Differential item function and adaptive tests of early language

Or How I Started to Worry about and Love Measurement

George Kachergis

July 24, 2024

"The history of science is the history of measurement."
—James M. Cattell (1893), founder of *Psychological Review*

Input:
language environment

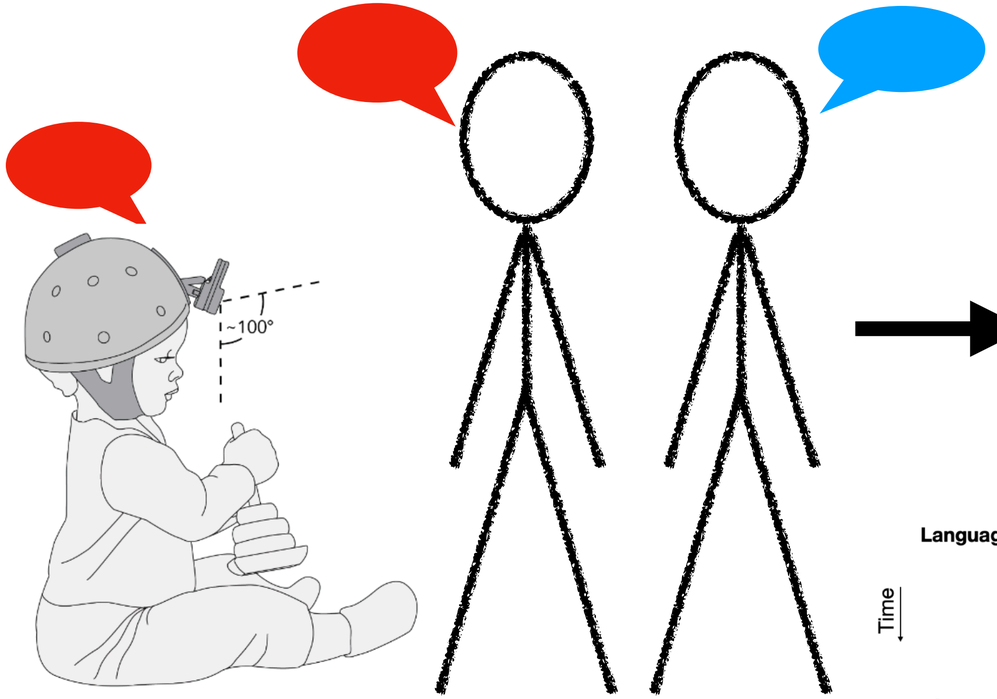
Measure early learning environments

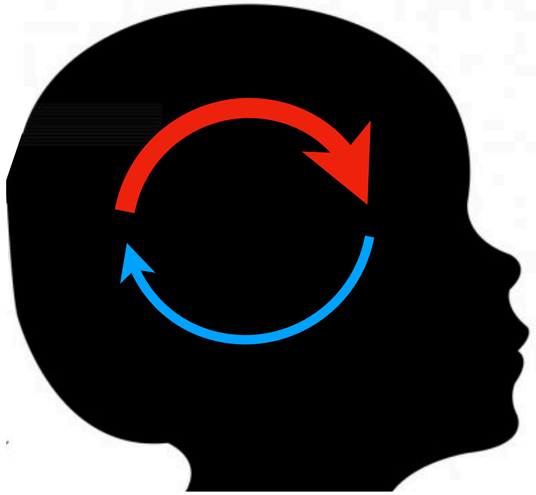
Mechanisms:
of processing and learning

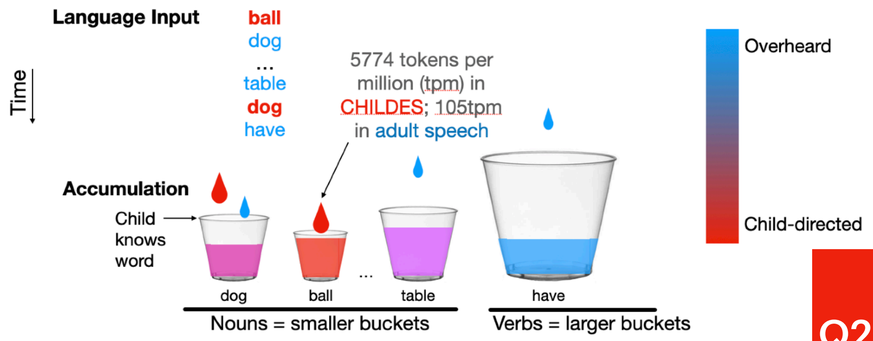
Test models (theories) of individual learning

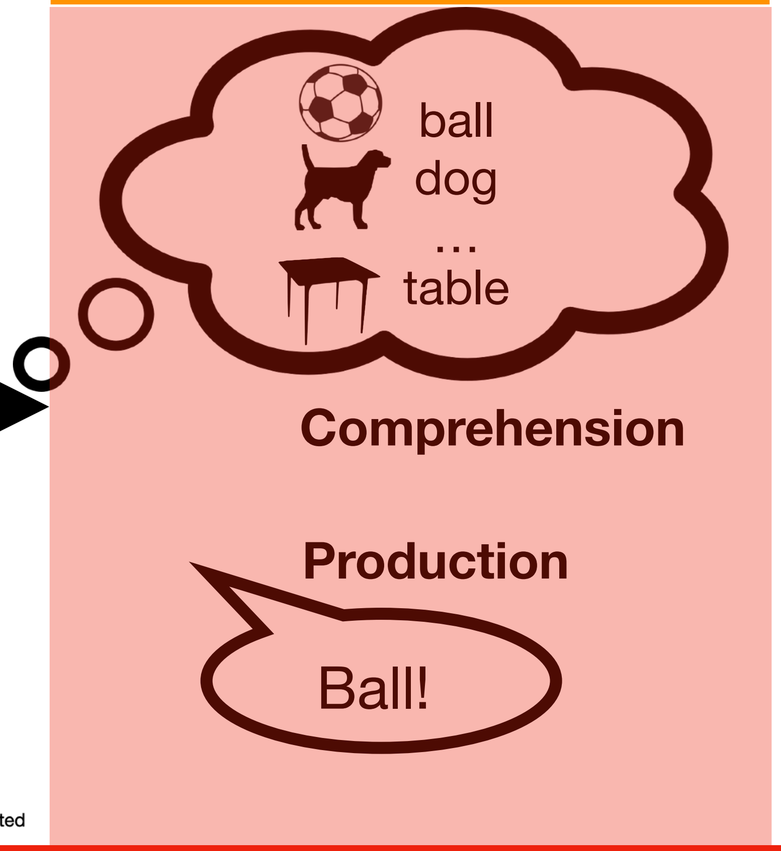
Uptake:
learning outcomes

Create short, valid, and fair tests of early language









Q1: How short can tests be?
Q2: How can we assess (& improve) fairness?

The MacArthur-Bates Communicative Development Inventory (CDI)

Parent-report measure of children's early language comprehension and production.

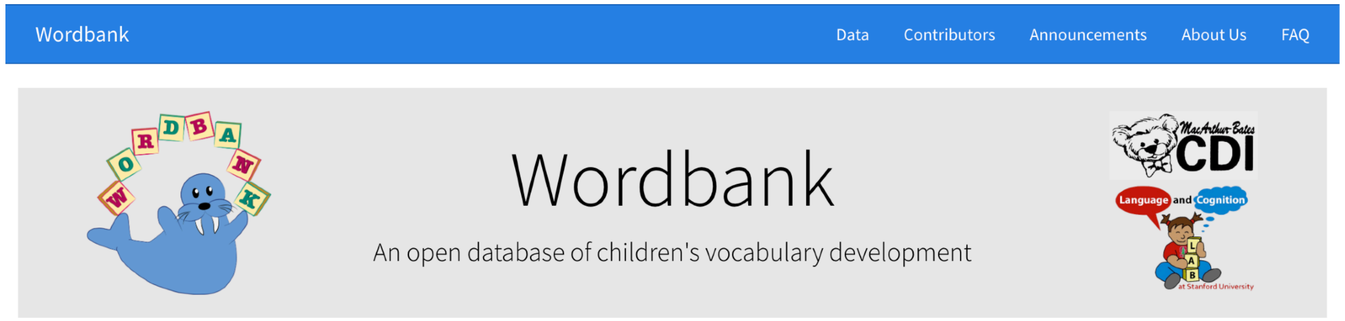
5. Food and Drink (30)

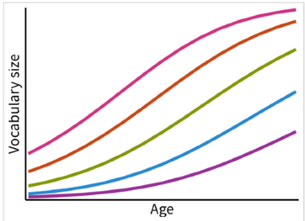
	Understands	Understands and says		Understands	Understands and says
apple	●	○	cheerios	○	○
banana	○	○	cheese	○	○
bread	○	○	chicken	○	○
butter	○	○	coffee	○	○
cake	○	●	cookie	○	○
candy	○	●	cracker	○	○
carrots	○	○	drink	○	○
cereal	○	○	egg	○	○

Impressive reliability and predictive validity

Fenson et al. (1994, 2007)

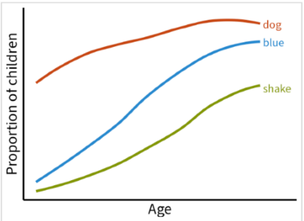
The MacArthur-Bates Communicative Development Inventory (CDI)





Vocabulary Norms

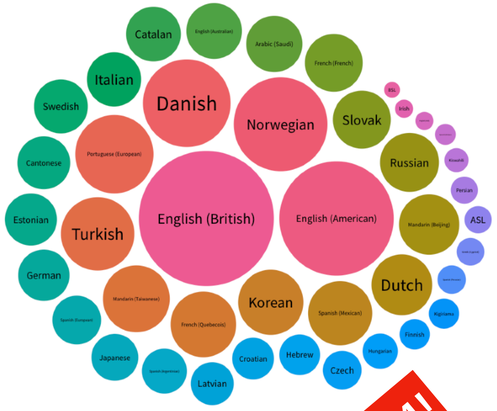
Explore vocabulary size growth curves for various languages and demographic groups.



Item Trajectories

Explore developmental trajectories of individual words in various languages.

Wordbank contains data from 92,771 children and 105,290 CDI administrations, across 42 languages and 89 instruments:



https://wordbank.stanford.edu

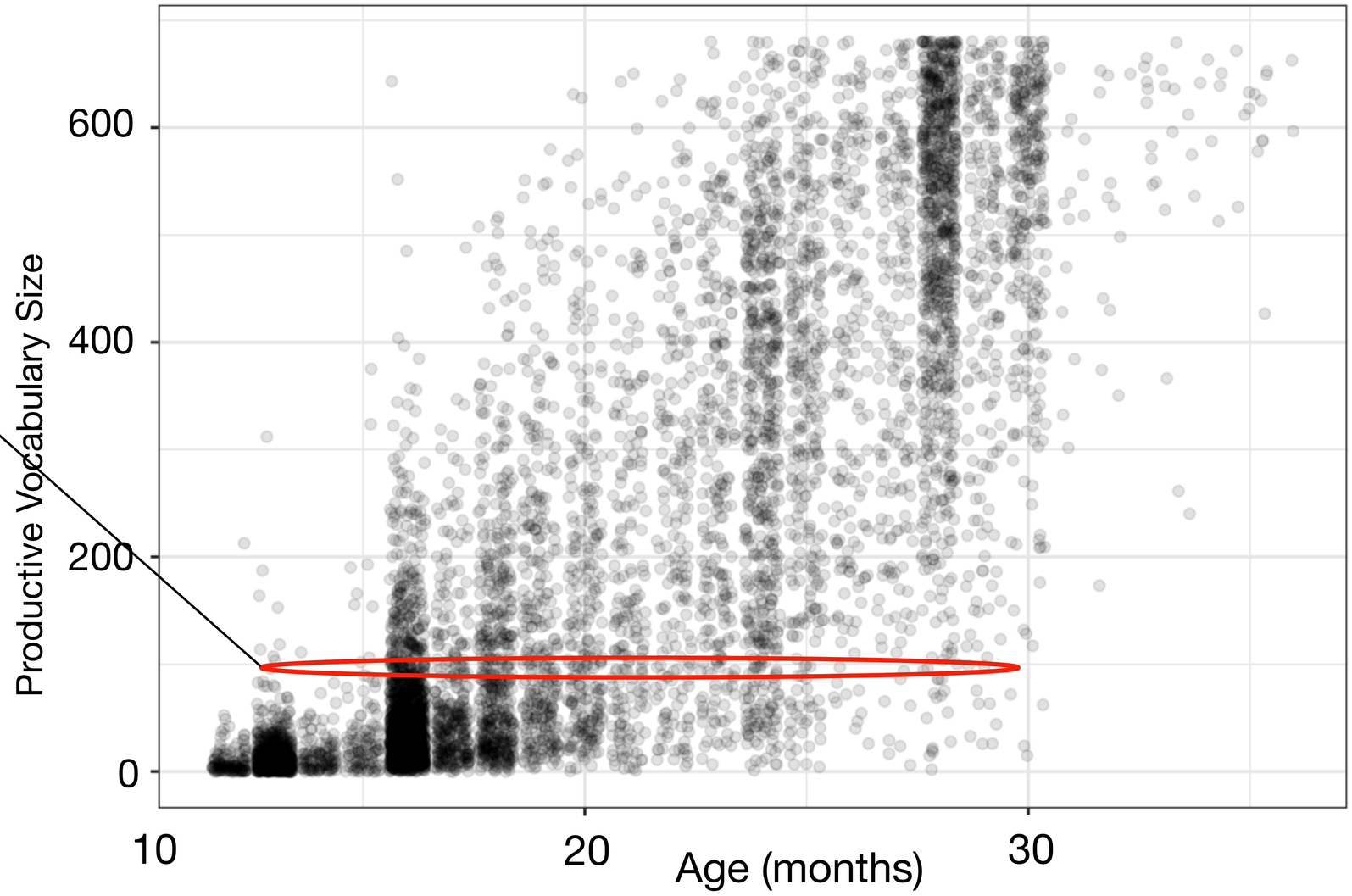
2024!

Frank, Braginsky, Yurovsky, & Marchman (2016)

Fenson et al. (1994, 2007)

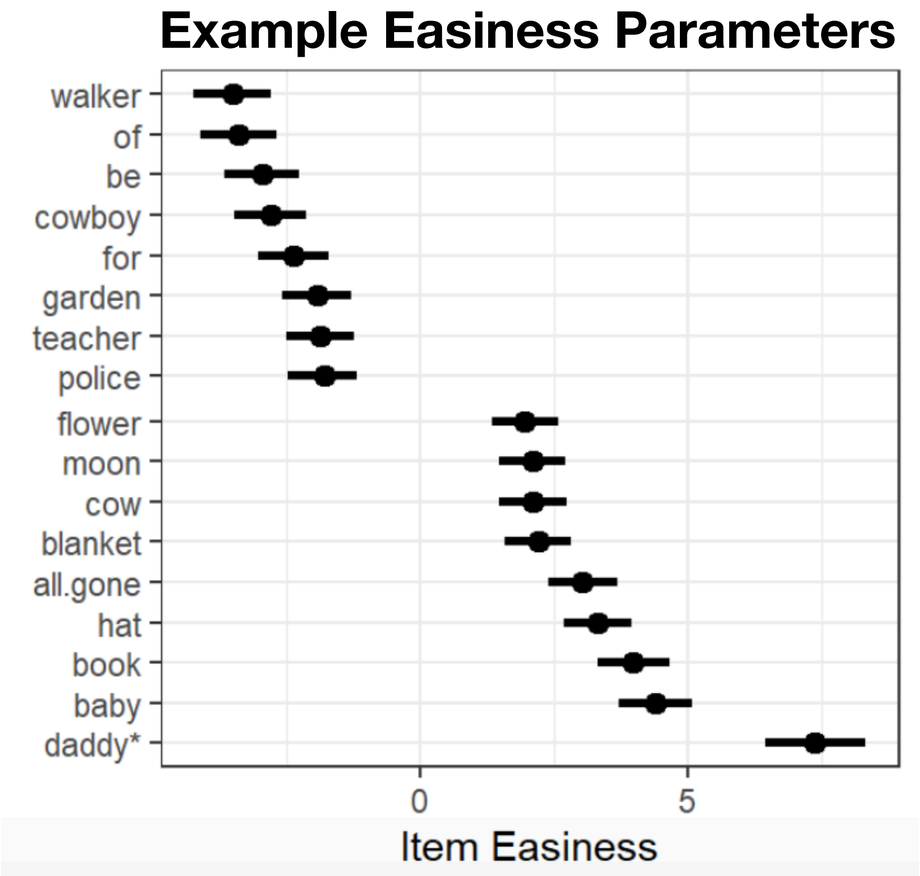
Beyond Vocabulary Size

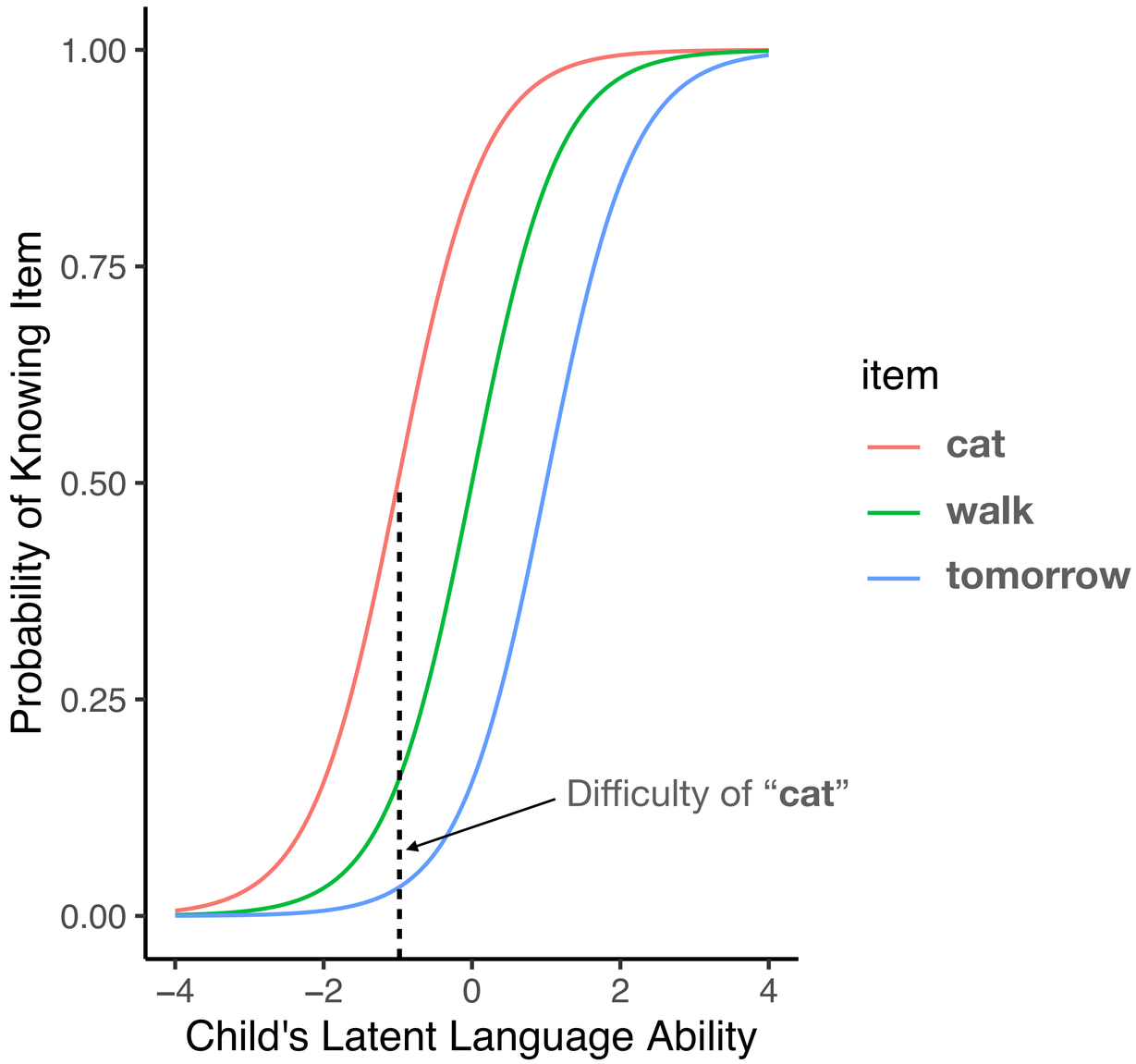
- Vocabulary size of 7,000 English-speaking children
- Do children with equal vocab size (e.g. 100) all have the same language ability?
- Words have different underlying difficulties (e.g., conceptual, phonological)
- Go beyond classical test theory ('sumscore') with psychometric models



Item-Response Theory (IRT)

- Jointly estimate 1) a latent ability θ, for each child *j*, and 2) a difficulty b_i for each item (word) *i*





Computerized Adaptive Test: CDI-CAT

()

Traditional CDI

680 items
>20 minutes

Q1 Q2 Q3 Q4 Q5

Q6 Q7 Q8 Q9 Q10

… Q680

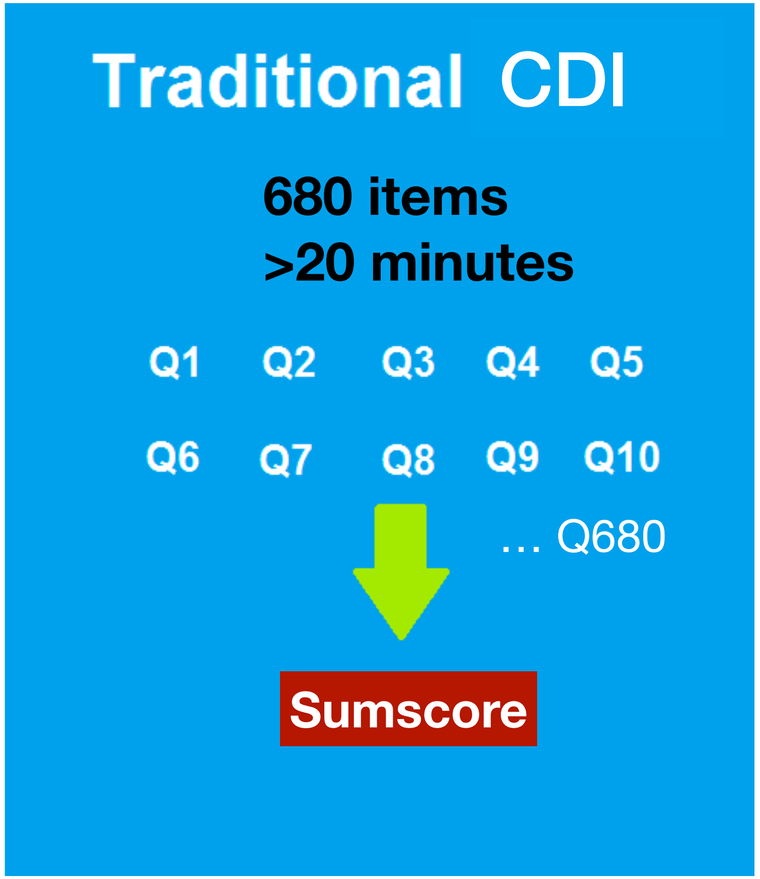
Sumscore

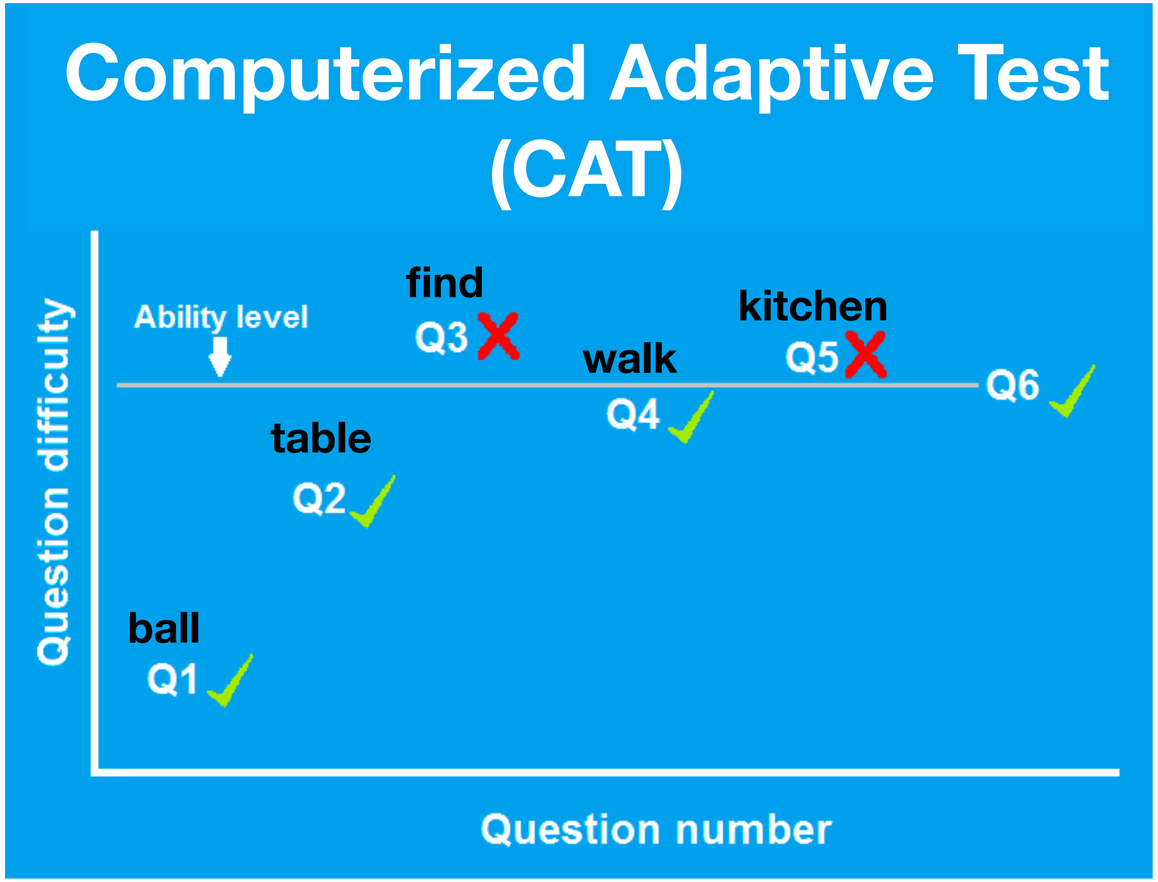
Can we create short, valid tests of early language?

Kachergis et al., 2022 *J. of Speech, Lang, & Hearing Research*

Computerized Adaptive Test: CDI-CAT

()

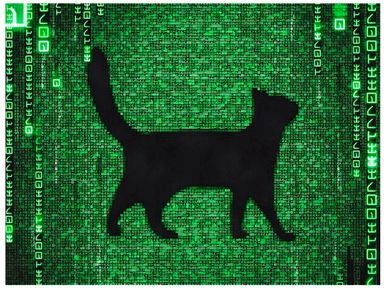




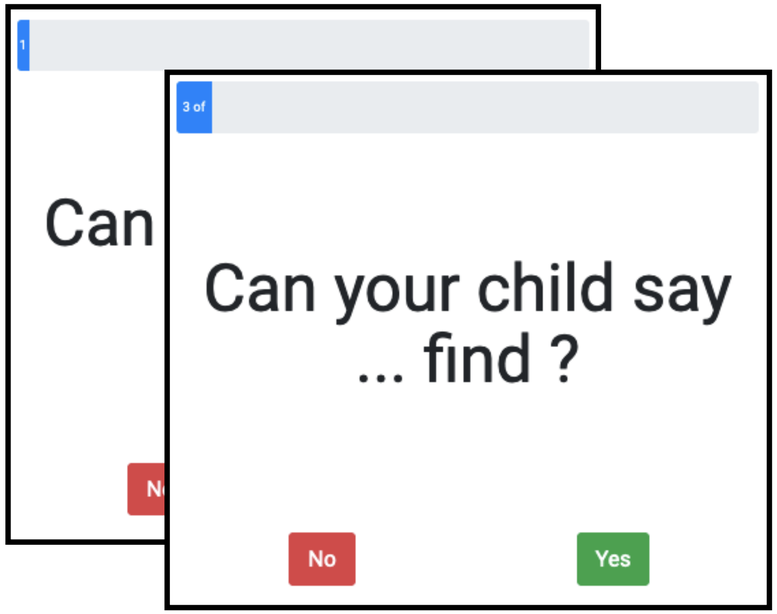
Kachergis et al., 2022 *J. of Speech, Lang, & Hearing Research*

Computerized Adaptive Test: CDI-CAT

()

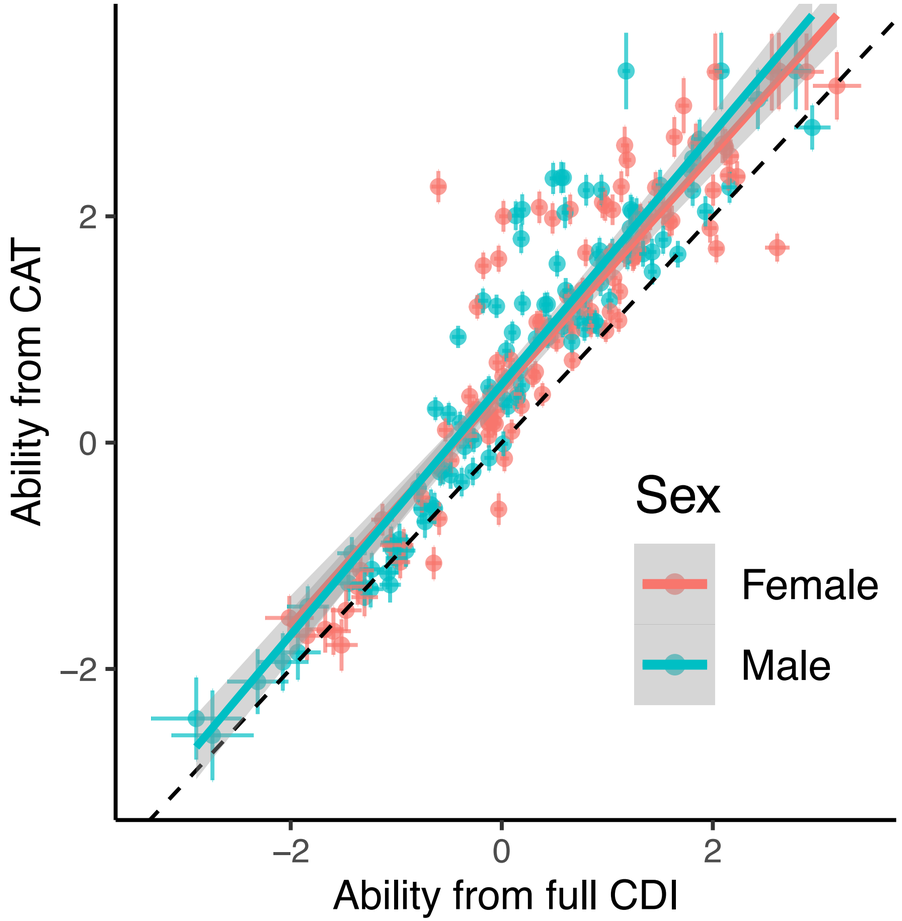


Computerized Adaptive Test (CAT) with 25-50 questions (vs. >650) reliable across a wide age range (12-36 mos)



https://webcdi.stanford.edu/

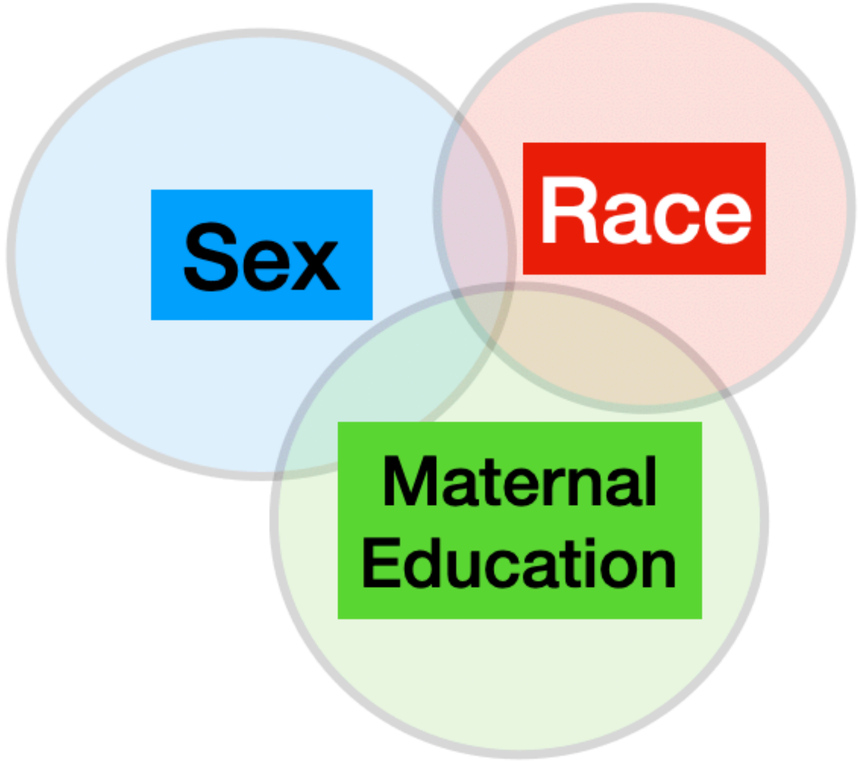
- Validation study (N=200) showed a strong association between full CDI & CAT ability ($r = 0.92$)
- Comprehension *and* production
- English & Spanish in NIH Baby Toolbox (& WebCDI)
- Added French (2023) & Japanese (2024)



Kachergis et al., 2022 *J. of Speech, Lang, & Hearing Research*

Identifying Measurement Bias

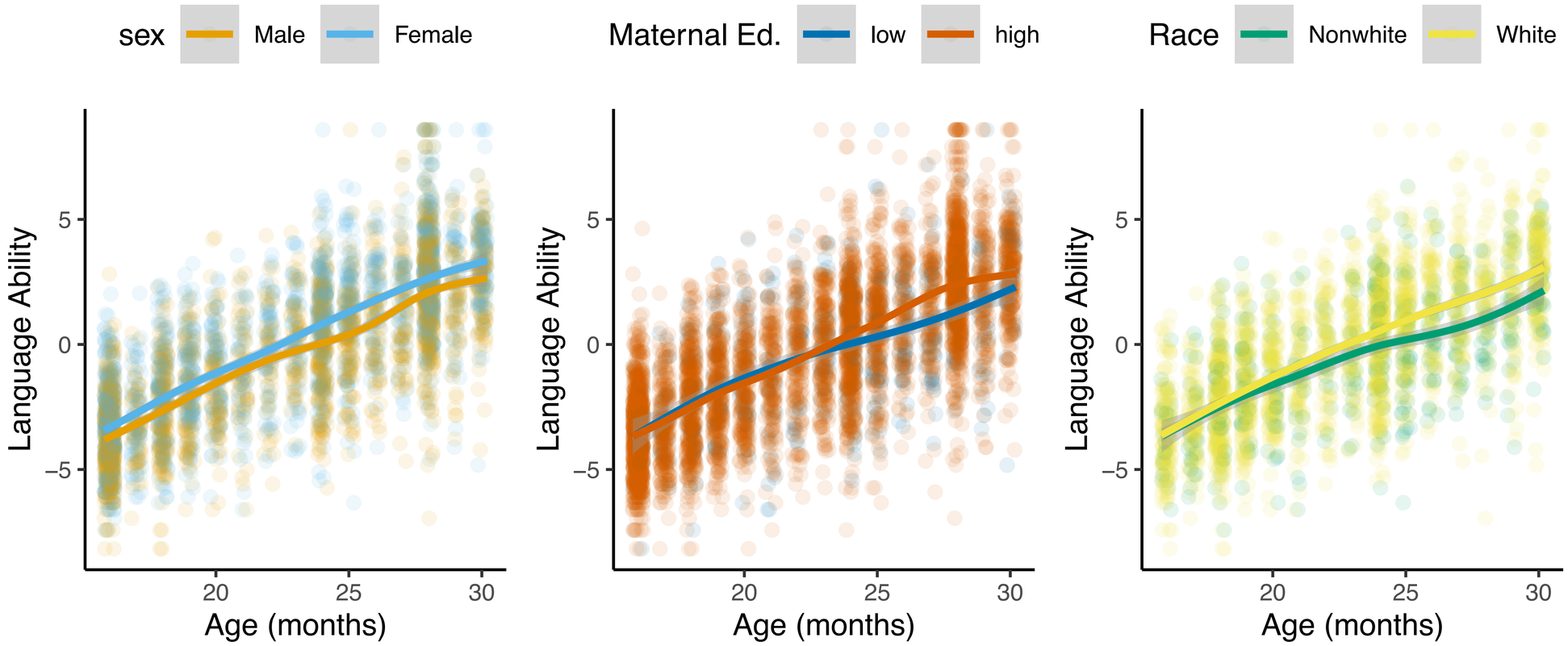
- Wordbank CDI data show demographic differences in vocabulary size advantaging 1) females, 2) white children, and 3) children of highly-educated mothers (a proxy for SES) (Eriksson et al., 2012 Frank et al., 2021)
- Sex-related differences in language skill that persist until high school (Peterson, 2018)
- Could the set of CDI items be biased?
- How to evaluate CDI items for potential bias?



Kachergis, Francis, & Frank, 2022 *Topics in CogSci*
Winner of Diversity and Social Inequality Award at *CogSci 2022*

Identifying Measurement Bias

- Ability vs. age by demographic group in a baseline Rasch (1PL) model



- Using a multigroup Rasch model, we estimate item difficulties that are allowed to vary by demographic group. DIF exists if the difference between group parameters is non-zero.
- The question becomes: How many items are (significantly) biased in favor of each group? Is the item bank representative of all possible items?

Kachergis, Francis, & Frank, 2022 TiCS, *CogSci*

Differential Item Function

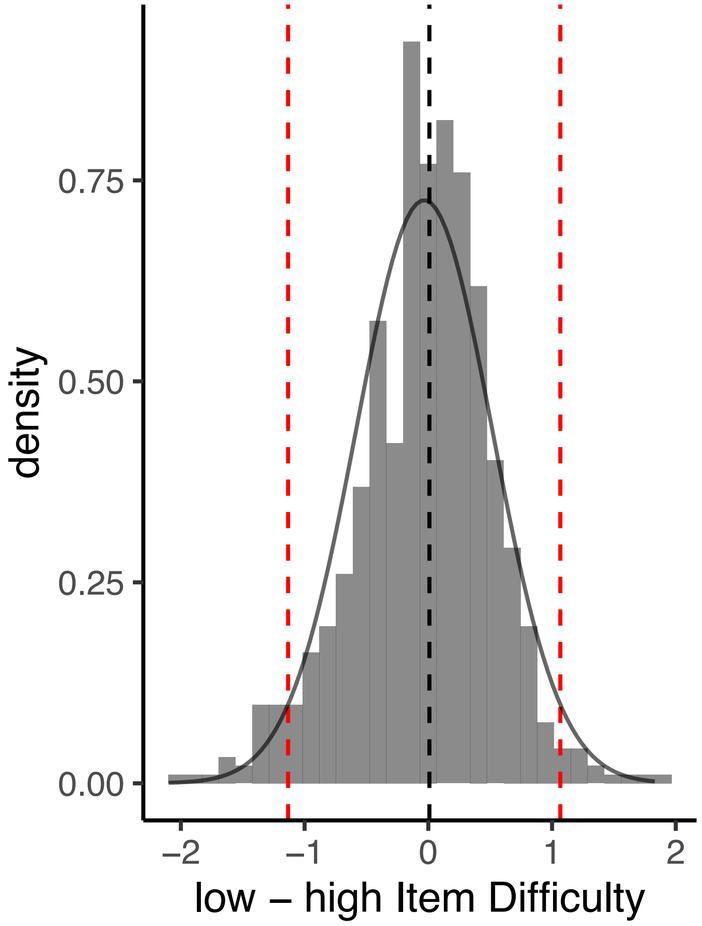
- DIF can decrease the validity of a test: imagine groups A & B have no mean difference in ability, but some items are easier for group A (e.g., farm equipment for rural children).

- If many of these items are selected to be on a test, the test will overestimate the ability of those in group A, and underestimate the ability of those in group B: the test is unfair.

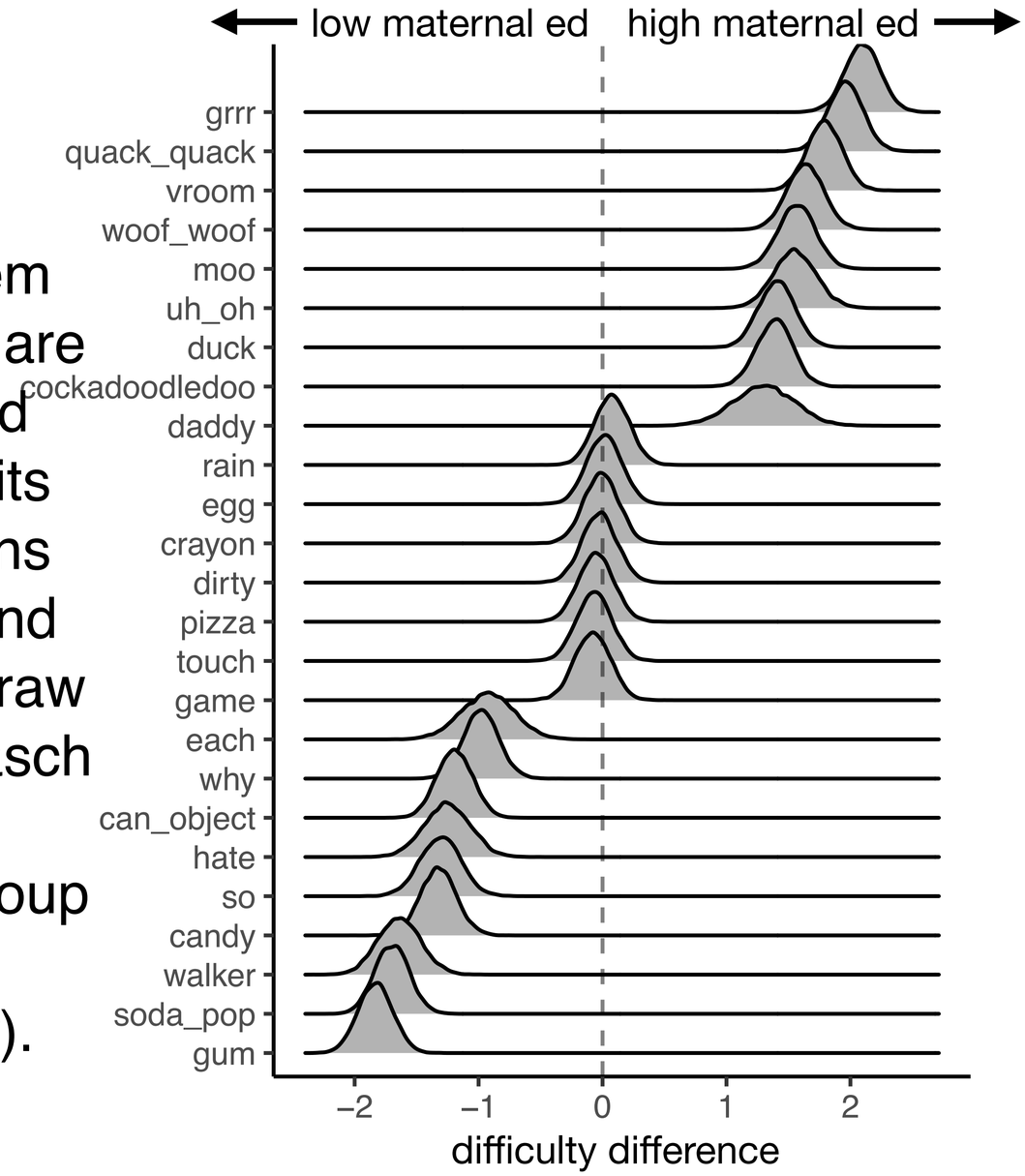
- Of course, a true ability difference may exist between groups — regardless, the selection of items can either inflate or deflate the actual group difference.

Identifying Measurement Bias

A distribution of item difficulty differences for low vs. high maternal education groups:

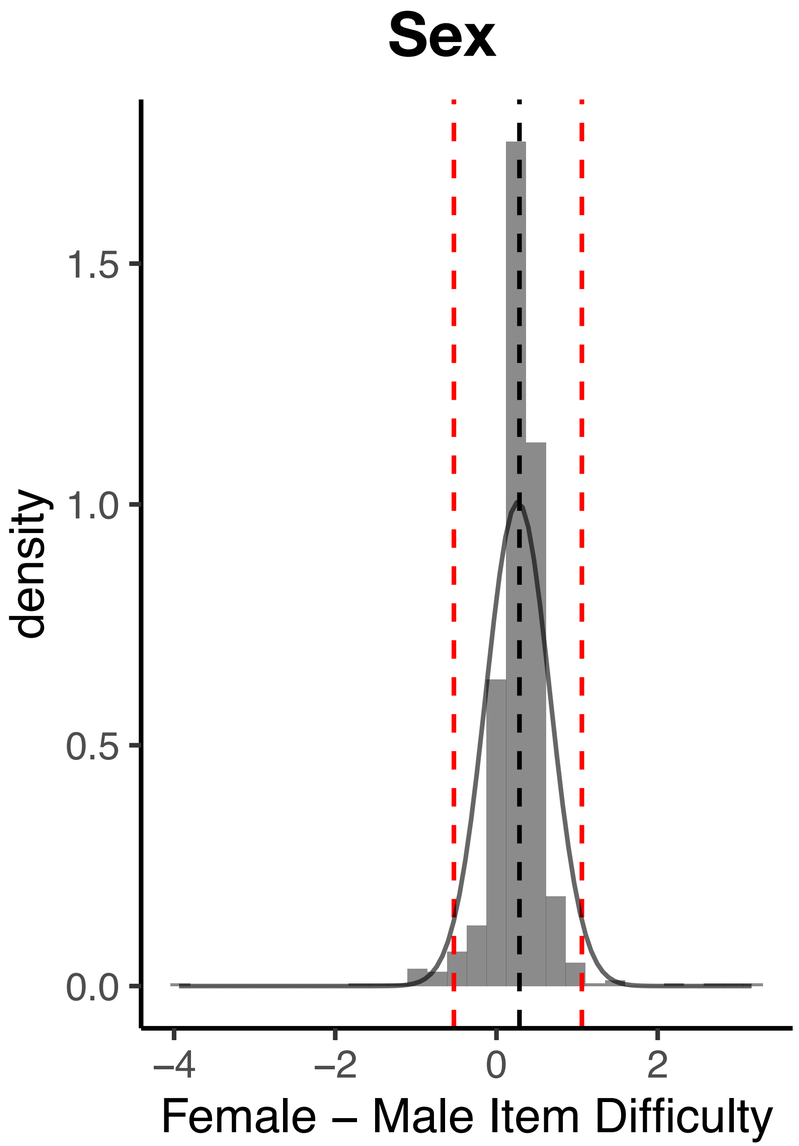


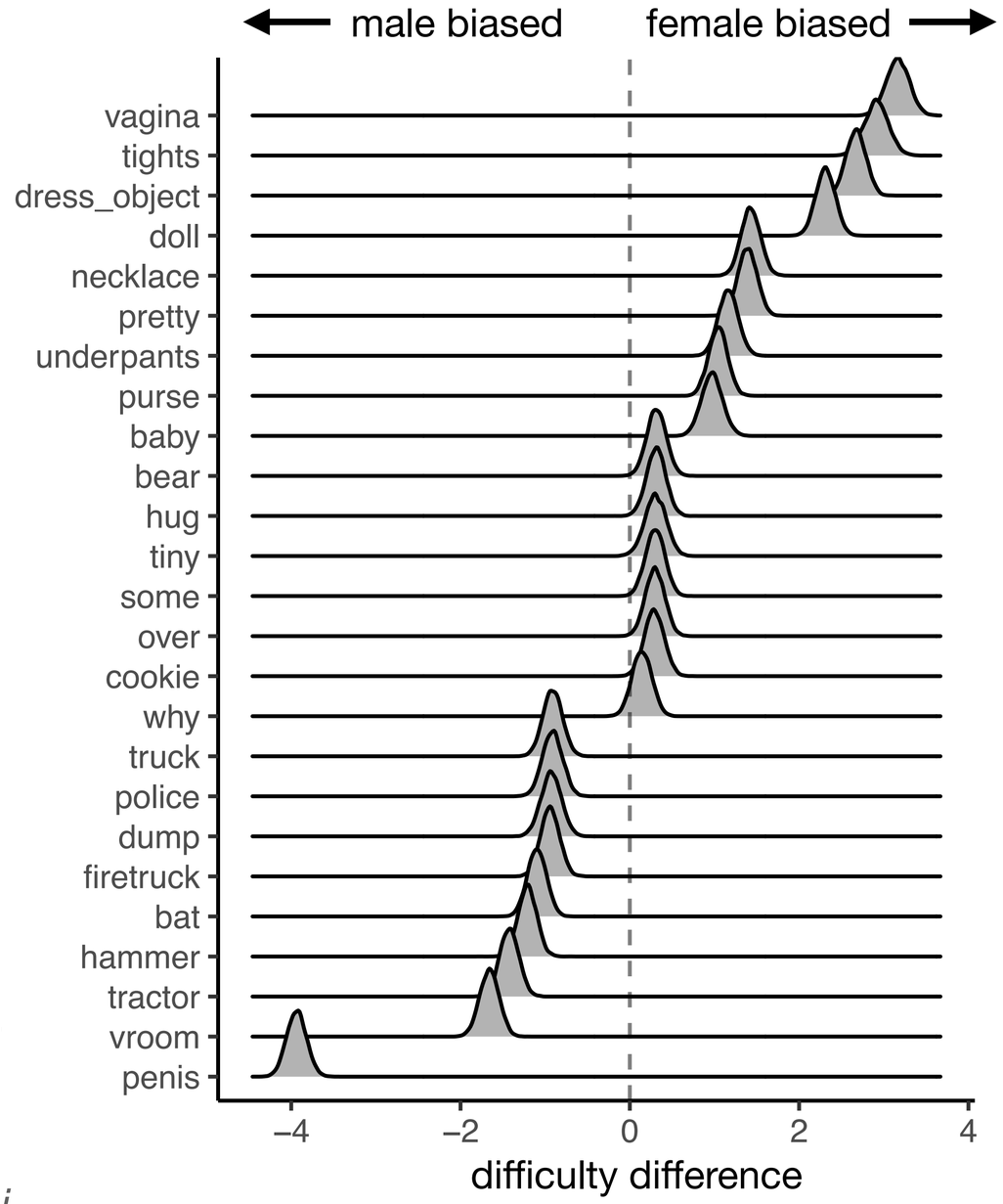
To understand between-item variation (i.e., which items are especially biased), we used GLIMMER (Graphs of Logits Imputed Multiply with Means Equal; Stenhaug, Frank, and Domingue, 2021), which draw from a fitted multigroup Rasch model that assumes mean language ability in each group is the same (pushing all variance into item difficulty).



Kachergis, Francis, & Frank, 2022 *Topics in CogSci*, *CogSci*

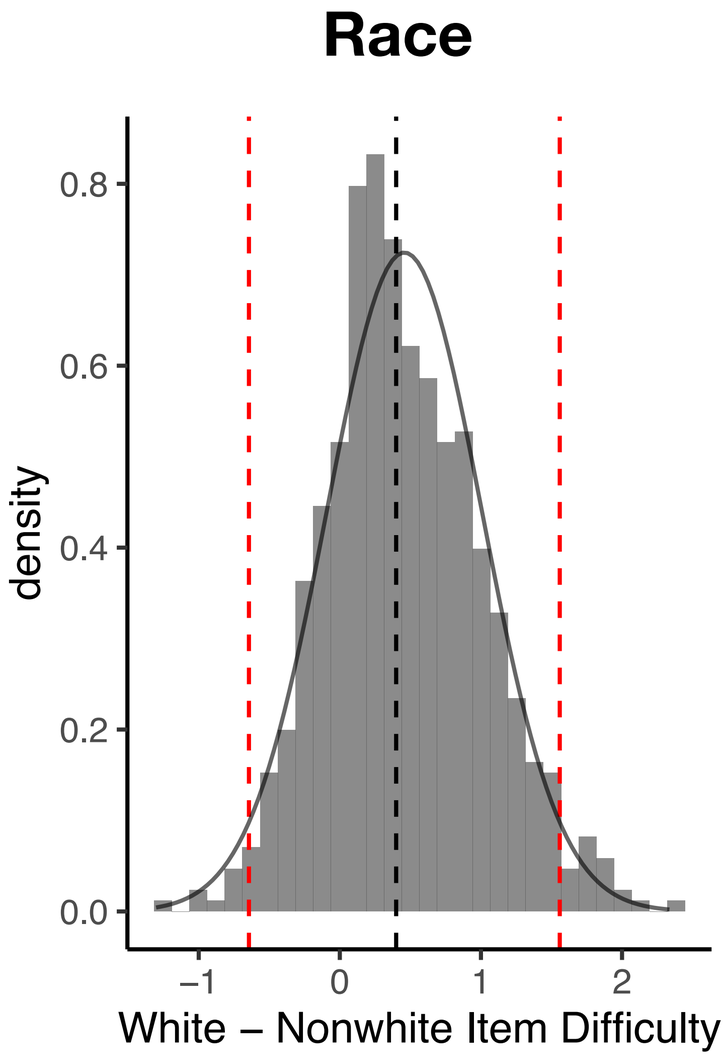
Identifying Measurement Bias

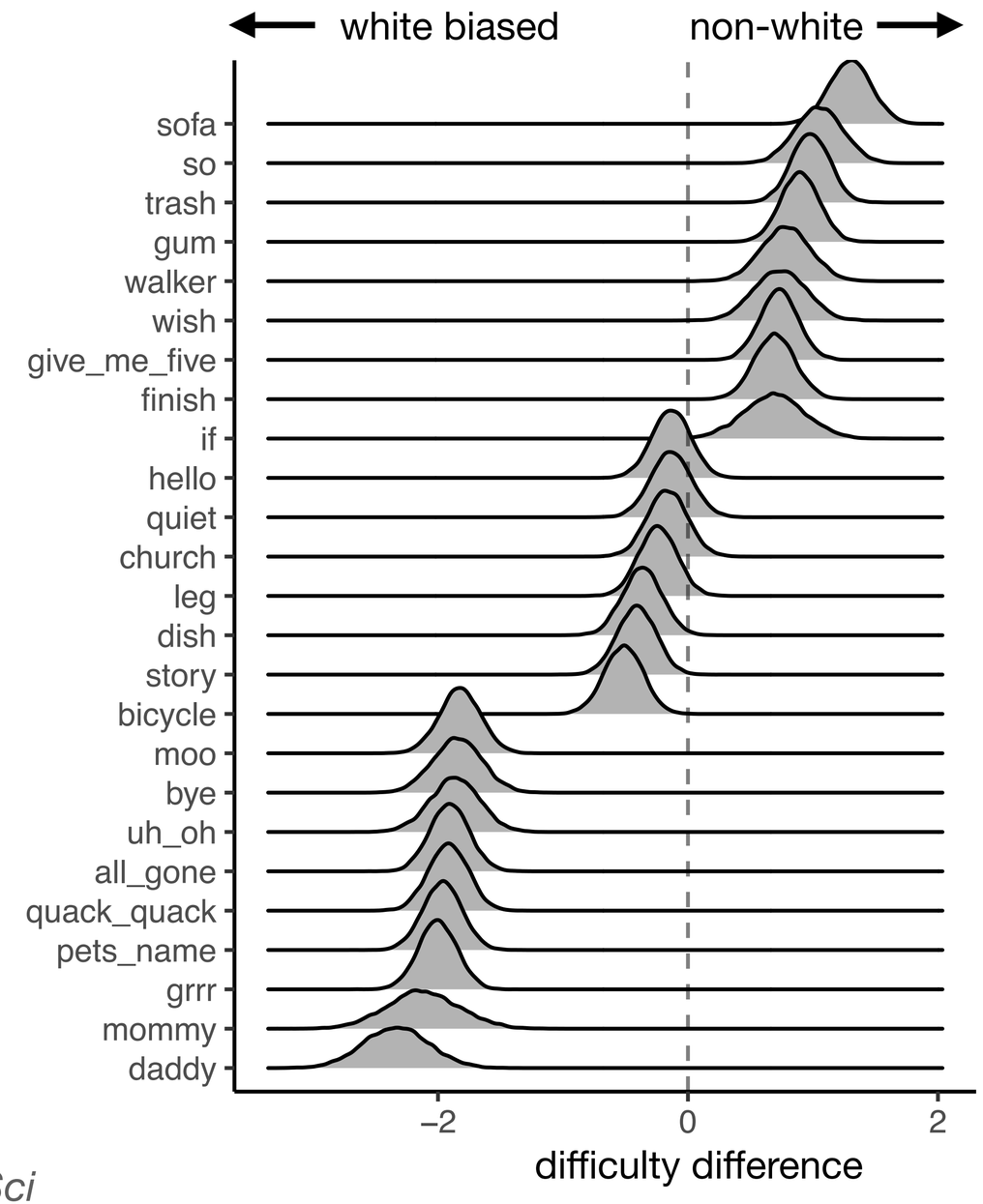




Kachergis, Francis, & Frank, 2022 *Topics in CogSci*, *CogSci*

Identifying Measurement Bias



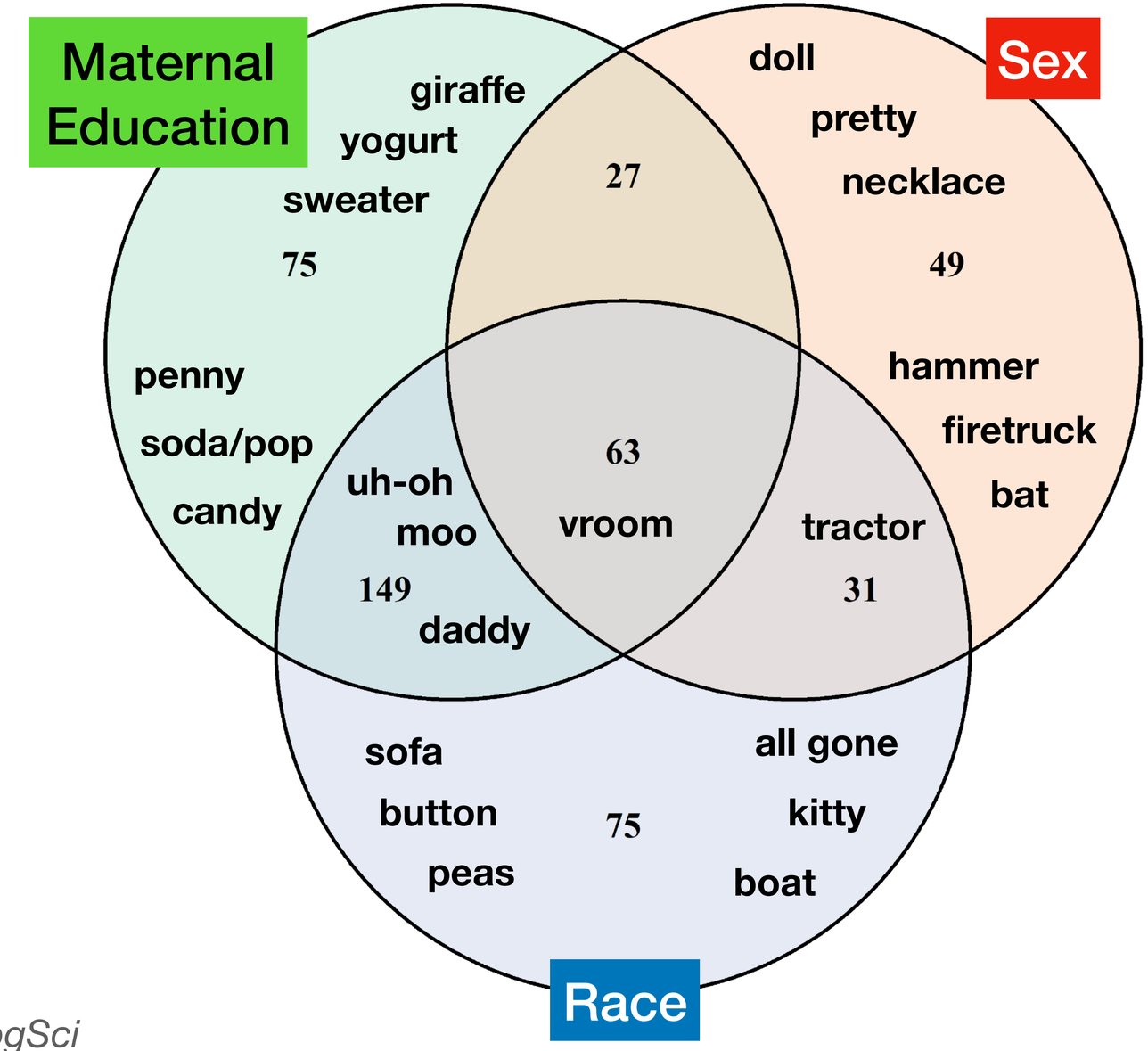


Kachergis, Francis, & Frank, 2022 *Topics in CogSci*, *CogSci*

Reducing Measurement Bias

Many items showed significant bias favoring one or more demographic groups.

But we don't want to get rid of all/most of these items: How does eliminating the extrema change demographic-based differences?

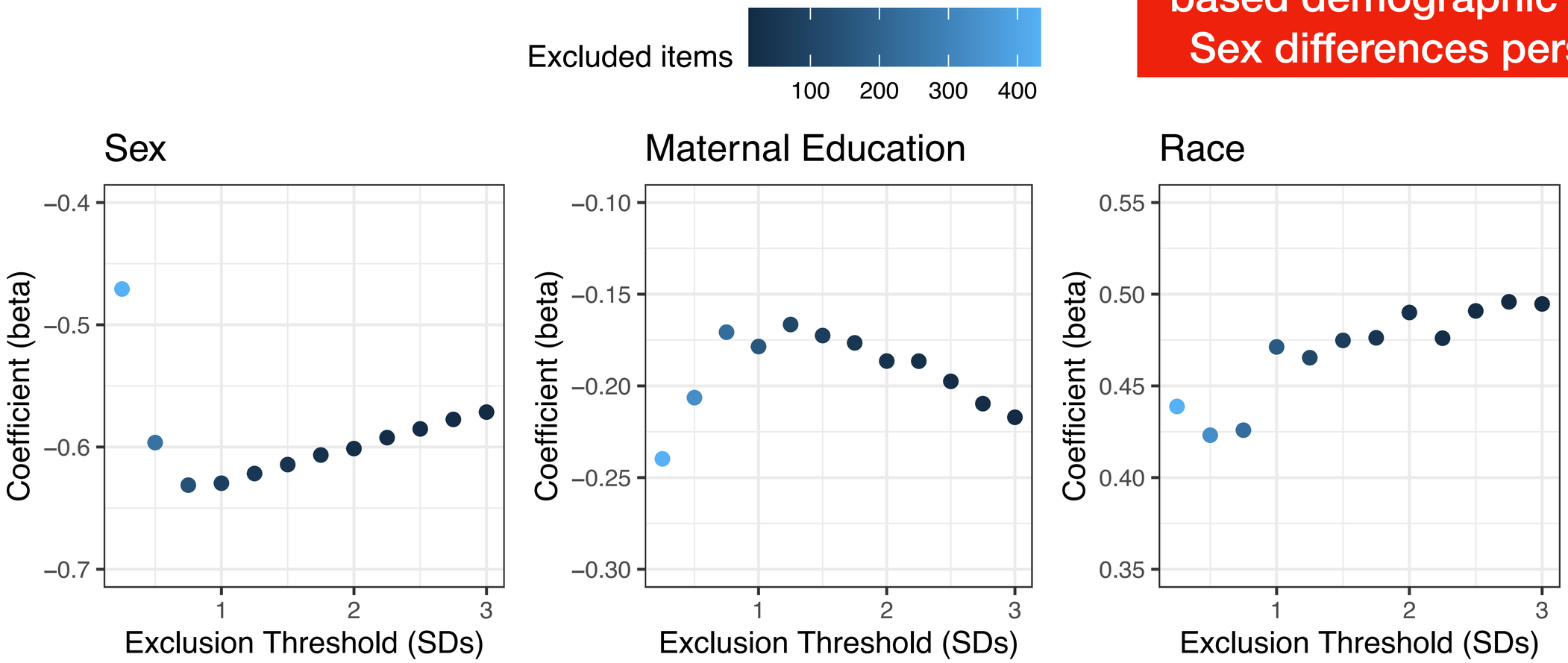


Kachergis, Francis, & Frank, 2022 *Topics in CogSci*, *CogSci*

Reducing Measurement Bias

Suggestion: prune items showing extreme disadvantage for any demographic group

Pruning 59 extrema reduces the size of SES- and race-based demographic effects. Sex differences persisted.



Kachergis, Francis, & Frank, 2022 TiCS, *CogSci*

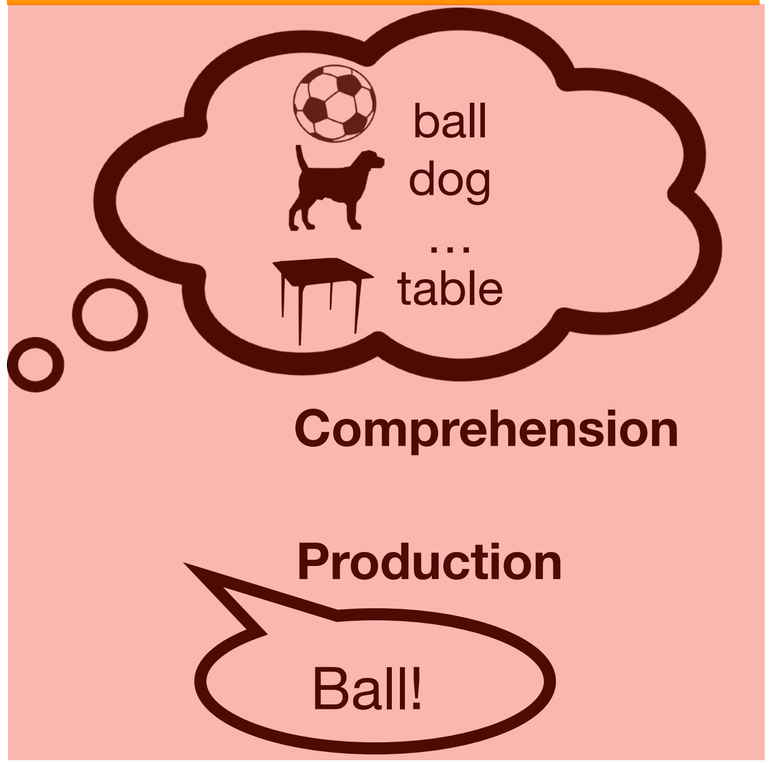
Conclusion

- Computerized Adaptive Tests (CATs) can evaluate children's language ability in a few minutes, with high validity
  - How short can tests of early language be? *~25 (and up to 50) words*
- Identified demographically-biased items on the English CDI, suggested removing a small number of extrema to mitigate SES and racial bias
  - How can we assess & improve test fairness? *Evaluate Differential Item Functioning, eliminate outliers, and (ToDo) consider replacements*

"The history of science is the history of measurement."
—James M. Cattell (1893), founder of *Psychological Review*

Uptake:
learning outcomes

Create short, valid, and fair tests of early language



Thank you! — Questions?

kachergis@stanford.edu

https://kachergis.com

Thanks to my collaborators on these projects:

Michael Frank

Virginia Marchman

Bria Long

Alvin Wei Ming Tan

Nathan Francis

…and thanks to the **Language & Cognition** lab.