Names). First we identify all characters that have a gendered title such as “Señora Cuervo,” “Uncle Robin,” or “Queen Swan.” We then use these gendered titles to predict the gender of that character (the list of gendered words we use here is the same list that we used in the token counts). For the remaining characters, we extract the first name and estimate the probability that the character is female using data on the frequency of names by gender in the US population from the US Social Security Administration. For example, if a character’s first name is “Cameron,” our estimated probability that the character is female is 9.16 percent because that is the proportion of people named “Cameron” in Social Security data who are female. We limit our sample of Social Security data to include only years which overlap with the years in our sample of children’s data. If the predicted probability that a character is female is greater than 50 percent, we label that character as female.

---

31 The Pantheon 2.0 curators run a classifier over the English text of the Wikipedia biographies to extract information such as place of birth and gender from each biography. Their classifier was trained on a data set called Pantheon 1.0 (Yu et al., 2016) which contains a subset of bios which had been more manually curated.

32 The entity categorization (e.g. person, location, etc.) is not always correct, so there may be entities misclassified or missed overall. We do not use this categorization when identifying famous characters.
female. Otherwise, the character is labeled as male. Using this method, we are able to make gender predictions for approximately 150,000 entities. To test how accurate these predictions are, we predicted the gender of each famous person in our data using their first names and compared these predictions to their gender identified using Wikipedia and found that our predictions were 96.35% accurate.

We are not able to make a prediction for the remaining named entities. For example, characters such as “New Yorker” which the spaCy NER engine identified and labeled as a person will not receive a prediction because “New” is not a name that appears in our Social Security data.

IV.C Text Analysis: All Gendered Mentions

We aggregate all gendered mentions (gendered tokens, predicted gender of character first names, and matched gender of famous characters) to generate a composite measure of gender representation in text.

V Measures of Representation used in the Analysis

When collapsing our data to the collection (or collection by decade) level, we first collapse each variable to the book level and then find the averages across books in a given collection. For example, to find the average probability that a detected face in a book belonging to the Mainstream collection is female presenting, we first find the average probability that a face is female presenting over all the faces in each book in the collection and then take the average across books. This way, our measures of race, gender, and age representation in each book are equally weighted – meaning books with more faces do not receive more weight in the collection averages. We describe these measures below and in Table 2.

V.A Variables of analysis

In this section, we describe the ways in which we measure race, gender, and age in images and text.

Race Representation. We measure racial constructs through: (1) skin color classification of detected character faces, (2) race classification of detected character faces, (3) manually coded race of famous figures, (4) birthplace of famous characters, and (5) counts of words relating to ethnicity and, separately, color word token counts. We do not attempt to classify the race of non-famous characters mentioned in our text data. Other recent text analysis has shown that conventional methods for classifying race using names fail to successfully distinguish between Black and White non-famous characters appearing in a different body of books (Garg et al., 2018).
Gender Representation. We measure representation of gender identity through: (1) gendered pronoun counts, (2) gendered word counts, (3) gender classifications of famous characters, (4) predicted gender of characters based on their first name, and (5) predicted gender of detected character faces.

Age Representation. We measure representation of age through: (1) age-by-gender word counts and (2) predicted age of detected character faces.

V.B Presenting our results

We present results summarizing our measurements in the following ways:

Proportion of entities with a given trait. For categorical variables, we show bar charts and other figures indicating the book-level average proportion of entities – such as famous people or images – with a given trait, or in a given category representing a given identity. For example, in images, we show the average proportion of faces in these images that are classified as female; in text, we show the average proportion of gendered words that refer to females in books.

Distributions. For continuous variables, we show probability density plots of the proportion of the observations – be they words, images, or books – in each collection across possible values of the variable. For example, we plot the proportion of faces across the gradient of skin colors, and we plot the proportion of books which have a given proportion of gendered words that are female.

Patterns over time. We also plot how collection-level proportion variables change over time, using line graphs to show the evolution of decade-specific estimates for these variables.

VI Results

In this section, we present our results characterizing the representation of race and gender in the images and text of the books in our collections. First, we present patterns of the representation of racial constructs (race, origin, and skin color). We then present patterns of the representation of gender identity. We conclude the section with a discussion of representation of race, gender, and age.

VI.A Representation of the Construct of Race

In this section, we discuss our measurement of a series of variables capturing the broader, hard-to-measure construct of race in images and text. We characterize three traits which characterize the latent, human-perceived construct of race – skin color, putative race, and birthplace – as discussed in Sections I, III, and IV. Our measures include: (1) the skin color of detected character faces in images, (2) the predicted race of detected character faces
in images, (3) classification of the putative race of famous figures named in the text, (4) the birthplace, or place of origin, of famous figures named in the text, and (5) token counts of color words.

Skin color of faces. We first report our estimates of the representation of race in images, focusing on the skin color of a character’s face. Perhaps the most direct form of representation is what a child sees in the visual portrayals of people in the images shown in a book, particularly before a child becomes textually literate.

In Figures 3 and 4, we show patterns in representation across collections, time, and across collections over time. This analysis reveals several important patterns. First, Figures 3a and 4a show the distribution of perceptual tint for the Mainstream and Diversity Collections. These figures show that, regardless of image type, the faces in the Diversity Collection have darker average skin tones than those in the Mainstream. Second, Figures 3b and 4b show that, over time, the proportion of medium and darker skin tones is increasing relative to that of lighter skin tones, both for the Mainstream and Diversity Collections.

Figures 3c and 4c show a different version of the distributions presented in Figures 3a and 4a, classifying each face into one of three terciles: Darker, Medium, or Lighter skin. For both Mainstream and Diversity Collections, the medium skin color tercile is the most represented, with more than 50 percent of faces in both collections falling in this tercile. In the Mainstream Collection, however, lighter skin is in the second most common skin color tercile (roughly 25 percent of faces), while in the Diversity Collection, darker skin comprises the second most common skin color tercile (roughly 35 percent of faces). This suggests that the Diversity collection is more representative of characters that have darker skin tones.

Figures 3d and 4d show the analog of Figures 3c and 4c, but for each of the collections in our data. This shows that the Mainstream Collection has the lowest proportion of faces falling in the darker skin color tercile of all collections, and that the African American Collection has the greatest proportion. The Mainstream collection has a much higher proportion of faces with medium skin colors, with a far more pronounced mass in the distribution around a modal beige tone than the Diversity collection, wherein the distribution of skin colors is more dispersed. This racial chromatic ambiguation is sometimes referred to as a “butterscotching,” which some may argue sends an assimilationist message regarding the representation of race (Yoon, Simpson and Haag, 2010).

We also show the skin colors of the individual faces we detect in the images in the books in each collection. We show these separately for polychromatic images and for monochromatic (e.g., black and white, sepia). In Appendix Figure A4 we show the total number
of faces detected in each collection-by-decade cell. We show our results for the polychromatic images, and in Appendix Figure A5 we show the monochromatic analogue. We find that in the earlier decades of the Mainstream Collection, there was a greater proportion of monochromatic images, with a general trend over time to have more polychromatic images. In the Diversity Collection, and in particular the People of Color Collection, there is a consistently high proportion of monochromatic images, perhaps representing the use of historical black-and-white photographs.

Figures 3d and 4d present the tercile-level results for all of the collections over time. We see increased representation of medium and darker skin tones in books in collections not only recognized for highlighting the experiences of people of color – and especially those that highlight the experiences of Black people – but also by those in the Female collection.

Race of detected characters. We then examine the predicted race of characters in the images. Figure 5 shows that within a given race, the Mainstream collection is likely to show characters within a given race as lighter than their counterparts in the Diversity collection, which is consistent with the notion of “butterscotching.” We see in Figure 6 that the characters that are detected are overwhelmingly classified as being White males or females.

Race of famous figures. We then examine the race of the famous figures mentioned in the text. In Appendix Figure A11 we show the proportion of famous figures in each collection identified as Asian, Black, Indigenous, Latinx, Multiracial, or White. We find that, in all collections, the famous figures mentioned are predominantly White. In the Mainstream Collection, over 90 percent of famous figures are White. The African American Collection is the only collection to have a majority identity other than White represented. Other groups appear far less frequently. Black people are the next-most represented, comprising 51 percent of the famous people in the African American collection, and eight to 24 percent in the other collections. Famous people of Asian, Latinx, Indigenous and Multiracial identities account for between two and 16 percent of famous people combined, a high level of inequality in

33 As we see in Appendix Figure A6, which shows these estimates for polychromatic “non-typical” skin colors, the method of classifying “human” vs. “non-typical” skin colors may underestimate the number of darker-skinned faces if the browns that are used do not follow the R > G > B rule. However, Appendix Figure A7 shows that this does not change the patterns in skin color representation by collection over time.

34 Appendix Figure A8 shows that most pictured characters are classified as being White. Appendix Figure A9 shows a substantial portion of pictured characters predicted to be female-presenting, and Figure 6 suggests that most of these pictured characters are White females. We map these shares on their respective shares of the US population in Appendix Figures A10a and A10b.

35 Recent work reporting conventional content analysis of the race of main characters in the Caldecott and, separately, Newbery award-winning books finds qualitatively similar results (Koss, Johnson and Martinez, 2018; Koss and Paciga, 2020).
representation relative even to population averages.  

We also examine how representation of race in famous figures varies by gender in Figure 7, which shows the proportion of famous figures in each race-by-gender category. We find that the vast majority of unique characters in all collections are White males. Even in the African American collection, there are more famous White males than there are Black males (the next most represented group in this collection). In all other collections, White males are 45 percent or more of all famous people mentioned. The next most represented groups are White females (10-32 percent of famous people) and Black males (4-36 percent of famous people). The representation of Black females (between one and 11 percent of famous people, except in the African American collection, where they comprise 16 percent) is consistently less than that of Black males, despite their roughly equal shares in the population. This highlights that even within collections of books curated to highlight a given racial identity, race and gender are often treated as mutually exclusive categories of experience, overlooking the representation of groups such as Black females whose identities lay at the intersection of multiple experiences of exclusion.

In Appendix Table A1, we plot the five most frequently mentioned famous people overall, listing their race and gender. The most commonly mentioned person in the Mainstream Collection is George Washington; in the Diversity Collection, it is Martin Luther King Junior. For the Mainstream collection, all five of the most commonly mentioned people are white males. For the Diversity Collection, all five are males and three are white (Abraham Lincoln, George Washington, and Martin Luther), with the other two are Black (Martin Luther King Junior and Langston Hughes). Even in the female collection, the three most commonly mentioned people are males (John F. Kennedy, Jimmy Carter, and Martin Luther King Junior) the fourth is a female (Betty Friedan) and the fifth is also a male (Richard Nixon).

We then explore whether these trends in racial representation of famous people track the US population share of different races in Figure 8. In the Mainstream collection, White people have been consistently overrepresented, and Black people and Latinx people underrepresented, relative to their population share in the US.

**Birthplace of famous figures.** We then examine representation of famous figures in terms of their place of origin, which is another contributor to the construct of race. Exposure

---

36 The US Census Bureau, for example, estimates that roughly 60 percent of the US population is non-Latinx White (US Census 2019).

37 Appendix Tables A2 and A3 show this for the top five females and top five males, respectively, uniquely mentioned in each collection.
to real people from different parts of the world can expand a child’s understanding about other countries beyond narrow stereotypes. We show the distribution of birthplaces of famous figures mentioned in the Mainstream and Diversity Collections in Figure 9, which presents a map with a dot for each birthplace to demonstrate the representation of national and subnational identities presented to children. This figure shows that Mainstream Collection books primarily feature famous figures from Europe and the eastern portion of the United States. By contrast, Diversity Collection books feature famous figures from across the world and, more precisely, an order of magnitude more famous people from South America, Africa, and Asia.

We next analyze how the birthplace of famous people presented in these books varies by gender, another way of studying the representation of intersectionality. We show these results in Appendix Figure A12. We find that males have more diverse representation in terms of birthplaces than females across both the Mainstream and Diversity collections. Females that are represented are far more likely to be from North America (primarily the United States) and Europe than males, who, particularly in the Diversity Collection, come from many more parts of the world.

Words related to ethnicity and color. We next examine the construct of race in text by counting the proportion of words related to color (e.g., black, white, and blue) in these collections. This method, while more straightforward than our other analyses, serves as a barometer for these other measures and helps us understand what a simpler approach to content analysis might have yielded. Appendix Figure A13 measures the incidence of ethnicity words such as Kenyan, Indian, and Canadian, in addition to color words such as black, white, and blue over time. The collection of books that recognize the Black or African-American experience is much more likely to mention the words black and white. We then look at mentions of non-race colors such as red and blue as a falsification exercise. They are consistently a negligible proportion of words overall; this remains relatively constant over time.

Our findings highlight an important pattern related to the notion of intersectionality. We observe two related phenomena: one, the failure of the collections focusing on race and ethnicity to be equitably gender-representative; and two, the relatively lower performance of the collections which focus on diversity not related to race and ethnicity to center the experiences of race and ethnicity in their books. These phenomena underscore a key pattern we find throughout this analysis: the low representation of intersectional experiences – for example, the experience of Black women – even in collections which are deliberately chosen for their diversity in representation.
VI.B Representation of Gender Identity

We characterize the representation of gender using (1) the numbers of gendered tokens in the text, (2) the predicted gender of character first names in the text, (3) the gender of famous figures in the text, and (4) the predicted gender classification of the detected character faces in images. We present patterns for these measures of representation, and conclude by showing how the representation of gender in images and text compares.

All gendered words. We first report the patterns for an aggregated measure of the textual representation of gender, which includes all gendered tokens and all gender classifications for mentioned names. In Figure 10, we present estimates of the book-level proportion of the gendered words and characters which are female. In Figure 10a, we show the distribution of these estimates for the Mainstream and Diversity collections. In Figure 10b, we show this distribution for each of the collections, including the smaller collections that highlight specific identities. The patterns here show that in all but the Female collection, the central tendency of the distribution is skewed towards more male representation. In both panels, we observe that the Mainstream collection is the most male-skewed of all the collections.

In Figures 11a, 11b, and 11c, we present a numerical accounting of the proportion of female words relative to all gendered words. The main pattern we observe is that, for all collections except those books specifically recognized for highlighting females, fewer female words are present than male words. Figure 11a shows that the proportion of female words in these collections is between 35 and 44 percent, as opposed to 59 percent in the Female collection. Figure 11b presents the ratio of male words to female words as opposed to just the proportion and documents that this pattern of greater male representation is consistent across time for collection and decade book averages. Figure 11c shows that this proportion is gradually increasing over time but remains below the US population share of females for all collections in every decade, except for the Female collection.

One might imagine that there could be differences by type of gendered word. For example, until recently, grammar rules dictated that male pronouns would be used as “gender-neutral” pronouns, which would then lead us to overstate the male representation in these books. However, the pattern holds when the analysis is restricted to each type of gendered word: pronouns, specific gendered tokens such as “princess” and “prince,” gendered first names of characters, or gender of famous people mentioned (Appendix Figure A3).

These patterns of discrepancy in the representation of gender in text are consistent

---

38 We describe the methods for each of these in Section IV (1 – 3) and Section III (4).
39 Recall that the Diversity collection is the aggregate of multiple smaller collections which focus on centering different diverse identities. The classification is shown in Figure 1.
across other measures of gender representation, whether they are represented as individuals or groups of females vs. males (Appendix Figure A14), or if they are represented as the subject (as opposed to the object) of a sentence (Appendix Figure A15).

**Famous figures.** A related but distinct parameter is the number of unique famous figures mentioned in these books. The specific people who are named in a book transmit more implicit information to a child beyond generic tokens. By naming these individuals, they take on a greater significance to children and can represent an entity to which a child might aspire, as in the role model effects studied in [Dee (2005)](Dee2005) and [Porter and Serra (2019)](PorterSerra2019), for example. We show our collection-specific estimates of this parameter in Figure 11d. On this dimension, inequality in representation of gender is much more severe. In the Mainstream Collection, on average 84 percent of the unique historical figures mentioned in a book were male, for example. Even the Female Collection ceases to be more representative of women than of men (not even one-third of the unique historical figures were women on average). Furthermore, two collections (LGBTQ, diversity) contain similar average proportions of unique characters who are female as the Female Collection.

**Gender of characters.** Next, we describe the representation of gender in the images of these books. We show the proportion of faces in each collection identified as female in Appendix Figure A9a. In the majority of the collections, fewer than half of the detected faces are classified as female-presenting. In the Female Collection, however, we classify 71 percent of the faces as female, and in the Ability Collection, we classify 64 percent of the faces as female. Appendix Figure A9b shows that, unlike for text, there is no obvious trend in gender representation in images, within collections over time.40

Next, we examine the representation of skin color by gender. In Appendix Figure A17, we show the perceptual tint of faces, separated by their detected gender. In images, we find no evidence of a (perceptible) difference between classified females and males in terms of the frequency of different skin tones represented.

**Seen more than heard?** We then compare representation of gender across images and text. In Figure 12a, we show a scatterplot of collection-by-decade average proportions of female words on the x-axis and the average proportion of female-presenting faces on the y-axis. It shows that for females, the representation of females in gendered words is less equal than the representation of females in images. In other words, females are more likely to be visualized (seen) than mentioned in the text (heard). This suggests that authors or illustrators may perfunctorily include additional females in pictures in order to give the

---

40In Appendix Figure A16, we show a similar pattern when using a continuous measure of the average probability that a face is classified as being female.
appearance of equity while not actually having them play an important role in the story. Importantly, it also shows that on average, females are represented less than half of the time in both images and text. Figure 12b shows the converse of this for males, underscoring that they are represented more than half of the time in both images and text.

VI.C Representation of Age

Finally, we briefly discuss the representation of people by age in the images and text of our books. In Figure 13a, we show the age classifications of gendered words (e.g., girl vs. woman and boy vs. man). This shows that, in most books, the distribution of young people by gender is roughly similar, though in the Female Collection, girls are roughly twice as likely to appear than boys. For words specific to gendered adults, however, men are always more likely to appear. This discrepancy is largest in the Mainstream and African American Collections, where adult men are roughly 60 percent of adult gendered mentions and adult women only 40 percent.

In Figure 13b we show detected character faces by age and gender. Similar to in text, images of males dominate in most collections, though in no case is the discrepancy as extreme as it is in gendered adults in text. In the Female and Ability Collections, there are more females than males. Regardless of gender, in both images and text, we show that there are more adults than children depicted in the books in each collection. We also see in Appendix Figure A10: that adults are overrepresented relative to their share in the census. This raises a question as to why adult experiences or depictions are privileged in books targeted to children.

We also study how the representation of skin tone across gender varies by age. In Figure 14a, we show plots of the distribution of skin color tints indicating that females are not depicted differently than males by age, patterns similar to what we see in Appendix Figure A17. We see, however, that when children are depicted in images, they are more likely to be shown with a lighter skin tone than adults, regardless of collection, as seen in Figure 14b. This pattern is inconsistent with evidence that skin color may actually lighten with age, the result of a biological decrease in melanin levels over one’s lifetime (Sarna et al., 2003). There are many potential interpretations of this pattern, but a particularly concerning one is that brightness may be used to connote innocence (e.g., of childhood), supernatural features (e.g., of angels), or another type of emphasis in order to separate the character from

---

41 One concern may be that the age classification algorithms are primarily trained on adult faces, and therefore overclassify adults; however, we see consistent ratios of adult presence to children presence in text and in images.

42 One concern could be that the algorithms are trained to classify faces as being more likely to be a child if the skin color of the detected face is lighter, which then would attenuate the number of children detected.
VII AI is Only Human

Our paper brings a set of artificial intelligence tools to bear on the field of content analysis. These tools are powerful, computer-driven methods. They are designed by humans and, in many cases, trained with initial human input. We use them because they offer a few key advantages. The first is scale: because algorithms are automated, they allow for analysis of a much larger set of content than would be possible using conventional, “by hand” methods. The second is adaptability: we can rapidly change one dimension of measurement and re-run the analysis at low cost. Were we to do this via hand-coding, the cost would increase linearly with each addition or adjustment (see Section VIII); with AI-based analysis, the marginal cost of such additions or adjustments is much lower.

Measuring representation in content via any means will generate some errors in measurement. In traditional content analysis, analysts may misclassify some images or text. If this occurs at random, this can be treated as standard measurement error, which would be captured via estimating inter-rater reliability (Neuendorf 2016 Krippendorff 2018). If, however, traits of the analyst systematically influence their coding, then error from misclassification may be non-classical, leading to a bias in expectation (Krippendorff 1980). This can arise, for example, if an analyst’s identity (e.g., one’s race and/or gender) causes them to classify content differently than analysts of different identities (Boer, Hanke and He 2018).

These same biases appear in AI models. Many AI models, including those we use, are trained using a set of data which are first labeled by humans. Furthermore, nearly all models are either fine-tuned, evaluated, or both, based on their performance relative to human classification. As a result, the bias in classical content analysis is “baked into the pie” for computer-driven content analysis (Das, Dantcheva and Bremond 2018).

Most face detection models are trained using photographs of humans - particularly White humans, which would lead us to undercount people of color and illustrated characters. To address this, we trained our own face detection model using approximately 7,000 faces from the Caldecott and Newbery corpora (discussed in Section III.A). This same issue could be repeated in the skin segmentation process where we did manually segment the skin region of faces but rather relied on the convolutional neural networks.

These issues persist when classifying features. In the case of gender, for example, all public data sets with labels for gender that we encountered have a binary structure, limiting classification to “female” or “male,” and neglecting to account for gender fluidity or nonbinary identities. Furthermore, intrinsic to these models is the general assumption that
we can predict someone’s gender identity using a region of interest in an image of their faces (Leslie 2020).

We work to minimize the risk of this by ensuring that our human classification is checked multiple times by analysts of different racial and gender identities. Nonetheless, in the absence of an extremely high volume of training data coded by multiple analysts, the scope for systematic misclassification remains.

However imperfect, we believe this analysis is crucial and, via deliberate attention to the potential sources and remedies for these types of error, we hope to minimize the risk of such bias influencing our results.

While AI is a product of and therefore reflects human biases, this problem is also intrinsic to traditional “by-hand” content analysis. Manual coding necessarily can only reflect the biases of the individual coders. We observed that the identities of the manual labelers on our team led to non-classical error, particularly in the classification of the race. We therefore use multiple measures for each identity to try to understand the extent of this potential measurement error. For example, in addition to manually coded putative race of famous figures, we examine two other constructs of race - birthplace of famous figures and skin color of detected characters.

While we use AI tools to study representation, we end this section by emphasizing that AI and manual coding need each other. The tools we use are meant to rapidly estimate how a human might categorize these phenomena. They are motivated by human perception and, ultimately, their performance is also evaluated based on how accurately they can determine how a human might perceive the representations in text and images. Our use of these tools depends crucially on human input at each stage, from the conception of tools and the labelling of training data, to the evaluation of the tools’ accuracy and the way that we interpret its results. We see our efforts to expand computer-driven content analysis as a natural extension of the rich history of human-driven analysis in this field, adding the strengths of recent advances in computational science to this crucial, human-driven endeavor.

VIII Tool Validation / Cost-Effectiveness

Drawing from validation theory, we conducted traditional manual content analysis to validate our measures (Kane 2013; Neuendorf 2016). To do so, we hand-coded representations in a sample of 30 short stories and poems for children written and illustrated by a variety of authors and illustrators. This helped us to evaluate the plausibility of our measures and also identify messages our tools fail to detect, clarifying limitations of computer-led content analysis.
The stories and poems that we hand-coded were drawn from a 3rd grade reading textbook published in 1987. It took approximately 33–40 hours to code the entire book (400 pages at an average of 5 to 6 minutes per page). While the length of time needed to code “by hand” varies with the grade level of the books in our sample, we estimate that it would have taken us approximately 13,585 – 16,302 hours to hand-code the 164,482 pages in our sample of children’s books. At an hourly wage of between $15 and $20, this work would have cost anywhere from $205,000 to $329,000.

While it was certainly much cheaper to measure the gender and race representations using computer vision/machine learning tools, the data collected by the hand-coders was much more detailed. For example, we collected data on the role of each character in a story such as whether they are a protagonist, antagonist, or support character. Appendix Figure A18 shows the breakdown of character roles by gender and race in our hand-coded children’s stories. Our team also hand-coded the actions/adjectives/adverbs associated with each character, number of times each character is pictured/illustrated, character impairments such as being blind or deaf, hair color/length, eye color, character sexuality, story genre & perspective (such as 1st person or 3rd person), along with many other variables. Using this data we can learn about not only the quantity of race and gender representation but also the quality of representation or how characters from each group are portrayed.

Regardless of whether we use manual coding or computer vision, the broad patterns are the same. Over 50% of the characters and gendered words are male and the skin colors depicted are skewed away from darker-skinned individuals. These results comparing the hand coded representations to the computer vision representations are shown in Appendix Figures A19 and A20.

IX Summary and Concluding Remarks

The books we use to educate our children teach them about the world in which they live. The way that people are – or are not – portrayed in these books demonstrates who can inhabit different roles within this world and can shape subconscious defaults. Historical and persistent inequality, both by race and gender and in other dimensions, can be either affirmed or challenged by what we teach children about the world. While many educators and schools wish to eliminate materials that have overt racial and gender bias and use content that promotes positive messages about all people, such efforts are necessarily piecemeal and the judgments behind them subjective. Per the adage “a picture is worth a thousand words,” images in particular convey numerous messages. We hope that our work to systematize the measurement of their content can help us gather important information about the messages implicitly and explicitly being sent to children through these visual depictions
which, previously, could not have been quantified systematically.

In this paper, we make two primary contributions. First, we introduce machine-led computer vision methods to convert images into data on skin color, gender, and age of pictured characters. Second, we apply these image analysis tools – in addition to established natural language processing methods that analyze text – to award-winning children’s books to document the representation to which children have been exposed over the last century. We analyze these books by the purposes of their award categories, broadly categorizing them as Mainstream collections if they were selected without explicit intention to highlight a specific identity and as Diversity collections if they were deliberately chosen for a given award because of their focus on underrepresented groups. We then further create smaller collections that highlight specific identities.

These image analysis tools show that books selected to highlight people of color or females increasingly depict characters with darker skin tones over time. However, books in the Mainstream collection primarily depict medium skin tones, potentially trying to appeal to the assumed preferences of the median reader. These Mainstream books have increased representation of lighter skin tones over the last two decades despite increased rhetoric about the importance of representation. Moreover, we see that children consistently have been more likely to be depicted with lighter skin than adults, which is contrary to the patterns present in actual society where there are not systematic differences in skin tones across ages.

We compare the patterns found in images to the patterns found in text. We see that females are more likely to be represented in images than in text over time, consistent with the maxim that women should “be seen but not heard.” This suggests there may be symbolic inclusion in pictures without substantive inclusion in the actual story. Males, especially White males, are persistently more likely to be represented by every measure, with little change over time despite substantial changes in female societal participation.

Our approach is not without its limitations. First, artificial intelligence tools reflect the biases of the human coders that trained the models, but not moreso than traditional content analysis completely done manually. Second, the measures of representation that we use are imperfect. The current measures of gender identity neglect measurement of non-binary and gender-fluid identities. Race is a multifaceted construct of human categorization that is ill-defined and difficult to measure. Third, the algorithms do not perfectly detect faces or isolate the skin from faces, thus leading to measurement error. Fourth, our analysis consists of a numerical accounting of different characters through simple representational statistics, i.e. whether characters are included. However, if a character is depicted in a reductive
or stereotypical manner, then solely the existence of representation will be insufficient and possibly counterproductive. Thus, future work should further develop tools that can measure how people are represented and help ascertain what content depicts characters in their full humanity.

The “optimal” level of representation is a normative question beyond the scope of this paper, but the actual representation in books is something that can be measured and, given some reasonable set of goals, improved upon. To achieve any progress toward such goals, practitioners and publishers require mechanisms to systematically measure and compare the amount and type of representation in the content they consider for inclusion in curriculum or even for prospective consideration for publication.

A systemic problem such as this requires a systemic solution. Our tools will directly contribute to lasting improvement of the practice of education by helping guide curriculum choices and prospectively assessing representation in the creation of new content. Our tools, we hope, will also allow social scientists across a wide range of fields to systematically use printed content – images, as well as text – as primary source data. This work could, for example, describe patterns of representation in other important bodies of content and, subsequently, study how variation in representation shapes human beliefs, behavior, and outcomes. Finally, these methods can be applied to the study of other text and visual media, from print-based and online news to television and film. As a whole, we hope that this work expands our understanding of the measurement, levels, and impacts of diversity in content; in so doing, we contribute to the crucial, ongoing work that aims to overcome the structural inequality that pervades society and our daily lives.
References


Han, Ted, and Amanda Hickman. 2019. “Our search for the best OCR tool, and what we found.”


Lu, Conny. 2018. “Github: Face segmentation with CNN and CRF.”


Steele, Claude M. 2010. Whistling Vivaldi: And other clues to how stereotypes affect us (issues of our time). WW Norton & Company.


Zoph, Barret, and Quoc V. Le. 2017. “Neural Architecture Search with Reinforcement Learning.”
Table 1. Summary Statistics

<table>
<thead>
<tr>
<th>Collection Totals</th>
<th>Mainstream</th>
<th>Diversity</th>
<th>People of Color</th>
<th>African American</th>
<th>Ability</th>
<th>Female</th>
<th>LGBTQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of Books</td>
<td>495</td>
<td>638</td>
<td>580</td>
<td>131</td>
<td>29</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>Averages Over All Books</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Pages</td>
<td>141</td>
<td>149</td>
<td>138</td>
<td>148</td>
<td>213</td>
<td>314</td>
<td>268</td>
</tr>
<tr>
<td>Number of Words</td>
<td>29,402</td>
<td>34,894</td>
<td>31,713</td>
<td>35,722</td>
<td>47,260</td>
<td>97,062</td>
<td>75,979</td>
</tr>
<tr>
<td>Number of Faces</td>
<td>44</td>
<td>52</td>
<td>53</td>
<td>38</td>
<td>30</td>
<td>30</td>
<td>73</td>
</tr>
<tr>
<td>% of Female People</td>
<td>11</td>
<td>13</td>
<td>13</td>
<td>16</td>
<td>7</td>
<td>36</td>
<td>14</td>
</tr>
<tr>
<td>% of Female Words</td>
<td>34%</td>
<td>44%</td>
<td>43%</td>
<td>41%</td>
<td>44%</td>
<td>61%</td>
<td>49%</td>
</tr>
<tr>
<td>% of Female Faces</td>
<td>48%</td>
<td>49%</td>
<td>48%</td>
<td>42%</td>
<td>64%</td>
<td>71%</td>
<td>44%</td>
</tr>
<tr>
<td>% of Female Famous People</td>
<td>14%</td>
<td>24%</td>
<td>22%</td>
<td>26%</td>
<td>27%</td>
<td>34%</td>
<td>42%</td>
</tr>
<tr>
<td>% of Asian Famous People</td>
<td>2%</td>
<td>6%</td>
<td>7%</td>
<td>0%</td>
<td>1%</td>
<td>9%</td>
<td>4%</td>
</tr>
<tr>
<td>% of Black Famous People</td>
<td>5%</td>
<td>21%</td>
<td>22%</td>
<td>51%</td>
<td>11%</td>
<td>24%</td>
<td>8%</td>
</tr>
<tr>
<td>% of Indigenous Famous People</td>
<td>0%</td>
<td>1%</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>% of Latinx Famous People</td>
<td>1%</td>
<td>8%</td>
<td>9%</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
<td>2%</td>
</tr>
<tr>
<td>% of White Famous People</td>
<td>92%</td>
<td>62%</td>
<td>59%</td>
<td>47%</td>
<td>86%</td>
<td>66%</td>
<td>83%</td>
</tr>
<tr>
<td>% of Multiracial Famous People</td>
<td>0%</td>
<td>2%</td>
<td>2%</td>
<td>1%</td>
<td>0%</td>
<td>1%</td>
<td>3%</td>
</tr>
</tbody>
</table>

Note: In this table we present summary statistics (described in the row titles) for each collection of books we analyze (named in the column titles).
<table>
<thead>
<tr>
<th>Measure</th>
<th>Image</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race</td>
<td>• Skin color</td>
<td>• Race of famous figures</td>
</tr>
<tr>
<td></td>
<td>• Predicted race of face</td>
<td>• Birthplace of famous figures</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Color token counts</td>
</tr>
<tr>
<td>Gender</td>
<td>• Predicted gender of face</td>
<td>• Pronoun counts</td>
</tr>
<tr>
<td></td>
<td>• Probability of gender of face</td>
<td>• Gendered token counts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Gender of famous figures</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Predicted gender of first names</td>
</tr>
<tr>
<td>Age</td>
<td>• Predicted age of face</td>
<td>• Age-by-gender token counts</td>
</tr>
</tbody>
</table>

Note: In this table we show the different methods we use to measure race, gender, and age in the text and, separately, in the faces in the images in children’s books.
Note: This figure shows the main sources of data we use for our analysis. In Panel A, we list the book awards whose books we collected and digitized, along with the collections into which we group them in our analysis. In Panel B, we show the sample size, in terms of the number of books we have in each collection, over time.
Figure 2. Workflow Pipeline

(a) Image Analysis Pipeline

Face Detection Using Google AutoML Vision (FDAl)

Skin Detection
Remove hair, clothes, background, etc.

Skin Colors
Five predominant skin colors (weighted by prevalence)

Skin Color Classification

FC-CNN

k-Means Clustering

Weighted Average

Dark

Gender Prediction

Female

Male

Age Prediction

Adult

Child

Race Prediction

Asian

Black

Latinx + Others

White

Note: In this figure, we show how we process scanned book pages into image and text data. In Panel A, we show how we extract image data and classify skin color, gender, age, and race. In Panel B, we show how we extract and isolate various dimensions of text, such as race of names or words related to gender.
Figure 3. Representation of Skin Colors in Images, Human Skin Colors

(a) Distribution of Skin Colors

(b) Average Percent of Each Skin Color Tercile

(c) Average Percent of Each Tercile, All Collections

(d) Proportion of Faces in Each Tercile

Note: This figure shows our analysis of the skin color in the faces we detected in the books we analyze, focusing on “polychromatic” (i.e., non-monochromatic) images. Panel A shows the distribution of skin color tint for images in books in the Mainstream and Diversity collections. In Panels B-D, we categorize each skin color into three terciles: darker, medium, or lighter skin based on perceptual tint. Panel B shows the overall collection-specific proportion of faces in each skin color tercile for these two collections. Panel C shows this same proportion for each of the seven collections. In Panel D, we show the proportion of faces, over time, for faces in the Mainstream and Diversity collections. Skin classification methods are described in Section [III] and the note to Figure [A4]. We limit this analysis to typical human skin colors. (polychromatic skin colors where R > G > B). We remove non-typical skin colors from this analysis; the analog of this analysis for monochromatic images is given in Figure [I].
Figure 4. Representation of Skin Colors in Images, Monochromatic

Note: This figure shows our analysis of the skin color in the faces we detected in the books we analyze, focusing on “monochromatic” images. Panel A shows the distribution of skin color tint for images in books in the Mainstream and Diversity collections. In Panels B-D, we categorize each skin color into three terciles: darker, medium, or lighter skin based on skin tint. Panel B shows the overall collection-specific proportion of faces in each skin tint tercile for these two collections. Panel C shows this same proportion for each of the seven collections. In Panel D, we show the proportion of faces, over time, for faces in the Mainstream and Diversity collections. Skin classification methods are described in Section III and the note to Figure A4. The analog of this analysis for polychromatic images is given in Figure 3.
Figure 5. Skin Color and Race Predictions of Pictured Characters

(a) Skin Color x Race, Human Skin Colors

(b) Skin Color x Race, Monochromatic

Note: This figure shows the distribution of skin color tint by predicted race of the detected faces in the mainstream and diversity collections. Skin tint is determined by a weighted average of the five most predominant colors of a detected face; the most predominant colors are extracted using $k$-means clustering with $k=5$. Non-typical skin colors removed.
Figure 6. Race and Gender Predictions of Pictured Characters

<table>
<thead>
<tr>
<th>Race</th>
<th>Asian</th>
<th>Black</th>
<th>Latinx + Others</th>
<th>White</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainstream</td>
<td>4.3%</td>
<td>2.1%</td>
<td>0.9% 1.1%</td>
<td>37.7%</td>
</tr>
<tr>
<td>Diversity</td>
<td>10.3%</td>
<td>5.4%</td>
<td>5.3% 8.1%</td>
<td>29.7%</td>
</tr>
<tr>
<td>People of Color</td>
<td>10.9%</td>
<td>5.8%</td>
<td>5% 8.8%</td>
<td>28.3%</td>
</tr>
<tr>
<td>African American</td>
<td>7.5%</td>
<td>3.5%</td>
<td>8% 14.9%</td>
<td>23.2%</td>
</tr>
<tr>
<td>Ability</td>
<td>3.1%</td>
<td>2.7%</td>
<td>7% 0.9%</td>
<td>47.2%</td>
</tr>
<tr>
<td>Female</td>
<td>9.1%</td>
<td>0.1%</td>
<td>17.3% 3.5%</td>
<td>43.4%</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>2%</td>
<td>2.6%</td>
<td>2.5% 1.2%</td>
<td>35.6%</td>
</tr>
</tbody>
</table>

Note: In this figure, we show the proportion of detected faces in all collections by race and gender predictions. Race and gender was classified by our trained AutoML model.
Figure 7. Race and Gender Classifications of Famous Figures in the Text

Note: In this figure, we show the percent of famous people in each collection by race and gender. We identify famous individuals and their predicted gender using methods described in Section IV.B.1. The race of these famous individuals was manually labeled. We collapse the following identities: East Asian, Middle Eastern, and South Asian into the Asian category; North American indigenous peoples and South American indigenous peoples into the Indigenous category; and African-American and Black African into the Black category. If an individual was coded as having more than one race, they were classified as multiracial.
Figure 8. Share of US Population and Unique Famous People in the Text, by Race

Note: In this figure, we show the percent of famous people in the Mainstream and Diversity collections by predicted race and the share of the US population by race. As described in Section IV.B.1, we classify famous people using a two step-process. The race of these famous people was manually labeled. We collapse the following identities: East Asian, Middle Eastern, and South Asian into the Asian category; North American indigenous peoples and South American indigenous peoples into the Indigenous category; and African-American and Black African into the Black category. If an individual was coded as having more than one race, they were classified as multiracial.
Figure 9. Birthplace of Famous Figures

Note: In this Figure, we show two maps, one for the Mainstream Collection and one for the Diversity Collection, plotting the distribution of the place of birth of the famous people in our books described in the note to Appendix Figure A11. We identify birthplace using a model run on scraped Wikipedia biographies collected by Pantheon (Yu et al., 2016). If the city/town they were born in was unavailable, we use birth country. Size of dots correspond to the number of unique famous characters born in a given location.
Figure 10. Distribution of All Female Words across Collections

(a) Mainstream vs. Diversity Collections

(b) All Collections

Note: In this figure we show the proportion of words which have gendered meaning that refer to females (i.e., “female gendered words”), by collection. Panel A shows this for the Mainstream and Diversity Collections; Panel B shows this for Mainstream Collection, Diversity Collection, and the separate collections which comprise the Diversity Collection. See Figure 1 for a description of which awards comprise each collection. We use a pre-specified list of gendered words that includes: pronouns; gendered titles; gendered references; the gender classifications of identified famous people; and gender predictions of first names of characters. A full list of the gendered words we use is given in the Data Appendix.
Figure 11. Female Words as a Percent of all Gendered Words

(a) Percent Female Words

(b) Male Words vs. Female Words

(c) Female Words & US Population Share

(d) Percent Famous Females

Note: In this figure, we show four different analyses of gendered words. Panel A shows the proportion of gendered words that are female in each collection. Panel B shows how this value varies by decade. Panel C shows the number of female words vs. the number of male words by collection and decade book averages, as opposed to just the proportion. Panel D shows a slightly different parameter: the proportion of famous people in the books who are female.
Figure 12. Women Should be Seen More Than Heard?

(a) Percent Female Faces Detected vs. Female Words & Characters

(b) Percent Male Faces Detected vs. Male Words & Characters

Note: In this figure we contrast the representation of females in the text of these collections of books with representation of females in the images of the same books. In Panel A, we plot collection-by-decade averages of female representation in images (on the y-axis) and female representation in text (on the x-axis). On the y-axis, we plot the average percent of female faces out of all faces detected. On the x-axis, we plot the average percent of gendered words which are female. Panel B shows the inverse to Panel A: the proportional representation of males in images and text. We detect faces using a Google Vision AutoML model trained on illustrations. Within these faces, we classify gender using an AutoML algorithm trained using a manually labeled random sample of our data, assigning the female value to all faces receiving the female label with a prediction value of greater than 50 percent.
Figure 13. More adults than children: Age and Gender Classifications of Detected Character Faces

(a) Percent of Faces Detected by Age Group and Gender

(b) Gendered Words by Age and Gender

Note: In this figure we show analysis of the representation of age and gender. In Panel A, we show analysis of age and gender in images. Specifically, we plot the proportion of identified faces classified in each age (adult vs. child) and gender (female vs. male) category. We detect faces using a Google Vision AutoML model trained on illustrations. Within these faces, we classify age and gender using an AutoML algorithm trained using a manually labeled random sample of our data. We assign the female value to all faces receiving the female label with a prediction value of greater than 50 percent. In Panel B, we show analysis of age and gender in text. Specifically, we plot the proportion of words that refer to specific gender-age combinations (e.g., female children or adult men) as a percent of all gendered words in the book. These words were populated with a pre-specified list given in the appendix.
Figure 14. Children are consistently depicted with lighter skin compared to adults: Skin Color by Age by Gender Classifications, All Collections

(a) Skin Color and Age, by Gender

(b) Skin Color and Age of Faces Detected

Note: In this figure, we show analysis of how the representation of skin color varies with the age and gender of the person being represented. In Panel A, we show the distribution of skin tint values of detected faces, across all seven collections, by the classified gender (female vs. male) and age (adult vs. child) of the face. In Panel B, we show a similar analysis, but we classify the skin tint values into three terciles: dark, medium, and light. We detect faces using a Google Vision AutoML model trained on illustrations. Within these faces, we classify age and gender using an AutoML algorithm trained using a manually labeled random sample of our data. We classify a face as being female if the prediction value was greater than 50 percent. We limit this analysis to typical human skin colors. (polychromatic skin colors where R > G > B). We remove non-typical skin colors from this analysis.
Appendices

A.1 Appendix Figures

Figure A1. Corpora Sample Size by Decade

Note: In this figure we show the number of scanned books in our sample, by decade, for each book award.
Figure A2. Skin Segmentation Challenges

(a) Examples of Skin Segmentation of Monochromatic Cartoon Faces

(b) Examples with Incorrect Skin Segmentation

Note: This figure demonstrates our skin segmentation techniques applied to a set of challenging faces from our sample.
Figure A3. Female Word Percentage Representation, by Type of Word Representation

(a) Pronouns
(b) Specific Words
(c) Character First Names
(d) Famous Figures

Note: In this Figure we show four analyses of the percent of female words across collections. These analyses are done separately by category of word. In Panel A, we show the percentage of gendered pronouns which are female. In Panel B, we show the percentage of words from a pre-specified list of gendered words (tokens) which are female; this includes pronouns, gendered titles, and gendered references. In Panel C, we show the percentage of character first names that are predicted to be female based on Social Security Administration data. In Panel D, we show the percentage of mentions of famous figures which are female.
Note: In this Figure, we show collection-by-decade distributions of the color of the skin in the faces we identify in the images in these books. As described in Section III, we use a Google Vision AutoML model trained on illustrations to classify faces in images. Skin color is determined by a weighted average of the five most predominant colors of a detected face; the most predominant colors are extracted using $k$-means clustering with $k=5$. This figure shows only “polychromatic” images of human skin colors; we remove non-typical skin colors (methods discussed in Section III) and all monochromatic (i.e. black and white) images. We plot the analog figure for monochromatic images in Figure A5.
Figure A5. Face Colors over Time, Monochromatic

Note: In this Figure, we show collection-by-decade distributions of the color of the skin in the faces we identify in the images in these books. This figure shows only monochromatic images of human skin colors (e.g., black and white); we remove non-typical skin colors (methods discussed in Section III) and plot the analog figure for non-monochromatic (i.e., “polychromatic”) images in Figure A4. See the note for that figure for further discussion of our methods.
Figure A6. Face Colors over Time, Polychromatic, Non-Typical Skin Colors

Note: In this figure, we plot the non-typical skin colors that were removed from Figures A4 and A5 by collection and decade.
Figure A7. Representation of Skin Colors in Images, Polychromatic Non-Typical Skin Colors

(a) Distribution of Skin Colors

(b) Average Percent of Each Skin Color Tercile

(c) Average Percent of Each Tercile, All Collections

(d) Proportion of Faces in Each Tercile

Note: In this figure, we plot the tint of non-typical skin colors in detected faces in our sample. These are colors which either do not follow $R > G > B$ or are not monochromatic. These are the colors that were removed from Figures 3 and 4 by collection and decade.
Figure A8. Most pictured characters classified as White

<table>
<thead>
<tr>
<th></th>
<th>Asian</th>
<th>Black</th>
<th>Latinx + Others</th>
<th>White</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainstream</td>
<td>6.4%</td>
<td>2%</td>
<td>3.9%</td>
<td>87.7%</td>
</tr>
<tr>
<td>Diversity</td>
<td>15.7%</td>
<td>13.4%</td>
<td>3.2%</td>
<td>67.7%</td>
</tr>
<tr>
<td>People of Color</td>
<td>16.7%</td>
<td>13.8%</td>
<td>3.1%</td>
<td>66.4%</td>
</tr>
<tr>
<td>African American</td>
<td>11%</td>
<td>22.9%</td>
<td>2.8%</td>
<td>63.2%</td>
</tr>
<tr>
<td>Ability</td>
<td>5.9%</td>
<td>8%</td>
<td>4%</td>
<td>82.1%</td>
</tr>
<tr>
<td>Female</td>
<td>9.2%</td>
<td>20.8%</td>
<td>0.6%</td>
<td>69.4%</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>4.6%</td>
<td>3.7%</td>
<td>5.2%</td>
<td>86.6%</td>
</tr>
</tbody>
</table>

Note: In this figure, we show our main analysis of race in images, reporting the proportion of faces in images which our model labels as a given race. We detect faces using a Google Vision AutoML model trained on illustrations. Within these faces, we classify race using an AutoML algorithm trained using a manually labeled random sample of our data, assigning the race value to all faces receiving the majority race label.
Note: In this figure, we show our main analysis of gender in images, reporting the proportion of faces in images which our model labels as female, as opposed to male. In Panel A, we show collection-level estimates of the percent of detected faces classified as female. In Panel B, we show these values over time. We detect faces using a Google Vision AutoML model trained on illustrations. Within these faces, we classify gender using an AutoML algorithm trained using a manually labeled random sample of our data, assigning the female value to all faces receiving the female label with a prediction value of greater than 50 percent.
Figure A10. Share of US Population and Pictured Characters, by Identity

(a) Race

(b) Gender

(c) Age

Note: In this figure, we show the share of the US population of specific identities mapped on the share of the pictured characters classified as a given identity over time. In Panel A, we show this by race. In Panel B, we show this by gender. In Panel C, we show this by age.
Figure A11. Race of Unique Famous People in the Text

<table>
<thead>
<tr>
<th></th>
<th>Asian</th>
<th>Black</th>
<th>Indigenous</th>
<th>Latinx</th>
<th>Multiracial</th>
<th>White</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainstream</td>
<td>2.4%</td>
<td>5.2%</td>
<td>0.1%</td>
<td>0.6%</td>
<td>0.2%</td>
<td>91.5%</td>
</tr>
<tr>
<td>Diversity</td>
<td>6.4%</td>
<td>20.8%</td>
<td>0.7%</td>
<td>8.4%</td>
<td>2%</td>
<td>61.7%</td>
</tr>
<tr>
<td>People of Color</td>
<td>6.8%</td>
<td>21.7%</td>
<td>0.8%</td>
<td>9.3%</td>
<td>2.1%</td>
<td>59.4%</td>
</tr>
<tr>
<td>African American</td>
<td>0.5%</td>
<td>51.4%</td>
<td>0%</td>
<td>0.2%</td>
<td>1.1%</td>
<td>46.9%</td>
</tr>
<tr>
<td>Ability</td>
<td>0.9%</td>
<td>11.1%</td>
<td>0%</td>
<td>1.3%</td>
<td>0.3%</td>
<td>86.3%</td>
</tr>
<tr>
<td>Female</td>
<td>8.7%</td>
<td>23.7%</td>
<td>0%</td>
<td>0.3%</td>
<td>1.2%</td>
<td>66.1%</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>4%</td>
<td>8.3%</td>
<td>0%</td>
<td>1.8%</td>
<td>3.2%</td>
<td>82.6%</td>
</tr>
</tbody>
</table>

Note: In this Figure, we show a separate measure of the representation of race in text: the proportion of famous people of different racial identities. We show these proportions for each collection, classified into six racial categories defined on the x axis. We identify famous individuals using methods described in Section IV.B.1. The race of these famous people was manually labeled. We collapse the following identities: East Asian, Middle Eastern, and South Asian into the Asian category; North American indigenous peoples and South American indigenous peoples into the Indigenous category; and African-American and Black African into the Black category. If an individual was coded as having more than one race, they were classified as multiracial.
Figure A12. Gender and Birthplace of Famous Figures

Note: In this figure, we show collection-specific measures of the birthplace of famous figures, separately for females and males. We identify famous individuals as well as their predicted gender and birthplace using methods described in Section [IV.B.1]. The size of the dots correspond to the number of unique characters born in a given location. We identify birthplace using a model run on scraped Wikipedia biographies collected by Pantheon [Yu et al., 2016]. If the city/town they were born in was unavailable, we use birth country. Size of dots correspond to the number of unique characters born in a given location.
Figure A13. Token-Based Proxies for Race: Ethnicity and Color

(a) Words Related to Ethnicity

(b) Color Words

Note: In this Figure, we show two measures of the representation of race in text: words related to ethnicity and words related to color. In Panel A, we show collection-specific averages of the proportion of words in a book that relate to ethnicity. In Panel B, we show collection-by-time averages of mentions of three color words: black, white, and blue – as a proportion of all words in our data. We generated the estimates using a pre-specified list of words (also known as “tokens,” as described in Section IV.A). We provide this list in the Data Appendix.
Figure A14. Females and Males as Individuals or Groups

Note: In this figure, we show how gender representation in text varies by whether it is a person or a group of people being referred to; in other words, whether the representation of gender varies by singular or plural representations. We show collection-specific averages by decade. In the top two plots, we show the percentage of singular female and male words as a proportion of all gendered words; the bottom two plots, we show the percentage of plural female and male words as a proportion of all gendered words.
Figure A15. Females and Males as Subjects and Objects in Sentences

Note: In this figure, we plot the representation of gender by its appearance in sentences. The top two plots show the average proportion of all gendered pronouns in a book that are uppercase and the bottom two plots show those that are lowercase. The left plots show the female-related pronouns, and the right plots show the male-related pronouns. We present these separately because an uppercase pronoun is more likely than a lowercase pronoun to be the subject, as opposed to the object, of the sentence in which it appears.
Figure A16. Female-Presenting Faces Detected Over Time, by Average Probability a Face is Female

Note: In this figure we a slightly different version of the analysis of gender in the images of books, across collections and decades. This shows the average probability that a face was classified as being female in a given collection by decade. This feature is given by the AutoML model. As mentioned in the note to Figure A9, in that figure we estimate that any face with a predicted value of more than 50 percent female is classified as female.
Figure A17. Gender and Skin Color: Distribution of Skin Color Tint by Gender

Note: In this figure we show the distribution of skin tint by gender in detected faces for each collection of books. Skin tint is determined by a weighted average of the five most predominant colors of a detected face; the most predominant colors are extracted using $k$-means clustering with $k=5$. Monochromatic skin colors and non-typical skin colors removed.
Figure A18. Hand-Coded Character Roles in a Sample of Children’s Stories & Poems

(a) By Character Gender

(b) By Character Race

Note: In this Figure we show hand-coded character roles from a sample of 30 short stories and poems written for children. In Panel A, we show the percentage of characters in each role by gender. In Panel B, we show the percentage of characters in each role by race (if they are a human character, otherwise they are coded as “Non-human.”)
Figure A19. Hand-Coded Measures of Representation

(a) Gender Representation

(b) Skin Color Representation

Note: In this Figure we show hand-coded measures of gender and race representations from a sample of 30 short stories and poems written for children. In Panel A, we show the percent of characters (either pictured/illustrated as in the top row or discussed in the text as in the bottom row) that were coded as male, female, or unspecified. In Panel B, we show the percentage of pictured human characters in each skin color classification based on the Fitzpatrick scale (Pathak et al., 1976).

Figure A20. Measures of Representation Collected Using AI & Natural Language Processing Tools

(a) Gender Representation

Note: In this Figure we show measures of gender representations collected using artificial intelligence and natural language processing tools from a sample of 30 short stories and poems written for children. We show female and male words as a percentage of all gendered words (e.g. pronouns, gendered tokens, and gendered character names).
### B.2 Appendix Tables

#### Table A1. Top Five Most Visible Famous People by Collection

<table>
<thead>
<tr>
<th>Collection</th>
<th>Rank</th>
<th>Name</th>
<th>Gender</th>
<th>Race</th>
<th>Mentions</th>
<th>Books</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainstream</td>
<td>1</td>
<td>George Washington</td>
<td>Male</td>
<td>White</td>
<td>182</td>
<td>32</td>
</tr>
<tr>
<td>Mainstream</td>
<td>2</td>
<td>Abraham Lincoln</td>
<td>Male</td>
<td>White</td>
<td>189</td>
<td>23</td>
</tr>
<tr>
<td>Mainstream</td>
<td>3</td>
<td>Thomas Jefferson</td>
<td>Male</td>
<td>White</td>
<td>74</td>
<td>15</td>
</tr>
<tr>
<td>Mainstream</td>
<td>4</td>
<td>Benjamin Franklin</td>
<td>Male</td>
<td>White</td>
<td>30</td>
<td>15</td>
</tr>
<tr>
<td>Mainstream</td>
<td>5</td>
<td>John Adams</td>
<td>Male</td>
<td>White</td>
<td>58</td>
<td>13</td>
</tr>
<tr>
<td>Diversity</td>
<td>1</td>
<td>Martin Luther King Junior</td>
<td>Male</td>
<td>Black</td>
<td>285</td>
<td>59</td>
</tr>
<tr>
<td>Diversity</td>
<td>2</td>
<td>Abraham Lincoln</td>
<td>Male</td>
<td>White</td>
<td>100</td>
<td>44</td>
</tr>
<tr>
<td>Diversity</td>
<td>3</td>
<td>Langston Hughes</td>
<td>Male</td>
<td>Black</td>
<td>187</td>
<td>41</td>
</tr>
<tr>
<td>Diversity</td>
<td>4</td>
<td>George Washington</td>
<td>Male</td>
<td>White</td>
<td>73</td>
<td>40</td>
</tr>
<tr>
<td>Diversity</td>
<td>5</td>
<td>Martin Luther</td>
<td>Male</td>
<td>White</td>
<td>63</td>
<td>36</td>
</tr>
<tr>
<td>People of Color</td>
<td>1</td>
<td>Martin Luther King Junior</td>
<td>Male</td>
<td>Black</td>
<td>271</td>
<td>56</td>
</tr>
<tr>
<td>People of Color</td>
<td>2</td>
<td>Abraham Lincoln</td>
<td>Male</td>
<td>White</td>
<td>99</td>
<td>43</td>
</tr>
<tr>
<td>People of Color</td>
<td>3</td>
<td>Langston Hughes</td>
<td>Male</td>
<td>Black</td>
<td>186</td>
<td>40</td>
</tr>
<tr>
<td>People of Color</td>
<td>4</td>
<td>George Washington</td>
<td>Male</td>
<td>White</td>
<td>70</td>
<td>38</td>
</tr>
<tr>
<td>People of Color</td>
<td>5</td>
<td>Martin Luther</td>
<td>Male</td>
<td>White</td>
<td>60</td>
<td>33</td>
</tr>
<tr>
<td>African American</td>
<td>1</td>
<td>Langston Hughes</td>
<td>Male</td>
<td>Black</td>
<td>80</td>
<td>21</td>
</tr>
<tr>
<td>African American</td>
<td>2</td>
<td>Martin Luther King Junior</td>
<td>Male</td>
<td>Black</td>
<td>107</td>
<td>19</td>
</tr>
<tr>
<td>African American</td>
<td>3</td>
<td>Rosa Parks</td>
<td>Female</td>
<td>Black</td>
<td>47</td>
<td>14</td>
</tr>
<tr>
<td>African American</td>
<td>4</td>
<td>Malcolm X</td>
<td>Male</td>
<td>Black</td>
<td>84</td>
<td>13</td>
</tr>
<tr>
<td>African American</td>
<td>5</td>
<td>Duke Ellington</td>
<td>Male</td>
<td>Black</td>
<td>22</td>
<td>13</td>
</tr>
<tr>
<td>Ability</td>
<td>1</td>
<td>Harold Pinter</td>
<td>Male</td>
<td>White</td>
<td>78</td>
<td>2</td>
</tr>
<tr>
<td>Ability</td>
<td>2</td>
<td>Andy Warhol</td>
<td>Male</td>
<td>White</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Ability</td>
<td>3</td>
<td>Marco Polo</td>
<td>Male</td>
<td>White</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Ability</td>
<td>4</td>
<td>Duke Ellington</td>
<td>Male</td>
<td>Black</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Ability</td>
<td>5</td>
<td>Emily Dickinson</td>
<td>Female</td>
<td>White</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Female</td>
<td>1</td>
<td>John F. Kennedy</td>
<td>Male</td>
<td>White</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Female</td>
<td>2</td>
<td>Jimmy Carter</td>
<td>Male</td>
<td>White</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>Female</td>
<td>3</td>
<td>Martin Luther King Junior</td>
<td>Male</td>
<td>Black</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>Female</td>
<td>4</td>
<td>Betty Friedan</td>
<td>Female</td>
<td>White</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>Female</td>
<td>5</td>
<td>Richard Nixon</td>
<td>Male</td>
<td>White</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>1</td>
<td>Alicia Keys</td>
<td>Female</td>
<td>Black</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>2</td>
<td>Britney Spears</td>
<td>Female</td>
<td>White</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>3</td>
<td>Julia Roberts</td>
<td>Female</td>
<td>White</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>4</td>
<td>Alexander Hamilton</td>
<td>Male</td>
<td>White</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>5</td>
<td>Sid Vicious</td>
<td>Male</td>
<td>White</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

Note: This Table shows the five most frequently mentioned famous people in each collection, along with their race, their gender, the number of times they were mentioned, and the number of books in which they appeared.
Table A2. Top Five Most Visible Famous Females by Collection

<table>
<thead>
<tr>
<th>Collection</th>
<th>Rank</th>
<th>Name</th>
<th>Race</th>
<th>Mentions</th>
<th>Books</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainstream</td>
<td>1</td>
<td>Eleanor Roosevelt</td>
<td>White</td>
<td>32</td>
<td>7</td>
</tr>
<tr>
<td>Mainstream</td>
<td>2</td>
<td>Queen Victoria</td>
<td>White</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Mainstream</td>
<td>3</td>
<td>Martha Washington</td>
<td>White</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Mainstream</td>
<td>4</td>
<td>Rosa Parks</td>
<td>Black</td>
<td>45</td>
<td>5</td>
</tr>
<tr>
<td>Mainstream</td>
<td>5</td>
<td>Emily Dickinson</td>
<td>White</td>
<td>30</td>
<td>5</td>
</tr>
<tr>
<td>Diversity</td>
<td>1</td>
<td>Rosa Parks</td>
<td>Black</td>
<td>189</td>
<td>32</td>
</tr>
<tr>
<td>Diversity</td>
<td>2</td>
<td>Eleanor Roosevelt</td>
<td>White</td>
<td>54</td>
<td>21</td>
</tr>
<tr>
<td>Diversity</td>
<td>3</td>
<td>Harriet Tubman</td>
<td>Black</td>
<td>44</td>
<td>21</td>
</tr>
<tr>
<td>Diversity</td>
<td>4</td>
<td>Emily Dickinson</td>
<td>White</td>
<td>63</td>
<td>17</td>
</tr>
<tr>
<td>Diversity</td>
<td>5</td>
<td>Harriet Beecher Stowe</td>
<td>White</td>
<td>70</td>
<td>14</td>
</tr>
<tr>
<td>People of Color</td>
<td>1</td>
<td>Rosa Parks</td>
<td>Black</td>
<td>184</td>
<td>30</td>
</tr>
<tr>
<td>People of Color</td>
<td>2</td>
<td>Harriet Tubman</td>
<td>Black</td>
<td>44</td>
<td>21</td>
</tr>
<tr>
<td>People of Color</td>
<td>3</td>
<td>Eleanor Roosevelt</td>
<td>White</td>
<td>53</td>
<td>20</td>
</tr>
<tr>
<td>People of Color</td>
<td>4</td>
<td>Harriet Beecher Stowe</td>
<td>White</td>
<td>70</td>
<td>14</td>
</tr>
<tr>
<td>People of Color</td>
<td>5</td>
<td>Emily Dickinson</td>
<td>White</td>
<td>60</td>
<td>14</td>
</tr>
<tr>
<td>African American</td>
<td>1</td>
<td>Rosa Parks</td>
<td>Black</td>
<td>47</td>
<td>14</td>
</tr>
<tr>
<td>African American</td>
<td>2</td>
<td>Harriet Tubman</td>
<td>Black</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>African American</td>
<td>3</td>
<td>Lena Horne</td>
<td>White</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>African American</td>
<td>4</td>
<td>Gwendolyn Brooks</td>
<td>Black</td>
<td>29</td>
<td>8</td>
</tr>
<tr>
<td>African American</td>
<td>5</td>
<td>Zora Neale Hurston</td>
<td>Black</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>Ability</td>
<td>1</td>
<td>Emily Dickinson</td>
<td>White</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Ability</td>
<td>2</td>
<td>Judy Blume</td>
<td>White</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Ability</td>
<td>3</td>
<td>Shirley Temple</td>
<td>White</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>Ability</td>
<td>4</td>
<td>Anna Lee</td>
<td>White</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Ability</td>
<td>5</td>
<td>Avril Lavigne</td>
<td>White</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Female</td>
<td>1</td>
<td>Betty Friedan</td>
<td>White</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>Female</td>
<td>2</td>
<td>Mary Pickford</td>
<td>White</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Female</td>
<td>3</td>
<td>Anne Brontë</td>
<td>White</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>Female</td>
<td>4</td>
<td>Gloria Steinem</td>
<td>White</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Female</td>
<td>5</td>
<td>Billie Jean King</td>
<td>White</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>1</td>
<td>Alicia Keys</td>
<td>Black</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>2</td>
<td>Britney Spears</td>
<td>White</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>3</td>
<td>Julia Roberts</td>
<td>White</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>4</td>
<td>Patsy Cline</td>
<td>White</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>5</td>
<td>Amy Winehouse</td>
<td>White</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Note: In this Table, we show the five most frequently mentioned famous females in each collection, along with their race, the number of times they were mentioned, and the number of books in which they appeared.
<table>
<thead>
<tr>
<th>Collection</th>
<th>Rank</th>
<th>Name</th>
<th>Race</th>
<th>Mentions</th>
<th>Books</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainstream</td>
<td>1</td>
<td>George Washington</td>
<td>White</td>
<td>182</td>
<td>32</td>
</tr>
<tr>
<td>Mainstream</td>
<td>2</td>
<td>Abraham Lincoln</td>
<td>White</td>
<td>189</td>
<td>23</td>
</tr>
<tr>
<td>Mainstream</td>
<td>3</td>
<td>Thomas Jefferson</td>
<td>White</td>
<td>74</td>
<td>15</td>
</tr>
<tr>
<td>Mainstream</td>
<td>4</td>
<td>Benjamin Franklin</td>
<td>White</td>
<td>30</td>
<td>15</td>
</tr>
<tr>
<td>Mainstream</td>
<td>5</td>
<td>John Adams</td>
<td>White</td>
<td>58</td>
<td>13</td>
</tr>
<tr>
<td>Diversity</td>
<td>1</td>
<td>Martin Luther King Jr.</td>
<td>Black</td>
<td>285</td>
<td>59</td>
</tr>
<tr>
<td>Diversity</td>
<td>2</td>
<td>Abraham Lincoln</td>
<td>White</td>
<td>100</td>
<td>44</td>
</tr>
<tr>
<td>Diversity</td>
<td>3</td>
<td>Langston Hughes</td>
<td>Black</td>
<td>187</td>
<td>41</td>
</tr>
<tr>
<td>Diversity</td>
<td>4</td>
<td>George Washington</td>
<td>White</td>
<td>73</td>
<td>40</td>
</tr>
<tr>
<td>Diversity</td>
<td>5</td>
<td>Martin Luther</td>
<td>White</td>
<td>63</td>
<td>36</td>
</tr>
<tr>
<td>People of Color</td>
<td>1</td>
<td>Martin Luther King Jr.</td>
<td>Black</td>
<td>271</td>
<td>56</td>
</tr>
<tr>
<td>People of Color</td>
<td>2</td>
<td>Abraham Lincoln</td>
<td>White</td>
<td>99</td>
<td>43</td>
</tr>
<tr>
<td>People of Color</td>
<td>3</td>
<td>Langston Hughes</td>
<td>Black</td>
<td>186</td>
<td>40</td>
</tr>
<tr>
<td>People of Color</td>
<td>4</td>
<td>George Washington</td>
<td>White</td>
<td>70</td>
<td>38</td>
</tr>
<tr>
<td>People of Color</td>
<td>5</td>
<td>Martin Luther</td>
<td>White</td>
<td>60</td>
<td>33</td>
</tr>
<tr>
<td>African American</td>
<td>1</td>
<td>Langston Hughes</td>
<td>Black</td>
<td>80</td>
<td>21</td>
</tr>
<tr>
<td>African American</td>
<td>2</td>
<td>Martin Luther King Jr.</td>
<td>Black</td>
<td>107</td>
<td>19</td>
</tr>
<tr>
<td>African American</td>
<td>3</td>
<td>Malcolm X</td>
<td>Black</td>
<td>84</td>
<td>13</td>
</tr>
<tr>
<td>African American</td>
<td>4</td>
<td>Duke Ellington</td>
<td>Black</td>
<td>22</td>
<td>13</td>
</tr>
<tr>
<td>African American</td>
<td>5</td>
<td>Martin Luther</td>
<td>White</td>
<td>21</td>
<td>13</td>
</tr>
<tr>
<td>Ability</td>
<td>1</td>
<td>Harold Pinter</td>
<td>White</td>
<td>78</td>
<td>2</td>
</tr>
<tr>
<td>Ability</td>
<td>2</td>
<td>Andy Warhol</td>
<td>White</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Ability</td>
<td>3</td>
<td>Marco Polo</td>
<td>White</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Ability</td>
<td>4</td>
<td>Duke Ellington</td>
<td>Black</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Ability</td>
<td>5</td>
<td>Mark Twain</td>
<td>White</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Female</td>
<td>1</td>
<td>John F. Kennedy</td>
<td>White</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Female</td>
<td>2</td>
<td>Jimmy Carter</td>
<td>White</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>Female</td>
<td>3</td>
<td>Martin Luther King Jr.</td>
<td>Black</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>Female</td>
<td>4</td>
<td>Richard Nixon</td>
<td>White</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Female</td>
<td>5</td>
<td>Barack Obama</td>
<td>Black</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>1</td>
<td>Alexander Hamilton</td>
<td>White</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>2</td>
<td>Sid Vicious</td>
<td>White</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>3</td>
<td>Alice Cooper</td>
<td>White</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>4</td>
<td>James Dean</td>
<td>White</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>LGBTQ</td>
<td>5</td>
<td>Michael Jackson</td>
<td>Black</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Note: In this Table, we show the five most frequently mentioned famous males in each collection, along with their race, the number of times they were mentioned, and the number of books in which they appeared.
C.3 Data Appendix

C.3.A Text Cleaning

The current data cleaning process standardizes the following names, which may have variants in how they are written: abraham lincoln, martin luther king junior, rosa parks, eleanor roosevelt, harriet tubman, george washington.

Our current data cleaning process also removes alpha consecutive lines with no more than beta words. Currently, we use a value of four for alpha and three for beta. This is meant to serve as a way to remove copyright pages or indices.

This process generates a list of words, also known as tokens, which we aggregate to generate “token counts,” or measures of frequency of the mention of specific words.

C.3.B All Gendered Mentions

The total number of female words in book $i$ is calculated as follows:

$$(\text{female words})_i = (\text{total number of female-specific tokens})_i + (\text{total number of mentions of famous female characters})_i + (\text{total number of characters with female first names})_i$$

C.3.C Specific Token Counts

The list of specific tokens (words) that we use in our text analysis are listed below.

Gendered Tokens. The gendered tokens we enumerate are as follows. Subset lists are used for the specific gendered token counts, gendered pronouns, singular/plural gendered token counts, younger/older gendered token counts and uppercase/lowercase pronouns.

Female. abuela, abuelita, actress, aunt, auntie, aunties, aunts, aunty, czarina, damsel, damsels, daughter, daughters, empress, empresses, empress, empresses, fairies, fairy, female, females, girl, girls, grandma, grandmas, grandmom, grandmother, grandmothers, her, hers, herself, housekeeper, housekeepers, ladies, lady, ma’am, madame, mademoiselle, mademoiselles, maid, maiden, maidens, maids, mama, mamas, mermaid, mermaids, miss, mlle, mme, mom, mommies, mommy, moms, mother, mothers, mrs, ms, nana, nanas, princess, princesses, queen, queens, she, sissie, sissy, sister, sisters, stepmother, stepmothers, tita, tsarevna, tsarina, tsaritsa, tzaritza, waitress, wife, witch, witches, wives, woman, women

43 This differs slightly from the notion of “lemmas,” which are measures of the count of word stems. To understand the difference, take the words “father,” “fatherly,” and “fathered.” If each word appeared once in a book, it would generate a token count of one for each word, but a lemma count of three for the lemma “father.”
Male. abuelito, abuelo, actor, boy, boys, bro, brother, brothers, butler, butlers, chap, chaps, czar, dad, daddies, daddy, dads, einstein, emperor, emperors, father, fathers, fellow, fellows, gentleman, gentlemen, granddad, granddads, grandfather, grandfathers, grandpa, grandpas, he, him, himself, his, hirsself, husband, husbands, king, kings, knight, lad, lads, lord, lords, male, males, man, master, masters, men, mermen, mermen, mr, paige, paiges, papa, papas, prince, princes, sir, sirs, son, sons, squire, squires, stepfather, stepfathers, tio, tsar, uncle, uncles, waiter, wizard, wizards

Racial Proxy Tokens. The tokens we use as proxies for race are as follows.

Color. black, blue, brown, gold, golden, green, orange, pink, purple, red, silver, violet, white, yellow

Ethnicity. afghan, african, african, albanian, algerian, american, andorran, angolan, antiguan, apache, argentinean, armenian, asian, australian, azerbaijani, bahamian, bahraini, bangladeshi, barbadian, barbudans, batswana, belarusian, belgian, belizean, beninese, bhutanese, bolivian, bosnian, brazilian, british, british, bruneian, bulgarian, burkinabe, burmese, burundian, cambodian, cameroonian, canadian, capeverdean, chadian, cherokee, chicana, chicano, chicani, chilean, chinese, choctaw, colombian, comoran, congolese, congolese, croatian, cuban, cypriot, czech, danish, djibouti, dominican, dutch, dutchman, dutchwoman, ecuadorean, egyptian, emirian, english, eritrean, estonian, ethiopian, fijian, filipino, finnish, french, gabonese, gambian, georgian, german, ghanaian, greek, grenadian, guatemalan, guineabissauan, guinean, guinean, guynese, haitian, herzegovinian, hispanic, honduran, hungarian, i-kiribati, icelander, indian, indonesian, iranian, iraqi, irish, irouquois, israeli, italian, ivorian, jamaican, japanese, jordanian, kazakhstani, kenyan, kittian, korean, kuwaiti, kyrgyz, lankan, laotian, latina, latino, latinx, latvian, lebanese, leonese, liberian, libyan, liechtensteiner, lithuanian, lucian, luxembourg, macedonian, malagasy, malawian, malaysian, maldivian, malian, maltese, marinese, marshallse, mauritian, mauritian, mexican, micronesain, moldovan, monacan, mongolian, mongols, moroccan, mosotho, motswana, mozambican, namibian, nauruan, navajo, nepalese, netherlander, nevisian, nivanuatu, nicaraguian, nigerian, nigerian, norwegian, ojibwe, omani, pakistani, palauan, panamanian, paraguayan, persian, peruvian, polish, portuguese, qatari, rican, romanian, russian, rwandan, salvadoran, samoan, saudi, scottish, senegalese, serbian, seychellois, singaporean, sioux, slovakian, somali, spanish, sudanese, surinamer, swazi, swedish, swiss, syrian, tajik, tanzanian, thai, timorese, tobagonian, togoese, tomean, tongan, trinidadian, tunisian, turkish, tuvaluan, ugdandan, ukrainian, uruguyan, uzbekistani, venezuelan, vietnamese, welsh, yemenite, zambian, zealander, zimbabwean