Children's books are an early source of gender knowledge

Ellen Converse[1], Molly Y. Lewis[1, 2, 3], Matt Cooper Borkenhagen[1], Gary Lupyan[1], & Mark S. Seidenberg[1]

University of Wisconsin - Madison[1], University of Chicago[2] , Carnegie Mellon University[3]

Reading books to children ("shared reading") is an important activity with numerous benefits. For the child-learner, books are an important source of information about reading, language, and the world. By age 5 children have developed expectations about characteristics associated with gender, which are thought to arise from parental modeling, and other sources. Given that books are an early source of knowledge of various types, the language they contain may contribute to the development of gender stereotypes, among others. **QUESTION:** ***Does the gender of the language in children's books vary across words and texts?***

METHODS

A corpus of 247 children's books (birth to 5 years old) was created to examine the language and content of the text, building off two previous databases (Hudson Kam & Matthewson, 2017; Montag, Jones, & Smith, 2015).

Books were transcribed yielding a database of words (N = 181,225).

For a subset (n = 2,275) content words and other common words, adult participants provided ratings of their gendered-ness ("very male" to "very female"). 40% coverage per book was normed.

Data were also collected to investigate other word-level semantics, age of acquisition, other properties in relation to words' gender.

RESULTS

Words rated as more feminine are more emotionally charged (valence), less concrete, and learned earlier in life (AoA) – see Table 1, below.

All five word-level measures predicted independent variance.

With R^2 = .17, the linear model (word level characteristics predicting word-level gender) suggests that while important, there is much left to be explained about what drives the word-level effects of gender.

Table 1

Pairwise Correlation Coefficients for Word-level Measures

Measure	1	2	3	4	5	6
1. Gender (femaleness)	--					
2. Arousal	-0.07*	--				
3. Valence	0.36*	-0.09*	--			
4. Concreteness	-0.12*	-0.18*	0.00	--		
5. Age of Acquisition	-0.08*	0.02	-0.19*	-0.25*	--	
6. Frequency (TASA)	0.00	-0.07*	0.15*	-0.21*	-0.4	--

Note. Pearson's r is shown. Correlations that were $r < .05$ are given with an asterisk (*). Word frequency was taken from the TASA norms (Zeno et al., 1995) and log transformed for the purposes of these calculations.

RESULTS (CONTINUED)

Words across texts co-occupied semantic clusters that were often reliably gendered (Table 2, below).

When gender ratings were aggregated for each book (average gender across all words within book), the overall gender bias varies dramatically across all books (M = 3.00, SD = .18, min = 2.59, max = 3.86) as shown in Figure 1, right. The dashed line indicates the overall mean across books, and color indicates the gender of the primary character.

Noteworthy variation was found within books (bars represent 95% confidence intervals) for word-level gender as well (Figure 1).

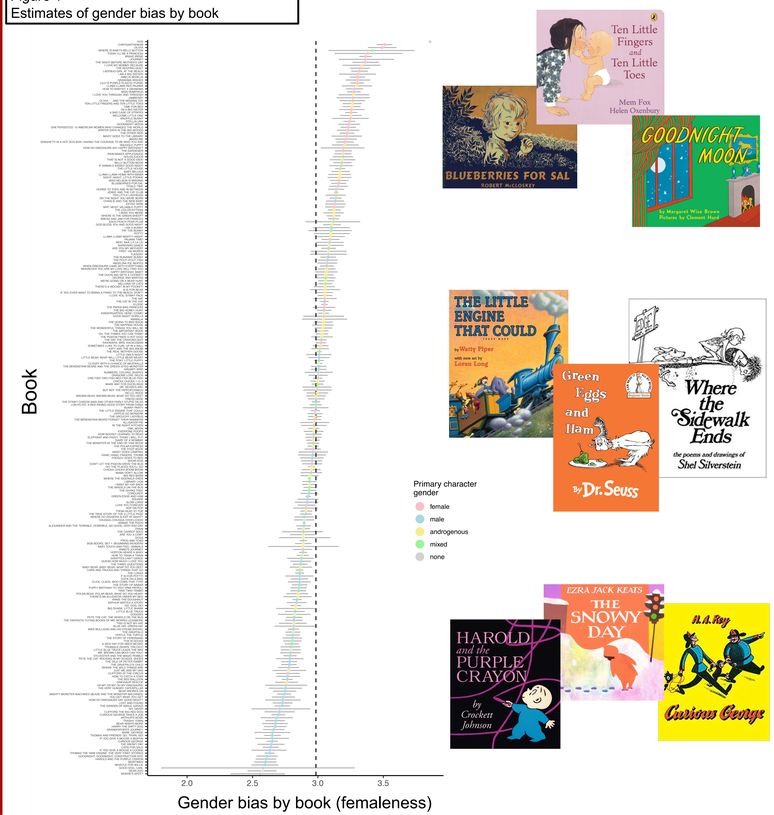
Table 2

Examples of Gendered Clusters from Multidimensional Embeddings

Category	# of words	Examples
Male clusters		
Animal	32	wolf, cow, goat, llama, dog, donkey, moose
Transmissions	20	burst, fast, instant, mail, signal, package
Self-movement	17	climb, feet, jump, ladder, steps, walking
Tools, work	15	blade, gun, knife, shovel, bulldozer, loader, wheelbarrow
Professions	27	fireman, doctor, captain, clerk, judge, mayor, principal
Female clusters		
Modifiers, abstract	77	absolutely, agree, certainly, doing, idea, imagine, else, exactly
Affection, family	31	baby, friend, girl, grandma, heart, jealous, kiss, tears
Food and taste	26	chocolate, coffee, flavors, gum, milk, sprinkles, sweet
Flowers, fruits	15	cherry, daisies, flower, garden, maple, peach, rose
Knitting, sewing	7	knit, rug, sew, yarn, silk, bobbin

Note. Sample of clusters derived from English fastText embeddings (www.fasttext.cc) for words in corpus. For "male" clusters, the mean gender rating of the words was significantly greater than the mean for all words. For "female," the mean rating was significantly smaller than the grand mean. Cluster labels serve as approximate descriptions.

Figure 1
Estimates of gender bias by book



DISCUSSION

Noteworthy variability exists with regards to the genderedness of words that are used throughout children's books. Also, we see that children's books when analyzed as units, contain dramatic variability in aggregated gender bias based on the words they contain.

This suggests that children's books are an early form of gender knowledge, and exposure to books will drive learning of gender biases in ways that are likely to vary as a function of the book being read.

Future work should examine the relationship between the text's language and illustrations, as well as higher order language structures that contribute to systematic gendered-ness beyond individual words

REFERENCES

Hudson Kam, C. L., & Matthewson, L. (2017). Introducing the infant bookreading database (IBDb). *Journal of child language*, *44*(6), 1289-1308.

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.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). The educators word frequency guide. New York: Touchstone. *Applied Science*.

ACKNOWLEDGEMENTS

This work has been supported by the Vilas Trust at UW-Madison and by 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. Additional funding provided by the University of Wisconsin-Madison L&S Honors Program through a Summer Senior Thesis Research Grant awarded by the L&S Honors Program.