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 contain hundreds of vocabulary items, some of which may not be heard (and thus learned) equally often by children of varying backgrounds. This study used a database of American English CDIs to identify words that showed strong bias for particular demographic groups of children, on dimensions of sex (male vs. female), race (white vs. non-white), and maternal education (high vs. low). For each dimension, we identified dozens of strongly biased items, and showed that eliminating these items reduced the expected ability difference between groups. Additionally, we investigated how well the relative frequency of words spoken to young girls vs. boys predicted sex-based word learning bias, and discuss possible sources of demographic bias in early word learning.

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

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. Children show both consistency in some skills across 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 maternal education, often used as a proxy for socioeconomic status (SES), has also been found to be associated with children's language processing and vocabulary by 18 months (Fernald, Marchman, & Weisleder, 2013), and to be predictive of later educational outcomes (Marchman & Fernald, 2008; see Schwab & Lew-Williams, 2016 for a review). 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 get a complete measure of young children's language skill: long recordings are prohibitively difficult to collect and transcribe (see Roy, Frank, DeCamp, Miller, & Roy, 2015, the exception that proves the rule), and yet any short recording (e.g., a 1-hour play session) will elicit only a small proportion of the words and constructions that children know. Thus, researchers of early word learning have constructed tests with hundreds of words, intentionally oversampling words that are more likely to be known by young children.

We focus on the MacArthur-Bates Communicative Development Inventories (CDIs; Fenson et al., 2007), a set of parent-reported measures of children's productive and receptive language skills, which offer 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 & Putnick, 2012; Duff, Reen, Plunkett, & Nation, 2015). Our primary focus is 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. 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 being reported on) have normally-distributed ability, but that items (words on the CDI) have normally-distributed

difficulty. The efficacy of these IRT-based models depends on words varying in difficulty, in order for the test to be more informative of the ability level of each individual. 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 produce the former, while only 47% produce the latter.

While it is quite reasonable to expect that some CDI words are "easier" (i.e., more likely to be known by children) 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 widely known) 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 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), maternal education (no more than secondary vs. at least some college), and race (white vs. non-white).

The outline of this paper is as follows. First, we introduce the Wordbank data and the IRT model, and use it to 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 maternal 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. Due to sparse data for many categories, we binarized participants' race/ethnicity as White (2202) or Non-white (513), recognizing that there may be important variation between groups that this will fail to capture. Data for the maternal education analysis included CDI data from children whose mother's had the following levels of education: 8 with no more than primary school education, 123 with some secondary school, 416 with no more than secondary school, 613 with some college, 870 with no more than a college degree, 162 with no more than some graduate school, and 584 with a graduate degree. Again due to data sparsity, we binarized the 4973 children whose mothers had at least some college or more as high maternal education (high-ME), and those whose mothers had at most high school (547 children) as low maternal education (low-ME).

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 Rasch model jointly estimates 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 by group. For example, in the sex-based multigroup model, item i's difficulty is b_i^{female} for females, and b_i^{male} for males. In the quest to identify DIF, a multigroup Rasch model will be fitted for each demographic dimension of interest (sex, maternal education, and ethnicity), and we will examine the between-group difficulty difference for each item (e.g., $d_i = b_i^{female} - b_i^{male}$). If there is no DIF for a given item, then $d_i \approx 0$ as the two groups find the item equally difficult.

Results

We will first examine the size of demographic effects on language ability in a baseline Rasch model fitted without regard to demographic group. We then fit a multigroup Rasch model for each examine evidence for DIF, assuming that Finally, we prune the CDI of items showing varying degrees of DIF

Demographic effects in baseline Rasch model

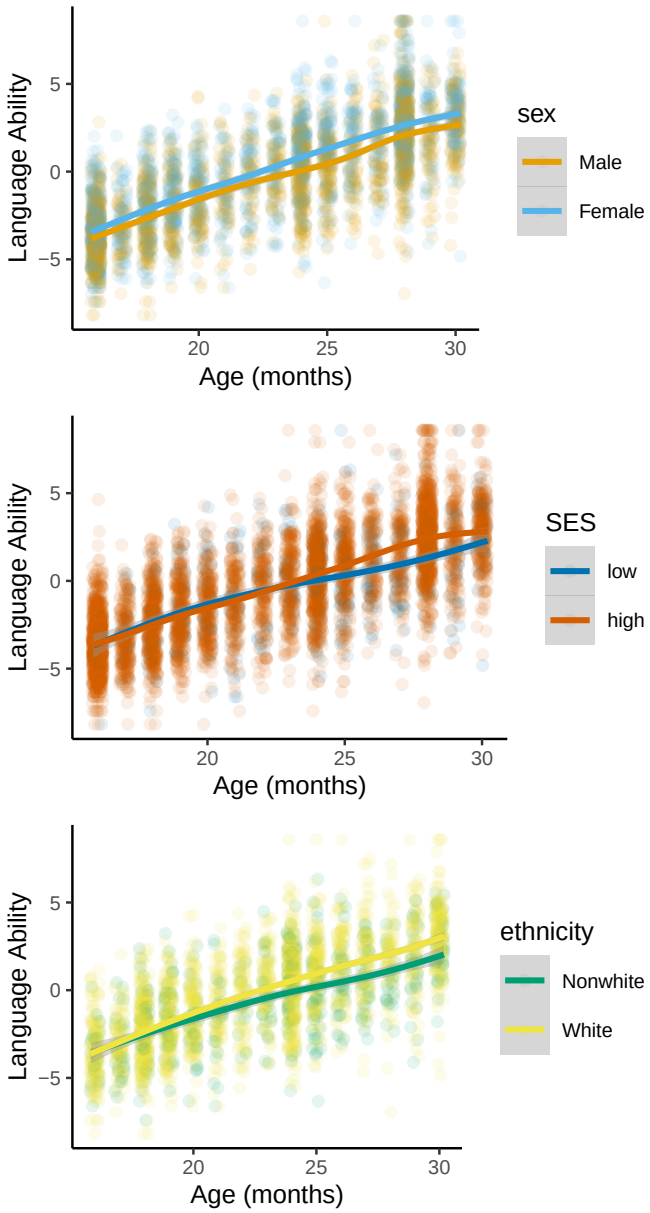

Figure 1: Language ability vs. age by demographic group, from the Rasch model.

To contextualize the later DIF results, we first fitted a baseline Rasch model to the entire dataset, without demographic information. 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. 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.

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, Frank, & Domingue (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 $b_i^{female} < b_i^{male}$, but we can still examine the distri-

bution of differences in item difficulty ($d_j = b_i^{male} - b_i^{female}$) 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.

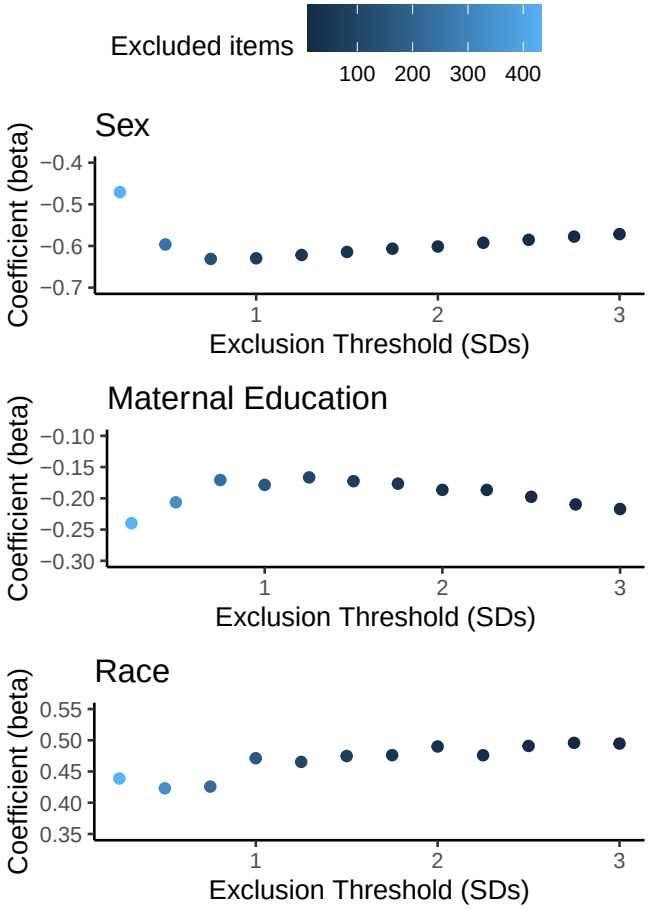

Figure 2: Size of demographic effects (regression coefficients) vs. the threshold of pruned items.

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 maternal education 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). Nonetheless, we describe the distributions of item-level group difficulty differences, and identify extreme values, flagging items with difficulty differences that were outside 2 standard deviations of the mean of that distribution.

Sex For sex, the median difficulty difference (male-female) was 0.28 (M=0.27, sd=0.4), with 593/680 items being easier for females than males. The fact that the bulk of this distribution favors females seems to confirm that the female language advantage is pervasive across the CDI:WS, and is thus likely to be a real advantage.

Maternal Education For maternal education (ME), the median difficulty difference (low-high) was -0.01 (M=0.03, sd=0.55), with 337/680 items being easier for high-ME than low-ME children. With the mean and median difficulty differences close to 0, and roughly half of the words favoring each ME group, it is tempting to conclude that the CDI:WS items are somewhat balanced with respect to ME.

Race For race, the median difficulty difference (nonwhite-white) was 0.4 (M=0.46, sd=0.55), with 549/680 items being easier for white than non-white children, revealing a fairly pervasive advantage for white children on CDI items.

This identified 25 items with extreme sex-based difficulty differences, only 7 of which were easier for females, with the other 18 items favoring males. For SES, 39 extrema were identified, only 13 of which were easier for low-SES children. For race, 31 extrema were identified, only 11 of which

were easier for non-white children. Thus, while if anything the bulk of the extrema in the sex-based model advantaged the lower language ability group (males), the extrema in the other two models mostly favored the advantaged groups: 26/39 extreme items in the SES model favored high-SES children, and 20/31 items in the race model favored white children. 17 CDI items were identified as extrema in more than one model: "tractor" and "vroom" were extreme in all three models, "choo choo" was extreme in both SES and sex models, "give me five!" was extreme in both race and sex models, and 13 other items were extreme in both SES and race models ("grrr", "moo", "quack quack", "uh oh", "duck", "owl", "candy", "gum", "walker", "daddy", "pet's name", "up", and "so"). In our final analysis, we explore the effect of pruning the 15 items that were extreme in both SES and race models.

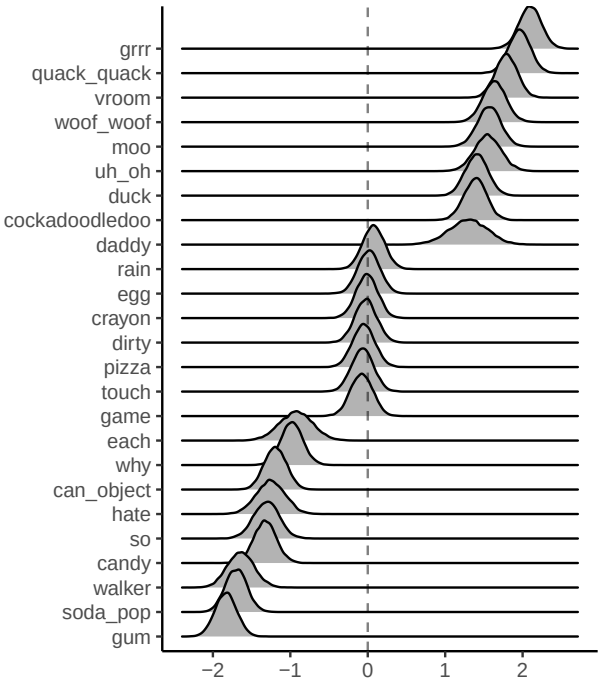

Figure 4: GLIMMER plot of a sample of CDI:WS words from the SES bias model. Words at the top are easier to learn for children from low-SES families, while those at the bottom are easier for those from high-SES families.

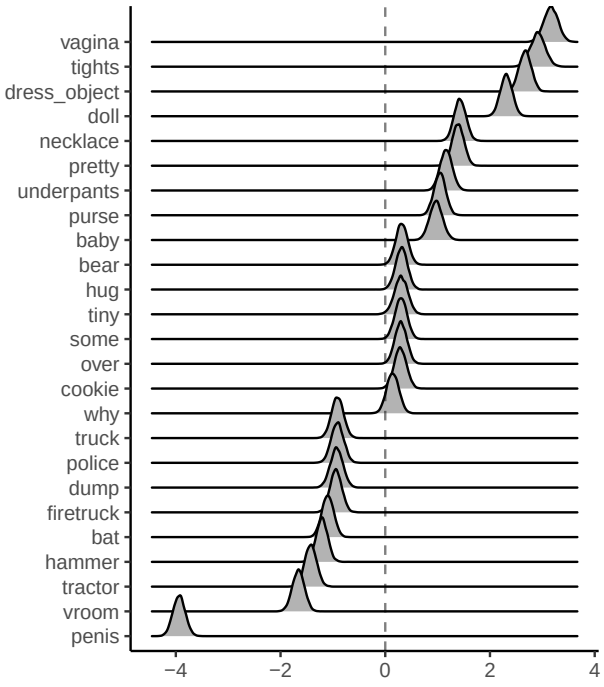

Figure 3: GLIMMER plot of a sample of CDI:WS words from the sex bias model. Words at the top are easier to learn for females, while those at the bottom are easier for males.

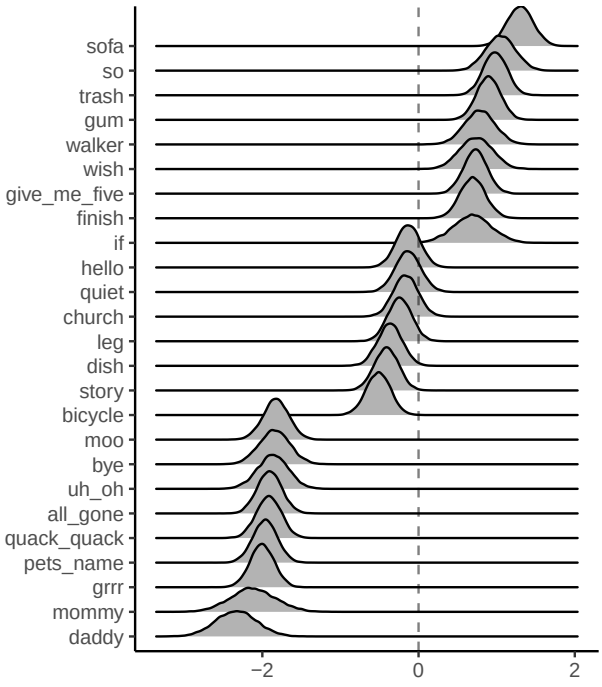

Figure 5: GLIMMER plot of a sample of CDI:WS words from the ethnicity bias model. Words at the top are easier to learn for children from low-SES families, while those at the bottom are easier for those from high-SES families.

Demographic effects after pruning

We identified 76 CDI items that function quite differently across demographic groups, and chose to prune just the 15 items with large difficulty differences for both SES and race. After pruning just these 15 items from the CDI:WS and refitting a Rasch IRT model, regressions predicting ability with each demographic variable (and age, centered) showed reduced main effects of SES ($\beta = 0.20$ vs. $\beta = 0.23$) and race ($\beta = 0.48$ vs. $\beta = 0.50$). Although these are small reductions in the SES and race effect sizes, we find them striking as we pruned only 2.2% of the CDI:WS items (and only 20% of the 76 extrema we identified), and did not specifically prune items that were biased against the disadvantaged groups (non-white, low-SES).

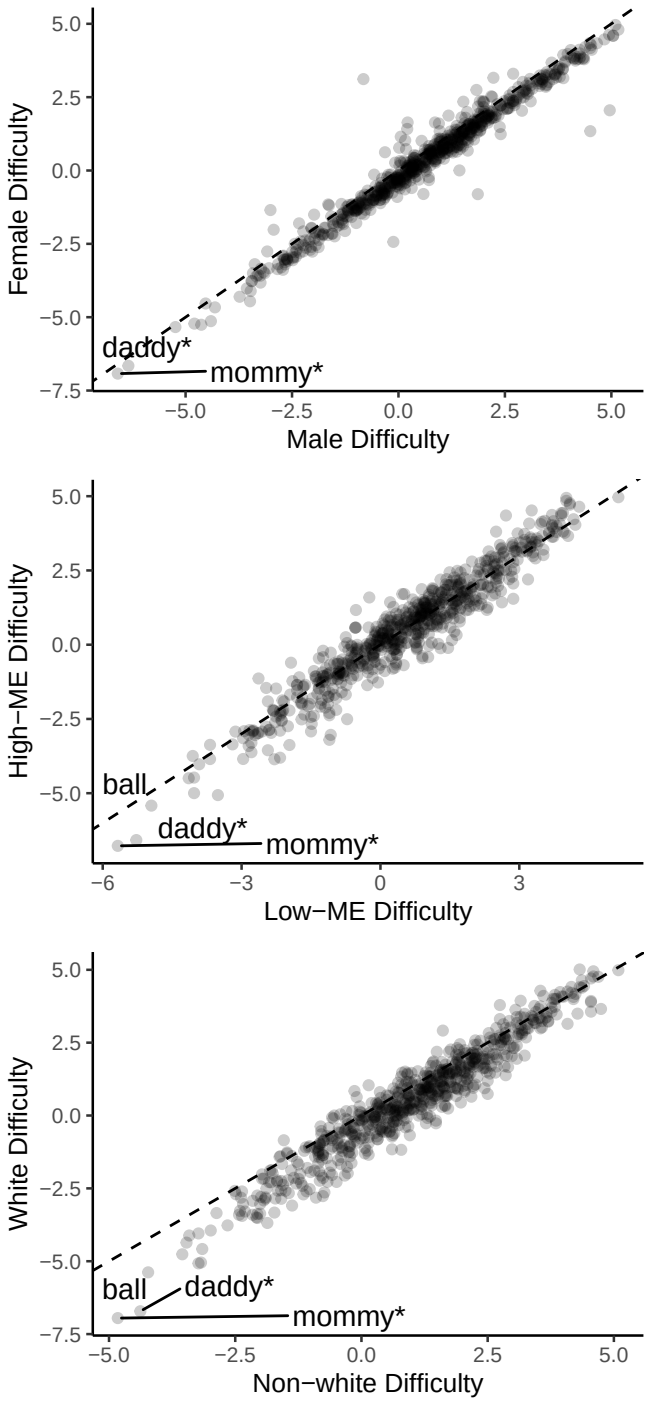

Figure 6: Scatterplots showing item difficulty per demographic group along each dimension.

Relating child-directed speech to demographic bias

Demographic differences in language ability are likely to be at least partially explained by differences in linguistic input received by children in different groups. Indeed, input quantity (total daily tokens) and some measures of quality (e.g., lexical diversity: ratio of word types vs. tokens) have been predictive of language learning outcomes in some demographic studies (Huttenlocher, Haight, Bryk, Seltzer, &

Lyons, 1991; Rowe & Goldin-Meadow, 2009). Here, we investigated the extent to which word frequency in child-directed speech to male vs. female children was predictive of the amount of DIF shown by CDI items. Similar to the approach taken by Braginsky, Meylan, & Frank (2016), we used the CHILDES corpus of transcripts from dyadic play sessions (MacWhinney, 2000), which are labeled with the sex of the target child, but not other demographic variables. (Number of males, number of females, total word counts, a couple examples...)

Overall, the correlation between the proportion of times a word was spoken to a female child and the size of the female (vs. male) advantage for that CDI word was modest, but significant ($r = 0.19$, $p < .001$).

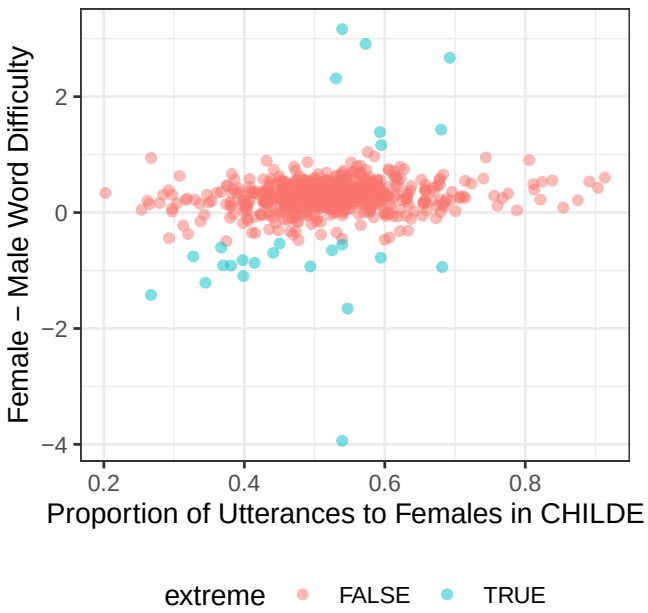

Figure 7: CHILDES frequency of input to male vs. female children.

Discussion

We investigated the CDI:WS, a popular parent-report measure 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. The IRT-based analysis revealed differential item functioning (DIF) for many items along each demographic dimension, but only in the case of sex were clear clusters of items that were more well-known to females (including feminine clothing and genitalia), and a clear item that was more well-known to males (male genitalia). For the rest of the items, and for SES- and race-based analysis, there was a smooth continuum of DIF, making the boundary of true DIF subjective, as this would rely on knowing the true difference in language ability between groups–which we reciprocally estimate from instruments like the CDI. To move forward, we identified candidate DIF items by looking at the extremes of each distribution, and found that for SES and race, the majority of these extrema (66% in both cases) were easier for the majority demographic group (i.e., high-SES, white), meaning that if these items were removed, SES and race effects on language ability as measured by the CDI:WS would decrease. In contrast, the majority of the extrema in the sex-based DIF analysis were easier for the disadvantaged group (males): if these extrema are removed, the female language advantage on the CDI:WS will likely increase. To confirm these implications, we pruned a very small number of CDI items (15 of 680) that were DIF extrema in both SES- and race-based analyses, and found that the effect of these dimensions on language ability was reduced.

The difficulty of DIF

However, DIF is fundamentally difficult to identify for multiple reasons (for an overview, see Stenhaug et al., 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 $\mu_{male}\ N(0,1)$, 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.

This investigation is only a first step in measuring demographic bias in the items on the CDI:WS. In aggregate, it seems to suggest that we may be slightly overestimating the magnitude of differences in language ability along the dimensions of SES and race, and perhaps underestimating the female advantage. However, the best way to estimate endogenous differences in ability (rather than exogenous differences, e.g. in the language environment) would be to conduct controlled, in-lab experiments measuring children's ability to learn novel words. Future research may also look into the specific items that are easier for one group than another, and determine whether these differences can be accounted for by different environmental contexts. For example, it is striking that many of the extrema favoring high-SES children are animals and animal sounds (e.g., "grrr", "quack quack", "woof

woof", "baa baa", "duck", "sheep", "giraffe", "zebra"): do high-SES households visit the zoo more, or do they often engage in other activities related to naming animals and noises they make? If certain high-SES activities are driving early word learning for these children, what are the activities (and associated vocabulary) that low-SES households are instead engaging in? A truly fair test of children's early vocabulary would contain a representative sample of words from all activities that children engage in, across demographic groups.

Acknowledgements

[Redacted for anonymous review.]

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.

Bornstein, M. H., & Putnick, D. L. (2012). Stability of language in childhood: A multiage, multidomain, multimeasure, and multisource study. *Developmental Psychology*, *48*(2), 477.

Braginsky, M., Meylan, S. C., & Frank, M. C. (2016). Gender differences in lexical input and acquisition. Boston.

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

Duff, F. J., Reen, G., Plunkett, K., & Nation, K. (2015). Do infant vocabulary skills predict school-age language and literacy outcomes? *Journal of Child Psychology and Psychiatry*, *56*(8), 848–856.

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.

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*(2), 234–248.

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.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*(2), 236.

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.

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk. Transcription format and programs* (Vol. 1). Psychology Press.

Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science*, *11*(3), F9–F16.

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.

Rowe, M. L., & Goldin-Meadow, S. (2009). Differences in early gesture explain ses disparities in child vocabulary size at school entry. *Science*, *323*(5916), 951–953.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663–12668.

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.