Measuring children's early vocabulary in low-resource languages using a Swadesh-style word list

Abstract

Early language skill is predictive of later life outcomes, and is thus of great interest to developmental psychologists and clinicians. The Communicative Development Inventories (CDIs), parent-reported inventories of early-learned vocabulary items, have proven to be valid and reliable instruments for measuring children's early language skill. CDIs have been painstakingly adapted to dozens of languages, and cross-linguistic comparisons thus far show both consistency and variability in language acquisition trajectories. However, thousands of languages do not yet have CDIs, posing a significant barrier to increasing the diversity of languages that are studied. Here, we propose a method for selecting candidate words to include on new CDIs, leveraging analysis of psychometric properties of translation-equivalent concepts that are frequently included on existing CDIs. Leveraging 32 datasets from existing CDIs, we propose a list of 100 concepts that have low variability in their cross-linguistic learning difficulty. This pool of common concepts—analogous to the "Swadesh" lists used in glottochronology—can be used as a starting point for future CDI adaptations. We test how well the proposed list generalizes to data from 10 additional languages.

Introduction

Tools that enable valid assessments of children's early language abilities are invaluable for researchers, clinicians and parents, as early language skill is predictive of educational outcomes years later (e.g., Bleses et al. 2016). The MacArthur-Bates Communicative Development Inventories (CDIs, Fenson et al. 2007; Marchman, Dale, and Fenson 2023) are parent report assessments that provide reliable and valid estimates of children's early vocabulary size and other aspects of early communicative development, such as use of gestures and of word combinations. Parent report is a relatively quick and low-cost method to assess early language skills as it takes advantage of the fact that parents are "natural observers" of their child's skills and does not depend on a child engaging with an unfamiliar experimenter.

Over the years, the CDIs have been adapted to dozens of languages, with forms now available in English, Spanish, French, Hebrew, and Mandarin, to name just a few. Recently, data from 97925 CDIs in 42 languages have been archived in a central repository (Wordbank, Frank et al. 2017). These data have revealed both cross-linguistic consistency and variability in early language skills, with insights from these patterns informing theories of early language learning (Frank et al. 2021). For example, cross-linguistic analyses indicate that measures of vocabulary size are tightly correlated with other aspects of early language skill, like gesture and grammatical competence. Thus, over development, the language system is "tightly woven" (Bates et al. 1994; Frank et al. 2021) and early vocabulary size serves as a good proxy measure of children's overall language skill.

On the CDIs, vocabulary size is assessed via a checklist format, which enables caregivers to scan and recognize words their child produces or understands, rather than relying on recall alone. For example, the American English CDI Words & Sentences (CDI:WS) form, targeting children 16–30 months of age, is comprised of 680 words from 22 semantic categories, including nouns (e.g., Body Parts, Toys, and Clothing), action words, descriptive words, and closed-class words such as pronouns. Items on this original CDI:WS were chosen to reflect a range of difficulty levels (i.e., easy, moderate, and more difficult), as well as capture the linguistic and societal contexts of (most) children living in the US. The CDI Words & Gestures form (CDI:WG) targets children 8–18 months of age, typically consisting of a subset of approximately 500 of the easier items from the CDI:WS of the same language, and which asks caregivers to report children's comprehension as well as

production of each word. Short versions of the CDI:WS forms are also available (e.g., Fenson et al. 2000), each consisting of a set of around 100 items, often selected to generate scores that strongly correlate with scores on the full forms, while retaining representation across a broad set of semantic categories.

Creating a new CDI requires a lot of effort and resources, presenting a daunting barrier to increasing the diversity of languages studied. Following the guidelines[1] from the MacArthur-Bates CDI Advisory Board, the process of adapting a CDI for a language other than American English goes well beyond simply translating items on these forms to that new language. While the process can begin with identifying translation equivalents (i.e., items that capture the same general concept in both languages, e.g., "dog" in English, and "perro" in Spanish), the final item set must then be filtered so that all items appropriately reflect the linguistic and sociocultural context of the children learning that language. This process usually requires considerable time and effort by researchers who are both native speakers of the language and who have experience with children, to first select and identify translation equivalents and to then iteratively add, refine, and pilot the new CDI in the target language (see (Jarůšková et al. 2023)). Because the goal is to obtain the set of items that best capture general trends and individual differences in that language, the items across CDIs in different languages do not necessarily overlap to a great extent. For example, the American English CDI:WS and Mexican Spanish CDI:WS forms each have 680 words, but only have 463 overlapping concepts (68%).

It is well-established that, all over the world, early-learned words reflect the people and things that children are likely to experience, that is, words for family members, animals, and common household objects (Tardif et al. 2008; Frank et al. 2021). Given this finding, it is reasonable to ask: Is there a set of translation equivalents that would meet the criteria for inclusion on CDIs from multiple languages? Identifying a small set of translation-equivalent items that function well for assessing early language development could lead not only to shorter assessments in languages that already have CDIs, but also to lowering the burden of creating CDIs in new languages. Our goal is to leverage the roughly 100,000 CDI administrations from 32 languages on Wordbank to choose and systematically evaluate sublists that meet researchers' criteria for creating a new CDI.

To facilitate this effort, it is useful to leverage Item-Response Theory (IRT, Embretson and Reise 2013) models. IRT models infer both the abilities of test takers and the difficulty of individual test items (i.e., words), along standardized dimensions. Recent work using IRT models has facilitated our understanding of the psychometric properties of specific CDI instruments. As such, they offer the potential to not only yield more accurate measures of children's language ability, but also to enable the construction of language-specific Computerized Adaptive Tests (CATs), which choose the next test item based on the responses to the previous items, and thus quickly hone in on the test taker's language ability. CAT-based CDIs presenting 50 or fewer items have been found to strongly correlate with scores on the full CDI:WS (Chai, Lo, and Mayor 2020; Mayor and Mani 2019; Makransky et al. 2016). A general method for creating CDI CATs that work well across a broader age range (12–36 months) has been proposed, and tested for American English and Mexican Spanish (Kachergis et al. 2022). However, the IRT model driving each CAT needs to be trained on a large and normative dataset (that is, a sample that is representative of a target population), which may not be available in a given language. To date, the IRT models are fitted separately for each language, and the fitted parameters (e.g., word difficulty) are likely to vary across languages. Importantly, this work has also revealed that scores on random subsets of items from a CDI form are highly-correlated with scores on the full CDI [e.g., for English, full CDI vs. 100 random items $r = 0.989$; Kachergis et al. (2022)]. However, random items from a single form are not guaranteed to be 1) relevant in other languages, 2) of similar difficulty in other languages, or 3) representative of the overall proportion of semantic categories present on the full CDI (for a given language, let alone all languages).

The goal of the current study is to use IRT modeling in conjunction with data from Wordbank to examine whether there might be a core set of concepts that are frequently included on CDIs, and—importantly—whether enough of them are of roughly equal difficulty across many languages to allow them to be used as candidate items in new languages. This work takes its inspiration from the fields of lexicostatistics and glottochronology, where researchers (notably, Swadesh 1971) have proposed lists of common concepts that exist in all catalogued languages, in order to quantify the genealogical relatedness and dates of divergence

[1] https://mb-cdi.stanford.edu/adaptations.htm

of languages. For example, the original Swadesh list contains 100 words, comprised of categories including common pronouns (*I*, *you*, *we*), animals (*man*, *fish*, *bird*, *dog*), objects (*tree*, *leaf*, *sun*, *mountain*), and verbs (*die*, *see*, *sleep*, *kill*). Extending this work to the development of a universal CDI, or "Swadesh CDI," would include many of the concepts that researchers have chosen to include on several CDI:WS adaptations, and which have relatively similar difficulty across many languages. If such a list were generalizable to other languages, it could serve as a helpful starting point for the development of new CDI adaptations, since the constituent words would already have good cross-linguistic difficulty estimates. It would also provide a method of approximating children's language abilities even in the absence of a large normative study.

In particular, our contributions are 1) to revise and extend a set of translation-equivalent concepts in Wordbank, 2) to fit IRT models to CDI:WG and CDI:WS data from 32 languages, 3) to evaluate candidate lists of Swadesh CDI items from a cross-linguistic comparison of concept difficulty and inclusion, 4) to identify and characterize the most consistent 100 items to compose a Swadesh CDI list, and 5) to evaluate its generalization to a set of 10 additional low-data languages. We then make a concrete proposal for how this Swadesh CDI list could be used to create future CDI adaptations to greatly expand the diversity of languages studied. We end by discussing the strengths and weaknesses of our approach. Our full analysis, the Swadesh CDI list, and other information valuable for developing a new CDI are openly available on OSF.

Methods

Item Response Theory

A variety of IRT models targeting different types of testing scenarios have been proposed (see Baker 2001 for an overview), but for the dichotomous responses that parents make for each item (word) regarding whether their child can produce a given word, we used the popular 2-parameter logistic (2PL) model that is best justified for CDI data out of four standard models (see Kachergis et al. 2022).

The 2PL model jointly estimates for each child j a latent ability θ_j (here, language skill), and for each item i two parameters: the item's difficulty b_i and discrimination a_i, described below. In the 2PL model, the probability of child j producing a given item i is

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

where D is a scaling parameter ($D = 1.702$) which makes the logistic more closely match the ogive function used in a standard factor analysis (Chalmers 2012; Reckase 2009). 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. The discrimination (a_i) adjusts the slope of the logistic (in the classic 1-parameter logistic "1PL" model, the slope is always 1). Items with higher discrimination (i.e., slopes) better distinguish children above vs. below that item's difficulty level, and hence are generally more useful. While other standard IRT models exist (e.g., the 3PL model adds a "guessing" parameter for each test item), a recent study found the 2PL model most appropriate for multiple Wordbank datasets (Kachergis et al. 2022).

Datasets

language	participants	items
English (British)	22751	538
English (American)	15267	682
Norwegian	11860	745
Danish	5506	727
Portuguese (European)	3804	658
Turkish	3208	746
Spanish (Mexican)	2703	742
Mandarin (Taiwanese)	2392	696
French (Quebecois)	2327	700
Korean	2179	703
Mandarin (Beijing)	1906	1131
Dutch	1806	1037
Russian	1720	761
French (French)	1700	698
Slovak	1616	687
English (Australian)	1483	558
Catalan	1302	722
Cantonese	1292	804
Swedish	1292	743
Italian	1267	715
Estonian	1235	631
German	1178	588
Japanese	1003	711
Spanish (European)	872	648
Arabic (Saudi)	865	925
Spanish (Argentinian)	781	699
Latvian	646	796
Hebrew	577	683
Croatian	576	768
Finnish	534	591
Czech	493	553
Hungarian	363	802
Kigiriama	216	747
American Sign Language	198	701
Greek (Cypriot)	176	815
Spanish (Peruvian)	174	615
British Sign Language	151	532
Persian	148	741
Kiswahili	109	729
English (Irish)	99	660
Irish	99	691
Spanish (Chilean)	51	455

Table 1: Total CDI items and participants per dataset (language). The final 10 datasets were used for a generalization test.

We pulled data for 42 languages from Wordbank (Frank et al. 2017). For each language, we extracted production data for all forms present on Wordbank (including WG and WS, as well as other language-specific forms). We then stitched the data across all forms within a language by matching items by their item definition (i.e., what was actually presented on the form for that item) and category; we used fuzzy matching with manual correction to allow for instances where essentially one item had slight variation in item definitions across forms (e.g., spelling differences, different ordering for multiple options). We used data from 32 languages with more than 300 participants as our training languages from which we constructed our candidate word list, as it was possible to fit reliable IRT models from these data. The remaining 10 languages had too few participants to be analyzed with IRT, and we used them as our generalization languages to test how well the word list could generalize to novel languages.

Unilemmas

Comparison across languages requires a method to map between words that correspond to broadly similar concepts across languages. As such, each item on the CDI:WS for each language was mapped onto a set of "universal lemmas" or "unilemmas", which are approximate cross-linguistic conceptual mappings of words. For example, "chat" (French) and "gato" (Spanish) both correspond to the same unilemma, *cat*. These mappings were recently updated to improve their quality and systematicity, and to increase coverage across items and languages. This new set of unilemmas was constructed based on glosses provided by the original contributors of the Wordbank datasets, which were then verified by native or advanced proficient speakers of the language, and cleaned to increase their consistency across languages. All unilemmas are accessible from Wordbank; details about the recent update can be found at https://github.com/langcog/update_unilemmas.

Participants

The Wordbank[2] datasets consisted of CDI:WG and CDI:WS production data for 97925 children aged 8–30 months 29874 items across 42 languages.[3] Note that the distributions of demographic variables (age, sex, etc.) of these datasets are not matched, so comparing overall language ability estimates across languages would be ill-advised. (See Frank et al. (2021) for a discussion of effects of demographic variables on vocabulary development.) Thus, we focused only on the estimated item parameters, and in particular the variability of item difficulty (b_i).

Instruments

When a CDI:WG form was administered, caregivers were asked to indicate for each vocabulary item whether their child 1) understands that word ("comprehends") or 2) both understands and says ("produces") that word. Leaving the item blank indicates that the child neither comprehends nor produces that word. When a CDI:WS forms was administered, caregivers were asked to indicate for each vocabulary item on the instrument whether or not their child can recognizably produce (say) the given word in an appropriate context.

"Produces" responses were coded as 1 and all other responses were coded as 0. Our datasets consisted of a dichotomous-valued response matrix for each language, of size N subjects $\times$ W words. All models, data, and code for reproducing this paper are available on OSF[4].

Results

Across the 42 IRT models for different languages' CDI forms, difficulty and discrimination parameters for a total of 28541 items were fitted. Of those items, 95.5% had unilemmas defined, with a median of 677 per language (range: 440 in Chilean Spanish to 1061 in Mandarin Chinese). A total of 2084 unique unilemmas were defined across the forms, but 478 of these were singletons, appearing on only one of the forms.[5] There was a significant relation between how often a unilemma appears and its difficulty: the more often a unilemma appears, the *easier* it tended to be ($r = 0.35$, $t(2082) = 16.83$, $p < .001$). Moreover, there was a weak but significant relation between the number of forms a unilemma appears on and its cross-linguistic variability ($r = 0.2$, $t(1604) = 8.16$, $p < .001$). It is perhaps intuitive that lower-variability items tend to be earlier-learned, and are thus often selected to be on CDI forms, echoing prior work characterizing the consistency of children's first words across several languages (Tardif et al. 2008). However, these modest but significant correlations were also important to keep in mind as we chose our Swadesh CDI candidates, as selecting too many easy items could result in older children being at ceiling.

[2]http://wordbank.stanford.edu/contributors

[3]OSF repository: https://osf.io/8swhb/.

[4]OSF repository: https://osf.io/8swhb/.

[5]These singletons were significantly more difficult than the 1606 unilemmas appearing more than once ($M_1 = 1.88$; $M_{>1} = 1.3$; $t(664) = -6.79$, $p < .001$).

Identifying Swadesh CDI Candidates

There were two key desiderata for a Swadesh CDI list:

1. Generalizability to new languages, and
2. Comprehensive measurement of a child's vocabulary.

To satisfy the generalizability criterion, we aimed to choose unilemmas with low variability in their cross-linguistic difficulty—that is, unilemmas that are similarly difficult to learn across languages, operationalized as having the lowest standard deviation in item difficulty. To satisfy the comprehensiveness criterion, we adopted three specific criteria that were most often used in previous short form constructions: diversity in item difficulties, semantic categories, or syntactic categories represented. These criteria were operationalized by stratifying all unilemmas by difficulty, semantic category, or syntactic category, and selecting the lowest variability unilemmas within each stratum.

We arbitrarily selected a list size of 100 items[6], noting that the same procedure could be followed to generate lists of larger or smaller sizes. We thus wanted to select the 100 unilemmas with the least variability in item difficulty; this process was either conducted across all unilemmas (i.e., unstratified), or stratified by difficulty, by semantic category, or by syntactic category. For difficulty stratification, we binned the unilemmas into 2–5 quantiles by difficulty, and the least variable items were equally drawn across all strata. For semantic category stratification, we considered the semantic categories that were the most common across all languages, while for syntactic category stratification, we classified all items as nouns, predicates (verbs and adjectives), function words, or other (e.g., onomatopoeia, routinized phrases, closed class adverbs). In each case, we calculated the mean proportion of each semantic or syntactic category across all languages. The mean proportions were then rounded to fit 100 items, resulting in quotas for each category; least variable items were drawn within each category to satisfy the quotas. For example, 9.7% of unilemmas are classified as food/drink, so a quota of 10 food/drink items would be selected for the semantic category stratified Swadesh lists. The 32 semantic categories and 4 syntactic categories used in these analyses are listed in Appendix A.

The method of choosing the lowest variability items tended to prefer items that appear on fewer forms, since it was more likely for two items to coincidentally have very similar difficulties than for 20 items to have similar difficulties, even though this measure of variability was likely to be an underestimate of true cross-linguistic variability in the former case. As such, we also used a threshold k, reflecting the minimum number of current languages which must contain the unilemma (i.e., for $k = 5$, we included only unilemmas that appeared on the forms of at least 5 languages).

In order to choose the optimum value of k, we conducted leave-one-out cross-validation over our training languages. Specifically, we held out one training language and conducted the selection procedure using the data from all other training languages for all values of $k \in [2, 31]$. With these candidate lists, we compared the item difficulties in the held-out language with the mean item difficulties in the remaining training languages, and selected the value of k for which the correlation was the highest for each method (unstratified, category stratified, and difficulty stratified).

Finally, we re-ran the selection procedure using the best k values across all training languages (without holding out any language) to arrive at a final Swadesh CDI list proposed for each method.

Choosing a Random Baseline

To determine whether a candidate Swadesh list functions well, it is necessary to have a baseline list of random items (of similar length) to compare to. However, there are many ways to select a random set of unilemmas to compare to, and each method could be argued to unfairly advantage either the Swadesh list or the random

[6]Though we note that most CDI short-forms are of this length: 100 items yields, on balance, a reliable measure of children's early vocabulary without being too time-consuming for use in a variety of studies of early development.

baseline. For example, if we simply selected 100 unilemmas uniformly at random from the unilemmas for all languages, many of the selected unilemmas would not appear on many other languages' CDI forms, and thus random would perform poorly. At the other extreme, if we sample 100 random unilemmas from each language's full CDI forms, the Swadesh list is unfairly disadvantaged, as some of the 100 Swadesh items may be missing from any given language's forms. To somewhat level the playing field, we selected the random comparison items uniformly at random from the list of unilemmas that appear in at least k languages – the same constraint we placed on choosing Swadesh candidates. This ensured that, in expectation, the same number of items per language would appear on the random baseline lists and the Swadesh candidate lists.

However, note that selecting N items uniformly at random from a full CDI list sets quite a high bar: if your other method of selecting items introduces any bias (e.g., selecting easier or more difficult items, on average), then the randomly-selected baseline will have the stronger correlation with the full set.[7] Thus, the goal for the Swadesh list is to generate scores as strongly correlated with the full CDI scores as the random baseline's correlation – while also selecting items that better generalize to out-of-distribution languages by having less variation in cross-linguistic difficulty.

Cross-validation Results

Table X shows the optimal values of k selected for each sublist selection method, along with how well each method performs with the chosen k. The syntactic category selection method with $k = 27$ had the highest overall correlation ($r = 0.856$), $and result edin an average overlap of 92.2 items on the resulting lists. There was nearly a 3-way tie for the next best selection method : 2 difficulty strata (r=0.844\$, $k = 27$), unstratified ($r = 0.841$, $k = 23$), and semantic category ($r = 0.841$, $k = 25$). Random selection showed the lowest correlation ($r = 0.776$, $k = 23$), but resulted in more overlap than semantic category selection (88.4 vs. 85.8 items). It is noteworthy that despite the wide range of k values that were evaluated, there was some consistency in the optimal k for all methods (range: 23-27). The similar correlations and overlap of the top methods also motivates us to investigate whether there is some convergence in the items that are being selected, which we will return to after examining the generalization results.

Table 2: The optimal k for each sublist selection method, and measures of average overlap and item difficulty correlations with held-out languages.

Selection Method	Best k	Avg. Overlap	Difficulty r
syntactic	27	92.188	0.856
2 strata	27	93.469	0.844
unstratified	23	90.688	0.841
semantic	25	85.812	0.841
4 strata	27	93.438	0.835
5 strata	26	93.000	0.831
3 strata	25	91.844	0.828
Random	23	88.426	0.776

For construction of the final Swadesh-CDI list, we chose the syntactic category subselection method with the optimal $k = 27$ due to it resulting in the highest correlation between S-CDI and full CDI scores during cross-validation. However, we were struck that the top four selection methods performed so similarly, and thus opted to explore whether these distinct methods were choosing similar items (a convergent solution), or were finding distinct subsets of words that merely achieved similar results (equivalent local maxima). If the methods are indeed finding convergent solutions, we may have greater hope that the solution may generalize well to other languages.

[7] In fact, random subsets of items work so well–ending up with representative numbers of words per semantic and syntactic categories, and of varying difficulty–that researchers initially start CDI short forms using a random subset of items (e.g., ToDo CITE).

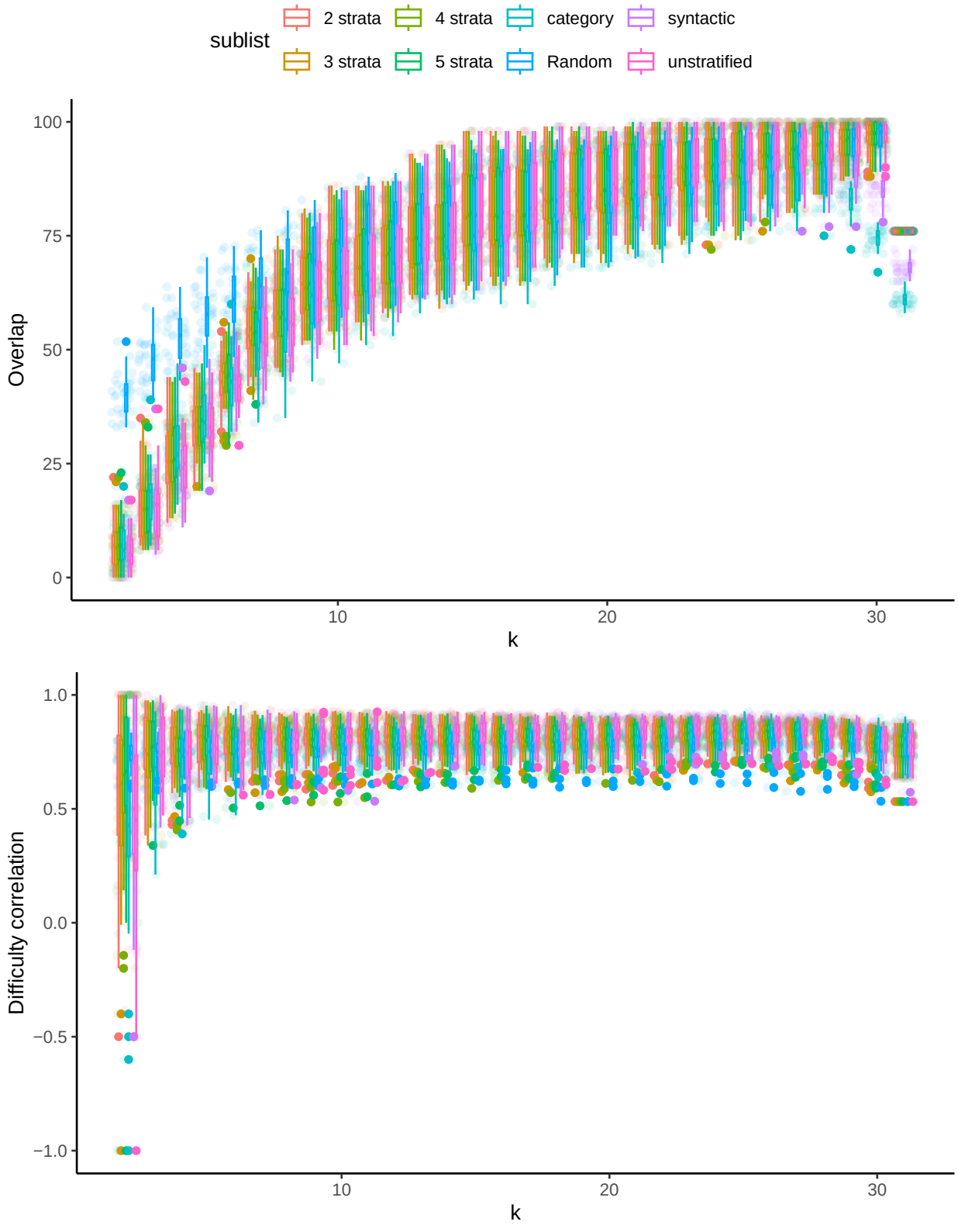

Figure 1: Difficulty correlation and overlap by k for the different sublist selection methods.

Examining the lists selected by each of the top four methods, we found that only 156 unique unilemmas were selected across the 400 items in the lists: 53 unilemmas were selected by all four methods – far higher than chance (which increases with k, but for $k = 27$ the expected overlap of 4 randomly selected lists would be ~ 1 unilemma out of the 483 unilemmas). A further 31 unilemmas were selected by 3 of the 4 methods (vs. chance: ~ 13 unilemmas).

Characterizing the Swadesh-CDI

```
## [1] 96.00058
```

semantic_category	n	EN_freq	Swad_freq
action_words	103	0.15	0.11
food_drink	67	0.10	0.05
descriptive_words	63	0.09	0.10
outside	52	0.08	0.02
household	49	0.07	0.08
animals	43	0.06	0.15
furniture_rooms	32	0.05	0.03
people	29	0.04	0.06
clothing	28	0.04	0.03
body_parts	27	0.04	0.06
locations	27	0.04	NA
games_routines	25	0.04	0.02
helping_verbs	21	0.03	NA
toys	18	0.03	NA
pronouns	16	0.02	0.06
quantifiers	16	0.02	NA
vehicles	14	0.02	0.04
sounds	12	0.02	0.06
time_words	12	0.02	0.04
question_words	7	0.01	0.06
connecting_words	6	0.01	0.03

The 100 S-CDI items, shown in Fig. 3, represented 17 of the 22 semantic categories present on the original American English CDI:WS form, with concrete nouns being most prevalent, followed by adjectives and verbs, and some categories entirely unrepresented: connecting words, helping verbs, quantifiers, pronouns, and toys (but see proposed extension below). 64% of the S-CDI concepts were nouns, 20% were predicates (verbs and adjectives), XX% were function words, and XX% belonged to other lexical categories. Compared to the relative frequency of items per lexical category on the 680-item English CDI:WS (46% nouns, 24% predicates, 15% function words, and 5% other), the S-CDI list tends to have more nouns (especially animals) and fewer function words. The S-CDI items were also present on more forms than typical in the selection set: on average, each item appeared on 0 forms, despite only being required to appear on at least 23 forms. Finally, 2 S-CDI unilemmas were not present on the American English CDI:WS: *fly (animal)* and *crocodile*.

Figure 2 shows the average cross-linguistic difficulty of CDI items by semantic category, for both Swadesh and non-Swadesh items. The difficulty of Swadesh items generally tracked with that of non-Swadesh items, although there were cases where one or the other was more or less difficult.

Comparing the IRT parameters of the S-CDI unilemmas to the rest of the items (across all CDI:WS forms) showed that the discrimination parameter (i.e., slope) of the Swadesh items did not significantly differ from the others, suggesting that the Swadesh items could measure ability as well as non-Swadesh unilemmas. However, S-CDI items were significantly easier than other unilemmas (mean S-CDI d = -0.18, others' mean d = -0.94, $t(6580) = 24.68$, $p < .001$).

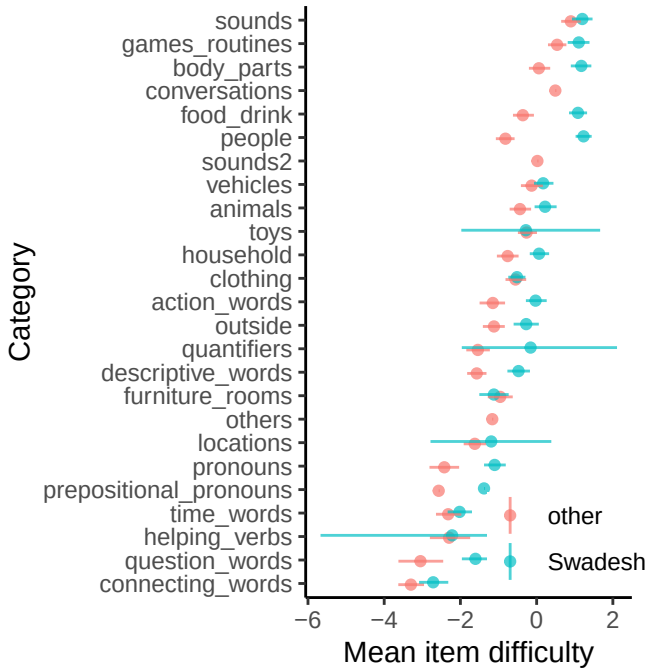

Figure 2: Mean cross-linguistic difficulty of CDI words by semantic category, showing that selection of Swadesh concepts broadly maintained representative difficulty. Bars represent bootstrapped 95% confidence intervals.

To limit the likelihood of a ceiling effect, we proposed extending the S-CDI with 10 additionanl unilemmas chosen from the original Swadesh (1971) list, which were selected to fill gaps in the S-CDI, including pronouns (*I, you, we, this, that*), quantifiers (*all, many, not*), and question words (*who, what*). These unilemmas are included on a mean of 21 of the 26 forms, but are of greater difficulty (mean $d = 0.84$), and cross-linguistic variability (mean $sd(d) = 1.36$).

Validating the Swadesh CDI

To validate the S-CDI, we first measured how well simulated raw scores from the S-CDI items correlated with full CDI:WS scores. This metric approximately reflects how reliable the S-CDI would be as a measure of a child's vocabulary. On average, for the 32 large CDI:WS datasets, the S-CDI's scores were strongly related to the full CDI:WS scores (mean $r = 0.996$; $min = 0.989$, $max = 0.998$; full table on OSF). We compared these correlations against a baseline that simulated an upper bound for how well the Swadesh CDI could be expected to perform: randomly sample N items from the actual CDI:WS of the target language, where N is the number of Swadesh items that are present on the form. On average, the 32 CDI forms included $N = 98$ of the Swadesh items.

Scores from a random subsample of CDI items tend to perform very well at predicting the overall CDI score, as there are no sampling biases related to item difficulty, or cross-linguistic variability in difficulty or inclusion. However, note that this is *not* a viable method to create a new CDI, as in a true CDI construction scenario rather than a simulation, the target CDI would not actually exist! Thus, if the S-CDI comes close to performing as well as a random sample from manually curated CDIs, we consider it a success. Indeed, the random tests had a mean correlation with full CDI scores of $r = 0.997$, only $\epsilon = 0.001$ higher than the S-CDI.

Next, we measured the total test information yielded by the two baselines (recalling that total test information was one criterion for the construction of the S-CDI). This metric reflects how well the S-CDI would be able

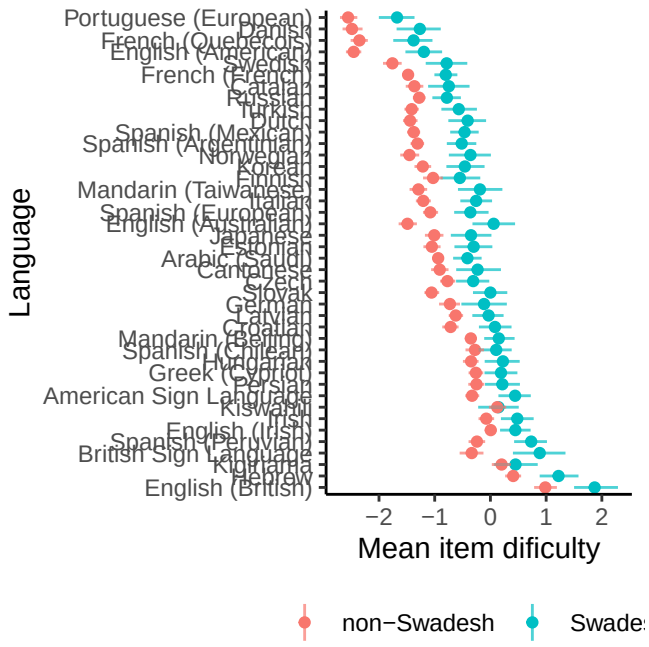

Figure 3: Mean item difficulty of Swadesh vs. non-Swadesh CDI items per language. Swadesh items are consistently easier than non-Swadesh items, but show the same difficulty trend as non-Swadesh items across languages. Bars represent bootstrapped 95% confidence intervals.

to differentiate the ability of children across different ability levels. As reported above, the S-CDI yielded a mean total test information of 66085, while the random unilemmas baseline yielded a total test information of 66374. Although test information was slightly lower for the S-CDI, the values were virtually indistinguishable.

Testing Generalization of the Swadesh CDI

We then evaluated the S-CDI's performance in a test of generalization to 10 more CDI datasets. For the 10 low-data languages, a comparison of simulated S-CDI scores to full CDI:WS revealed that the S-CDI's raw scores were again strongly related (mean $r = 0.990$), with an average of $N = 91$ S-CDI items appearing per list. Table 2 shows the results of this comparison, alongside the upper bound random baseline. Once again, the S-CDI performed nearly as well as a random sample of the actual CDI (random mean $r = 0.994$). With the 10-item extension, the S-CDI's correlation rose to mean $r = 0.993$, demonstrating the value of including items from the more difficult (and variable) categories that were underrepresented on the original list.

language	2 strata	Random	category	syntactic	unstratified
American Sign Language	0.97	0.98	0.98	0.98	0.97
British Sign Language	0.99	0.99	0.99	0.99	0.99
English (Irish)	0.97	0.98	0.97	0.96	0.97
Greek (Cypriot)	0.98	0.99	0.99	0.98	0.98
Irish	0.98	0.98	0.99	0.98	0.97
Kigiriama	0.96	0.94	0.93	0.96	0.96
Kiswahili	0.98	0.98	0.99	0.99	0.98
Persian	0.95	0.95	0.96	0.96	0.96
Spanish (Chilean)	0.97	0.90	0.90	0.97	0.97
Spanish (Peruvian)	0.98	0.97	0.96	0.98	0.96

Table 4: Generalization test results (r vs. full CDI scores).

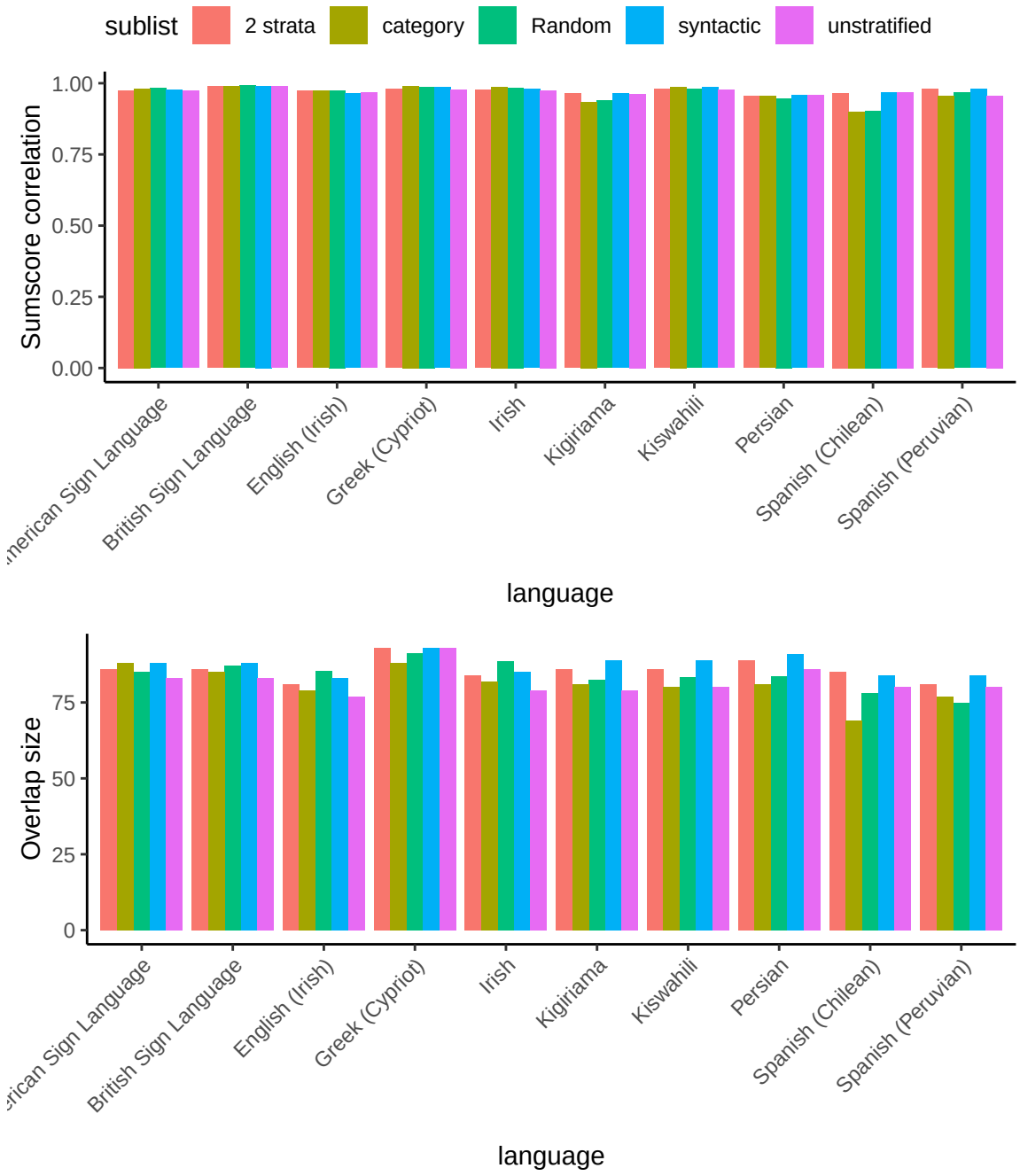

Figure 4: Correlation of each sublist's ability scores vs. full CDI sumscores and overlap in the generalization test.

Discussion

This study compared psychometric models fitted to 32 CDI datasets in order to find concepts that had low variability in their cross-linguistic difficulty, and that were frequently included on CDI:WS forms. We identified 100 concepts that appeared on at least 23 of the CDIs, and which had more consistent cross-linguistic difficulty than other concepts appearing on multiple CDIs. Using real-data simulations, we showed that administering this set of Swadesh CDI items would generate scores that were strongly related to full CDI scores, both for the original 32 datasets, and in a generalization test to 10 low-data languages. Moreover, the Swadesh CDI items resulted in comparable total test information to tests of the same length composed of randomly-selected unilemmas from the target test—a challenging baseline to beat, and a construction method impossible to use when creating a new CDI. The Swadesh CDI contains items with relatively stable cross-linguistic difficulty estimates, and in the absence of access to researchers who are familiar with relevant local cultures and concepts, they may serve as a rapid, simple means of approximating children's ability levels, even in the absence of a large norming dataset.

However, the Swadesh CDI items were also significantly easier than other items, meaning that older children may perform at ceiling if given only the Swadesh CDI items. (This may be unsurprising from the perspective that Swadesh words are meant to be universal, and are therefore more frequent and basic—both within and across individual children's experiences.) Thus, our suggested use case for the Swadesh CDI list is as a starting point for researchers seeking to develop a CDI in a new language, rather than as a complete short-form CDI based on existing long-form CDI data. In particular, researchers should seek to add relevant unilemmas from categories that were less well-represented on the S-CDI, including question words, quantifiers, helping verbs, and pronouns. These categories also tended to be more difficult, so adding items from them is likely to increase the difficulty ceiling of the form. Indeed, the inclusion of 10 such items drawn from the original Swadesh (1971) list increased generalization performance.

Another potential limitation of this work is that most existing CDIs (and most datsets available in Wordbank) target languages in the Indo-European language family. It is not clear to what extent this bias in the existing data might interfere with generalizing to non-Indo-European languages. Nonetheless, our original 32 datasets include 7 FixMe non-Indo-European languages (3 Sino-Tibetan, 1 Afro-Asiatic, 1 Uralic, 1 Koreanic, 1 Turkic), and the generalization datasets include 1 Uralic and 2 Niger-Congo languages (a language family not represented in the original datasets); the broad consistency across language families thus suggests that the effectiveness of the S-CDI may be sufficiently robust. Such a robustness also corroborates with results from Floccia et al. (2018), who demonstrated that a bilingual vocabulary model based on a set of 30 unilemmas had good predictive validity on non-target languages, even those from other language families.

Developing a list of appropriate vocabulary words is not the only challenge researchers face when seeking to develop and use parent-report measures in a new language and culture. The pragmatics of language between children and adults can differ greatly across cultures, and has been found to interfere with administration of parent-report measures of early vocabulary, for example in Kiswahili (Alcock 2017) and Wolof (Weber et al. 2018). As such, local cultural knowledge remains essential in appropriately developing and administering novel CDI adaptations.

ToDo: discuss bi/multilingualism - e.g. generalizing approach used in DLL EN-ES (Tamis-Lemonda2023)...

Despite the myriad challenges that remain in creating new measures of early language development, we believe that the proposed Swadesh CDI list will give researchers a solid foundation to start from, lowering the barrier to the adaptation of CDI forms in new languages, since these are often time-consuming and challenging to construct. Expanding the number of languages with effective vocabulary measures would be a critical step in addressing issues related to the under-representation of linguistic diversity in language acquisition research (Kidd and Garcia 2022). Certainly, increasing the diversity of languages studied is a critical step towards developing a truly general understanding of how young children learn language.

Acknowledgements

We would like to thank all of the contributors to Wordbank, from the researchers who created and adapted the CDIs to those who collected the data (as well as the participants), to those who have created and maintained Wordbank over the years.

References

Alcock, Katherine J. 2017. "Production Is Only Half the Story—First Words in Two East African Languages." *Frontiers in Psychology* 8: 1898.

Baker, Frank B. 2001. *The Basics of Item Response Theory*. ERIC.

Bates, Elizabeth, Virginia Marchman, Donna Thal, Larry Fenson, Philip S Dale, J Steven Reznick, Judy Reilly, and Jeff Hartung. 1994. "Developmental and Stylistic Variation in the Composition of Early Vocabulary." *Journal of Child Language* 21 (1): 85–123.

Bleses, Dorthe, Guido Makransky, Philip S Dale, Anders Højen, and Burcak Aktürk Ari. 2016. "Early Productive Vocabulary Predicts Academic Achievement 10 Years Later." *Applied Psycholinguistics* 37 (6): 1461–76.

Chai, Jun Ho, Chang Huan Lo, and Julien Mayor. 2020. "A Bayesian-Inspired Item Response Theory-Based Framework to Produce Very Short Versions of MacArthur-Bates Communicative Development Inventories." *Journal of Speech, Language, and Hearing Research* 63 (10): 3488–3500.

Chalmers, R. Philip. 2012. "mirt: A Multidimensional Item Response Theory Package for the R Environment." *Journal of Statistical Software* 48 (6): 1–29. https://doi.org/10.18637/jss.v048.i06.

Embretson, Susan E, and Steven P Reise. 2013. *Item Response Theory*. Psychology Press.

Fenson, Larry, V. A. Marchman, D. J. Thal, P. S. Dale, Reznick J. S., and E. Bates. 2007. *MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual (2nd Ed.)*. Baltimore, MD: Brookes.

Fenson, Larry, S. Pethick, C. Renda, J. L. Cox, P. S. Dale, and J. S. Reznick. 2000. "Short-Form Versions of the MacArthur Communicative Development Inventories" 21: 95–116.

Floccia, Caroline, Thomas D Sambrook, Claire Delle Luche, Rosa Kwok, Jeremy Goslin, Laurence White, Allegra Cattani, et al. 2018. "Vocabulary of 2-Year-Olds Learning Learning English and an Additional Language: Norms and Effects of Linguistic Distance."

Frank, Michael C, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. 2017. "Wordbank: An Open Repository for Developmental Vocabulary Data." *Journal of Child Language* 44 (3): 677.

———. 2021. *Variability and Consistency in Early Language Learning: The Wordbank Project*. MIT Press.

Jarůšková, Lucie, Filip Smolík, Kateřina Chládková, Zuzana Oceláková, and Nikola Paillereau. 2023. "How to Build a Communicative Development Inventory: Insights From 43 Adaptations." *Journal of Speech, Language, and Hearing Research*. https://doi.org/10.1044/2023_JSLHR-22-00591.

Kachergis, G., V. A. Marchman, P. S. Dale, J. Mankewitz, and M. C. Frank. 2022. "Online Computerized Adaptive Tests of Children's Vocabulary Development in English and Mexican Spanish." *Journal of Speech, Language, and Hearing Research* 65 (6): 2288–308.

Kidd, Evan, and Rowena Garcia. 2022. "How Diverse Is Child Language Acquisition Research?" *First Language* 42 (6): 703–35. https://doi.org/10.1177/01427237211066405.

Makransky, G., P. S. Dale, P. Havmose, and D. Bleses. 2016. "An Item Response Theory-Based, Computerized Adaptive Testing Version of the MacArthur-Bates Communicative Development Inventory: Words & Sentences (CDI:WS)." *Journal of Speech, Language, and Hearing Research* 59 (2): 281–89.

Marchman, Virginia A, Philip S Dale, and Larry Fenson. 2023. *MacArthur-Bates Communicative Development Inventories User's Guide and Technical Manual, 3rd Edition*. Baltimore, MD: Brookes Publishing Co.

Mayor, Julien, and Nivedita Mani. 2019. "A Short Version of the MacArthur-Bates Communicative Development Inventories with High Validity." *Behavior Research Methods* 51 (5): 2248–55.

Reckase, Mark D. 2009. "Multidimensional Item Response Theory Models." In *Multidimensional Item*

Response Theory, 79–112. Springer.

Swadesh, M. 1971. *The Origin and Diversification of Language*. Edited by Joel Sherzer. Chicago, IL: Aldine.

Tardif, Twila, Paul Fletcher, Weilan Liang, Zhixiang Zhang, Niko Kaciroti, and Virginia A Marchman. 2008. "Baby's First 10 Words." *Developmental Psychology* 44 (4): 929.

Weber, Ann M., Virginia A. Marchman, Yatma Diop, and Anne Fernald. 2018. "Validity of Caregiver-Report Measures of Language Skill for Wolof-Learning Infants and Toddlers Living in Rural African Villages." *Journal of Child Language* 45 (4): 939–58. https://doi.org/10.1017/S0305000917000605.