What might books be teaching young children about gender?

XXX[1], XX[2], XXX[2], XXX[2], & XXX[2]

[1] XXX

[2] XXX

Abstract

We investigated how gender is represented in children's books using a 200,000 word corpus comprising 247 popular, contemporary books for young children (0-5 years). Using human judgments and word co-occurrence data, we quantified the gender biases of words within the corpus and within individual books. We find that children's books contain large numbers of words that adults judge as gendered. Semantic analyses based on co-occurrence data yielded word clusters related to gender stereotypes (e.g., feminine: emotions; masculine: tools). Co-occurrence data also indicate that books instantiate gender stereotypes found in other research (e.g., girls are better at reading and boys at math). Finally, we used large-scale data to estimate the gender distribution of the audience for individual books, and find that children tend to be exposed to gender stereotypes for their own gender. Together the data suggest that children's books may be an early source of gender stereotypes.

STATMENT OF RELEVANCE: Gender stereotypes such as that girls are better at reading and boys are better at math have origins in early childhood. We examined the extent to which popular children's books may expose young 0-5 year old children to information about gender stereotypes. Our data suggest that children's books have rich information about gender stereotypes, and that some of these stereotypes are more strongly represented in children's books than in adult fiction. These findings suggest that popular children's books may be unintentionally teaching young children about gender stereotypes.

Keywords: reading, gender, language development

Word count: 2051 (excluding methods and results)

What might books be teaching young children about gender?

Many gender stereotypes held by adults have origins in early childhood. Beginning in their second year, children already exhibit knowledge of behaviors that are stereotypically feminine (e.g., vacuuming), masculine (e.g., building), and neutral (e.g., sleeping; Poulin-Dubois, Serbin, Eichstedt, Sen, & Beissel, 2002). By age three, children distinguish individuals by gender, race, and age (Shutts, Banaji, & Spelke, 2010). By age five, children have developed "a constellation of stereotypes about gender (often amusing and incorrect) that they apply to themselves and others" (Martin & Ruble, 2004). For example, preschoolers act in accordance with the stereotype that girls are better at reading while boys are better at math (Cvencek et al., 2011b), and that girls are less likely than boys to be "very, very smart" (Bian, Leslie, & Cimpian, 2017).

The *sources* of this knowledge are less well understood. Certainly, some of what children know about gender characteristics comes from their own and observed interactions with adults (Hilliard & Liben, 2010). Toys and activities are often gender stereotyped in home, day care, and preschool social settings (Weisgram, Fulcher, & Dinella, 2014). Another source of information is language. Children commonly receive feedback from adults about gender-normative activities, e.g., girls more often receive adult linguistic feedback for dress and helping behaviors, whereas boys receive comments on their size and physical skills (Chick, Heilman-Houser, & Hunter, 2002), and children are surprisingly sensitive to seemingly small differences in linguistic descriptions of gender-related information (e.g., Chestnut & Markman, 2018; Cimpian & Markman, 2011; Moty & Rhodes, 2019). For example, Cimpian and Markman (2011) found that when a novel game was introduced to children using a generic subject ("Girls are really good at a game called 'gorp'") they were more likely to associate it with a gender than when the game was introduced with a specific subject ("There is a girl who is good at...").

Here, we examine a particular potentially pervasive source of information about

gender: books for young children. The practice of reading to young children (also called "shared reading") has been widely encouraged because of its numerous benefits (Bus, Van Ijzendoorn, & Pellegrini, 1995; Duursma, Augustyn, & Zuckerman, 2008; High & Klass, 2014). Shared reading marks the child's entré to literacy and facilitates its development (Snow, Burns, & Griffin, 1998). Unlike everyday speech to children, linguistic input from books exposes children to information beyond their immediate experience, and therefore may be an especially powerful way of conveying gender stereotypes.

Much of the past work on how gender is represented in books has used "content analysis" methods that emphasize detailed analyses of a small number of texts. For example, Diekman and Murnen (2004) presented college students with books for middle-schoolers previously categorized as "sexist" or "nonsexist". Students rated them on a 72-item questionnaire. Questions probed whether books conveyed common gender stereotypes and inequalities, such as "Males, but not females, are shown as dominant" and "The book depicts female characters as the natural servants of male characters." The results suggested that gender differences and inequalities were expressed even in books intended to be nonsexist.

Our goal was to conduct a broader analysis of gender representation in books aimed at young (0-5 year old) children and to better understand who is being exposed to them. We operationalize the notion of gender representations by examining the extent to which individual concepts are differentially associated with a particular gender. We refer to this as "gender bias." Some of these gender biases reflect well-known gender stereotypes, such as the tendency to associate boys with math and girls with reading. By using human norms and automated analyses of distributional semantics, our approach provides a scalable and reproducible method of estimating gender bias without requiring explicit judgments of pre-specified properties of texts (as in studies such as Diekman & Murnen, 2004).

We begin by describing the creation of the Wisconsin Children's Book Corpus

(WCBC). We then quantify the biases in individual books and the corpus as a whole using two different methods. In Study 1, we quantify gender bias by measuring the genderedness of words in the books as judged by adults. In Study 2, we quantify gender bias in these books based on the statistical co-occurrence of the book's text. Finally, in Study 3, we use automated analyses of book reviews to estimate whether each book is being read primarily to boys or girls. To the extent that children are more likely to imitate those that are like them (Bussey & Bandura, 1999), finding that books containing boy- and girl-biased concepts are more often read to boys and girls, respectively, would suggest that books that are prominent in young children's experience may offer extensive opportunities for learning about gender.

Children's Book Corpus

Method

The Wisconsin Children's Book Corpus (WCBC) consists of 247 books marketed for children 5 years old and under. These are books that caregivers commonly read to children; some are also read independently by older children. Books were selected from four sources: (1) the top selling books for children in this age range from Amazon.com at the time of collection; (2) titles collected by Hudson Kam and Matthewson (2017) based on a survey of Canadian respondents; (3) Time Magazine's "100 best children's books of all time" (https://time.com/100-best-childrens-books); and (4) books in the corpus compiled by Montag, Jones, and Smith (2015). The union of these four sets yielded 247 books. The corpus includes the complete text of each book and basic metadata (author, title, etc.). In total, the corpus includes 202,445 word tokens ($M = 819.62$ per book; min $= 7$; max $= 23,352$; $SD = 2,082.69$) and 10,174 types (distinct orthographic forms; $M = 222.11$ per book; min $= 2$; max $= 2,575$; $SD = 283.47$). Arrangements for public access to the corpus are under negotiation.

Study 1: Measuring gender bias with human judgments

Study 1a: Gender bias in words

As a first step in understanding the genderedness of words in the book corpus, we had adult English speakers rate words on genderedness using a 5-point scale from masculine to feminine (Scott, Keitel, Becirspahic, Yao, & Sereno, 2019). This procedure yields systematic data with good face-validity: words such as *cuddle* and *pink* were rated as feminine, *axe* and *engine* as masculine, and *exactly* and *nose* as neutral.

Method. Participants ($N = 426$) were recruited on Amazon Mechanical Turk. Participants who answered any of 6 performance integrity check items incorrectly (e.g., "The word red has two letters") were excluded ($N = 80$). We also excluded 1 participant who responded with the midpoint on almost all items, and 6 non-native English speakers. The final sample included 339 participants (174 who identified as male, 162 female, and 3 other), with a mean age of 36.40 years ($SD = 10.70$).[1]

Because it was not feasible to collect gender norms for all 10,174 unique words, ratings were obtained for a large subset of the most important words ($N = 2,373$). The normed word set excluded stop words ($N = 30$), and was largely comprised of nouns (51.75%) and verbs (25.96%). We also included the names of all the characters (e.g. "Grover," "Amelia", "Yertle"). A short context was provided to indicate a specific meaning of homonymous words, e.g., "pin (hold down)," "creep (move slowly)," "act (part of a play)," "act (to take action)." The full norming included 82.48% of the tokens in the corpus and at least 30% of the tokens in each book were normed ($M = 83.25\%$; $SD = 9.54\%$; types: $M = 78.44\%$; $SD = 10.75\%$; excluding stop words).

Participants were instructed to rate the gender of each word on a 1-5 scale with the intervals labeled "Very masculine," "Somewhat masculine," "Neither masculine nor

---

[1] All data and code available in a public repository: https://github.com/mllewis/WCBC_GENDER

feminine," "Somewhat feminine," and "Very feminine". The instructions did not provide explicit definitions of masculine or feminine; raters were encouraged to base ratings on their own intuitions. Each participant rated between 90 and 97 words. Words were quasirandomly assigned to participants to ensure that each word received at least 10 ratings; mean number of ratings per word was 13.58 ($SD = 1.79$).

Results. The overall mean gender rating was 2.98 ([2.96, 3.01]), i.e., very close to the midpoint. 30% of the words were significantly female biased, 24% significantly male biased, and the remaining did not differ from the overall mean gender rating. There was a small main effect of participant gender: female participants ($M = 2.99$ [2.96, 3.03]) rated words as more feminine on average compared to male raters ($M = 2.98$ [2.95, 3.01]; paired t-test: $t(2372) = 1.98$; $p = 0.05$; $d = 0.02$ [-0.03, 0.08]). Gender ratings for 1,001 of our words were also obtained by Scott et al. (2019) and the two sets of ratings are highly correlated, $r = 0.91$ [0.89, 0.92], $p < .001$. Data can be explored interactively at https://mlewis.shinyapps.io/SI_WCBC_GENDER/. See SI for analyses of the relationship between word gender ratings and other word properties, such as concreteness and frequency.

To examine the kinds of words that tend be rated as masculine or feminine we identified the semantic neighborhoods of words in our sample using a word embedding model (Mikolov, Chen, Corrado, & Dean, 2013). Word embedding models provide a semantic representation of a word based on patterns of co-occurrence in a corpus of text, characterizing words as similar to the extent that they occur in similar contexts ("distributional statistics"; Landauer & Dumais, 1997). Semantic representations extracted in this way capture important aspects of meaning and correlate with human judgments of semantic similarity (Hill, Reichart, & Korhonen, 2015), though not without limitations (Chen, Peterson, & Griffiths, 2017). We obtained semantic coordinates for each word in our sample (a 300 dimensional vector) from a word embedding model pre-trained on English Wikipedia (Bojanowski, Grave, Joulin, & Mikolov, 2016), and then reduced the

Table 1

Examples of Clusters from Multi-Dimensional Embeddings

Category	Effect Size	N	Examples
Female-Biased Clusters			
affection	1.33 [0.9, 2.1]	21	kisses, loved, smile, tears, heart, care
modifiers	0.79 [0.49, 1.27]	34	probably, whenever, truly, likely, completely, yet
communication verbs	0.74 [0.43, 1.14]	25	spoke, listened, heard, explained, asked, answered
school	0.54 [0.12, 1.12]	20	learning, practicing, school, students, writing, book
food	0.44 [0.15, 0.8]	43	meatballs, soup, eggs, milk, pie, salad
Neutral Clusters			
family relationships	0.19 [-0.18, 0.63]	29	children, brother, sister, uncle, aunt
body parts	0.14 [-0.16, 0.48]	41	eye, knee, ankle, hair, bone
house parts	0.08 [-0.24, 0.4]	40	bedroom, floor, lamp, roof, window
quantifiers	0.05 [-0.29, 0.4]	36	few, almost, many, most, whole
spatial terms	-0.31 [-0.71, 0.02]	39	across, long, low, through, close
Male-Biased Clusters			
zoo animals	-0.53 [-1.27, -0.07]	23	giraffe, elephant, gorilla, lion, monkey, zebra
airborne actions	-0.83 [-1.21, -0.54]	37	climbed, tossed, jumped, knocked, pulled, swung
tools	-0.89 [-1.42, -0.52]	20	axe, blade, knife, bow, stick, wood
transportation (ground)	-1.23 [-1.62, -0.93]	40	car, bicycle, trains, ambulance, engine, traffic
professions	-1.35 [-2.19, -0.92]	23	judge, policemen, guard, sailor, mayor, clerk

Note: Effect size measure is Cohen's d based on a one-sample t-test comparing the mean gender of words in a cluster to the overall word gender mean. Brackets give bootstrapped 95 percent confidence intervals. N indicates number of words in each cluster.

dimensionality of these coordinates to two using the t-sne algorithm (an algorithm similar to PCA but better suited for high-dimensional spaces; Maaten & Hinton, 2008). We then clustered the words into 100 clusters based on their coordinates using k-means clustering. Clustering is an unsupervised machine learning method for dividing observations into k clusters by minimizing within-cluster distance and maximizing across-cluster distance. We determined the gender bias of each cluster by comparing the mean rated genderedness of the words in a cluster to the mean rated genderedness of all words in our sample.

The clustering procedure yielded semantically coherent clusters with an average of 23.21 words ($SD = 8.94$) per cluster. Of the 100 clusters, 21 were female-biased, 19 were male-biased, and the remaining 60 were neutral. Figure 3 shows examples of female-biased, male-biased and neutral clusters along with representative words (see SI for complete results). Many of the gendered clusters differ in ways that reflect gender stereotyping. For

example, female clusters were associated with mental states (feelings, beliefs) and interactions with others (communicating, caregiving). Male clusters, in contrast, tended to be more closely associated with events in the physical realm (e.g., sports, tools, transportation). These findings suggest that gender is an organizing dimension of semantic space for words found in children's books: Meanings semantically similar to each other tend to have similar gender associations.

Study 1b: Gender bias in books

We next use the word gender bias judgments derived in Study 1a to quantify the genderedness of individual books.

Method and Results. Using the adult judgments of gender bias from Study 1a, we calculated overall gender bias score for each book as the mean gender bias score of all the words (tokens) it contained. On average, there were gender norms for 79.11% ([77.60%, 80.46%]) of all tokens in the books (see SI for details and additional analyses). The average gender score did not exhibit a strong bias ($M = 2.98$ [2.96, 3.01]), but there was substantial variability ($SD = 0.20$): some books contained many more "masculine" words, other books contained many more "feminine" words.

Figure 1 includes data from a subset of books, the 20 with the highest feminine bias scores, the 20 with the highest masculine bias scores, and 20 from the neutral range. Data for all books are available in the SI. Measured in this way, the books clearly vary in genderedness, falling along a continuum. Books at the feminine end include *Olivia*, *Brave Irene*, and *Amelia Bedelia*; the masculine end includes *Dear Zoo*, *Curious George*, and *Goodnight, Goodnight, Construction Site*; neutrals include *In the Night Kitchen*, *Hippos Go Berserk*, and *Everyone Poops*.

Differences in overall gender bias may be due to the distribution of gendered content but also to differences in the occurrence of intrinsically gendered words such as names (e.g.,

Jill), pronouns (e.g., *her*), and relational/generic gender terms (e.g., *mom, lady*). We therefore calculated two additional bias measures: one based on the gender words referring to the book's characters (the character gender score) and the remaining content words
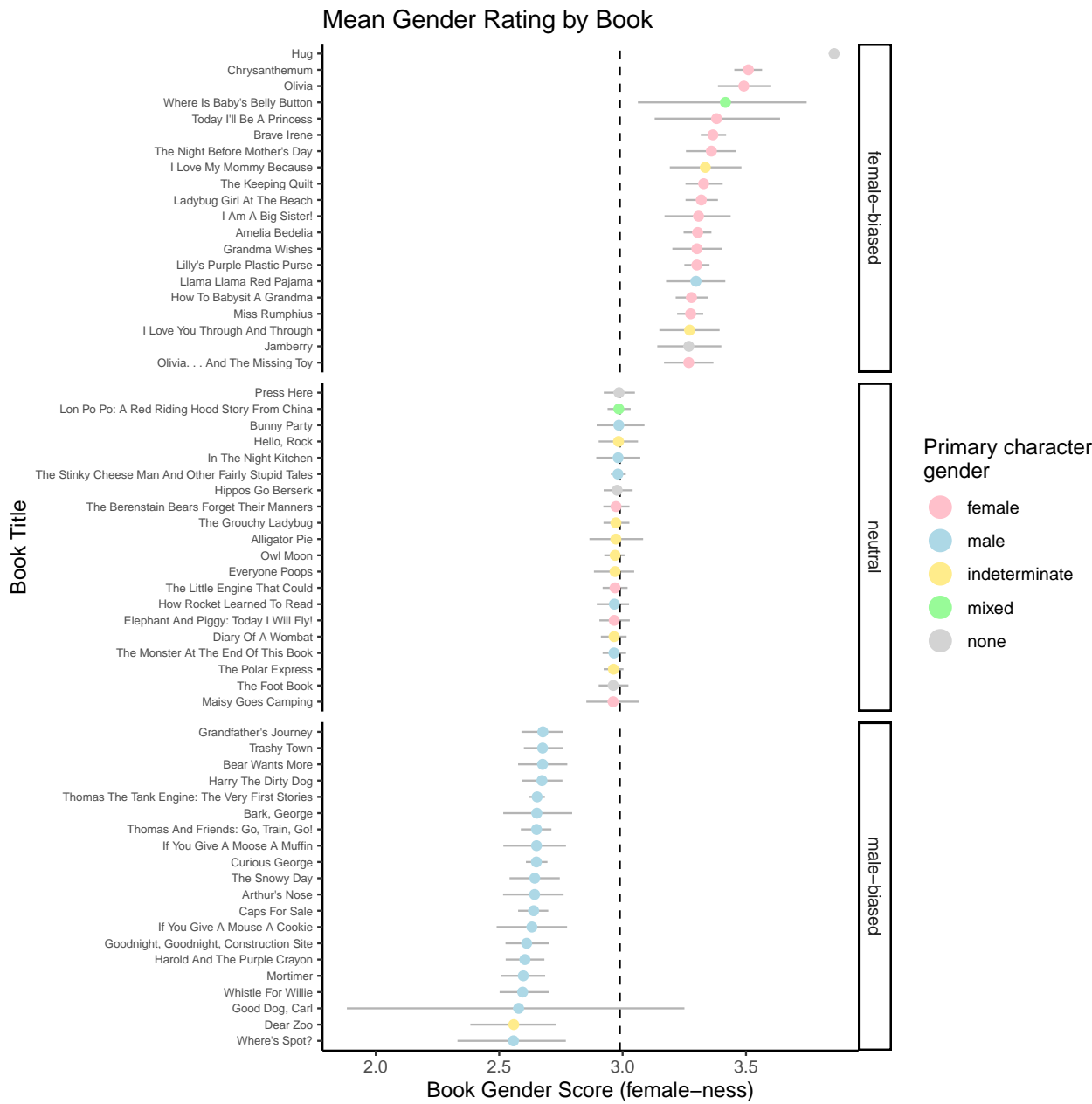
Mean Gender Rating by Book

Figure 1. Overall gender rating of a subset of books, the 20 with the highest feminine bias scores, the 20 with the highest masculine bias scores, and 20 from the neutral range. Bias scores are calculated from the mean gender ratings of words in each book (tokens). The dashed line indicates the overall mean across books, and color indicates the gender of the primary character. Error bars are bootstrapped 95% CIs.

Table 2

Examples of Clusters from Multi-Dimensional Embeddings

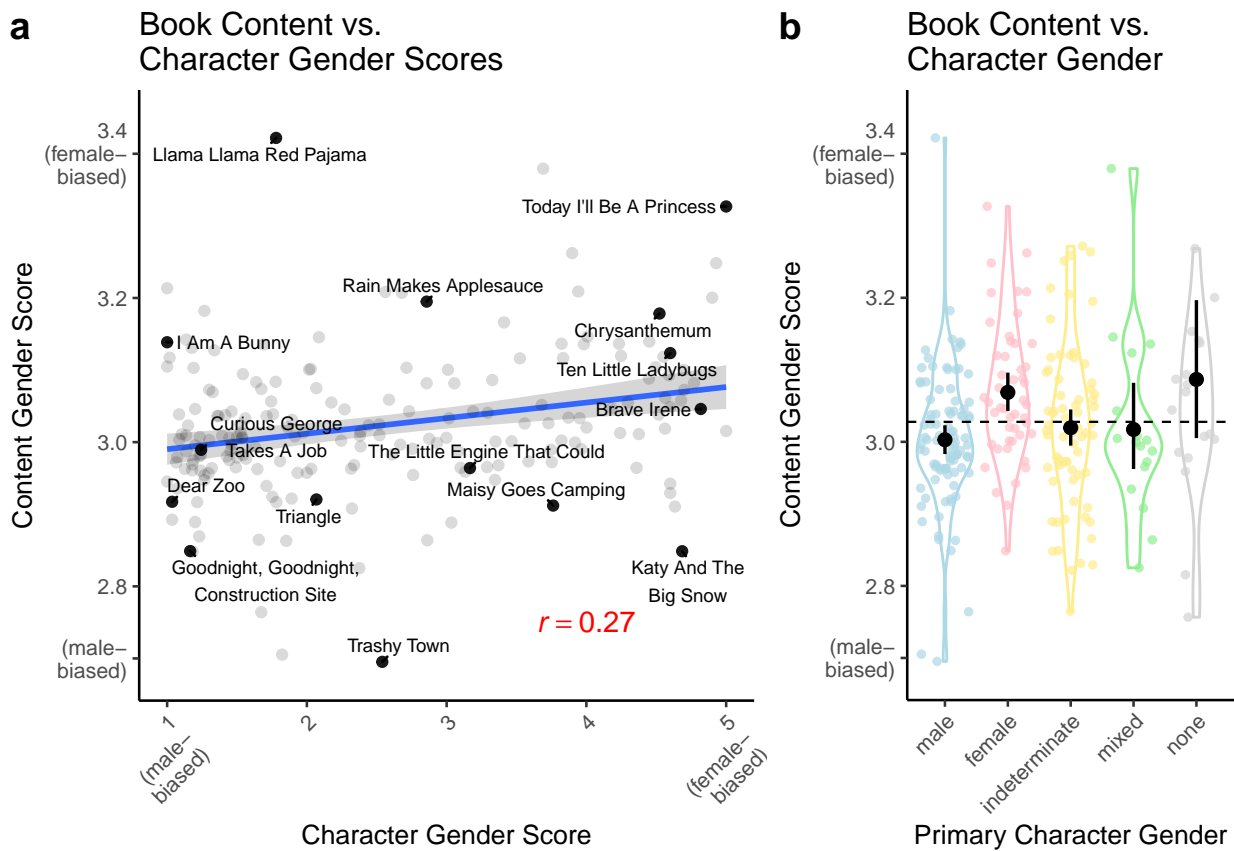| Category	Effect Size
Title	Chrysanthemum
Plot	The main character, Chrysanthemum, is ridiculed at school for her unusual
Main Character Gender	female
25 most freq. nouns/verbs	chrysanthemum, said, name, twinkle, father, mother, flower, named, though

Figure 2. (a) Mean content gender score for each book as a function of mean character gender score. Error bar shows the standard error of the linear model fit. (b) Distribution across books of content gender score as a function of primary character gender. Colored points show individuals books (one point excluded for visibility). Dashed line shows content gender score grand mean. Black points and error bars show mean and bootstrapped 95% percent confidence intervals for books of each primary character gender type.

(content gender score). Character and content scores had a moderate positive correlation with each other ($r = 0.27$ [0.13, 0.4], $p < .001$): books with more feminine-biased (less masculine-biased) content tended to have more female names, pronouns, and kinship terms

(Figure 2a). This finding suggests that the word gender biases reported by adults, could in principle, be learned from a word's associations with gendered terms in the book text.

Whereas the character gender score reflects the extent to which males and females are directly mentioned in a book, the gender of the story protagonist may be particularly salient for children. For each book, we manually coded the name of the primary protagonist character(s) and their gender as determined from text (i.e., pronouns). Text rather than illustrations was used to determine character gender because it was less ambiguous. A character was considered a protagonist if they were the primary agent of the story, in some cases in a collaborative fashion with another protagonist. The main character(s) were classified as either female, male, mixed, or indeterminate (Wagner, 2017). A book was coded as "mixed" if there was more than one primary character, and their gender composition was heterogeneous and as "indeterminate" if a given primary character had a gender that could not be determined from the text. Two research assistants and the second author coded character gender. Coders agreed on the protagonist type for 97% of books. Discrepancies were resolved through discussion.

About half of the books (142/248; 57%) had gendered primary characters that were exclusively male or exclusively female. Two-thirds of these books had male primary characters ($N = 94$; $\chi^2(1) = 14.9$, $p < .001$; $d = 0.68$ [0.34, 1.03]). Of the remaining books, 69 (28%) had main characters(s) of indeterminate gender, 17 (7%) had main characters of mixed genders, and 20 (8%) had no main character(s). These results are broadly consistent with those previously in a smaller sample of books (Wagner, 2017). We then examined book genderedness as a function of the gender of the primary character, using both content and character scores. Books with female primary characters tended to have higher female content scores ($M = 3.07$ [3.04, 3.1]; $t(47) = 2.96$, $p = 0.005$; $d = 0.43$ [0.15, 0.71]), compared to the overall averages, whereas books with male primary characters tended to have relatively higher male content scores ($M = 3$ [2.98, 3.02]; $t(93) = -2.52$, $p = 0.01$; $d = $

-0.26 [-0.48, -0.06]; Figure 2b). Notably, however, there was a large degree of variability in content scores across books with characters of a certain gender (female: $SD = 0.72$; male: $SD = 0.69$).

Our findings suggest that books vary appreciably along the dimension of gender in terms of both their content and characters. Together, the gender character and content gender data provide converging evidence that the text of children's books is significantly gendered: Books with female characters tend to have content typically associated with females, whereas books with male characters tend to have content typically associated with males.

Study 1c: Validation of book gender bias measure

One potential limitation of our method for estimating book bias is that the text is treated as a context-free "bag of words," yielding a simple average of the gender bias of words. It is possible that such words could occur in contexts that modulate their bias. For example, the gender bias of the word "brave" would be the same whether it occurred in the sentence "Sally is brave" or "Sally is not brave". To address this concern, we asked a new group of adult participants to provide information about main characters after reading the complete text of a book. We could then determine whether these participant-generated descriptions exhibited the gender biases identified using the simpler word-based measure. The two should diverge if the biases of individual words are modulated by the sentence and broader story context.

Method. We recruited 152 participants from Amazon Mechanical Turk. Eighty-one identified as female, 65 identified as male, 6 did not provide a response.

We divided the books in our corpus into quintiles based on the gender score described in Study 1b, and selected 15 books each from the first (female-biased: $M = 3.23$; $SD = 0.06$), third (neutral: $M = 2.96$; $SD = 0.03$), and fifth quintiles ($M = 2.64$; $SD = 0.03$) to

norm further. We excluded books that were either very short or very long (less than 100 words, or more than 900 words), or did not have a one or more named, gendered main character.

Participants were presented with the complete text of an individual book, and told that they would be asked questions about the characters in the story. After reading the text, participants were asked to list 2-5 main activities of particular character in the story in a free response form (e.g., "List 2-5 main activities Thomas does in the story."). The book text was displayed on the same page that responses were elicited, such that participants did not have to rely on their memory to answer the question. Next, participants were asked to complete a similar procedure but instead provide descriptions of a character, rather than associated activities (e.g., "List 2-5 words to describe what Thomas does in the story."). This procedure was repeated for all main and secondary characters in a book. Each participant provided responses for both character activities and character descriptions for a sample of three books.

On average, participants generated 3.83 responses per question ($SD = 1.24$). Participants' responses were lemmatized, corrected for spelling, and, in cases where a multi-word phrase (e.g., "builds a castle") was listed, the first word was selected for analysis. We identified the part of speech for the target word and excluded responses of the wrong class, analyzing only words that could be a verb for the activity question and an adjective, adverb, or noun for the description question. We also excluded responses that were very long (more than 35 characters), as these were likely to be full sentences rather than activity or description words. In total, 4% of responses were excluded, leading to a final sample of 4,889 responses and 947 unique lemmas. We then analyzed the gender bias of the activity and description words using previously-collected human judgments of word gender bias which covered 67% of the word tokens used to describe characters and their activities in this study. We collected an additional set of human judgments ($N = 251$; $M =$

11.33 ratings/word; $SD = 0.95$) such that gender bias estimates were available for all words produced more than once in Study 1c (93% of tokens; see SI).

Results.   Do the descriptions and actions of the book's characters produced by participants who read the books match the gender biases computed from the averaged word gender (the method used in Study 1b)? To find out, we fit a mixed-effect linear regression predicting the gender biases of characters' descriptions and actions from the averaged word gender of a book. The averaged word gender of a book was treated as a continuous fixed effect, and book and participant were included as random intercepts. The averaged word gender of a book predicted the gender bias of both activity ($\beta = 0.43$; $SE = 0.16$; $t = 2.74$) and description words ($\beta = 1.15$; $SE = 0.23$; $t = 5.12$; see SI for full model results), suggesting that the averaged word gender measure described in Study 1b captures aspects of book gender bias, even after taking into account the broader context of the book text.

Study 2: Measuring gender bias through co-occurrence statistics

So far we have presented findings about gendered information in children's books based on adult gender norms and semantic representations derived from adult text. The results are relevant to the beliefs of adults who read books with children, which they may convey in conversation during shared reading. We also sought to understand what a child could learn about gender from the content of children's books independent of other information sources. We therefore trained word embedding models on the WCBC to examine gender biases in these texts. Despite the relatively small size of the children's book corpus, the word embeddings yield coherent patterns and clear evidence for gender biases similar to those identified from adult texts and norms. Overall, children's books exhibited stronger gender stereotypes than comparable adult texts.

Study 2a: Word gender associations in the Children's Book Corpus

Method and Results.   A word embedding model was trained on the full corpus of text from all 247 books (see SI for training details). We then estimated the gender

association for each word by calculating its mean semantic similarity (cosine distance) to a set of unambiguously female anchor words ("woman," "girl," "sister," "she," "her," and "daughter"), and a corresponding set of male words ("man," "boy," "brother," "he," "him," and "son"; Caliskan, Bryson, & Narayanan, 2017; Lewis & Lupyan, 2020). A female gender score was calculated for each word as the mean female similarity minus the mean male similarity. For comparison, we also estimated a female score from models trained on an identically sized corpus of adult fiction published from 1990 to 2017 (Corpus of Contemporary American English; Davies, 2008), and a much larger corpus of Wikipedia (Bojanowski et al., 2016). We then examined how these estimates of word gender bias derived from language statistics compared to the gender norms we had previously collected from participants.

Table 3
Relationships between word gender biases

	Human gender ratings	Distributed semantics (WCBC)	Distributed semantics (COCA)
Distributed semantics (WCBC)	0.27		
Distributed semantics (COCA)	0.39	0.36	
Distributed semantics (Wikipedia)	0.66	0.32	0.42

Note:
Correlation values are Pearson's r. All correlations are significant at the $p < .001$ level. WCBC = model trained on Wisconsin Children's Book Corpus; COCA = Davies, 2008; Wikipedia = Bojanowski et al., 2016.

There were 1,893 words common across the word embedding models and human gender norms dataset. Estimates of gender bias from the WCBC were correlated with our adult judgments of word bias ($r = 0.27$ [0.23, 0.31], $p < .001$). Estimates of gender bias from the WCBC were also correlated with word level gender bias estimates from a model trained on adult fiction ($r = 0.36$ [0.32, 0.4], $p < .001$), as well as the model trained on Wikipedia ($r = 0.32$ [0.28, 0.36], $p < .001$; see Table 3 for all pairwise correlations). This pattern suggests that the word-level gender biases reported by adults could, at least partially, be learned from the co-occurrence language statistics in the WCBC.

Study 2b: Specific gender biases in the Children's Book Corpus

Method and Results. We next asked whether specific gender stereotypes, beyond individual word biases, are also represented in children's books. We focused on four gender stereotypes that have been reported in adults and children in the social psychology literature: (1) Women as "good", men as "bad"; (2) Women as better at language skills, men as better at math skills; (3) Women as better at art skills, men as better at math skills, and (4) Women as family-oriented, men as career-oriented. Each of these stereotypes has been demonstrated behaviorally in prior work using both explicit measures (e.g., asking "How strongly do you associate career and family with males and females?") and implicit measures, such as the Implicit Association Test (IAT; Greenwald, McGhee, and Schwartz, 1998; Table 4). The IAT quantifies these associations using reaction time in a word categorization task (e.g., women-good/men-bad vs. women-bad/men-good), though not without criticism about its validity (Greenwald et al., 2020; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). Faster responses in this task are taken to indicate that two categories are more closely cognitively associated.

Previous work has shown that the same biases demonstrated in the IAT are also present in the distributional semantics of language (Caliskan et al., 2017; Lewis & Lupyan, 2020). A bias can be quantified in a word embedding model as an effect size, using the same set of word items as in the behavioral IAT. The effect size is calculated as the relative (cosine) similarity of male words (e.g., "men") to male-stereotyped words (e.g., "work"), compared to the relative similarity of female words (e.g., "women") to female-stereotyped words (e.g., "family"; see SI for formal effect size description). Stereotypes that are revealed in the IAT as measured through reaction time (e.g., men-work; women-family) tend to be reflected in word embedding models, as measured by cosine distance.

Table 4
Four IATs used to study gender bias

Psychological Bias	Target Words	Behavioral Studies
women as good; men as bad	"good": good, happy, gift, sunshine, heaven "bad": bad, awful, sick, trouble, hurt	Cveneck, Meltzoff, & Greenwald (2011b, C); Skowronski & Lawrence (2001, C/A); Greenwald et al. (2002, A); Rudman & Goodman (2004, A)
women and family; men and career	"family": family, parents, children, home, cousins, wedding "career": job, work, money, office, business, desk	Nosek, Banaji, & Greenwald (2002, A)
women and language; men and math	"language": books, read, write, story, letters, spell "math": numbers, count, sort, size, shapes, different	Cveneck, Meltzoff, Greenwald (2011a, C); Nosek, Banaji, & Greenwald, (2002, A)
women and arts; men and math	"art": art, paint, draw, books, dance, story "math": numbers, count, sort, size, shapes, different	Nosek, Banaji & Greenwald (2002, A)

Note: The left column describes the bias; the middle column lists the actual words tested for the target categories; the right column cites behavioral studies measuring the psychological bias. The words for the "female" and "male" categories were identical across all tests (see Main Text). Note that the words differ slightly from the stimuli used in the behavioral studies. "C" and "A" in citations indicate whether participants were children or adults, respectively.

We used this same method to examine whether these gender-related biases are also present in the language statistics of the WCBC. Target category items are listed in Table 4, along with citations for the corresponding behavioral IAT experiments with children and adults. Gender category word items were identical to those used in Study 2a. Other items were taken from the corresponding behavioral experiments, replacing items with more child-friendly alternatives in cases where the target word did not occur in the WCBC (e.g., "algebra" was changed to "numbers"). We conducted this analysis on a model trained on the WCBC, as well as models trained on a sample of the adult fiction section of COCA matched in size to the WCBC (Davies, 2008) and a model trained on Wikipedia (Bojanowski et al., 2016). The starting point for the text from the adult fiction book was randomly determined. We trained 10 models each on the COCA and WCBC corpora and estimated the average effect size for each IAT type.
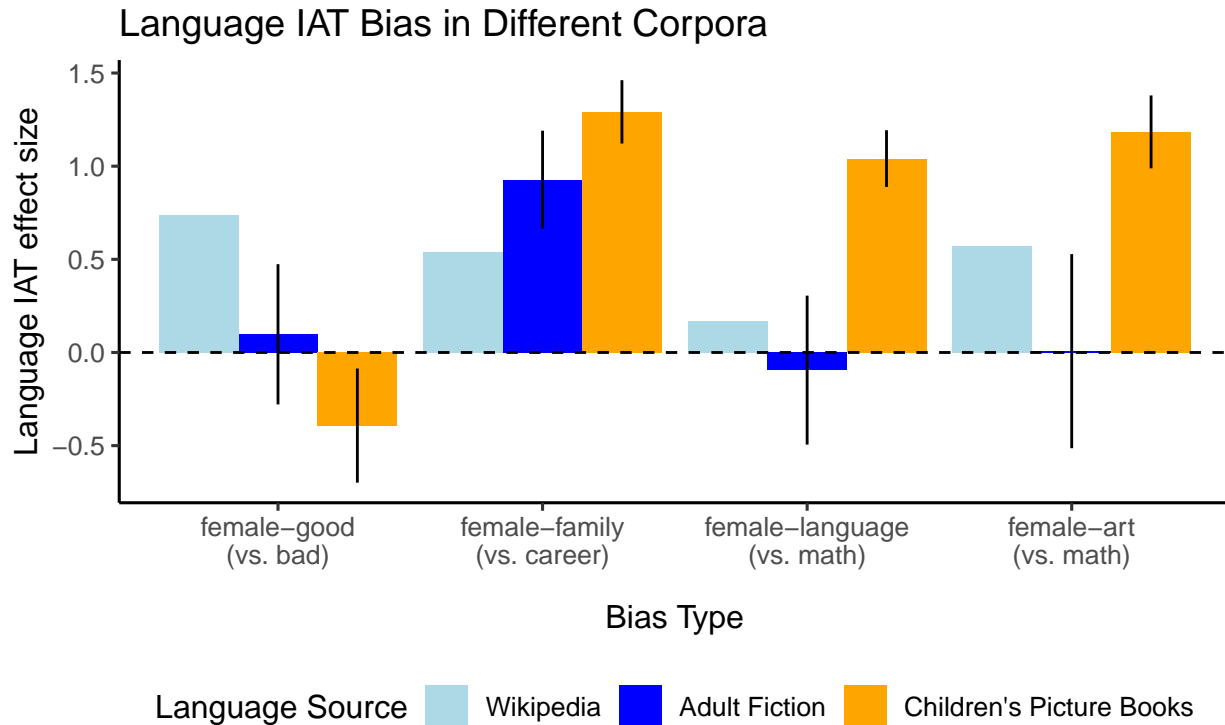
Language IAT Bias in Different Corpora

Figure 3. Estimates of the magnitude of gender biases in word embedding models trained on the Wisconsin Children's Book Corpus (orange), adult fiction corpus (COCA; dark blue), and Wikipedia (light blue). Positive effect sizes indicate a bias to associate women with the stereotypical category (e.g., 'family'); negative effect sizes indicate a bias to associate women with the non-stereotypical category (e.g., 'career'). Ranges indicate 95% confidence intervals across models. Biases are described more fully in Table 3.

Figure 3 shows the effect size for each of the four biases from models trained on each of the three corpora types. Positive values indicate a bias to associate women with the stereotypical female category (e.g. women-family). Three of the four gender biases were present in the co-occurrence statistics of the WCBC – Language-Math, Arts-Math, and Family-Career. Importantly, these biases were larger in children's books than in corpora containing mostly adult-directed language. This finding that behaviorally measurable gender biases are present in an exaggerated form in books for young children provides additional evidence that these books instantiate gender stereotypes that may influence children's learning of gender stereotypes.

In summary, in Study 2, we that adult word gender associations and specific gender

stereotypes that have been demonstrated behaviorally in adults and children, such as the bias to associate girls with language and boys with math, are both reflected in the co-occurrence statistics of the children's book corpus.

Study 3: Book Gender and Child Gender

The results so far suggest that the text of popular children's books contains rich information about gender. In this final study, we sought to better understand the processes through which this information might influence children's socialization into gender stereotypes by examining who is being exposed to these books. We created a novel measure based on the content of book reviews on a large online bookstore and validated this measure using existing survey data directly measuring the audience of a book. These data indicate that children's books more frequently read to girls tend to have both more female content and more female characters, and children's books more frequently read to boys tend to have both more male content and more male characters.

Method. For each book in the WCBC we collected a sample of the most recent reviews on Amazon.com. There were reviews for all but two books, with an average of 473.96 reviews per book ($SD = 194.53$; min $= 1$; max $= 1,290$). The content of each review was coded for the presence of 16 gendered kinship terms (e.g., "son", "daughter", "nephew", "niece"; see SI for full list). We selected these target words because they had a high likelihood of referring to the child for whom the book was purchased (e.g., "My son loves *Goodnight Moon.*"), rather than referring to a book character. All but two books had reviews containing at least one of our target gendered kinship terms. Overall, 27.63% of reviews per book contained at least one target gendered kinship term ($SD = 0.08$). For each review, we calculated an audience gender score as the proportion of female kinship terms (tokens) present relative to all target kinship words, and then averaged across reviews from the same book to get a book-level estimate of the gender of book addressees ($M = 0.49$; $SD = 0.19$; see SI for supplemental models predicting book gender at the review level).
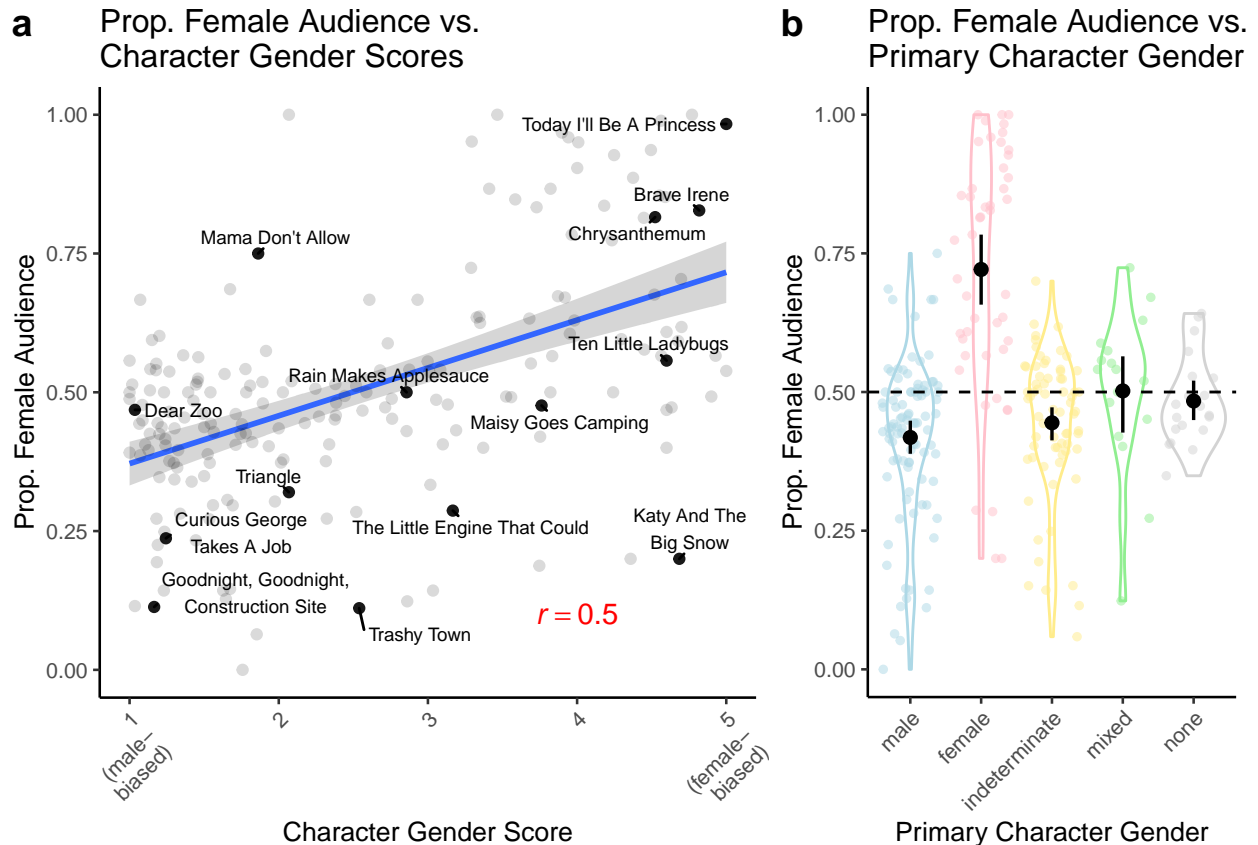
Figure 4. (a) Estimated audience gender for each book as a function of mean character gender score. Error bar shows the standard error of the linear model fit. (b) Distribution across books of audience gender as a function of primary character gender. Colored points show individuals books. Dashed line shows grand mean of proportion female audience. Black points and error bars show mean and bootstrapped 95% percent confidence intervals for books of each primary character gender type.

We validated our computed audience gender score by comparing it to survey data collected by Hudson Kam and Matthewson (2017), who asked a sample of 1,107 Canadian caregivers to list the five books most frequently read to their male or female child. Of the books with at least 5 survey responses, 103 were also in the WCBC. Our review-based gender measure was positively correlated with Kam and Matthewson's survey based measure ($r = 0.58$ [0.44, 0.7], $p < .001$), suggesting that book reviews can be used to estimate whether a given book is primarily read to boys or girls.

Results. We compared our audience gender score for each book to the measures of book genderedness described above. Both the content gender scores ($r = 0.37$ [0.26, 0.48],

$p < .001$) and book character gender scores ($r = 0.53$ [0.41, 0.62], $p < .001$; Figure 4a) were correlated with audience gender scores: Books that contained more female-biased content words and more female characters tended to be read more often to girls. In an additive linear model predicting audience gender with both types of gender scores, both content ($\beta = 0.67$; $SE = 0.12$; $Z = 5.47$; $p < .001$) and character gender scores ($\beta = 0.07$; $SE = 0.01$; $Z = 7.32$; $p < .001$) predicted independent, and roughly equal, variance. Together, they accounted for 37% of the total variance in audience gender.

Consistent with this general pattern, books with female primary characters also tended to be more often read to girls, compared to the overall average ($t(46) = 7.02$, $p < .001$; $d = 1.02$ [0.68, 1.56]; Figure 4b). Books with male ($t(92) = -5.12$, $p < .001$; $d = -0.53$ [-0.71, -0.37]) or gender indeterminate primary characters ($t(68) = -3.24$, $p = 0.002$; $d = -0.39$ [-0.59, -0.2]) tended to be more often read to boys. Notably, the effect size for girls was more than twice that of boys, suggesting that there was a stronger bias to read books with female characters to girls, relative to books with male characters to boys. There was no bias in audience gender for books with multiple primary characters of different genders ($t(16) = 0.24$, $p = 0.81$; $d = 0.06$ [-0.35, 0.82]) or books without primary characters ($t(19) = -0.56$, $p = 0.58$; $d = -0.13$ [-0.68, 0.3]).

In sum, these findings suggest that children's books featuring a particular gender and content associated with that gender tend to be read disproportionately to children of that same gender.

General Discussion

What gender messages are conveyed by popular children's books and who is being exposed to them? We constructed a corpus of 247 contemporary children's books and analyzed the extent to which the books contain biased gender associations. Using adult judgments of individual words, we found that over half of the words in the corpus tended

to be associated with a particular gender, and tended to cohere in gender stereotypical categories (e.g., social interaction for females; physical interaction for males). At the book level, we found that books varied in their gender associations, and the associations tended to reflect gender stereotypes (e.g., girl characters tended to do stereotypically girl activities). Further, the language statistics of the corpus itself reflected word gender biases elicited from adult judgments and more specific gender stereotypes (e.g., boys are relatively better at math, and girls are relatively better at reading). These biases were more exaggerated in the children's book corpus, relative to adult fiction. Finally, we derived a novel metric for measuring the gender distribution of a book's audience using automated analysis of book reviews. Children tended to be exposed to books that conveyed gender stereotypes about their own gender. Our work provides the first quantitative assessment of how gender is represented in contemporary children's books and reveals that they contain many statistical regularities that could inform children's understanding of gender stereotypes.

There are several reasons to think that the statistical regularities we identified in children's books may be shaping children's gender stereotypes. First, many of the stereotypical patterns that we report are implicit in the distributional statistics of the text, rather than conveyed via explicit statements ("boys are better at math than girls"). The implicit nature of these messages may make them particularly difficult for adult readers to track or explicitly contradict. Second, children are exposed to books with a caregiver (compared to, e.g., watching TV). The caregiver's presence may signal implicit endorsement of these stereotypes as correct or desirable and lead the child to make to stronger inferences (Lewis & Frank, 2016; Xu & Tenenbaum, 2007). Third, our data suggest that children tend to be exposed to books that contain gender stereotypes of their own gender presenting children with more information about own-gender-consistent associations. This may make gender-inconsistent preferences less familiar to children and therefore more difficult to emulate (Bussey & Bandura, 1999). Filtered through children's

cognitive and social biases, children's books may therefore be a potent means of teaching children about gender stereotypes.

Our work characterizes the messages in the text of children's books and begins to address the role they play in socialization, but there are a number of open questions about the causal link between the statistical regularities we observe and the gender stereotypes that children form. Importantly, little is known about how children themselves perceive the messages contained within these books. In the work presented here, we primarily measure word gender bias via adult judgments, yet children do not have the extensive knowledge and experience that underlies adult judgments. The fact that word embedding models trained exclusively on the statistics of the children's book corpus reflect adult-like word gender biases suggests that the adult gender biases could in principle be learned from sources like children's book text, but whether they are remains an open question. Future work could more directly address these questions by eliciting child ratings of word gender, and by experimentally manipulating the statistics of children's linguistic input about gender.

An unanswered question from our data is how children learn stereotypes about other genders, given that they are largely read storybooks containing stereotypes aligning with their own gender. One possibility is that children are exposed to information about other genders from other sources, such as other kinds of media and direct interactions. Alternatively, children may in fact receive more information about their own gender, relative to other genders, and consequently have less precise intuitions about stereotypes related to other genders. It is also an open question whether the tendency for children to be read books matching their own gender is due to caregiver or child preferences. This question is important in light of recent data on gender development in transgender children (Gülgöz et al., 2019). Transgender children show strong identity with the gender they feel they are by age three. If transgender children play an active role in their own socialization

(Martin & Ruble, 2004), our data suggest that children's books could be an early source of gender information for transgender children.

There is no doubt that shared reading has numerous benefits. However, our data show that embedded within contemporary children's books are pervasive gender stereotypes — indeed stronger than those found in adult-directed literature. Exposure to these language-embedded biases may lead to beliefs that help entrench gender disparities. Our work suggests caregivers may be able to influence children's gender stereotype development through their choices in which books they read to their children.

References

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

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv Preprint arXiv:1607.01759.*

Bus, A. G., Van Ijzendoorn, M. H., & Pellegrini, A. D. (1995). Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy. *Review of Educational Research*, *65*(1), 1–21.

Bussey, K., & Bandura, A. (1999). Social cognitive theory of gender development and differentiation. *Psychological Review*, *106*(4), 676.

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

Chen, D., Peterson, J. C., & Griffiths, T. L. (2017). Evaluating vector-space models of analogy. *arXiv Preprint arXiv:1705.04416.*

Chestnut, E. K., & Markman, E. M. (2018). "Girls are as good as boys at math" implies that boys are probably better: A study of expressions of gender equality. *Cognitive Science*, *42*(7), 2229–2249.

Chick, K. A., Heilman-Houser, R. A., & Hunter, M. W. (2002). The impact of child care on gender role development and gender stereotypes. *Early Childhood Education Journal*, *29*(3), 149–154.

Cimpian, A., & Markman, E. M. (2011). The generic/nongeneric distinction influences how children interpret new information about social others. *Child Development*, *82*(2), 471–492.

Cvencek, D., Greenwald, A. G., & Meltzoff, A. N. (2011a). Measuring implicit attitudes of 4-year-olds: The Preschool Implicit Association Test. *Journal of Experimental Child Psychology, 109*(2), 187–200.

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011b). Math–gender stereotypes in elementary school children. *Child Development, 82*(3), 766–779.

D'Addario, Daniel, Nathan, G., & Rayman, N. (n.d.). The 100 best children's books of all time. Retrieved from http://time.com/100-best-childrens-books/

Davies, M. (2008). The Corpus of Contemporary American English: 450 million words, 1990-present. *https://www.english-Corpora.org/Coca/.*

Diekman, A. B., & Murnen, S. K. (2004). Learning to be little women and little men: The inequitable gender equality of nonsexist children's literature. *Sex Roles, 50*(5-6), 373–385.

Duursma, E., Augustyn, M., & Zuckerman, B. (2008). Reading aloud to children: The evidence. *Archives of Disease in Childhood, 93*(7), 554–557.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109*(1), 3–25.

Greenwald, A. G., Brendl, M., Cai, H., Cvencek, D., Dovidio, J. F., Friese, M., . . . al. (2020). The implicit association test at age 20: What is known and what is not known about implicit bias. PsyArXiv. https://doi.org/10.31234/osf.io/bf97c

Gülgöz, S., Glazier, J. J., Enright, E. A., Alonso, D. J., Durwood, L. J., Fast, A. A., . . . others. (2019). Similarity in transgender and cisgender children's gender development. *Proceedings of the National Academy of Sciences, 116*(49),

24480–24485.

High, P. C., & Klass, P. (2014). Literacy promotion: An essential component of primary care pediatric practice. *Pediatrics*, *134*(2), 404–409.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695.

Hilliard, L. J., & Liben, L. S. (2010). Differing levels of gender salience in preschool classrooms: Effects on children's gender attitudes and intergroup bias. *Child Development*, *81*(6), 1787–1798.

Hudson Kam, C. L., & Matthewson, L. (2017). Introducing the Infant Bookreading Database (IBDb). *Journal of Child Language*, *44*(6), 1289–1308.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Lewis, M. L., & Frank, M. C. (2016). Understanding the effect of social context on learning: A replication of Xu and Tenenbaum (2007b). *Journal of Experimental Psychology: General*, *145*(9), e72–e80.

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, 1–8.

Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605.

Martin, C. L., & Ruble, D. (2004). Children's search for gender cues: Cognitive perspectives on gender development. *Current Directions in Psychological Science*, *13*(2), 67–70.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word
representations in vector space. *arXiv Preprint arXiv:1301.3781*.

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture
books and the statistics for language learning. *Psychological Science*, *26*(9),
1489–1496.

Moty, K., & Rhodes, M. (2019). The unintended consequences of the things we say: What
generics communicate to children about unmentioned categories. Retrieved from
https://psyarxiv.com/zkjyr/

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group
attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory,
Research, and Practice*, *6*(1), 101.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting
ethnic and racial discrimination: A meta-analysis of iat criterion studies. *Journal of
Personality and Social Psychology*, *105*(2), 171.

Poulin-Dubois, D., Serbin, L. A., Eichstedt, J. A., Sen, M. G., & Beissel, C. F. (2002).
Men don't put on make-up: Toddlers' knowledge of the gender stereotyping of
household activities. *Social Development*, *11*(2), 166–181.

Rudman, L. A., & Goodwin, S. A. (2004). Gender differences in automatic in-group bias:
Why do women like women more than men like men? *Journal of Personality and
Social Psychology*, *87*(4), 494.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow
Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, *51*(3),
1258–1270.

Shutts, K., Banaji, M. R., & Spelke, E. S. (2010). Social categories guide young children's preferences for novel objects. *Developmental Science*, *13*(4), 599–610.

Skowronski, J. J., & Lawrence, M. A. (2001). A comparative study of the implicit and explicit gender attitudes of children and college students. *Psychology of Women Quarterly*, *25*(2), 155–165.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children.* National Academies Press.

Wagner, L. (2017). Factors influencing parents' preferences and parents' perceptions of child preferences of picture books. *Frontiers in Psychology*, *8*, 1448.

Weisgram, E. S., Fulcher, M., & Dinella, L. M. (2014). Pink gives girls permission: Exploring the roles of explicit gender labels and gender-typed colors on preschool children's toy preferences. *Journal of Applied Developmental Psychology*, *35*(5), 401–409.

Xu, F., & Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, *10*(3), 288–297.