Measuring a Basic Developmental Vocabulary Across Many Languages

Anonymous CogSci submission

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), including a parent-reported inventory of early-learned vocabulary items, has proven to be a valid and reliable instrument for measuring children's early language skill. CDIs have been adapted to dozens of languages, and cross-linguistic comparisons thus far show both consistency and variability in language acquisition trajectories. Here, we use item-response theory models trained on 26 languages to examine the psychometric properties of translation-equivalent concepts that are frequently included on CDIs, with the goal of identifying a short list of concepts that are of similar learning difficulty in the majority of the languages, in order to propose a pool list of 'universal' words that can be used as a starting point for future CDI adaptations. After identifying a list of 96 items, we test how well this list generalizes to an additional 8 languages.

Keywords: early language learning; CDI; psychometrics; cross-linguistic comparison; Swadesh vocabulary;

Introduction

Assessing children's early language skill is important for researchers, clinicians and parents, as early language skill is predictive of educational outcomes years later. The MacArthur-Bates Communicative Development Inventories [CDIs; Fenson et al. (2007)] are a set of parent report forms that offer a holistic assessment of children's productive and receptive language skills. CDIs are low-cost to administer and produce reliable and valid estimates of early vocabulary and other aspects of early language (Fenson et al., 1994). The CDIs offer more comprehensive data than a short interaction in the lab with a child (Fenson et al., 2000), because they ask parents to report on vocabulary comprehension as well as production, and other milestones, such as communicative gesture use and use of word combinations. Vocabulary size is assessed via a checklist format, which allows caregivers to quickly scan and recognize words their child produces or understands, rather than relying on recall alone. Because of these properties, CDI forms have been adapted to dozens of languages. Data from CDIs are archived in a central repository [Wordbank; Frank, Braginsky, Yurovsky, & Marchman (2017)], and insights from these data have been used to inform theories of early language learning (Frank, Braginsky, Yurovsky, & Marchman, 2021).

Although CDIs measure several constructs related to early language, we focus here on vocabulary assessment. Across languages, measures of vocabulary on the CDIs are very tightly correlated with other aspects of early language like gesture and grammatical competence (Bates et al., 1994; Frank et al., 2021). These studies indicate that the language system looks to be "tightly woven" (Frank et al., 2021), in the sense that early vocabulary size serves as a good proxy measure of children's overall language skill.

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 representing common early-learned words, including nouns (subdivided into e.g., Body Parts, Toys, and Clothing), action words (e.g. verbs), descriptive words (e.g. adjectives), and closed-class words such as pronouns. Researchers have adapted the CDI:WS to many other languages: to date, Wordbank contains contributed datasets from 34 CDI:WS adaptations, and 28 adaptations of the CDI Words & Gestures (CDI:WG) form, which target younger children (8- to 16-month-olds) and are typically $\sim$ 400 words.

Although the CDIs have many advantages, one drawback is that it is difficult for researchers to adapt a CDI for a new language. Indeed, there is no standardized process for adapting a CDI, but the process usually requires years of effort by researchers who are native speakers of a language for which there is no CDI, to first select and translate words that are appropriate for the intended language, and to then iteratively add, refine, and pilot the nascent CDI. [any testimonial on how long this can take, or how many pilot subjects might be needed?] Due to the need to include vocabulary that is appropriate to the target language context, CDIs in different languages do not even necessarily overlap to a great extent. For example, the American English CDI:WS and Mexican Spanish CDI:WS forms – two of the first CDIs created – each have 680 words, but only have 463 overlapping concepts (68%). While it is possible that there is no set of early-learned words that are universal across all languages, it does seem reasonable to expect that some words are more universal than others.

Shortened versions of the CDI have been proposed, and might seem a good place to start when looking for essential words to include on a CDI. The 100-item short-form CDIs (Fenson et al., 2000, 2007) are derived from the 680-item American English CDI:WS form, with items selected based on difficulty and item-to-full score correlations, while also attempting to represent the diversity of semantic and linguistic categories. While the scores of the short-form CDI:WS are highly correlated with scores on the full CDI:WS, there is evidence for a ceiling effect for children older than 27 or 28 months of age. Moreover, it is unclear that the short-form items would generalize well to languages beyond English, as this was not a considered during their selection.

Other efforts to create short CDIs have leveraged Item-Response Theory [IRT; Embretson & Reise (2013)] models, which infer both the abilities of test takers and the difficulty of individual test items (i.e., words), along standardized dimensions. IRT models have the potential to not only yield more accurate measures of children's language ability, but also to enable the construction of Computerized Adaptive Tests (CATs), which choose the next test item based on the responses to the previous items, and thus queickly hone in on the test-takers language ability. CAT-based CDIs presenting 50 or fewer items have been found strongly correlate with scores on the full CDI:WS (Chai, Lo, & Mayor, 2020; Makransky, Dale, Havmose, & Bleses, 2016; Mayor & Mani, 2019), and 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 & Frank, 2022). However, the IRT model driving each CAT needs to be trained on a large and normative dataset, which may not be available in a given language. To date, the IRT models are also fitted separately for each language, and the fitted parameters (e.g., word difficulty) are likely to vary across languages.

The goal of the current study is to examine whether there might be a core set of concepts that are frequently included on CDIs, and–importantly–whether many of them are of roughly equal difficulty across many languages. Taking inspiration from the fields of lexicostatistics and glottochronology, where researchers (notably, Swadesh, 1971) have proposed a list of 'universal concepts'–concepts that exist in all catalogued languages–in order to quantify the genealogical relatedness and dates of divergence 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 ("kill", "die", "see", "sleep"). An analogous Swadesh CDI would include all of the concepts that researchers have chosen to include on the vast majority of CDI:WS adaptations, and which have relatively similar difficulty across many languages.[1] More broadly, examining which types of concepts have more or less similarity across languages in terms of the ease with which they are learned may reveal commonalities as well as idiosyncrasies in children's early experiences: Which semantic categories are most consistent across languages: animals, household objects, food and drink, or another category? Which are more variable? These cross-linguistic comparisons may give new insight into theoretical questions surrounding the similarity and differences between language experience and development in different cultural and linguistic contexts.

In particular, our contributions are 1) to fit IRT models to 28 CDI:WS datasets, 2) to identify 96 candidate "Swadesh CDI" items from a cross-linguistic comparison of concept difficulty and inclusion, and 3) to test how well this Swadesh CDI generalizes to a set of 8 additional low-data languages. We end by making a concrete proposal for how this Swadesh CDI list could be used in creating future CDI adaptations, and by discussing the strengths and weaknesses of our approach.

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 that word, we will use the popular 2-parameter logistic (2PL) model that we have previously found is best justified for CDI data (see Kachergis & Frank, 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 or Rasch) 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 3-parameter logistic model adds a 'guessing' parameter for each test item), we have found elsewhere (Kachergis & Frank, 2022) that the 2PL model is best supported by the Wordbank data.

[1] Low variability in difficulty is important: consider for example 'hat', which is higher difficulty in Spanish ('sombrero'), which would lead to underestimation of language ability Spanish and overestimation in English.

Datasets

Language	WS items	WS N
Norwegian	731	9304
English (American)	680	8828
Danish	725	3714
Portuguese (European)	639	3012
Turkish	711	2422
Mandarin (Taiwanese)	696	1897
Spanish (Mexican)	680	1853
English (Australian)	558	1520
Korean	641	1376
Cantonese	804	1295
German	588	1181
Slovak	609	1066
Mandarin (Beijing)	799	1056
Russian	728	1037
French (Quebecois)	664	929
Swedish	710	900
Spanish (Argentinian)	699	784
Italian	670	752
French (French)	690	665
Spanish (European)	588	593
Hebrew	605	518
Latvian	723	500
Czech	553	493
Croatian	717	377
Hungarian	802	363
Dutch	704	303
Greek (Cypriot)	815	176
Spanish (Peruvian)	600	105
Kigiriama	696	100
English (Irish)	660	99
Irish	691	99
Kiswahili	705	90
Finnish	581	70
Persian	558	50

Table 1: Number of CDI:WG and CDI:WS items and subjects (N) per language.

We report IRT analyses for twenty-six languages from Wordbank (Frank et al., 2017), comprising production data from CDI:WS vocabulary checklists.[2] Data from the first twenty-six rows of Table 1 (Norwegian through Dutch) will be used to select a pool of words with approximately equal cross-linguistic difficulty. CDI:WS production data from an additional eight languages (Table 1: Greek (Cypriot) through Persian) had too few participants to be analyzed with IRT, but will be used to test how well the selected pool of generalize to new languages.

Participants The CDI:WS production dataset consists of the combined Wordbank production data for 47527 children aged 16-30 months on 23020 items across 34 forms. Figure 1C and 1D shows children's production scores vs. age for the CDI:WG and CDI:WS datasets respectively. Note that the distributions of demographic variables (age, sex, maternal education, etc.) of these datasets are not matched, so comparing overall language ability estimates across languages would be impossible. (See Frank et al. (2021) for a discussion of effects of demographic variables on vocabulary development.) Thus, we will focus only on the estimated item parameters, and in particular the variability of item difficulty (b_i).

Instruments 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.

"Produces" responses were coded as 1 and all other responses were coded as 0. Our datasets consist of a dichotomous-valued response matrix for each language, of size N subjects × W words.

Results

All models, data, code for reproducing this paper are available on OSF[3]. Across the 26 IRT models for different CDI:WS forms, parameters for a total of 17715 items were obtained. Figure 1 shows the average difficulty of items in each language, but as mentioned above these differences should not be interpreted, as they may be related to differences in the samples. However, it is useful to show them to emphasize once more that our main focus will be on the cross-linguistic *variability* in words' difficulty (which should be immune to random effects from sample differences across languages), rather than differences in mean difficulties.

Figure 2 shows the average cross-linguistic difficulty of CDI items by semantic category. Sounds (e.g., animal noises), body parts, and common nouns tend to be early-learned, and thus have higher easiness values, while abstract words such as time words and morphologically complex helping verbs are later-learned (i.e., difficult).

Cross-linguistic similarities

We look at the Spearman correlation between the item difficulty of each language compared to each other language. We might expect this to recapitulate the historical relationship between languages, with more similar languages having more similar item difficulties (e.g., Quebecois and European French). (cut/move to appendix?)

Identifying Swadesh CDI Candidates CDI:WS forms have a median of 693 uni-lemmas defined (range: 553 (Czech) to 804 (Cantonese)), and there are a total of 1840 uni-lemmas defined across the 26 languages. Only 61 uni-lemmas appear on all 28 CDI:WS forms: too few to comprise a short form. To expand the pool of items, we consider thresholds on both 1) the variability of an item's difficulty across languages (lower is better), and 2) the number of CDI:WS forms

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

[3] OSF repository: `https://osf.io/XXX/`.

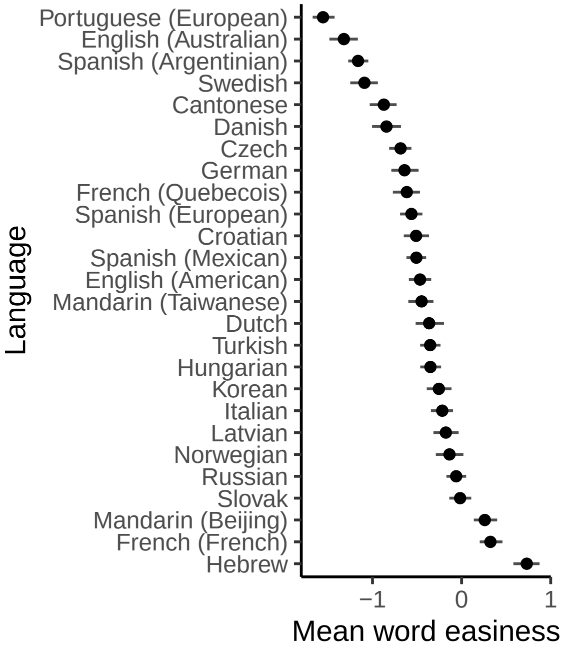


Figure 1: Mean word difficulty per language. Bars represent bootstrapped 95% confidence intervals.

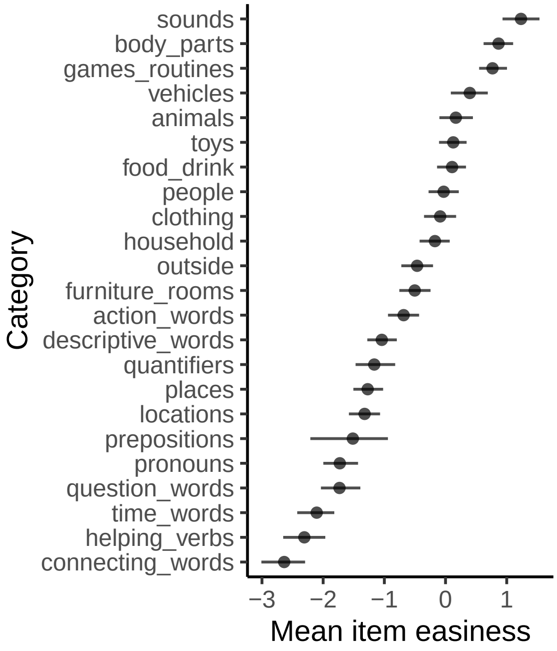
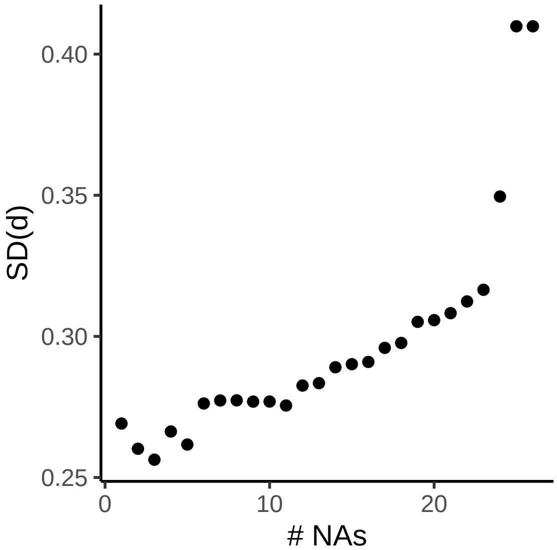


Figure 2: Mean cross-linguistic difficulty of CDI words by semantic category. Bars represent bootstrapped 95% confidence intervals.

on which it appears (more is better). Below we examine the standard deviation of uni-lemmas' cross-linguistic difficulty as a function of how many languages that uni-lemma is missing from. For now we consider items with less than median variability in difficulty that are included on 10 or more of the 26 CDI:WS forms.



Of these, a total of 1312 uni-lemmas are included in more than one language, and only 656 uni-lemmas are included in 10 or more of the languages. We will start by considering this more restricted list.

To evaluate how variable items are in their cross-linguistic difficulty, we calculate the standard deviation (SD) of each uni-lemma's difficulty. The median SD is 1.17 (SD=0.41), so we consider the items with SD less than half the median SD (i.e., SD < 0.96). These 106 candidate "Swadesh CDI" items are summarized by semantic category below (see OSF for full list). 52% of Swadesh CDI words are nouns, 21% are predicates, 22% are function words, and 5% are other. This breakdown is comparable to the lexical category percentages on the 680-item English CDI:WS (46% nouns, 24% predicates, %15 function words, and 15%), minimally suggesting that this method of selecting candidate items is not biased by lexical category. The candidate items, ultimately selected by lower variability, were also present on more forms than typical in the selection set: on average, each item appeared on 20.0754717 forms.

Comparing the IRT parameters of the Swadesh CDI words to the rest of the items showed that the candidate Swadesh items are significantly easier (mean Swadesh $d = 0.03$, other items' mean $d = -0.51$, $t(2737) = 12.74$, $p < .001$). If the Swadesh list is too focused on early-learned items, this may result in ceiling effects for older children. It may be prudent to consider expanding the Swadesh list to include some more difficult items. The discrimination parameter (i.e., slope) of the Swadesh items did not significantly differ from the other items.

Comparing Swadesh CDI to Full CDI:WS

How well do sumscores from the Swadesh CDI items correlate with full CDI:WS scores? On average, for the 26 languages the IRT analysis was based on, the Swadesh CDI's sumscores were strongly related to the full CDI:WS scores (mean $r = 0.990$; $min = 0.976$, $max = 0.996$). However, these correlations were slightly but significantly lower than scores based on randomly-selected items (mean $r = 0.993$, $min = 0.990$, $max = 0.996$, paired $t(25) = 5.67$, $p < .001$). This is likely due to some ceiling effects for older children on the Swadesh CDI, since the randomly-selected items are likely somewhat easier on average than the Swadesh items (see above results).

For the eight low-data languages, a similar comparison revealed that the Swadesh CDI's sumscores were again strongly related to the full CDI:WS scores (mean $r = 0.978$; $min = 0.959$, $max = 0.992$). As with the training set, however, the Swadesh list correlations were slightly but significantly lower than scores based on randomly-selected items (mean $r = 0.985$; $min = 0.972$, $max = 0.992$; $t(7) = 3.22$, $p = .01$).

(Maybe show plot of Swadesh vs. full CDI scores, by age? color by language? might be overwhelming..)

ToDo: use average IRT parameter per item to calculate ability rather than sumscore? propose a way of picking more difficult items? or maybe the value here is in finding the universal (easy) words, and it is up to expert native speakers to choose the more difficult (and idiosyncratic) words for each language.

If I had to find another way of getting the best cross-linguistic words, I would 1) run simulated CATs in all of the languages, and 2) pick the words that cropped up the most often across all languages.

Discussion

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.

Actions	dance, sing, blow, draw, swim, wait, catch, tickle, lick, stay, knock
Animals	cat, cow, *dog*, lion, animal, ant, chicken, spider, goat, turkey
Body Parts	*nose*, *tongue*, *tooth*, head, lip, shoulder, penis, vagina
Clothing	button, necklace, belt, skirt
Connectives	and
Descriptives	blue, green, yellow, black, first, heavy, sticky, dirty, awake
Food/Drink	milk, yogurt, peas, potato chips, lollipop, hamburger
Furniture/ Rooms	Door, washing machine, TV, drawer, garage
Games/ Routines	yes, shh, breakfast, shopping
Household	key, spoon, broom, hammer, radio, toothbrush, paper, glasses, nail, pacifier
Locations	in, under, behind, back
Outside	moon, *star*, sky, hose
People	mommy, daddy, doctor, pet's name, child's name, police, mailman
Places	school, church, gas station
Sounds	meow, cockadoodledoo, grrr, choo choo
Time Words	yesterday, tonight, tomorrow
Vehicles	airplane, helicopter, firetruck

Figure 3: The 96 proposed Swadesh CDI words by semantic category.

References

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

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P. S., Reznick, J. S., … Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, *21*(1), 85–123.

Chai, J. H., Lo, C. H., & Mayor, J. (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. 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

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

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.

Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & Reznick, J. S. (2000). Short-form versions of the MacArthur Communicative Development Inventories, *21*, 95–116.

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.

Kachergis, M., G., & Frank, M. C. (2022). Online computerized adaptive tests of children's vocabulary development in english and mexican spanish. *Journal of Speech, Language, and Hearing Research*, *65*(6), 2288–2308.

Makransky, G., Dale, P. S., Havmose, P., & Bleses, D. (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), 281–289.

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.

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

Swadesh, M. (1971). *The origin and diversification of language*. (J. Sherzer, Ed.). Chicago, IL: Aldine.