What might books be teaching young children about gender?

Molly Y. Lewis[1], Matt Cooper Borkenhagen[2], Ellen Converse[2], Gary Lupyan[2], & Mark S. Seidenberg[2]

[1] Department of Psychology, Carnegie Mellon University

[2] Department of Psychology, University of Wisconsin, Madison

Author Note

MCB was supported by the Vilas Trust at UW-Madison and the Institute of Education Sciences, US Department of Education, through Award #R305B150003 to UW-Madison. The opinions expressed are those of the authors and do not represent views of the US Department of Education. EC was supported by the Summer Senior Thesis Research Grant awarded by the UW-Madison L&S Honors Program. Additional support was provided by a Vilas research award to MSS and the Deinlein Language and Literacy Research fund.

Correspondence concerning this article should be addressed to Molly Y. Lewis, 4909 Frew St, Pittsburgh, PA 15213. E-mail: mollylewis@cmu.edu

Abstract

We investigated how gender is represented in children's books using a 200,000 word corpus comprising 249 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 more masculine or feminine. 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.

Keywords: reading, gender, language development

Word count: 1998 (excluding methods and results)

What might books be teaching young children about gender?

The gender stereotypes that are pervasive among adults have their origins in early childhood. Starting in their second year, children already exhibit knowledge of behaviors that are stereotypically feminine (e.g., vacuuming), masculine (e.g., shaving), 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. 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…"). This sensitivity may in part arise from an essentialist bias—a tendency to treat categories such as male and female as distinct with respect to visible, inferred, and assumed characteristics (Gelman & Taylor, 2000).

Here, we examine a particular potentially pervasive source of information about gender—books directed at 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 every day speech to children, linguistic input from books exposes children to information beyond their immediate experience, and therefore may be an especially potent means for conveying gender stereotypes.

The question of how gender is represented in books is not new. Much of the existing evidence derives from "content analysis" which emphasizes detailed analyses of a small number of texts. For example, Diekman and Murnen (2004) examined gender information in books for middle-schoolers they categorized as "sexist" or "nonsexist" (10 books per category), using a 72 item questionnaire completed by college students. 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 children—from infancy through 5 years old—and to better understand who is being exposed to them. By using human norms and automated analyses of distributional semantics, our approach provides a scalable and reproducible method of estimating gender biases 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 present three studies characterizing aspects of gender in the corpus. In Study 1, we assess the gender bias of individual words. We measure word gender bias using

adult judgments and describe how word gender bias relates to other word properties (e.g., age of acquisition; concreteness; Study 1A). We next characterize the semantics of word gender biases using word embedding models, and compare the word biases in our corpus to those in adult fiction (Study 1B). In Study 2, we quantify the gender bias of individual books. In the last study (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 stereotypes are read to boys and girls, respectively, would suggest that gender stereotypes in books may present particularly potent opportunities for learning.

Children's Book Corpus

The Wisconsin Children's Book Corpus (WCBC) consists of 249 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 249 books. The corpus includes the complete text of each book and basic metadata (author, title, etc.). In total, the corpus includes 203,433 word tokens (M = 817 per book; min = 7; max = 23,352; SD = 2,075) and 10,289 types (distinct orthographic forms; M = 222.25 per book; min = 2; max = 2,575; SD = 282.7). Arrangements for public access to the corpus are under negotiation.

Study 1A: Measuring word gender bias

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 feminine to

masculine (Scott, Keitel, Becirspahic, Yao, & Sereno, 2019). These ratings were quite systematic; 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 10,289 unique words, ratings were obtained for a large subset of the most important words ($N = 2{,}327$). The normed word set excluded stop words ($N = 30$), and was largely comprised of nouns (51.75%) and verbs (25.96%). 82.38% of the tokens in the corpus and at least 30% of the tokens in each book were normed ($M = 83.03$; $SD = 9.70$; excluding stop words). 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)."

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 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

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

of ratings per word was 13.58 ($SD = 1.79$).

A further question is whether genderedness (as rated by adults) is related to properties of words potentially relevant to the development of gender stereotypes: *valence* (degree of pleasantness), *arousal* (intensity of emotion), *concreteness* (whether a word refers to something that can be experienced directly or is more abstract), *age of acquisition* (*AoA*, an estimate of the age at which a word is learned), and *word frequency* (how often a word occurs in a language sample). Valence and arousal are implicated in common gender stereotypes (e.g., girls nice, boys aggressive); age of acquisition and word frequency provide evidence about children's exposure to words with these properties; the concrete-abstract dimension reflects the conceptual complexity of words.

We assessed correlations between rated gender and other lexical properties using existing norms. Warriner, Kuperman, and Brysbaert (2013) provide valence ratings on a 1 (happy) to 9 (unhappy) point scale and arousal ratings on a 1 (excited) to 9 (calm) scale. For age of acquisition (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), participants estimated the age in years at which they learned each word. For concreteness, participants rated the extent to which the meaning of a word can be experienced "directly through one of the five senses", rating each word on a 1 (abstract) to 5 (concrete) scale. Word frequency estimates depend on properties of the language sample that is used. We therefore conducted the correlational analyses using frequencies from three sources: (1) our corpus of children's books, (2) the cumulative frequency measure from the TASA norms (Zeno, Ivens, Millard, & Duvvuri, 1995) derived from a much larger sample of books from a broad range of reading levels, and (3) a large corpus of movie subtitles (Subtlex-US: Brysbaert & New, 2009). All frequency measures wore log transformed. Because word sense was not disambiguated in these norms, we averaged across words with the same word forms (but different senses) in our dataset for these analyses. Frequency measures from all three sources were available for 1,954 words in the WCBC. The three frequency measures were correlated (TASA-Subtlex: r

Table 1
Pairwise correlation between all word measures.

	Gender (fem.)	Arousal	Valence	Concreteness	AoA
Arousal	-0.08*				
Valence	0.35*	-0.09*			
Concreteness	-0.11*	-0.18*	0.00		
AoA	-0.08*	0.01	-0.20*	-0.24*	
Log Frequency (TASA)	0.00	-0.08*	0.14*	-0.21*	-0.41*

Note. Values are Pearson's r. Asterisks indicate statistical significance at the .01 level. AoA = Age of acquisition; TASA = Zeno et al., 1995 Corpus.

= 0.84 [0.82, 0.85], $p < .001$; TASA-WCBC: $r = 0.76$ [0.74, 0.78], $p < .001$; WCBC-Subtlex: $r = 0.72$ [0.69, 0.74], $p < .001$; all reported ranges indicate 95% confidence intervals), and the magnitudes are similar to ones reported previously (e.g., Zevin & Seidenberg, 2004). They also yielded very similar correlations with the other lexical measures. Below we report the results using the TASA frequencies and the 1,241 words for which there are data for the 4 additional measures. Results using the other frequency measures are included in the SI (https://mlewis.shinyapps.io/SI_WCBC_GENDER/).

Results

The overall mean gender rating was 2.98 ([2.95, 3.01]), i.e., very close to the midpoint. 30% of the words were significantly female biased, 30% significantly male biased, and the remaining did not differ from the overall mean gender rating. There was a numerically small, marginal effect of participant gender. Female participants ($M = 2.99$ [2.96, 3.02]) 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/.

Table 2
Model parameters predicting word gender association

Term	Std. Beta	SE	Z	p
(Intercept)	0.00	0.03	0.00	>.99
Arousal	-0.09	0.03	-3.26	0.001
Valence	0.34	0.03	12.87	<.001
Concreteness	-0.19	0.03	-6.40	<.001
AoA	-0.12	0.03	-3.88	<.001
Log Frequency (TASA)	-0.15	0.03	-4.85	<.001

Note. Larger gender values indicate greater association with females. AoA = Age of acquisition; TASA = Zeno et al., 1995 Corpus.

Table 1 shows the pairwise correlation between all word measures. Words that were rated as more feminine tended to be more positively valenced (r = 0.35 [0.3, 0.4], $p < .001$). More feminine words were also associated with lower arousal (r = -0.08 [-0.13, -0.02], p = 0.008), less concrete (r = -0.11 [-0.16, -0.05], $p < .001$), and learned earlier (r = -0.08 [-0.14, -0.03], p = 0.003). Word frequency was not correlated with word gender (r = -0.01 [-0.06, 0.05], p = 0.8).

We next fit an additive linear model to estimate the independent variance in gender explained by the other word measures. All five measures predicted independent variance in gender ratings (R^2 = 0.16), with valence being the strongest predictor of a word's gender association (i.e, more positively valenced words tend to be rated as more feminine; β = 0.34, SE = 0.03, Z = 12.87, p <.001; Table 2).

In summary, many of the most frequent content-bearing words in children's books have strong gender associations (54%), according to adult judgments. Words judged as more feminine were associated with more positive valence and lower arousal. More feminine words are also higher in frequency more concrete, and learned somewhat earlier (i.e., have a lower age of acquisition, holding frequency and the other variables listed in Table 2 constant).

Study 1B: Characterizing the semantic structure of word gender bias

The results of Study 1A indicate that children's books include many words that adults perceive as gendered. Here we investigate in more detail the semantic neighborhoods of these words by using word embedding models (Mikolov, Chen, Corrado, & Dean, 2013), a method for deriving lexical semantic representations based on patterns of co-occurrence. Word embedding models characterize words as similar to the extent that they occur in similar contexts ("distributional statistics"; e.g., 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 first used a pre-trained word embedding model to examine the semantic clusters of words in the children's books, and their relationship to the gender biases based on the adult judgments collected in Study 1A. We then trained word embedding models on the WCBC itself to understand what gender biases are present in the texts themselves, independent of other information sources. We find that the gender bias of words—estimated only from word co-occurrences in the corpus itself—is correlated with adult judgments of gender association. We also find that specific gender biases that have been demonstrated behaviorally in adults and children, such as the bias to associate girls with language and boys with math, are also present in the co-occurrence statistics of the children's book corpus.

Method and Results

Computing Genderedness from Language Statistics. Semantic coordinates for each word in our sample were obtained from a model trained on English Wikipedia (Bojanowski, Grave, Joulin, & Mikolov, 2016). We then reduced the dimensionality of these coordinates to two using the t-sne algorithm (Maaten & Hinton, 2008), and clustered the words into 100 clusters based on their similarity using k-means clustering. This procedure yielded semantically coherent clusters with an average of 23.21 words ($SD = 8.94$) per cluster

Table 3
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, walked, 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.

(Table 3; see SI for complete results).

The average rated genderedness of the words in these clusters was calculated using the gender norms. For each word cluster, we tested whether the mean gender rating of words in that cluster significantly differed from the overall mean gender rating of words. Of the 100 clusters, 21 were female-biased, 19 were male-biased, and the remaining 60 were neutral. Table 3 shows examples of female-biased, male-biased and neutral clusters along with representative words. 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).

Gender Associations derived from the Children's Book Corpus

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. However, we also want to understand what a child may learn about gender from children's books alone.

We estimated the extent to which gender associations were encoded in the language co-occurrence statistics in our corpus by training a word embedding model on the full corpus of text from all 249 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 female words ("woman," "girl," "sister," "she," "her," and "daughter"), and a set of male words ("man," "boy," "brother," "he," "him," and "son"). 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 4
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 judgement 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 4 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.

Specific Gender Biases derived from the Children's Book Corpus

The prior analysis suggests that stereotypical gender associates of individual words can be derived from the co-occurrence of words in the children's book corpus. Next, we asked whether specific gender stereotypes are also present in the statistics of children's books. We focused in particular on four gender stereotypes that have been demonstrated 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 through 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 5). The IAT quantifies these associations using reaction time in a word categorization task (e.g., women-good/men-bad vs. women-bad/men-good). 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, Bryson, & Narayanan, 2017; Lewis & Lupyan, 2019). A bias can be quantified in a word embedding model by measuring

the pairwise distance between words using the same set of word items as in the behavioral IAT. Categories that are closely associated in the IAT as measured by reaction time (e.g., women-family) tend to be closely associated in semantic space, as measured by cosine distance.

Table 5
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 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)
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)

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 measure the extent to which gender-related psychological biases were also present in the language statistics of the WCBC (see SI for method details). Target category items are listed in Table 5, along with citations for the corresponding behavioral IAT experiments with children and adults. Gender category word items were identical to the female score measure above. Other items were taken from the corresponding behavioral experiments, replacing items with more child-friendly alternatives in cases where

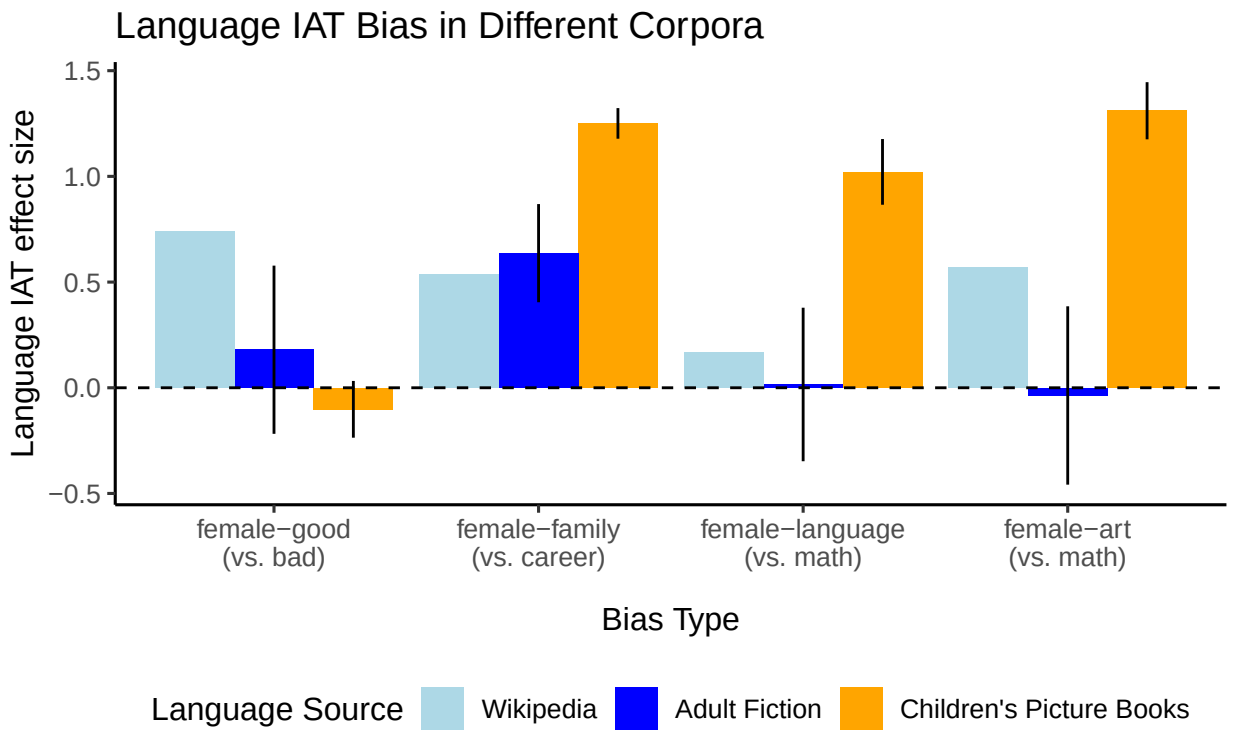

Figure 1. Estimates of the magnitude of gender biases in word embedding models trained on the Wisconsin Children's Book Corpus (orange), adult fiction corpus (COCA; dark blue), and Wikipedia (light blue). Positive effect sizes indicate a bias to associate women with the stereotypical category (e.g., '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 5.

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). We trained 10 models each on the COCA and WCBC corpora and estimated the average effect size for each IAT type.

Figure 1 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.

Study 2: Quantifying Gender Bias in Books

We next consider the genderedness of individual books.

Method and Results

Using estimates of word gender bias from adult judgments in Study 1, 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 0.78 ([0.77, 0.79]) 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 2 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 *Good Dog, Carl*; neutrals include *In the Night Kitchen*, *Hippos Go Berserk*, and *Everyone Poops*. The feminine titles include more references to family members (mommy, sister, grandma); the masculine titles include more references to animals and non-familial characters.

Differences in overall gender bias may be due to the distribution of content words without "intrinsic" gender (e.g., *beautiful*, *fight*) but also to differences in the occurrence of intrinsically gendered words such as names (*Jill*), pronouns (*her*), and relational/generic

gender terms (e.g., *mom*, *lady*). We therefore calculated two additional bias measures, one including only the intrinsically-biased words (the character gender score) and the other including all words except the intrinsically-biased ones (the content gender score). Both the character ($M = 2.53$ [2.35, 2.72]; $r = 0.77$ [0.7, 0.82], $p < .001$) and content scores ($M = 3.03$ [3.01, 3.04]; $r = 0.7$ [0.63, 0.75], $p < .001$) were correlated with the overall gender score. Thus, both gendered content words and intrinsically gendered words contribute to the overall gender differences between books.

Character and content scores had a moderate positive correlation with each other ($r = 0.27$ [0.14, 0.4], $p < .001$). Books with more feminine-biased (less masculine-biased) content words do tend to have more female names, pronouns and other intrinsically gendered words (Figure 3). This finding suggests that gender biases reported by adults for content words are potentially inferable from the character associations of the content words in the book texts.

Whereas the character gender score above reflects the extent to which males and females are mentioned in a book, the gender of the story protagonist may be particularly influential for children. For each book, we manually coded the name of the primary protagonist character(s) and the character's gender as determined from the text (i.e., not based on the illustrations). A character was considered a protagonist if that character was the primary agent of the story, even if in a collaborative fashion with another protagonist. The main character(s) were classified as either female, male, mixed, or indeterminate (Wagner, 2017). If there was more than one primary character, and their gender composition was heterogeneous, that group was classified as mixed. If a given primary character had a gender that could not be determined, no gender attribute was assigned ("indeterminate"). 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 (140/249; 56%) had gendered primary characters that were exclusively male or exclusively female. Two-thirds of these books had male primary

characters ($N = 92$; χ^2; $d = 0.66$ [0.31, 1.01]). Of the remaining books, 71 (29%) had main characters(s) of indeterminate gender, 17 (7%) had main characters of mixed genders, and 21 (8%) had no main character(s). We then examined book genderedness as a function of the

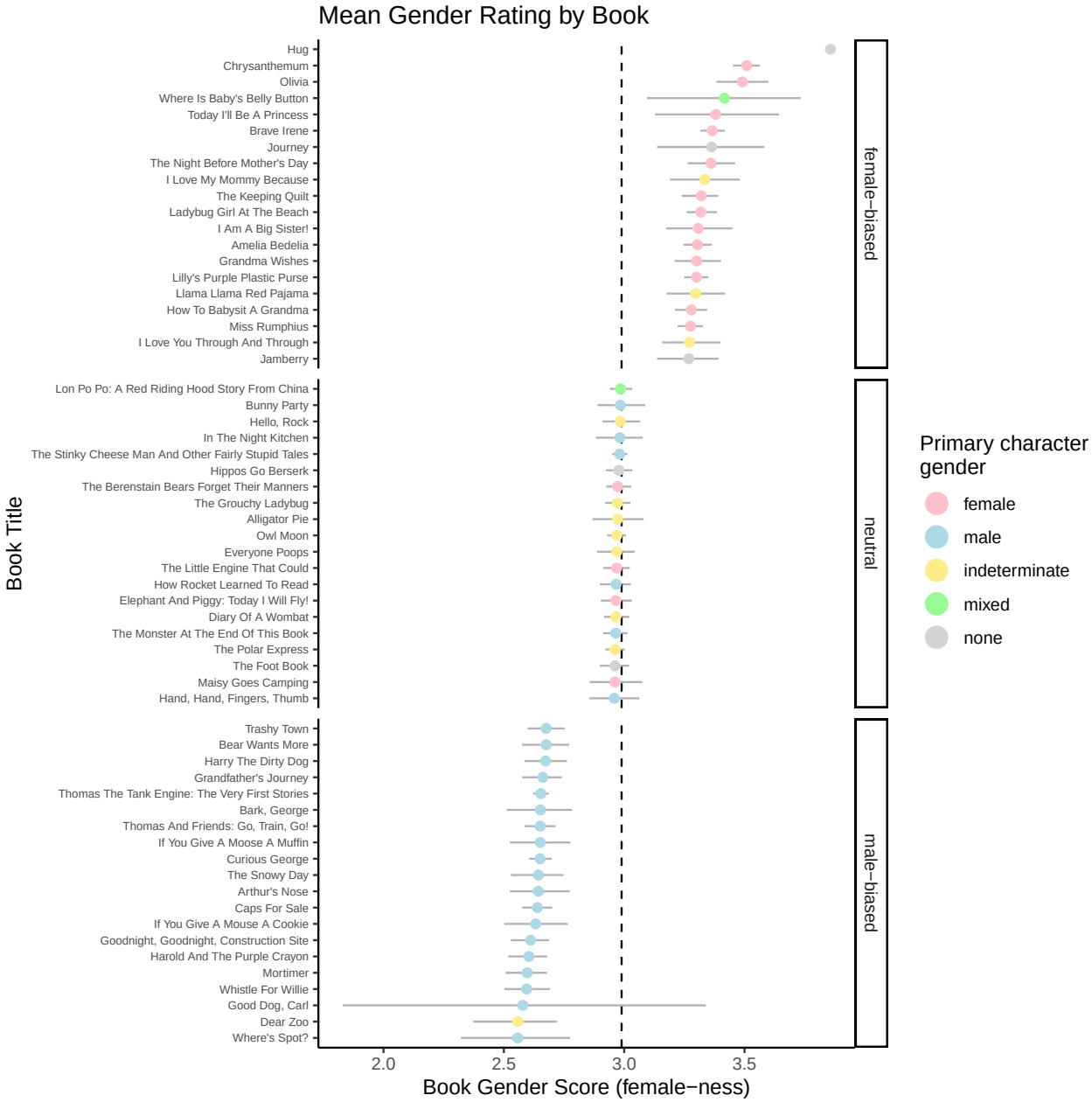

Figure 2. Overall gender rating of a subset of books, the 20 with the highest feminine bias scores, the 20 with the highest masculine bias scores, and 20 from the neutral range. Bias scores are calculated from the mean gender ratings of words in each book (tokens). The dashed line indicates the overall mean across books, and color indicates the gender of the primary character. Ranges are bootstrapped 95% CIs.

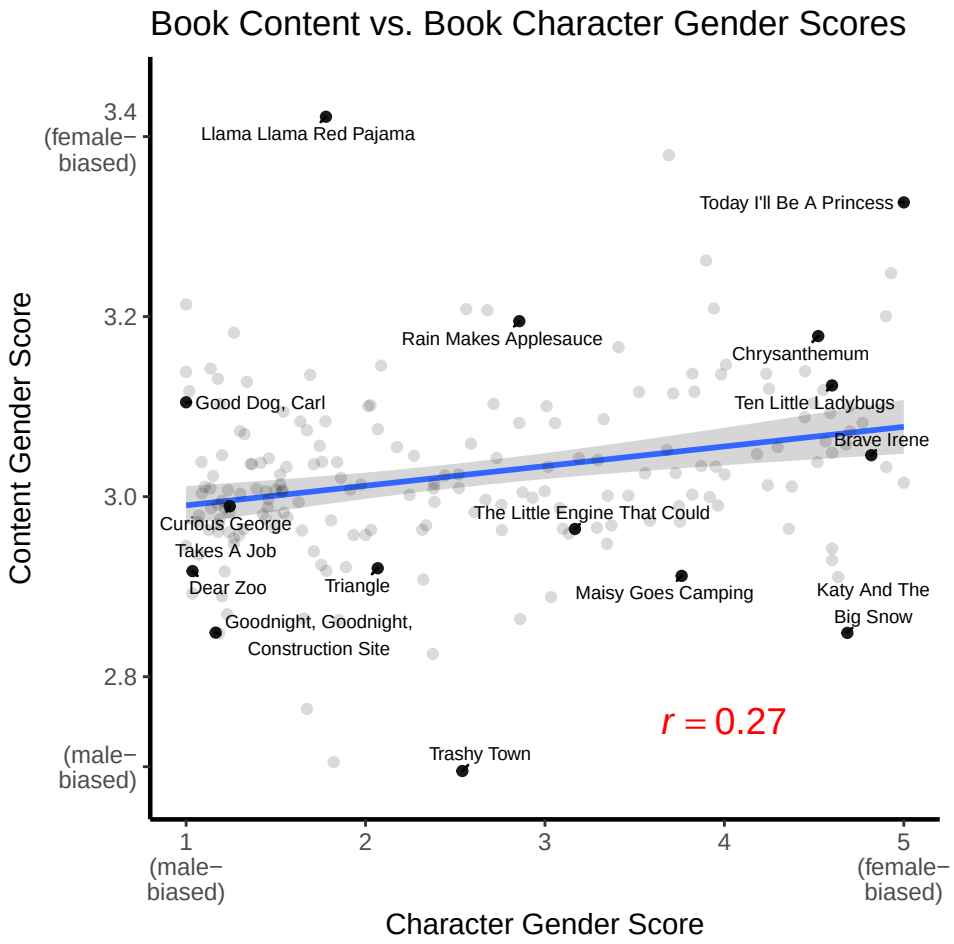

Figure 3. Mean content gender score for each book as a function of mean character gender score. Higher values indicate a greater propensity for female-biased words. Error bar shows the standard error of the linear model fit.

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, 3.09]; $t(47)$ = 2.9, p = 0.006; d = 0.42 [0.14, 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]; $t(91)$ = -3.41, $p < .001$; d = -0.36 [-0.53, -0.16]; Figure 4a).

We observed similar results for the character gender scores. Books with female primary characters tended to have higher (more female) character gender scores (M = 3.91 [3.68, 4.11]; $t(47)$ = 12.1, $p < .001$; d = 1.75 [1.29, 2.5]) compared to the overall average. Conversely, books with male character leads tended to have lower (less female) character

gender scores ($M = 3.91$ [3.68, 4.11]; $t(47) = 12.1$, $p < .001$; $d = 1.75$ [1.29, 2.5]). The magnitude of the effect for females was nearly twice that of males ($d = 1.75$ vs .96), suggesting that books with female primary characters tended to have text more heavily focused on same-gender characters (females), relative to books with a male primary character.

Our findings suggest that books vary appreciably along the dimension of gender in terms of both their content and characters. The gender distribution of characters we observe is comparable to that reported previously in a smaller sample of books (as in Wagner, 2017). Together, the gender character and gender content data provide converging evidence that information about gender associates of content words is present in the text of children's books: Books with female characters tend to have content stereotypically associated with females, whereas books with male characters tend to have content stereotypically associated with males.

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 begin to 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 = 194.53; max = 1,290.00. 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).

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 = 0.38 [0.27, 0.48], $p <$ .001) and book character gender scores (r = 0.52 [0.41, 0.62], $p < .001$) were correlated with audience gender scores: Books that contained more female-biased content words and more

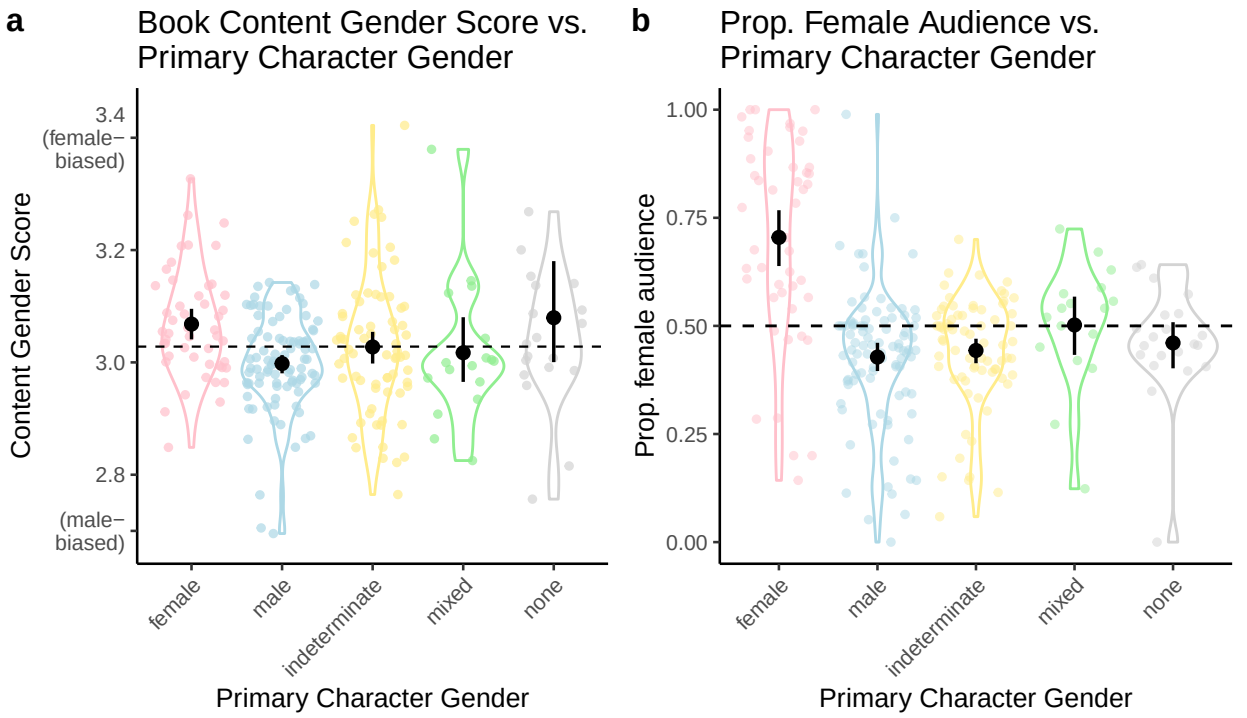

Figure 4. (a) Distribution across books of content gender score as a function of primary character gender. Colored points show individuals books (one point excluded for visibility). Dashed line shows content gender score grand mean. (b) Distribution across books of audience gender as a function of primary character gender. Colored points show individuals books. Dashed line shows grand mean of proportion female audience. Black points and ranges show mean and bootstrapped 95% percent confidence intervals for books of each primary character gender type.

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 ($\beta = 0.69$; $SE = 0.12$; $Z = 5.63$; $p < .001$) and character gender scores ($\beta = 0.07$; $SE = 0.01$; $Z = 7.29$; $p < .001$) predicted independent, and roughly equal, variance. Together, they accounted for 38% 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 ($t(46) = 6.19$, $p < .001$; $d =$ 0.9 [0.56, 1.4]). Books with male ($t(90) =$ -4, $p < .001$; $d =$ -0.42 [-0.63, -0.22]) or gender indeterminate primary characters ($t(70) =$ -3.27, $p = 0.002$; $d =$ -0.39 [-0.58, -0.19]; Figure 4b) tended to be more often read to boys. Notably, the effect size for girls was nearly

twice that of for boys, suggesting that there was a stronger bias to read books with female content to girls, relative to books with male content to boys. There was no bias in audience gender for books with multiple primary characters of different genders($t(16) = 0.3$, $p = 0.77$; $d = 0.07$ [-0.35, 0.77]) or books without primary characters ($t(20) = -1.09$, $p = 0.29$; $d = -0.24$ [-0.62, 0.17]).

In sum, these findings suggest that children's books tend to communicate to children information about how to behave as normative members of their own 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 249 contemporary children's books and analyzed the extent to which they contain gender stereotypes. 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. We then used word embedding models to explore the semantic associates of words in the corpus, and found that gender-biased words formed gender stereotypical categories (e.g., social interaction for females; physical interaction for males). Further, 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) were both reflected in the language statistics of the corpus itself, and were more exaggerated than in comparable adult fiction. At the book level, we found that books varied in their gender associations, and contained statistical regularities reflecting gender stereotypes (e.g., girl characters tend to do stereotypically girl activities). These statistical regularities were stronger for female stereotypes, relative to male stereotypes. Finally, we derived a novel metric for measuring the gender distribution of a book's audience using automated analysis of book reviews and found that children tended to be exposed to books that conveyed gender stereotypes about their own gender. Our work provides the first quantitative assessment of the nature of gender messages within 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 oppose. 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 reflect gender stereotypes for their own gender. This means that children tend to have more access to information that is biased toward gender-consistent preferences, thereby making gender-inconsistent preferences less familiar to children (and therefore more difficult to emulate). Further, children are more likely to imitate same-gender models (Bussey & Bandura, 1999). These factors suggest that children's books, coupled with children's cognitive and social biases, may 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 it is an open question whether they actually do. 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 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 three years of age. 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 in domains like STEM fields (Bian et al., 2017). Changing the books that children read with their caregivers is a relatively straight-forward intervention that could potentially have a large impact on children's gender stereotypes.

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*.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

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*.

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.

Gelman, S. A., & Taylor, M. G. (2000). Gender essentialism in cognitive development. *Toward a Feminist Developmental Psychology*, 169–190.

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.

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.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990.

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. (2019). What are we learning from language? Associations between gender biases and distributional semantics in 25 languages. *https://psyarxiv.com/7qd3g*.

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.

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.

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.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207.

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.

Zeno, S., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.

Zevin, J. D., & Seidenberg, M. S. (2004). Age-of-acquisition effects in reading aloud: Tests of cumulative frequency and frequency trajectory. *Memory & Cognition*, *32*(1), 31–38.