Estimating demographic bias on tests of children's early vocabulary

Anonymous CogSci submission

Abstract

Children's early language skill has been linked to later educational outcomes, making it important to accurately measure early language. Parent-reported instruments such as the Communicative Development Inventories (CDIs) have been shown to provide valid, consistent measures of children's aggregate early language skill. However, CDIs are predominantly comprised of hundreds of vocabulary items, some of which may not be heard (and thus learned) equally often by children of varying backgrounds. Here, we use a database of American English CDIs to identify words that show strong bias for particular groups of children, on dimensions of sex (male vs. female), race (white vs. non-white), and socioeconomic status (high vs. low). For each dimension, we identify dozens of strongly biased items, and show that eliminating these items reduces the expected ability difference between groups. For sex, we consider how to propose replacement words that may show less bias, on the basis of their relatively equal frequency in adult speech directed to male and female children.

Keywords: language acquisition; word learning; measuring instrument bias; development;

Introduction

Researchers, clinicians, and parents have long been fascinated with the surprising speed and variability in the growth of young children's vocabulary. Children's early vocabulary growth is assumed to reflect not only their exposure to child-directed speech, but also the varying difficulty of different types of words, and individual differences in the aptitude of the child – including potential language deficits. Research has uncovered both within-child consistency in development, as well as significant influence from external factors. For example, Bornstein, Hahn, & Putnick (2016) found stability in core language skills across 10 years of children's development, despite changes in maternal income and education over the study period. Yet socioeconomic status (SES) has also been found to be predictive of children's early language skill, and of later educational outcomes (for a review, see Schwab & Lew-Williams, 2016). Demographic factors are also predictive of language skill: first-born children tend to outpace their siblings, and female children tend to have better language skills than their age-matched male counterparts (Fenson et al., 1994) – a sex-based verbal advantage that continues through high school (see Petersen, 2018 for a review). However, it is also difficult to measure language skill: in any short recording children are unlikely to use all of the words and constructions that they know, and any comprehensive battery of language measures will likely exhaust children's attention span.

The MacArthur-Bates Communicative Development Inventories (CDIs; Fenson et al., 2007) are a set of parent-reported measures of children's productive and receptive language skills, which produce a low-cost and reliable way to estimate children's early language skills (Fenson et al., 1994). CDIs have shown good predictive validity [e.g., Fenson et al. (1994); Bornstein et al. 2012, Duff et al. 2015]. Our focus will be on the vocabulary checklist portion of the CDI Words & Sentences (CDI:WS) form, comprised of 680 early-learned words across 22 categories (e.g., animals, vehicles, action words, pronouns) selected to assess the productive vocabulary of children 16 to 30 months of age. For each item on the CDI:WS, caregivers are asked to respond whether the target child has been heard to say (i.e. produce) the given item.[1] Children's total vocabulary score on the CDI:WS is tightly correlated with other facets of early language (e.g., grammatical competence and gesture), suggesting that the language system is "tightly woven" (Frank, Braginsky, Yurovsky, & Marchman, 2021). Due to these desirable properties, CDIs have been adapted to dozens of languages, and a central repository of CDI data contributed from all over the world has been created (Wordbank; Frank, Braginsky, Yurovsky, & Marchman, 2017; Frank et al., 2021).

Inspired by the utility and widespread use of the CDI, researchers have recently been using psychometric models on CDI data to construct short, adaptive tests to reliably assess language ability using only a small subset of the CDI items (e.g., Mayor & Mani, 2019; Kachergis, Marchman, Dale, Mankewitz, & Frank, 2021). These psychometric models typically come from the Item-Response Theory (IRT) framework (Baker, 2001), which assumes that not only test-takers (here, children) have normally-distributed ability, but that items (words on the CDI) have normally-distributed difficulty. The very efficacy of these IRT-based models depends on words varying in difficulty, and hence being more/less informative of the ability level of different individuals. For example, asking whether a 22-month-old produces the word "ball" is far less informative of that child's language ability than asking whether they produce "table", as 96% of 22-month-olds can

[1] The CDI: Words and Gestures form includes a subset of the CDI:WS items and targets children 12 to 18 months of age, measuring both comprehension and production.

produce the former, while only 47% produce the latter.

While it is quite reasonable to expect that some CDI words are easier than others, and even to use these varying difficulties to predict variation in children's language ability, the use of psychometric models highlights the possibility that some CDI items may function differently (i.e., be more/less difficult) for different groups of children. The idea that some items on a test may show bias, favoring one group over another, is known as Differential Item Function (DIF; Holland & Wainer (1993)]. On any given test, it is clearly undesirable to have more items favoring one group (say, children from rural households) over another (urban children), as the test will overestimate the ability of test-takers in the former (rural) group – and not because of any underlying mean difference in ability between the groups, but simply because the test is unfair (Camilli, 2013). A variety of statistical methods for detecting DIF have been proposed, and investigations have in several instances identified DIF for many items on tests favoring particular one group over another (e.g., rural vs. urban). Our goal here is to test the items on the American CDI:WS for DIF along three main axes: sex (male vs. female), socioeconomic status (low- vs. high-SES), and race (white vs. non-white).

However, DIF is fundamentally difficult to identify for multiple reasons (for an overview, see Stenhaug, Frank, & Domingue, 2021). First, most techniques to identify DIF rely on defining a set of "anchor" test items that are assumed to be equally difficult (i.e., unbiased) for both groups of interest. Identifying anchor items is at best fraught when there may in fact be a difference in ability between groups (e.g., the female language advantage), and is further confounded when the magnitude of this ability difference is unknown. A reason that DIF is particularly tricky in measuring children's early language ability is that there is a finite universe of early-learned words to choose from–and we may expect many of them to be biased for various environmental reasons (e.g., children in Florida may not use mittens or skis). Hence, the presence of DIF on a wide variety of items may not indicate the presence of bias; it could indicate that one demographic has a higher average ability level than the other. For example, let male language ability be drawn from a standard normal μ_{male} $N(0,1)$, with female language ability slightly higher, on average ($\mu_{female} = \mu_{male} + 0.1$). Then we would expect items to be an average of 0.1 easier for females than for males, and we might identify items that are instead easier for males than females as showing undesirable DIF. And yet, without knowing the actual ability difference between two demographic groups–for which we also rely upon our tests, it is difficult to adjudicate which items show DIF, and which do not.

The outline of this paper is as follows. First, we introduce the Wordbank data and the IRT model, and use it examine the overall size of demographic differences in early word learning. We then fit the IRT model to each group along each demographic dimension, and examine the item parameters for evidence of DIF, noting in particular how many items are significantly biased in favor of each group. We identify a set of suspect items to eliminate (or in the future, replace), and provide updated estimates of the effect size of these demographic variables, were the biased items to be eliminated. Finally, we provide recommendations for next steps to be taken to identify and replace biased items on the CDI.

Methods

Vocabulary Data

Participants We analyze parent-reported Wordbank data from 5520 American English CDI: Words & Sentences administrations for children 16 to 30 months of age (Frank et al., 2017, 2021). Full demographic data are not reported in some datasets contributed to Wordbank: sex was available for 4094 children, race/ethnicity was available for 2715 children, and mother's education (a proxy for socioeconomic status; SES) was available for 5520 children.

The analysis of sex-based differences included CDI administrations from 1989 female and 2105 male children. The analysis of race-based differences included data from 2202 white, 67 Asian, 222 Black, 131 Hispanic, and 93 "Other" children. For this analysis we categorized participants' race/ethnicity as White (2202) or Non-white (513). For the SES-based analysis, we categorized the 4973 children whose mothers had at least some college or more as high-SES, and those whose mothers had at most high school (547 children) as low-SES.

Rasch Model

The Rasch model, also known as the 1-parameter logistic (1PL) model, is the simplest Item Response Theory model, and is thus the easiest to use to investigate potential differences in item function across different groups of participants. The goal of the Rasch model is to jointly estimate for each child j a latent ability θ_j, and for each item i a difficulty parameter b_i. In the model, the probability of child j knowing (i.e., producing or understanding) a given item i is

$$P_i(x_i = 1|b_i, \theta_j) = \frac{1}{1 + e^{-D(\theta_j - b_i)}}$$

where D is a constant scaling parameter ($D = 1.702$) which makes the logistic closely match the ogive function in traditional factor analysis (Chalmers, 2012; Reckase, 2009). Child ability (θ) and item difficulty (b) distributions are standardized (i.e., mean of 0), and expected to be normally-distributed. Children with high latent ability (θ) will be more likely to produce any given item than children with lower latent ability, and more difficult items will be produced by fewer children (at any given θ) than easier items.

In the multigroup Rasch model, an item's difficulty is allowed to vary depending on which group an individual child belongs to. For example, in the sex-based multigroup model, item i's difficulty is b_i^{female} for females, and b_i^{male} for males.

We fitted a multigroup Rasch model for each demographic dimension (sex, socioeconomic status, and ethnicity). To contextualize these results, we also first fitted a baseline Rasch model

Results

Demographic effects in baseline Rasch model

Figure 1 shows children's language ability vs. age by demographic group from the baseline Rasch model, which assumes no DIF (i.e., equal item parameters for all groups). A linear regression for each demographic group, with age (centered) and its interaction, showed significant effects in the expected directions, as can also be seen in the figure. Female children had higher language ability than male children ($\beta = 0.56$, $p < .001$), with no significant interaction with age ($\beta = 0.02$, $p = .10$. High-SES children had higher language ability than low-SES children ($\beta = 0.23$, $p = .02$), an advantage that grew with age ($\beta = 0.11$, $p < .001$). White children had higher language ability than non-white children ($\beta = 0.50$, $p < .001$), an advantage that grew with age ($\beta = 0.08$, $p < .001$). We will re-examine these demographic regressions after trimming items showing extreme DIF, which is what we turn to now.

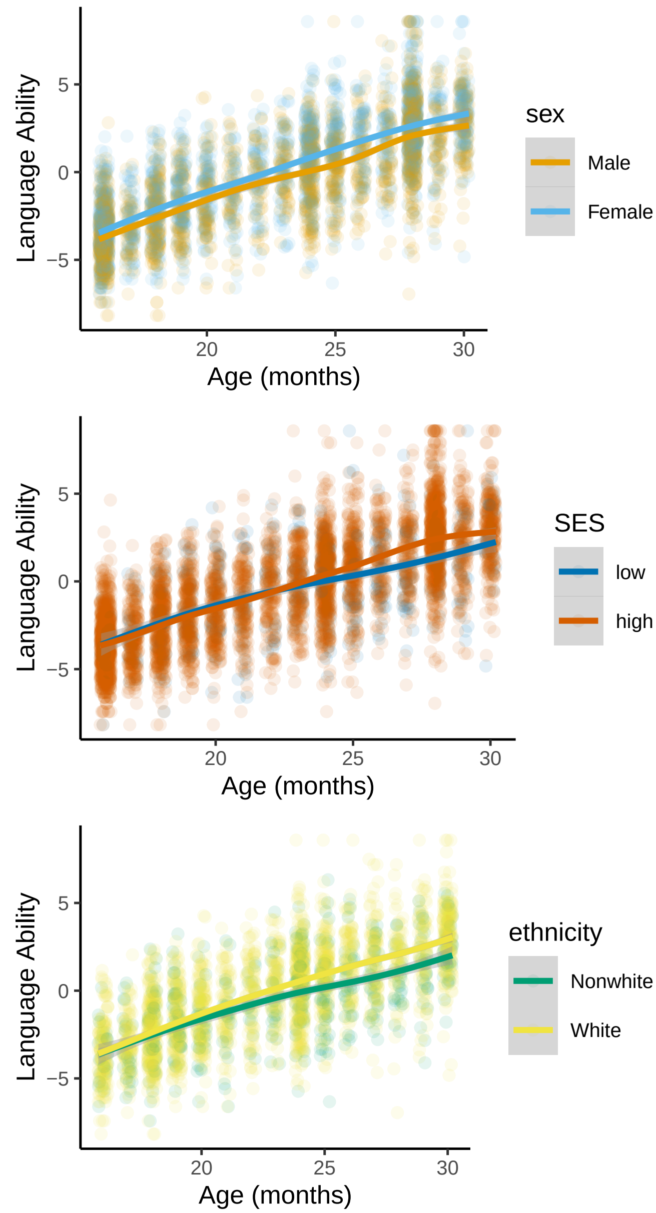

Figure 1: Language ability vs. age by demographic group, from the Rasch model.

Identifying biased CDI items

The first step we take to identify CDI items with DIF is to create GLIMMERs (Graphs of Logits Imputed Multiply with Means Equal; Stenhaug et al. (2021)), which visualize between-item variation in group performance differences. These parameters are drawn from a fitted multigroup Rasch model for each demographic variable (sex, SES, and race), with the assumption that the mean language ability in each group is the same (e.g. for sex, $\mu_{male} = \mu_{female} = 0$), thus pushing all between-group variation into the item difficulty parameters. For the case of sex, where we believe $\mu_{female} > \mu_{male}$, this means we may expect to find many items with difficulty $b_j^{female} < b_j^{male}$, but we can still examine the distribution of differences in item difficulty ($d_j = b_j^{male} - b_j^{female}$) for outliers. To give analysts a sense of the uncertainty about the existence of DIF, GLIMMER plots show distributions of parameter differences rather than point estimates. These distributions are generated by drawing 10,000 imputations from the item parameter covariance matrix.

Figures 2-4 show GLIMMERs for a selection of CDI items for sex, SES, and race, respectively. The full GLIMMERs, with all 680 CDI:WS items, are available on OSF, but were too large to include here. Nonetheless, it is important to inspect the full plots, for if there is a cluster of items in a GLIMMER, the analyst may conclude that these items are strong candidates for DIF on that dimension. For example, there is some clustering at the top and bottom of Figure 2: at the top, "vagina", "tights", "dress (object)", and "doll" form a cluster of items that are much easier for females, while at the bottom, "penis" stands out as much easier for males. For the SES and race GLIMMERs (Figures 3 and 4)–and in the rest of the full sex GLIMMER, there are not clusters, but rather a continuum of smoothly varying differences with overlap. This makes identifying items with DIF quite difficult, as different methods are likely to yield inconsistent results (Stenhaug et al., 2021).

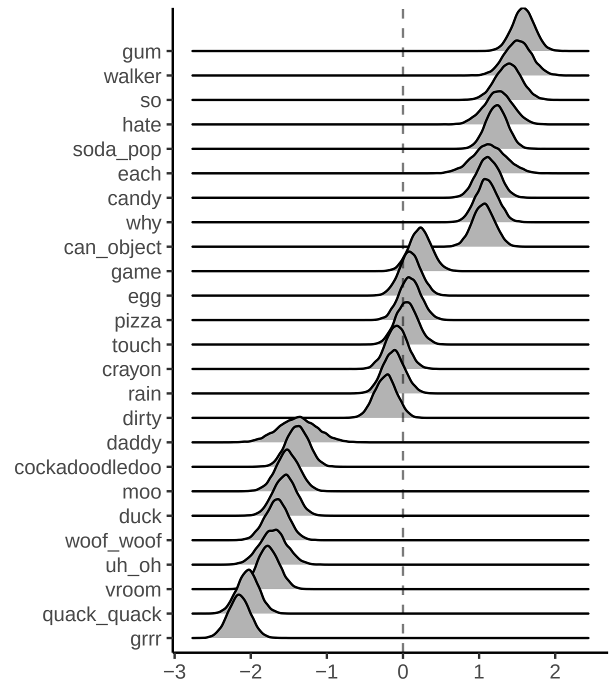

Figure 3: GLIMMER plot of a sample of CDI:WS words from the SES bias model. Words at the top are easier to learn for children from low-SES families, while those at the bottom are easier for those from high-SES families.

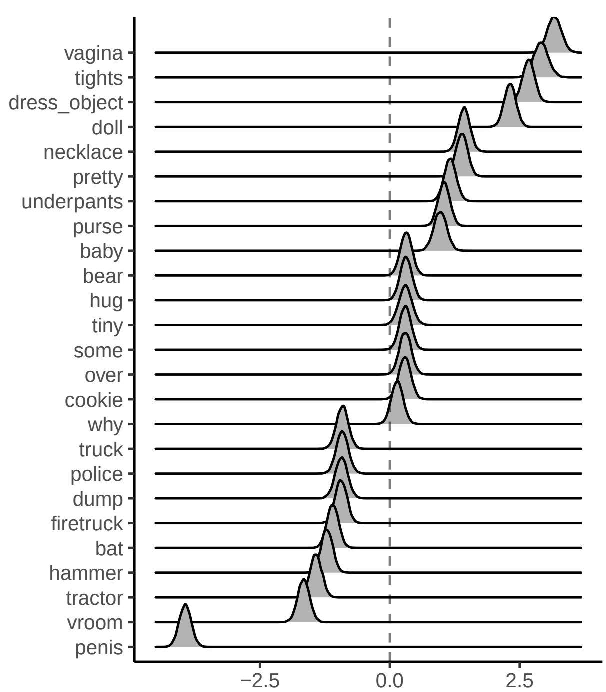

Figure 2: GLIMMER plot of a sample of CDI:WS words from the sex bias model. Words at the top are easier to learn for females, while those at the bottom are easier for males.

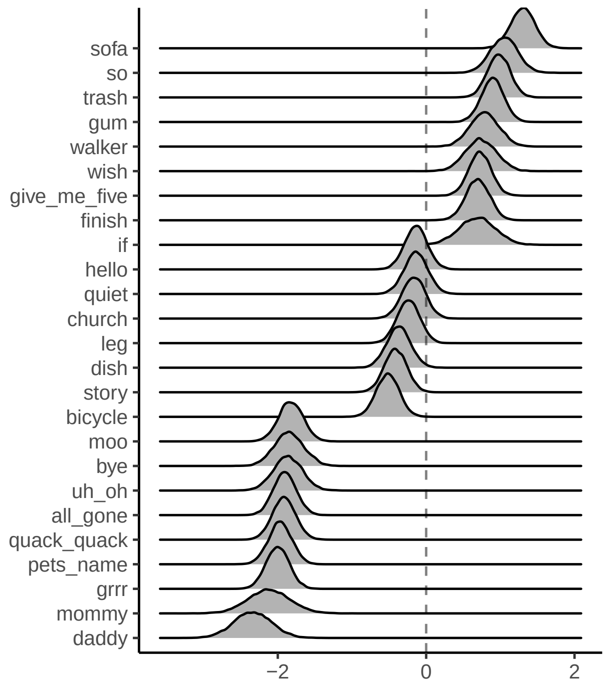

Figure 4: GLIMMER plot of a sample of CDI:WS words from the ethnicity bias model. Words at the top are easier to learn for children from low-SES families, while those at the bottom are easier for those from high-SES families.

Demographic effects after trimming

Having identified 76 CDI items that function differently across demographic groups – including 19 that

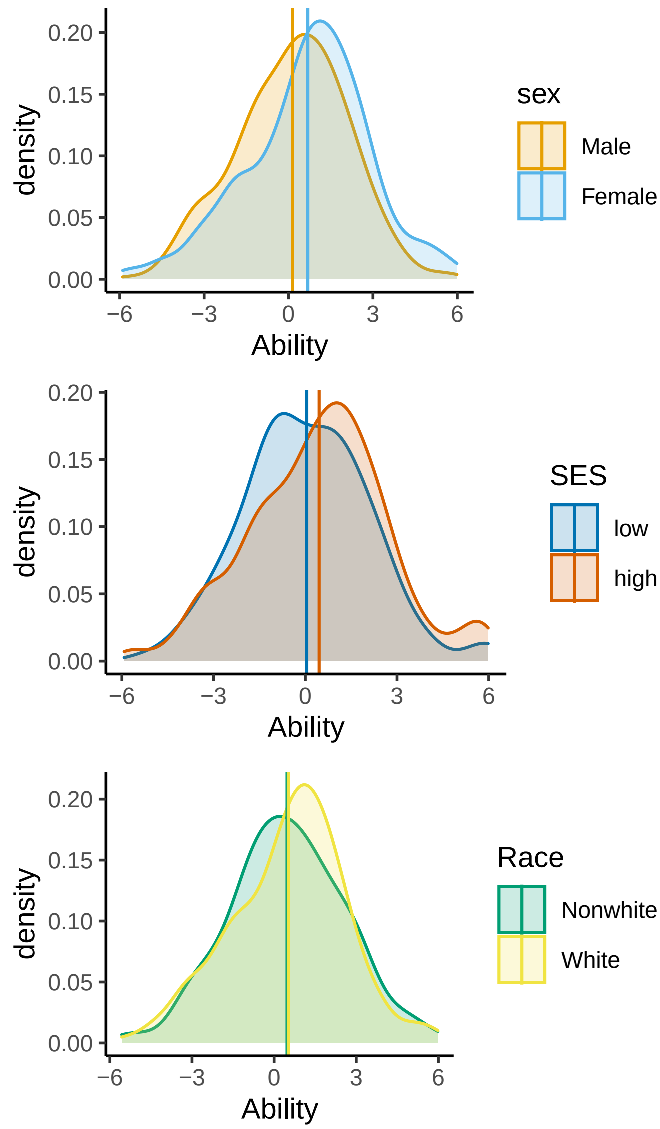

Figure 5: Density plots of language ability for children aged 23-25 months, based on the group models.

Discussion

We investigated a popular test of children's early vocabulary for potential demographic bias, examining the distribution of words' estimated difficulties for high- vs. low-SES children, females vs. males, and for white vs. non-white children.

Acknowledgements

Place acknowledgments (including funding information) in a section at the end of the paper.

References

Baker, F. B. (2001). *The basics of item response theory*. ERIC.

Bornstein, M. H., Hahn, C.-S., & Putnick, D. L. (2016). Stability of core language skill across the first decade of life in children at biological and social risk. *Journal of Child Psychology and Psychiatry*, *57*(12), 1434–1443.

Camilli, G. (2013). Ongoing issues in test fairness. *Educational Research and Evaluation*, *19*(2-3), 104–120.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. `http://doi.org/10.18637/jss.v048.i06`

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., … Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, i–185.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories: User's guide and technical manual (2nd ed.)*. Baltimore, MD: Brookes.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The wordbank project*. MIT Press.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Routledge.

Kachergis, G., Marchman, V. A., Dale, P., Mankewitz, J., & Frank, M. C. (2021). Online computerized adaptive tests (cat) of children's vocabulary development in english and mexican spanish.

Mayor, J., & Mani, N. (2019). A short version of the MacArthur-Bates Communicative Development Inventories with high validity. *Behavior Research Methods*, *51*(5), 2248–2255.

Petersen, J. (2018). Gender difference in verbal performance: A meta-analysis of united states state performance assessments. *Educational Psychology Review*. Springer.

Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory* (pp. 79–112). Springer.

Schwab, J. F., & Lew-Williams, C. (2016). Language learning, socioeconomic status, and child-directed speech. *WIREs Cognitive Science*, *7*, 264–275. `http://doi.org/10.1002/wcs.1393`

Stenhaug, B., Frank, M. C., & Domingue, B. (2021). Treading carefully: Agnostic identification as the first step of detecting differential item functioning.