

1 Web-CDI: A system for online administration of the MacArthur-Bates Communicative
2 Development Inventories

³ Benjamin deMayo¹, Danielle Kellier², Mika Braginsky³, Christina Bergmann⁴, Cielke
⁴ Hendriks⁴, Caroline Rowland^{4,6}, Michael Frank⁵, & Virginia Marchman⁵

⁵ ¹ Princeton University

⁶ ² University of Pennsylvania

⁷ ³ Massachusetts Institute of Technology

⁴ Max Planck Institute for Psycholinguistics

⁹ ⁵ Stanford University

¹⁰ ⁶ Radboud University

11

Abstract

12 Understanding the mechanisms that drive variation in children's language acquisition
13 requires large, population-representative datasets of children's word learning across
14 development. Parent report measures such as the MacArthur-Bates Communicative
15 Development Inventories (CDI) are commonly used to collect such data, but the traditional
16 paper-based forms make the curation of large datasets logistically challenging. Many CDI
17 datasets are thus gathered using convenience samples, often recruited from communities in
18 proximity to major research institutions. Here, we introduce Web-CDI, a web-based tool
19 which allows researchers to collect CDI data online. Web-CDI contains functionality to
20 collect and manage longitudinal data, share links to test administrations, and download
21 vocabulary scores. To date, over 3,500 valid Web-CDI administrations have been
22 completed. General trends found in past norming studies of the CDI (e.g., Feldman et al.,
23 2000) are present in data collected from Web-CDI: scores of children's productive
24 vocabulary grow with age, female children show a slightly faster rate of vocabulary growth,
25 and participants with higher levels of educational attainment report slightly higher
26 vocabulary production scores than those with lower levels of education attainment. We
27 also report results from an effort to oversample non-white, lower-education participants via
28 online recruitment ($N = 243$). These data showed similar demographic trends to the full
29 sample but this effort resulted in a high exclusion rate. We conclude by discussing
30 implications and challenges for the collection of large, population-representative datasets.

31 *Keywords:* vocabulary development, parent report

32 Word count: X

33 Web-CDI: A system for online administration of the MacArthur-Bates Communicative
34 Development Inventories

35 Children vary tremendously in their vocabulary development (Fenson et al., 1994;
36 Frank, Braginsky, Yurovsky, & Marchman, 2021). Characterizing this variability is central
37 to understanding the mechanisms that drive early language acquisition, yet capturing this
38 variation in broad, diverse samples of children has been a significant challenge for cognitive
39 scientists for decades. The MacArthur-Bates Communicative Development Inventories
40 (MB-CDI, or CDI for short) are a set of commonly-used parent report instruments for
41 assessing vocabulary development in early childhood (Fenson et al., 2007) that were
42 introduced in part to create a cost-effective method for measuring variability across
43 individuals.

44 In this paper, we introduce a web-based tool, Web-CDI, which was developed to
45 address the need for collecting CDI data in an online format. Web-CDI allows researchers
46 to increase the convenience of CDI administration, further decrease costs associated with
47 data collection and entry, and access participant samples that have traditionally been
48 difficult to reach in language development research. Our purpose in this paper is twofold:
49 first, we describe Web-CDI as a platform which streamlines the process of collecting CDI
50 data and collates the data in a way that facilitates the creation of large-scale, multisite
51 collaborative datasets. Second, we profile usage of Web-CDI thus far, with a particular
52 focus on broadening the reach of traditional paper-based methods of collecting vocabulary
53 development data.

54 **The Importance of Parent Report Data**

55 Gaining empirical traction on variation in children's early language requires reliable
56 and valid methods for measuring language abilities, especially in early childhood (8 to 30
57 months). Parent report is a mainstay in this domain. Parents' reports are based on their

58 daily experiences with the child, which are much more extensive than a researcher or
59 clinician can generally obtain. Moreover, they are less likely to be influenced by factors
60 that may mask a child's true ability in the laboratory or clinic (e.g., shyness). One widely
61 used set of parent-report instruments is the MacArthur-Bates Communicative Development
62 Inventories, originally designed for children learning American English (Fenson et al.,
63 2007). The American English CDIs come in several versions, two of which are Words &
64 Gestures (WG) for children 8 to 18 months, focusing on word comprehension and
65 production, as well as gesture use, and Words & Sentences (WS) for children 16 to 30
66 months, focusing on word production and sentence structure. Both the WG and WS
67 measures come in short forms with vocabulary checklists of approximately 90-100 words,
68 and long forms, which contain vocabulary checklists of several hundred items each. (An
69 additional shorter form of the Web-CDI for children 30-37 months, CDI-III, also exists.)
70 For our purposes here, we focus on the American English WG and WS long forms.
71 Together, the CDI instruments allow for a comprehensive picture of milestones that
72 characterize language development in early childhood. A substantial body of evidence
73 suggests that these instruments are both reliable and valid (e.g., Fenson et al., 2007, 1994)
74 leading to their widespread use in thousands of research studies over the last few decades.
75 Initial large-scale work to establish the normative datasets for the American English CDI
76 not only provided key benchmarks for determining children's progress, but also
77 documented the extensive individual differences that characterize early language learning
78 during this critical period of development (Bates et al., 1994; Fenson et al., 1994).
79 Understanding the origins and consequences of this variability remains an important
80 empirical and theoretical endeavor (e.g., Bates & Goodman, 2001; Bornstein & Putnick,
81 2012; see also, Frank, Braginsky, Yurovsky, & Marchman, 2021).

82 The popularity of CDI instruments has remained strong over the years, leading to
83 extensions of the methodology to alternative formats and cross-language adaptations
84 (Fenson et al., 2000). Many teams around the world have adapted the CDI format to the

85 particular language and community (Dale, 2015). Importantly, these adaptations are not
86 simply translations of the original form but rather incorporate the specific features of
87 different languages and cultures, since linguistic variability exists even among cultures that
88 share a native language. As an example of this phenomenon, the word “Cheerios” is more
89 common in the United States than it is in the United Kingdom; as a result, it might be
90 expected that caregivers would report children’s knowledge of this word in the U.S. and not
91 the U.K., even though English is the most common language in both countries. To date
92 there are more than 100 adaptations for languages around the globe. Moreover, several
93 research groups have developed shorter versions of the CDI forms by randomly sampling
94 items from the full CDI and comparing participants’ responses to established norms
95 (Mayor & Mani, 2019) or by developing computer adaptive tests (CATs) that use item
96 response theory or Bayesian approaches to guide the selection of a smaller subset of items
97 to which participants respond (Chai, Lo, & Mayor, 2020; Kachergis et al., 2021;
98 Makransky, Dale, Havmose, & Bleses, 2016).

99 While the reliability and validity of the original CDI instruments is well-established
100 for the American English versions of the forms, existing norming samples are skewed
101 toward families with more years of formal education and away from non-white groups
102 (Fenson et al., 2007). Representation in these norming samples is generally restricted to
103 families living on the U.S. east and west coasts. Further, although paper survey
104 administration is a time-tested method, increasingly researchers and participants would
105 prefer to use an electronic method to administer and fill CDI forms, obviating the need to
106 track (and sometimes mail) paper forms, and the need to key in hundreds of item-wise
107 responses for each child.

108 Here, we report on our recent efforts to create and distribute a web-based version of
109 the CDIs in order to address some of the limitations of the standard paper versions. Online
110 administration of the CDI is not a novel innovation – a variety of research groups have
111 created purpose-build platforms for administering the CDI in particular languages. For

example, Kristoffersen et al. (2013) collected a large normative sample of Norwegian CDIs using a custom online platform. Similarly, the Slovak adaptation of the CDI uses an online administration format (Kapalková & Slanèova, 2007). And many groups have used general purpose survey software such as Qualtrics and Survey Monkey to administer CDIs and variants online (e.g., Caselli, Lieberman, & Pyers, 2020). The innovation of Web-CDI is to provide a comprehensive researcher management interface for the administration of a wide range of CDI forms, allowing researchers to manage longitudinal administrations, download scores, and share links easily, all while satisfying strong guarantees regarding privacy and anonymity. Moreover, a key benefit of a unified data collection and storage system such as Web-CDI is that data from disparate sources are combined into a single repository. This substantially reduces the overhead efforts associated with bringing together data collected by researchers across the world and allows for the analysis of large comparative datasets with the power to detect general trends in vocabulary development that may emerge across languages. Finally, due to an agreement between the CDI Advisory Board and Brooks Publishing, the publisher of the print versions of the CDI suite, Web-CDI is free of charge for researchers to use.

128 Introducing Web-CDI

129 Web-CDI is a web-based platform for CDI administration and management.
130 Web-CDI allows researchers to communicate with families by sharing URLs (web links that
131 contain individual users' own administration of the Web-CDI) via email or social media,
132 facilitating access to families in areas distant from an academic institution and eliminating
133 costly mailings and laboratory visits. Web-CDI also standardizes electronic administration
134 and scoring of CDI forms across labs and institutions, making possible the aggregation of
135 CDI data for later reuse and comparison across administrations by different labs. Indeed,
136 users of Web-CDI grant the CDI Advisory Board permission to access and analyze the
137 resulting data on an opt-out basis, providing a path towards continual improvement of CDI

138 instruments. Since 2018, more than 3,500 CDIs have been collected by 15 research groups
139 throughout the U.S. who are using Web-CDI, demonstrating the potential for large-scale
140 data collection and aggregation.

141 Below, we outline how Web-CDI is used. We begin by detailing the consent obtention
142 process and participant experience. Second, we describe the interface that researchers use
143 to collect data using Web-CDI, specifying a number of common use cases for the platform.

144 **Participant interface**

145 Participants can complete the Web-CDI on a variety of devices, including personal
146 computers and tablets. Web-CDI can be administered on a smartphone, although the
147 experience is not as ideal for the user due to the length of the survey. As Web-CDI moves
148 in the future to incorporate more short forms and adaptive forms (e.g., Chai, Lo, & Mayor,
149 2020; Makransky, Dale, Havmose, & Bleses, 2016; Mayor & Mani, 2019),
150 smartphone-responsive design will become a priority.

151 When a participant clicks a URL shared by a researcher, they are directed to a
152 website displaying their own personal administration of the Web-CDI. In some cases, they
153 may be asked to read and accept a waiver of consent documentation, depending on
154 whether the researcher has chosen to use that feature (see also Researcher Interface below).

155 *Instructions.* After completing the first demographics page, participants are provided
156 with detailed instructions that are appropriate for either the Words & Gestures or Words
157 & Sentences version (see Figure 1). In addition, there are more detailed instructions for
158 completing the vocabulary checklist. Unlike the traditional paper versions, instructions on
159 how to properly choose responses are provided both in written and pictorial form. The
160 pictorial instructions (Figure 1) aim to further increase caregivers' understanding of how to
161 complete the checklist. For example, these instructions clarify that the child's
162 understanding of a word requires them to have some understanding of the object that the

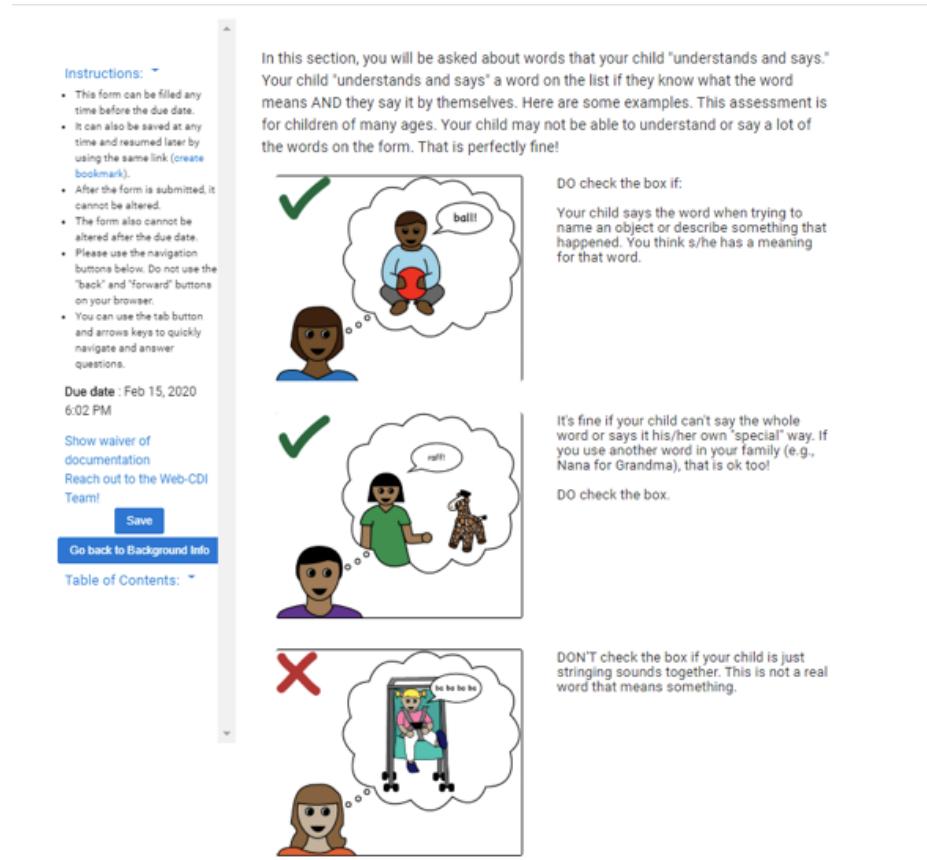


Figure 1. Pictorial instructions in the Web-CDI Words and Sentences instrument.

163 word refers to or some aspect of the word's meaning. In addition, caregivers are reassured
 164 that "child-like" forms (e.g., "raff" for "giraffe") or family- or dialect-specific forms (e.g.,
 165 "nana" for "grandma") are acceptable. Lastly, caregivers are reminded that the child
 166 should be able to produce the words "on their own" and that imitations are not acceptable.
 167 These general "rules of thumb" for completing the form should be familiar to researchers
 168 who are distributing the forms to caregivers so they can field any questions that may arise.
 169 While this is not possible for certain use-cases (e.g., social media recruitment), these
 170 instructions should ideally also be reviewed either in writing (e.g., via email) or verbally
 171 (e.g., over the phone), so that these pictured instructions serve merely as a reminder to
 172 caregivers when completing the form.

173 *Completing the instrument.* The majority of the participant's time is spent

A**PART 1: Early Words****Vocabulary checklist**

The following is a list of typical words in young children's vocabularies. For words your child UNDERSTANDS but does not yet say, place a mark in the first column ("understands"). For words that your child both understands and also SAYS, place a mark in the second column ("understands and says"). You only need to mark one column. If your child uses a different pronunciation of a word (for example, "raffe" for "giraffe" or "sketti" for "spaghetti") or knows a different word that has a similar meaning as the word listed here (e.g., "nana" for "grandma"), go ahead and mark it. Remember, this is a "catalogue" of words that are used by many different children. Don't worry if your child knows only a few right now.

Hide/Show Instructions: ▾

1. Sound Effects And Animal Sounds

baa baa	<input type="checkbox"/> understands	<input type="checkbox"/> understands and says
choo choo	<input type="checkbox"/> understands	<input type="checkbox"/> understands and says
cockadoodledoo	<input type="checkbox"/> understands	<input type="checkbox"/> understands and says

B**PART 1: Words Children Use****A: Vocabulary Checklist**

Children understand many more words than they say. We are particularly interested in the words your child both understands and SAYS. Please go through the list and mark the words you have heard your child SAY on their own. If your child uses a different pronunciation of a word (for example, "raffe" instead of "giraffe" or "sketti" for "spaghetti") or says a different word that has a similar meaning as the word listed here (e.g., "nana" for "grandma"), go ahead and mark it. Remember that this is a "catalogue" of all the words that are used by many different children. Don't worry if your child only says a few right now.

Hide/Show Instructions: ▾

1. Sound Effects And Animal Sounds

<input type="checkbox"/> baa baa	<input type="checkbox"/> choo choo
<input type="checkbox"/> cockadoodledoo	<input type="checkbox"/> grr
<input type="checkbox"/> meow	<input type="checkbox"/> moo
<input type="checkbox"/> ouch	<input type="checkbox"/> quack quack
<input type="checkbox"/> uh oh	<input type="checkbox"/> vroom

Figure 2. (A) Sample items from the American English Words and Gestures form. (B) Sample items from the American English Words and Sentences form.

174 completing the main sections of the instruments. As shown in Figure 2, on the American
 175 English Words and Gestures form, the vocabulary checklist portion (396 items) asks
 176 caregivers to indicate whether their child can "understand" or "understand and say" each
 177 word; they can also indicate that their child neither understands nor says the word by
 178 checking neither box. Additionally, gesture communication and other early milestones are
 179 assessed. In the American English Words and Sentences form, the vocabulary checklist
 180 (680 items) only asks caregivers to indicate which words their child "says." Additional
 181 items assess children's production of their three longest sentences, as well as morphological
 182 and syntactic development more broadly. All of these items are broken up across multiple
 183 screens for easier navigation through the form.

184 At the completion of the form, a graph is displayed illustrating the proportion of

185 words from each semantic category that the child currently produces or understands.

186 Participants can select to download their own responses. In addition, data from the

187 norming studies are used to estimate the ‘hardest’ (i.e., most advanced based on previous

188 work on age of acquisition of individual words, Frank, Braginsky, Yurovsky, and Marchman

189 (2021)) word that the child currently understands or produces. This feedback to caregivers

190 is intended to provide caregivers with a fun “thank you” and intentionally avoids any

191 information which frames their child’s progress relative to other children or any normative

192 standard. The closing page also reminds caregivers that their participation does not

193 constitute a clinical evaluation and that they should contact their pediatrician or primary

194 care physician if they have any concerns about their child’s development.

195 **Researcher interface**

196 One of the main goals of Web-CDI is to provide a unified CDI platform to the child

197 language research community. To that end, researchers request an account by contacting a

198 member of the CDI Advisory Board. Once they have registered an account they can create

199 studies to distribute to participants. One rationale for this personalized registration

200 process is that we ask that researchers allow fully anonymized data from their participants

201 to be shared with the CDI Advisory Board, so that it can be added to Wordbank

202 [<http://wordbank.stanford.edu/>; Frank et al. (2017)] and shared with the broader research

203 community. However, if particular participants indicate in the consent process that they do

204 not want their data to be shared more broadly, then researchers can indicate this in the

205 Web-CDI dashboard to prevent data from specific administrations being contributed to any

206 analyses conducted by the CDI Advisory Board and/or Wordbank. Data currently in

207 Web-CDI, which have not yet been added to the Wordbank repository, will be vetted before

208 being added to ensure that all data being added to Wordbank from Web-CDI are drawn

209 from families with typically-developing children who meet similar inclusion criteria to the

210 ones we describe below in the *Dataset 1* section. Additionally, date of form completion will
211 be preserved when adding Web-CDI data into Wordbank, so that researchers can choose to
212 filter out data that may be affected by the particular point in time at which they were
213 collected (for example, the COVID-19 pandemic, Kartushina et al., 2021).

214 A study in the context of the Web-CDI system is a set of individual administrations
215 created by a researcher that share certain specifications. Table A1 in the Appendix gives
216 an overview of the customizable features that are available at the study level in Web-CDI.
217 These features are set when creating a study using the “Create Study” tool, and most of
218 the features can be updated continuously during data collection using the “Update Study”
219 tool. While some of these features are only particularly relevant to specific use cases (e.g.,
220 longitudinal research and social media data collection, described below), others are relevant
221 to all researchers using Web-CDI.

222 There are currently several CDI forms available for distribution on Web-CDI,
223 including multiple versions of the English WG and WS forms and forms in other languages
224 (see Cross-linguistic research, below). When creating a study, researchers choose one of the
225 forms that they would like to distribute to participants; only one can be used in a given
226 study. Researchers who wish to send multiple forms to participants simultaneously (e.g.,
227 those conducting multilingual research) should create multiple studies, each with a single
228 instrument associated with it.

229 Researchers can download participant data in two formats. Both formatting options
230 output a comma-separated values file with one row per participant; the full data option
231 includes participant-by-item responses, and allows researchers to explore item-level trends,
232 while the summary data option omits item-level data and only provides summary scores
233 and normative information, including total number of words understood/produced and
234 percentile scores by age in months and gender. Percentile scores based are calculated to a
235 single percentile resolution using norms from Fenson et al. (2007).

236 Below, we outline several possible use cases of Web-CDI, as well the features which

237 may facilitate them from a researcher's perspective.

238 *Individual recruitment.* One possible workflow using Web-CDI is to send unique

239 study URLs to individual participants. Researchers do so by entering numerical participant

240 IDs or by auto-generating a specified quantity of participant IDs, each with its own unique

241 study URL, using the "Add Participants" tool in the researcher dashboard. New

242 participants can be added on a continual basis so that researchers can adjust the sample

243 size of their study during data collection. Unique links generated for individual participants

244 expire, by default, 14 days after creation, though the number of days before link expiration

245 is adjustable, which may be an important consideration for some researchers depending on

246 their participant populations and specific project timelines. Workflows that involve

247 generating unique links are most suitable for studies which pair the CDI with other

248 measures, or when researchers contact specific participants from an existing database.

249 *Longitudinal studies.* Web-CDI also facilitates longitudinal study designs in which

250 each participant completes multiple administrations. Researchers wishing to design

251 longitudinal studies can do so by entering a list of meaningful participant IDs using the

252 "Add Participants" tool in the researcher dashboard. If a certain participant ID is added

253 multiple times, Web-CDI will create multiple unique study URLs in the study dashboard

254 that have the same specified ID. In addition, when creating studies, researchers can select

255 whether they would like the demographics information, vocabulary checklist, or no sections

256 at all to be pre-filled when a participant fills out a repeat administration of the instrument.

257 Unless researchers are interested in cumulative vocabulary counts, it is strongly

258 recommended that they do not use the option to pre-fill the vocabulary checklist portion of

259 the instrument in longitudinal administrations as caregivers should complete the

260 instrument at each time point independently. In the case that researchers do choose this

261 option, this is recorded in the Web-CDI database so that, when the data are added to

262 WordBank, researchers can choose to filter out any pre-filled questionnaires.

263 *Social media and survey vendors.* Web-CDI contains several features designed to
264 facilitate data collection from social media recruitment or through third-party
265 crowd-sourcing applications and vendors (e.g., Amazon Mechanical Turk, Prolific). First,
266 rather than creating unique survey links for each participant, researchers can also use a
267 single, anonymous link. When a participant clicks the anonymous link, a new
268 administration with a unique subject ID is created in the study dashboard. Additionally,
269 Web-CDI studies have several customizable features that are geared towards anonymous
270 online data collection. For example, researchers can adjust the minimum amount of time a
271 participant must take to fill out the survey before they are able to submit; with a longer
272 minimum time to completion, researchers can encourage a more thorough completion of the
273 survey. This feature is typically only relevant in research designs in which participants are
274 not vetted by the researcher or those in which there is no direct communication between
275 participants and researchers, as might be the case when recruiting respondents on social
276 media. Responses collected via personal communication with participants show low rates of
277 too-fast responding, mostly removing the need for the minimum time feature. Even in the
278 case of anonymous data collection, however, it is recommended that researchers not raise
279 the minimum completion time higher than 6 minutes, since some caregivers of very young
280 children may theoretically be able to proceed through the measure quickly if their child is
281 not yet verbal. Aside from the minimum time feature, researchers can ask participants to
282 verify that their information is accurate by checking a box at the end of the survey, and
283 can opt to include certain demographic questions at both the beginning and end of the
284 survey, using response consistency on these redundant items as a check of data quality.

285 *Paid participation.* If researchers choose to compensate participants directly through
286 the Web-CDI interface, Web-CDI has built-in functionality to distribute redeemable gift
287 codes when a participant reaches the end of the survey. Web-CDI contains several features
288 to facilitate integration with third-party crowdsourcing applications and survey vendors
289 should they choose to handle participant compensation through another platform. For

example, when creating studies, researchers can enter a URL to redirect participants to when they reach the end of the survey. Researchers using the behavioral research platform Prolific can configure their study to collect participants' unique Prolific IDs and pre-fill them in the survey.

Cross-linguistic research. Web-CDI forms are currently available in English (U.S. American and Canadian), Spanish, French (Quebecois), Hebrew, Dutch and Korean. We are looking to add more language forms to the tool, as the paper version of the forms has been adapted into more than 100 different languages and further ongoing adaptations have been approved by the MB-CDI board (<http://mb-cdi.stanford.edu/adaptations>).

System Design

Web-CDI is constructed using open-source software. All of the vocabulary data collected in Web-CDI are stored in a standard MySQL relational database, managed using Django and Python and hosted either by Amazon Web Services or by a European Union (GDPR) compliant server (see below). Individual researchers can download data from their studies through the researcher interface, and Web-CDI administrators have access to the entire aggregate set of data from all studies run with Web-CDI. Website code is available in a GitHub repository at <https://github.com/langcog/web-cdi>, where interested users can browse, make contributions, and request technical fixes.

Data Privacy and GDPR Compliance

Web-CDI is designed to be compliant with stringent human subjects privacy protections across the world. First, for U.S. users, we have designed Web-CDI based on the United States Department of Health and Human Services "Safe Harbor" Standard for collecting protected health information as defined by the Health Insurance Portability and Accountability Act (HIPAA). In particular, participant names are never collected, birth

314 dates are used to calculate age in months (with no decimal information) but never stored,
315 and geographic zip codes are trimmed to the first 3 digits. Because of the architecture of
316 the site, even though participants enter zip codes and dates of birth, these are never
317 transmitted in full to the Web-CDI server. Since no identifying information is being
318 collected by the Web-CDI system, this feature ensures that Web-CDI can be used by
319 United States labs without a separate Institutional Review Board agreement between
320 users' labs and Web-CDI (though of course researchers using the site will need Institutional
321 Review Board approval of their own research projects).¹

322 In the European Union (EU), research data collection and storage is governed by the
323 Generalized Data Protection Regulation (GDPR) and its local instantiation in the legal
324 system of the member states. Some of the questions on the demographic form contain
325 information that may be considered sensitive (e.g., information about children's
326 developmental disorders), and in some cases, the possibility of linking this sensitive
327 information to participant IDs exists, particularly when researchers draw on local databases
328 that contain full names and addresses for recruitment and contacting. As a result, issues
329 regarding GDPR compliance arise when transferring data outside the EU, namely to
330 Amazon Web Services servers housed in the United States. Following GDPR regulations,
331 these issues would make a data sharing agreement between data collectors and Amazon
332 Web Services necessary. In addition, all administrators who can access the collected data
333 would have to enter such an agreement, which needs updating whenever personnel changes
334 occur. To overcome these hurdles, and in consultation with data protection officers, we
335 opted to leverage the local technical expertise and infrastructure to set up a sister site
336 housed on GDPR-compliant servers, currently available at <http://webcdi.mpi.nl>. This site

¹ Issues of de-identification and re-identifiability are complex and ever changing. In particular, compliance with DHHS "safe harbor" standards does not in fact fully guarantee the impossibility of statistical re-identification in some cases and if potential users have questions, we encourage them to consult with an Institutional Review Board.

337 is updated synchronously with the main Web-CDI website to ensure a consistent user
338 experience and access to the latest features and improvements. This site has been used in
339 135 successful administrations so far and is the main data collection tool for an ongoing
340 norming study in the Netherlands. We are further actively advertising the option to use
341 the European site to other labs who are following GDPR guidelines and are planning
342 adaptations to multiple European languages, where copyright allows.

343 **Current data collection**

344 We now turn to an overview of the data collected thus far using Web-CDI. First, we
345 examine the full sample of all of the Web-CDI administrations collected as of autumn 2020
346 (Dataset 1); we then focus in on a specific subset of Dataset 1 which is comprised of data
347 from recent efforts to oversample non-white, less highly-educated U.S. participants
348 (Dataset 2). Across both datasets, we show that general trends from prior research on
349 vocabulary development are replicated using Web-CDI, and we discuss the potential for
350 using Web-CDI to collect vocabulary development data from diverse communities online.

351 **Dataset 1: Full Current Web-CDI Usage**

352 In this section, we provide some preliminary analyses of Dataset 1, which consists of
353 the full sample of American English Web-CDI administrations collected before autumn
354 2020. At time of writing, researchers from 15 universities in the United States have
355 collected over 5,000 administrations of the American English CDI using Web-CDI since it
356 was launched in late 2017, with 2,868 administrations of the WG form before exclusions
357 and 3,594 administrations of the WS form before exclusions. We excluded participants from
358 the subsequent analyses based on a set of stringent criteria intended for the creation of
359 future normative datasets. We excluded participants if it was not their first administration
360 of the survey; if they were born prematurely or had a birthweight under 5.5 lbs (< 2.5 kg);
361 reported more than 16 hours of exposure to a language other than English per week on

Table 1

Exclusions from Dataset 1: full Web-CDI sample

Exclusion	WG	% of full	WS	% of full
	exclusions	WG sample	exclusions	WS sample
		excluded		excluded
Not first administration	163	5.68%	444	12.35%
Premature or low birthweight	37	1.29%	67	1.86%
Multilingual exposure	449	15.66%	492	13.69%
Illnesses/Vision/Hearing	191	6.66%	203	5.65%
Out of age range	88	3.07%	200	5.56%
Completed survey too quickly	319	11.12%	274	7.62%
System error in word tabulation	1	0.03%	4	0.11%
Total exclusions	1248	44%	1684	47%

362 average (amounting to > 10% exposure to English); had serious vision impairments,
 363 hearing deficits or other developmental disorders or medical issues²; were outside of the
 364 correct age range for the survey; or spent less time on the survey than a pre-specified
 365 timing cutoff. Timing cutoffs were determined by selecting two studies within Dataset 1
 366 that, upon a visual inspection, appeared to contain high-quality responses (i.e., did not
 367 contain a disproportionate number of extremely quick responders), and using these to
 368 estimate the 5th percentile of completion time by the child's age in months with a quantile
 369 regression. Thus, for each age on the WG and WS measures, we obtained an estimate of
 370 the 5th percentile of completion time and used this estimate as the shortest amount of time
 371 participants could spend on the Web-CDI without being excluded from our analyses here.

372 The exclusion criteria we used were designed to be generally comparable with those
 373 used in Fenson et al. (2007), who adopted stringent criteria to establish vocabulary norms

² Exclusions on the basis of child health were decided on a case-by-case basis by author V.M. in consultation with Philip Dale, Donna Thal, and Larry Fenson.

Table 2

Minimum times to completion, WG measure

Age in months	Minimum time to completion (minutes)
8	3.496
9	4.057
10	4.619
11	5.181
12	5.743
13	6.305
14	6.867
15	7.429
16	7.991
17	8.553
18	9.115

³⁷⁴ that reflect typically developing children's vocabulary trajectories. A complete breakdown
³⁷⁵ of the number of participants excluded on each criterion is in Table 1. Of the completed
³⁷⁶ WG forms, 1,248 were excluded, leading to a final WG sample size of 1,620 administrations,
³⁷⁷ and 1,694 WS administrations were excluded, leading to a final WS sample size of 1,900.

³⁷⁸ **Demographic distribution and exclusions.** Figure 3 shows the distribution of
³⁷⁹ participant ethnicities in Dataset 1 as compared with previously reported numbers in a
³⁸⁰ large scale norming study of the paper-based CDI form by Fenson et al. (2007). Several
³⁸¹ issues pertaining to sample representativeness are appreciable. First, as shown in Figure
³⁸² 3A, white participants comprised nearly three quarters of Dataset 1, which is comparable
³⁸³ to U.S Census estimates in 2019 of U.S. residents between the ages of 15 and 34 in 2019;

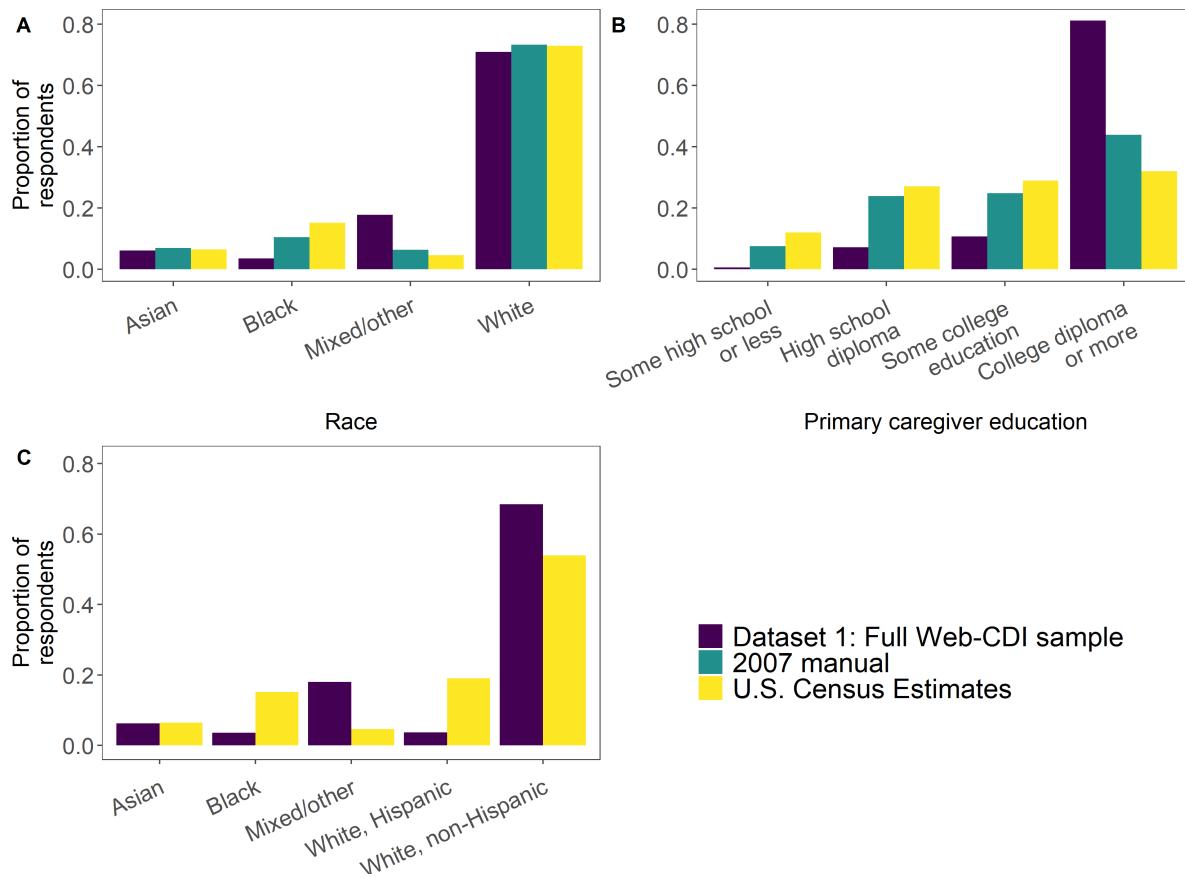


Figure 3. Top row: Proportion of respondents plotted by child race (A) and educational level of primary caregiver (B) from full Web-CDI sample (Dataset 1) to date ($N = 3,520$), compared with norming sample demographics from Fenson (2007) and U.S. Census data (American Community Survey, 2019; National Center for Education Statistics, 2019). Bottom row (C): Participant breakdown by race in Dataset 1 as compared with U.S. Census data, splitting white participants into those who are Hispanic and those are not.

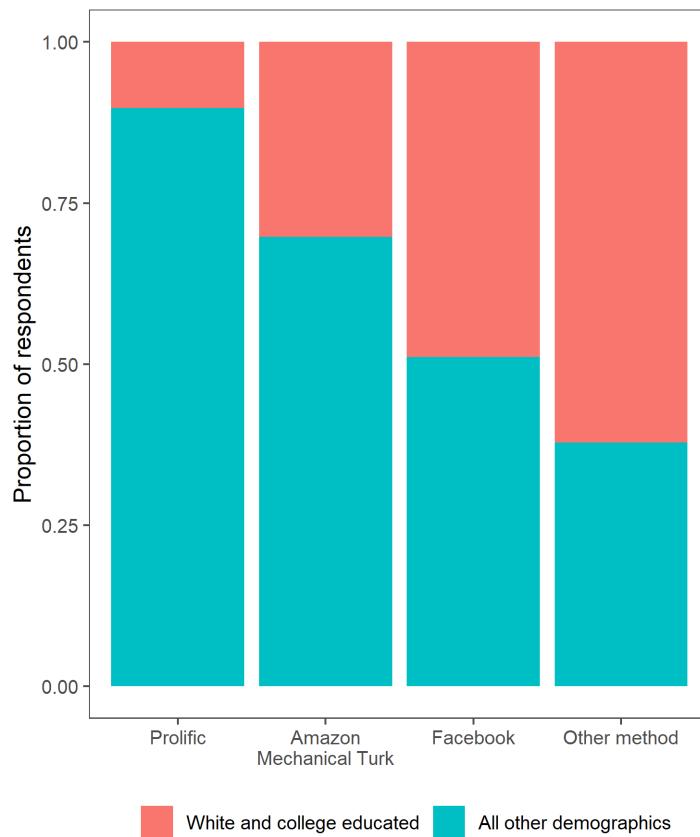


Figure 4. Proportion of participants from Dataset 1 who were white, college educated and not Hispanic, plotted by recruitment method.

384 however, Figure 3C shows that, compared with U.S. Census estimates, many more white
 385 participants in Dataset 1 were non-Hispanic than is true of the U.S. population in general,
 386 indicating that Web-CDI is significantly oversampling white, non-Hispanic individuals (the
 387 breakdown of white participants into Hispanic and non-Hispanic is not reported in the
 388 2007 norms). Moreover, few participants identified as Hispanic/Latinx: 6.4% of WG
 389 participants and 5.2% of WS participants reported Hispanic or Latinx heritage. The low
 390 percentage of Hispanic/Latinx participants was due in part to our exclusion of children
 391 with substantial exposure to languages other than English: before exclusions, 8.4% of WG
 392 participants were Hispanic/Latinx, and 8.1% of WS participants were Hispanic/Latinx.
 393 Finally, representation of Black participants is generally lower in Dataset 1 (3.5%) than in

³⁹⁴ the 2007 norms (10.5%), which is in turn lower U.S. Census estimates (15.2%). This
³⁹⁵ indicates that both Web-CDI data and existing norming samples tend to underrepresent
³⁹⁶ Black participants.

³⁹⁷ Participants' educational attainment level, as measured by the primary caregiver's
³⁹⁸ highest educational level reached³, was similarly skewed. In Dataset 1, 81.2% of responses
³⁹⁹ came from families with college-educated primary caregivers compared to 43.8% from the
⁴⁰⁰ same group in the 2007 norms and 32.0% (Figure 3). Furthermore, less than 1% of
⁴⁰¹ participants report a primary caregiver education level less than a high school degree,
⁴⁰² compared to 7% from the same group in the 2007 norms. The overrepresentation of white,
⁴⁰³ non-Hispanic Americans and those with high levels of education attainment points to a
⁴⁰⁴ general challenge encountered in vocabulary development research, which we return to
⁴⁰⁵ when we detail our efforts to recruit more diverse participants. Figure 4 shows that, of the
⁴⁰⁶ recruitment methods used in Dataset 1, the studies conducted using the platform Prolific
⁴⁰⁷ (which we detail in the *Dataset 2* section) contributed the least to the high proportion of
⁴⁰⁸ white, non-Hispanic, college educated participants. Respondents not known to be recruited
⁴⁰⁹ through an online channel or crowdsourcing platform (labeled "Other method" in Figure 4)
⁴¹⁰ showed the most overrepresentation of white, college educated participants, suggesting that
⁴¹¹ reliance on university convenience samples may be driving the demographic skewness of
⁴¹² Dataset 1 most acutely.

⁴¹³ **Results: Dataset 1.** Although the CDI instruments include survey items intended
⁴¹⁴ to measure constructs other than vocabulary size, such as gesture, sentence production and
⁴¹⁵ grammar, we focus exclusively on the vocabulary measures here. We also visualize key
⁴¹⁶ analyses from Dataset 1 alongside the analogous analyses on the American English CDI
⁴¹⁷ administrations from the WordBank repository (Frank, Braginsky, Yurovsky, & Marchman,

³ Maternal education level is a common measure of family socioeconomic status, but we probe *primary caregiver* education level here to accommodate family structures in which child-rearing may not primarily be the responsibility of the child's mother.

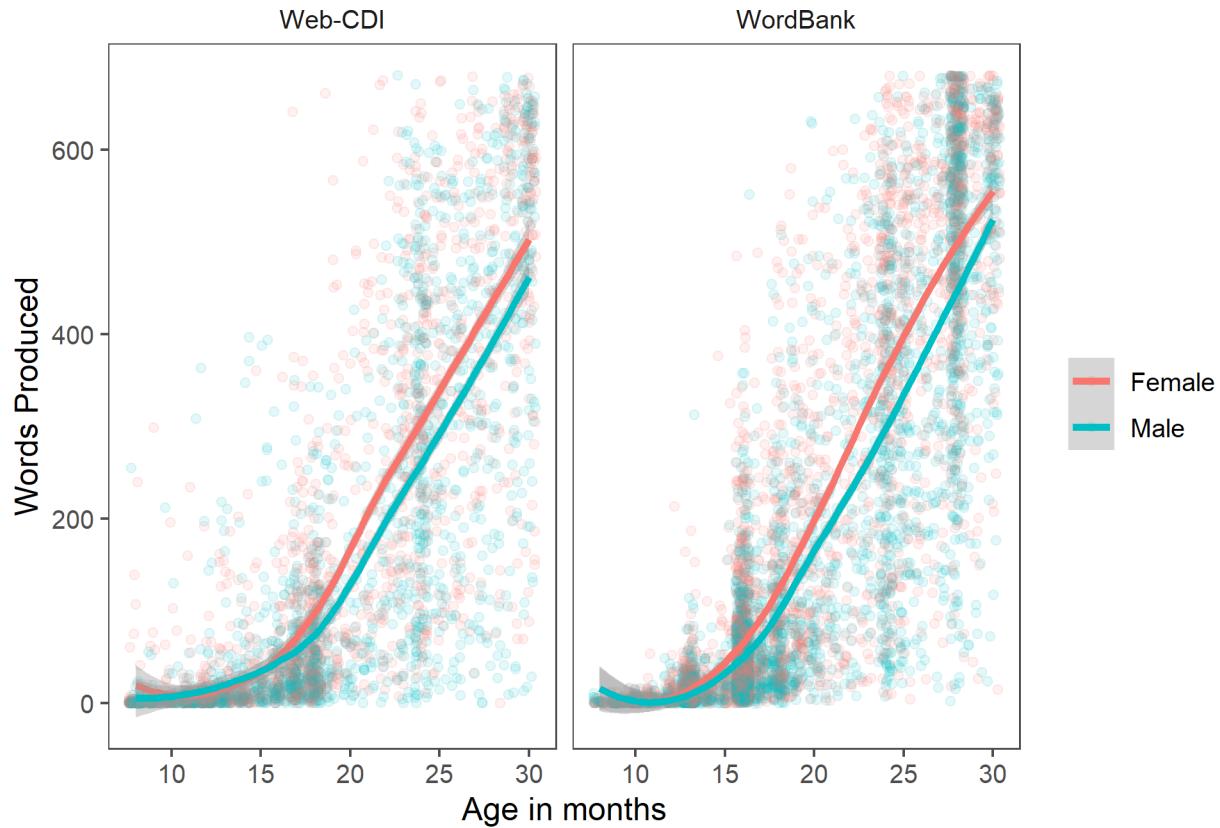


Figure 5. Individual children's vocabulary production scores plotted by children's age and gender (both WG and WS). Left panel: Dataset 1 (full sample of Web-CDI administrations, N = 3,510, with 1,673 girls). Right panel: American English CDI administrations in the WordBank repository (Frank et al., 2021), including only those administrations for which the child's gender was available (N = 6,486, with 3,146 girls). Lines are locally weighted regressions (LOESS) with associated 95% confidence intervals. Children with a different or no reported gender (N = 10) are omitted here.

418 2021) that include the relevant demographic information needed to provide a comparison
419 dataset of traditional paper-and-pencil forms. Across both the WG and WS measures,
420 Dataset 1 shows greater reported vocabulary comprehension and production for older
421 children. Moreover, data from both the WG and WS measures in Dataset 1 replicate a
422 subtle but reliable pattern such that female children tend to have slightly larger vocabulary
423 scores than male children across the period of childhood assessed in the CDI forms (Frank,
424 Braginsky, Yurovsky, & Marchman, 2021), though in these data this difference does not
425 appear until around 18 months (Figure 5).

426 On the WG form, respondents' reports of children's vocabulary comprehension and
427 production both increased with children's age (Figure 6). We replicate overall patterns
428 found by Feldman et al. (2000) in that, on both the "Words Understood" and "Words
429 Produced" measures, vocabulary scores were slightly negatively correlated with primary
430 caregivers' education level, such that those caregivers without any college education
431 reported higher vocabulary scores on both scales. A linear regression model with robust
432 standard errors predicting comprehension scores with children's age and primary caregivers'
433 education level (binned into categories of "High school diploma or less," "Some college
434 education" and "College diploma or more"⁴) as predictors shows main effects of both age
435 ($\beta = 20.05, p < 0.001$) and caregiver primary education ($\beta_{highschool} = 21.86, p = 0.05$).
436 Similarly, a linear regression model with robust standard errors predicting production
437 scores by children's age and primary caregivers' education level shows main effects of age
438 ($\beta = 7.60, p < 0.001$) and caregiver primary education ($\beta_{highschool} = 20.46, p = 0.008$).
439 These analyses were not preregistered, but generally follow the analytic strategy in Frank,
440 Braginsky, Yurovsky, and Marchman (2021); additionally, we fit linear models with robust
441 standard errors to account for heteroskedasticity in the data (Astivia & Zumbo, 2019).
442 Generalized linear model predictions for Web-CDI shown in Figure 6 differ somewhat from

⁴ "High school diploma or less" corresponds to 12 or fewer years of education; "Some college" corresponds to 13 - 15 years of education; "College diploma or more" refers to 16 or more years of education.

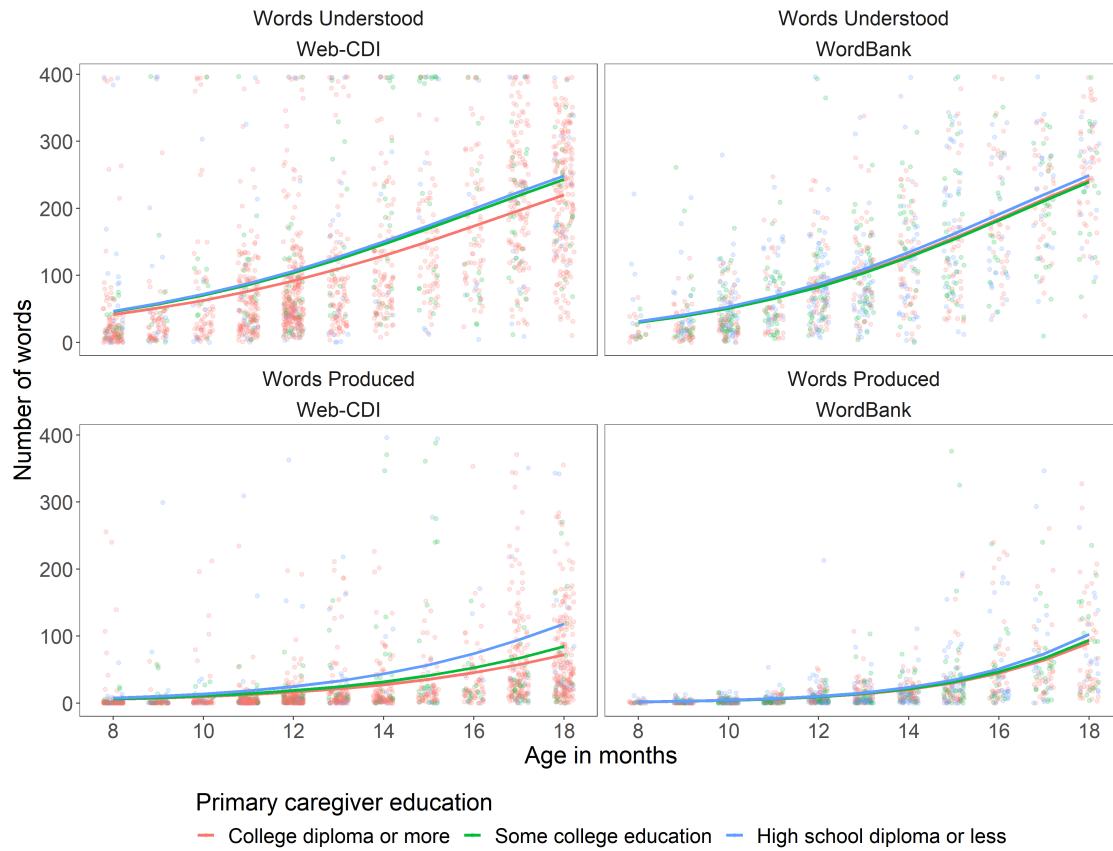


Figure 6. Individual children's word production (top panels) and comprehension (bottom panels) scores from Dataset 1 (full Web-CDI sample) plotted by age and primary caregiver's level of education (binned into "High school diploma or less," "Some college education," and "College diploma or more"). Left panels show results from the sample of Words and Gestures Web-CDI administrations collected as of November 2020 ($N = 1,620$), and right panels show the subset of American English administrations from Wordbank (Frank et al., 2020) that contain information about caregiver education ($N = 1,068$) for comparison. Curves show generalized linear model fits.

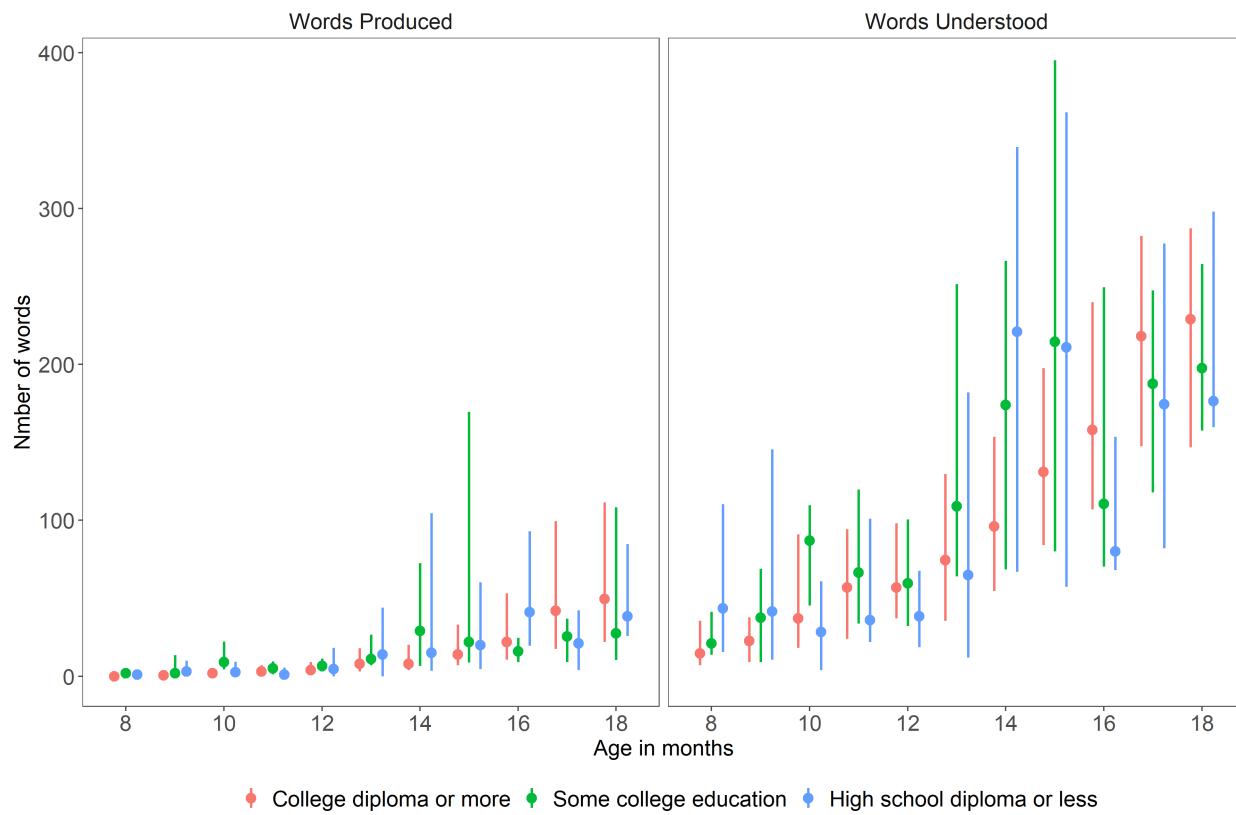


Figure 7. Median vocabulary production (left) and comprehension (right) scores from Dataset 1 (full Web-CDI sample) by age and primary caregiver's level of education attainment on the WG form. Lines indicate span between first and third quartiles for each age.

443 those for WordBank; prediction curves for caregivers of different education attainment
 444 levels diverge slightly more in the Web-CDI sample than in the WordBank sample.

445 The pattern of results seen in the WG subsample of Dataset 1 is consistent with prior
 446 findings indicating that respondents with lower levels of education attainment report
 447 higher vocabulary comprehension and production on the CDI-WG form (Feldman et al.,
 448 2000; Fenson et al., 1994). Although caregivers with lower levels of education attainment
 449 report higher mean levels of vocabulary production and comprehension, median vocabulary
 450 scores (which are more robust to outliers) show no clear pattern of difference across
 451 primary caregiver education levels (Figure 7). This discrepancy between the regression

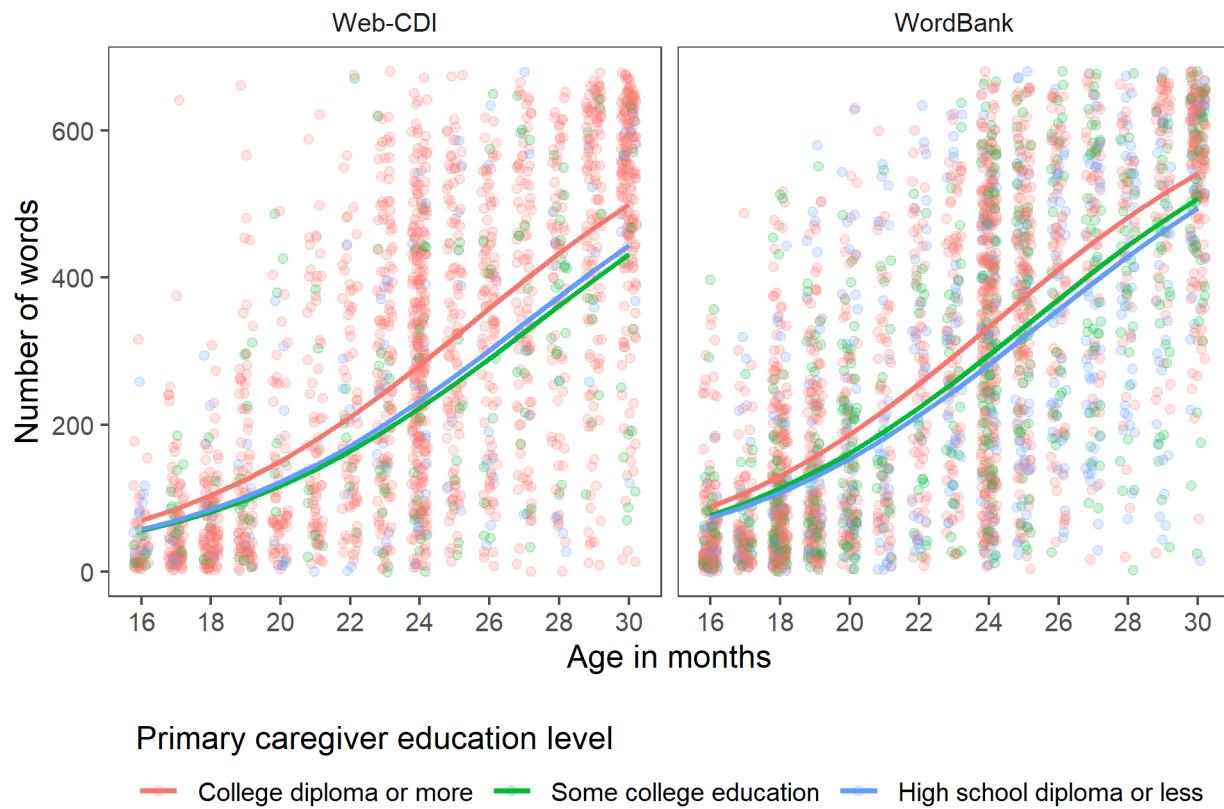


Figure 8. Individual children's vocabulary production scores from Dataset 1 (full Web-CDI sample) plotted by children's age and primary caregiver education level of primary caregiver education on as reported in the sample of Words and Sentences Web-CDI administrations collected as of November 2020 ($N = 1,900$, right panel) and in the WordBank repository ($N = 2,776$, left panel). Curves show generalized linear model fits.

452 effects and a group-median analysis suggests that the regression effects described
 453 previously are driven in part by differential interpretation of the survey items, such that a
 454 few caregivers with lower levels of education attainment are more liberal in reporting their
 455 children's productive and comprehensive vocabularies, especially for the youngest children,
 456 driving up the mean scores for this demographic group.

457 Vocabulary production scores on the WS form show the expected pattern of increase
 458 with children's age in months; in addition, scores replicate the trend reported in Feldman

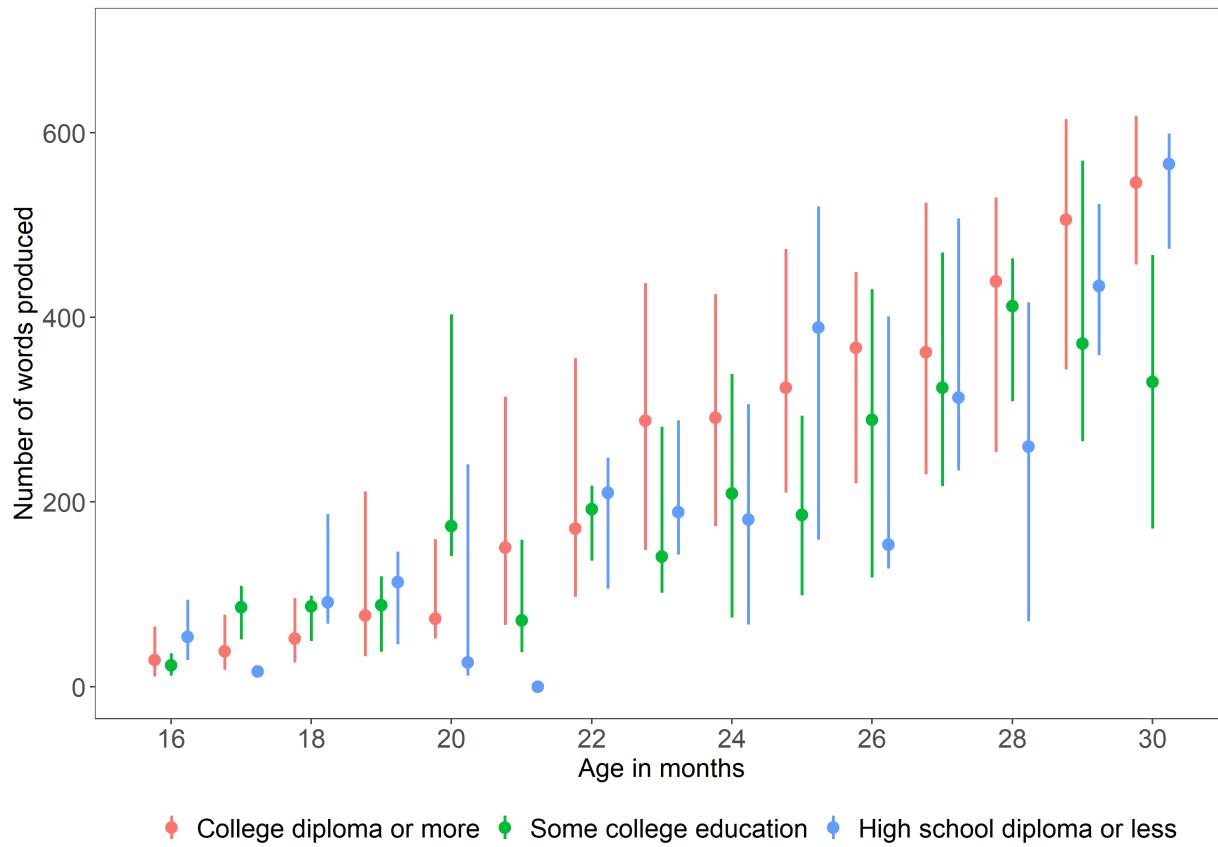


Figure 9. Median vocabulary production scores from Dataset 1 (full Web-CDI sample) by age and primary caregiver's level of education attainment on the WS form. Lines indicate span between first and third quartiles for each age.

et al. (2000) and Frank, Braginsky, Yurovsky, and Marchman (2021) such that primary caregiver education is positively associated with children's reported vocabulary size (Figure 8). Because representation of caregivers without a high school diploma is scarce ($N = 6$ out of a sample of 1,900), interpretation of the data from this group is constrained. Nevertheless, as shown in Figure 8, a small but clear positive association between primary caregiver education and vocabulary score exists such that college-educated caregivers report higher vocabulary scores than those of any other education level. Notably, this association is not the result of outliers and is still appreciable in median scores (Figure 9), unlike the data from the WG measure shown in Figure 7. The implications from these data

468 converge with previous findings which indicate that parental education levels, often used as
469 a metric of a family's socioeconomic status, are related to children's vocabulary size
470 through early childhood.

471 **Discussion: Dataset 1.** In general, the full sample of Web-CDI data after
472 exclusions (Dataset 1) replicates previous norming datasets used with the standard
473 paper-and-pencil form of the MB-CDI. We find that vocabulary scores grow with age and
474 that females hold a slight advantage over males in early vocabulary development.
475 Moreover, Dataset 1 replicates a previously documented relationship between primary
476 caregiver education level and vocabulary scores: on the WG form, primary caregiver
477 education shows a slight negative association with vocabulary scores, whereas the trend is
478 reversed in the WS form. Taken together, these data illustrate that Web-CDI and the
479 standard paper-and-pencil form of the CDI give similar results, and thus that Web-CDI
480 can be used as a valid alternative to the paper format.

481 The data discussed above have stemmed from efforts by many researchers across the
482 United States whose motivations for using the Web-CDI vary. As a result, they reproduce
483 many of the biases of standard U.S. convenience samples. In the next section, we describe
484 in more detail our recent efforts to use the Web-CDI to collect vocabulary development
485 data from traditionally underrepresented participant populations in the United States,
486 attempting to counteract these trends.

487 **Dataset 2: Using Web-CDI to Collect Data from Diverse U.S.-based
488 Communities**

489 Despite the large sample sizes we collected in the previous section, Dataset 1 is, if
490 anything, even more biased towards highly-educated and white families than previous
491 datasets collected using the paper-and-pencil form. How can we recruit more diverse
492 samples to remedy this issue? Here, we discuss and analyze Dataset 2, which consists of
493 those administrations from Dataset 1 which were part of recent data-collection efforts

494 (within the past year and a half) that were specifically aimed towards exploring the use of
495 online recruitment as a potential way to collect more diverse participant samples than are
496 typical in the literature. In other words, the following data from Dataset 2 were included in
497 the previous discussion and analysis of Dataset 1, but we examine them separately here to
498 give special attention to the issue of collecting diverse samples online.

499 While understanding that the performance of standard measurement tools like the
500 CDI among multilinguals is of immense import to the field of vocabulary development
501 research [Gonzalez et al., in prep; Floccia et al. (2018); De Houwer (2019)], we focused in
502 Dataset 2 only on vocabulary development in monolingual children, because collecting data
503 from multilingual populations introduces additional methodological considerations (e.g.,
504 how to measure exposures in each language) that are not the focus of our work here.
505 However, it will be imperative in future to collect large-scale datasets of vocabulary data in
506 bilingual children, both to better calibrate standard tools such as the CDI, as well as to
507 reduce the bias towards monolingual families in the existing literature on measuring
508 vocabulary development.

509 **Online data collection.** Online recruitment methods, such as finding participants
510 on platforms such as Amazon Mechanical Turk, Facebook and Prolific, represent one
511 possible route towards assembling a large, diverse sample to take the Web-CDI. These
512 methods allow researchers to depart from their typical geographical recruitment area much
513 more easily than with paper-and-pencil administration. Online recruitment strategies for
514 vocabulary development data collection have been used in the United Kingdom (Alcock,
515 Meints, & Rowland, 2020), but their usage in the U.S. context remains, to our knowledge,
516 rare. In a series of data collection efforts, we used Web-CDI as a tool to explore these
517 different channels of recruitment.

518 Dataset 2 consists of data that were collected in two phases. In the first phase, we
519 ran advertisements on Facebook which were aimed at non-white families based on users'
520 geographic locations (e.g., targeting users living in majority-Black cities) or other profile



Figure 10. Example Facebook advertisement in Phase 1 of recent data collection.

521 features (e.g., ethnic identification, interest in parenthood-related topics). Advertisements
522 consisted of an image of a child and a caption informing Facebook users of an opportunity
523 to fill out a survey on their child's language development and receive an Amazon gift card
524 (Figure 10). Upon clicking the advertisement, participants were redirected to a unique
525 administration of the Web-CDI, and they received \$5 upon completing the survey. This
526 open-ended approach to recruitment offered several advantages, namely that a wide variety
527 of potential participants from specific demographic backgrounds can be reached on
528 Facebook. However, we also received many incomplete or otherwise unusable survey
529 administrations, either from Facebook users who clicked the link and decide not to
530 participate, or those who completed the survey in an extremely short period of time (over
531 half of all completed administrations, Table 2).

532 In the second phase, we used the crowdsourcing survey vendor Prolific
533 (<http://prolific.co>) in the hopes that some of the challenges encountered with Facebook
534 recruitment would be addressed. Prolific allows researchers to create studies and post them
535 to individuals who are in the platform's participant database, each of whom is assigned a
536 unique alphanumeric "Prolific ID." Importantly, Prolific maintains detailed demographic
537 information about participants, allowing researchers to specify who they would like to
538 complete their studies. Prolific further has a built-in compensation infrastructure that
539 handles monetary payments to participants, eliminating the need to disburse gift cards
540 through Web-CDI.

541 In the particular case of Web-CDI, the demographic information needed to determine
542 whether an individual was eligible to complete our survey (e.g., has a child in the correct
543 age range, lives in a monolingual household, etc.) was more specific than the information
544 that Prolific collects about their participant base. We therefore used a brief pre-screening
545 questionnaire to generate a list of participants who were eligible to participate, and
546 subsequently advertised the Web-CDI survey to those participants. Given that we were
547 interested only in reaching participants in the United States who were not white or who
548 did not have a college diploma, our data collection efforts only yielded a sample that was
549 small ($N = 68$) but much more thoroughly screened than that which we could obtain on
550 Facebook.

551 Across both phases (Facebook and Prolific recruitment), we used the same exclusion
552 criteria as in the full Web-CDI sample to screen participants. A complete tally of all
553 excluded participants is shown in Table 2. In both the WG and WS surveys, exclusion
554 rates in Dataset 2 were high, amounting to 58% of participants who completed the survey.
555 The high exclusion rates were notably driven by an accumulation of survey administrations
556 which participants completed more quickly than our time cutoffs allow (Tables 2 and 3).
557 Many of the survey administrations excluded for fast completion had missing demographic
558 information reported: Among WG participants excluded for too-fast completions, 93% did

not report ethnicity, and among WS participants excluded for the same reason, 97% did not report ethnicity. Absence of these data prevents us from drawing conclusions about the origin or demographic profile of administrations that were excluded. After exclusions, full sample size in Dataset 2 was N = 128 WG completions and N = 115 WS completions.

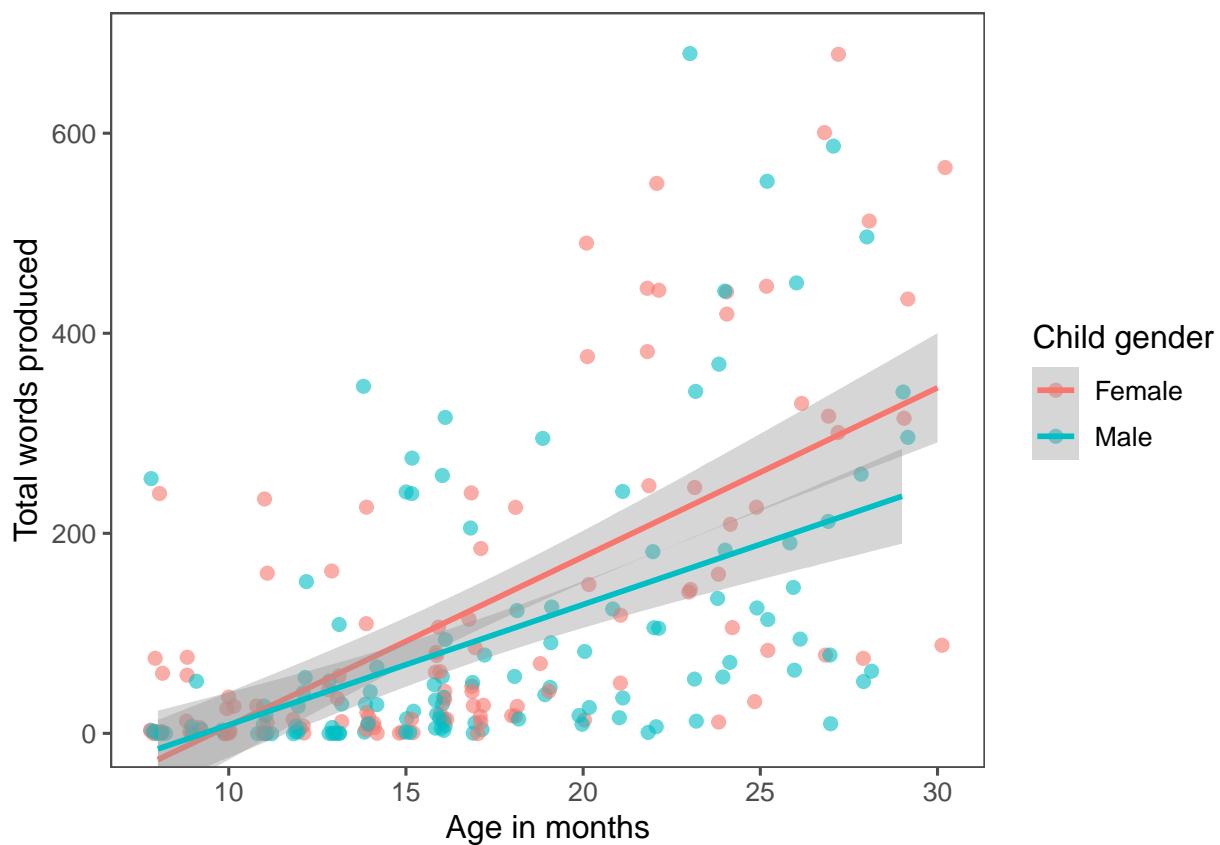


Figure 11. Individual children's vocabulary production scores from Dataset 2 (recent data collection efforts) plotted by children's age and gender (both WG and WS, N = 240, with 114 girls). Lines are best linear fits with associated 95% confidence intervals. Children with a different or no reported gender (N = 3) are omitted here.

The results from Dataset 2 show overall similar patterns to the full Web-CDI sample in several regards. Word production scores from both the WG and WS administrations reflect growing productive vocabulary across the second and third years, with a very small gender effect such that female children's vocabularies are higher across age than males'

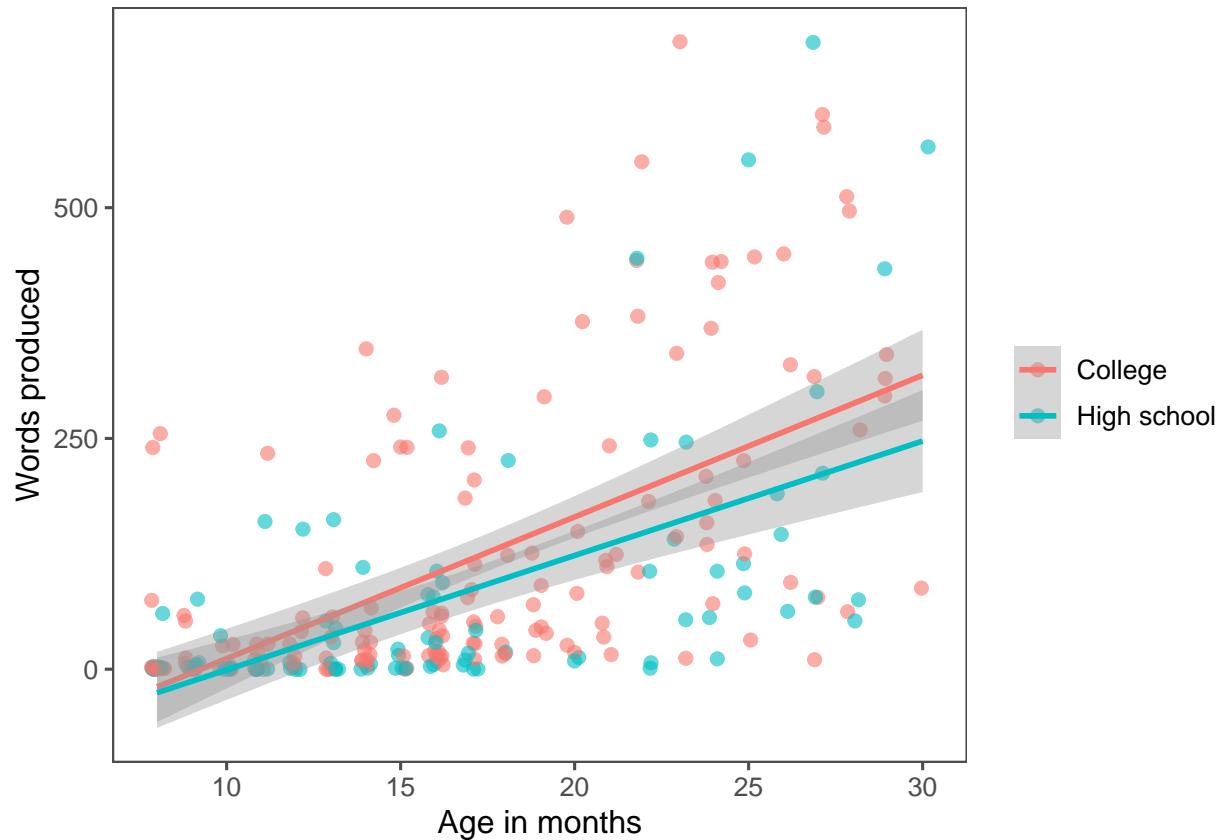


Figure 12. Individual children's vocabulary production scores from Dataset 2 (recent data collection efforts) plotted by age and level of primary caregiver education, binned into those with a high school diploma or less education and those with some college education or a college diploma ($N = 243$). Lines show best linear fits and associated 95% confidence intervals.

567 (Figure 11). The relationship between caregivers' reported levels of education and child's
 568 vocabulary score is not as clear as it is in the full Web-CDI sample (Figure 12); however,
 569 children of college-educated caregivers reported generally higher vocabulary scores across
 570 age than did children of caregivers without any college degree. These patterns suggest that
 571 our data show similar general patterns to other CDI datasets with other populations
 572 (Frank, Braginsky, Yurovsky, & Marchman, 2021).

573 Importantly, Dataset 2 showed a substantial improvement in reaching non-white or

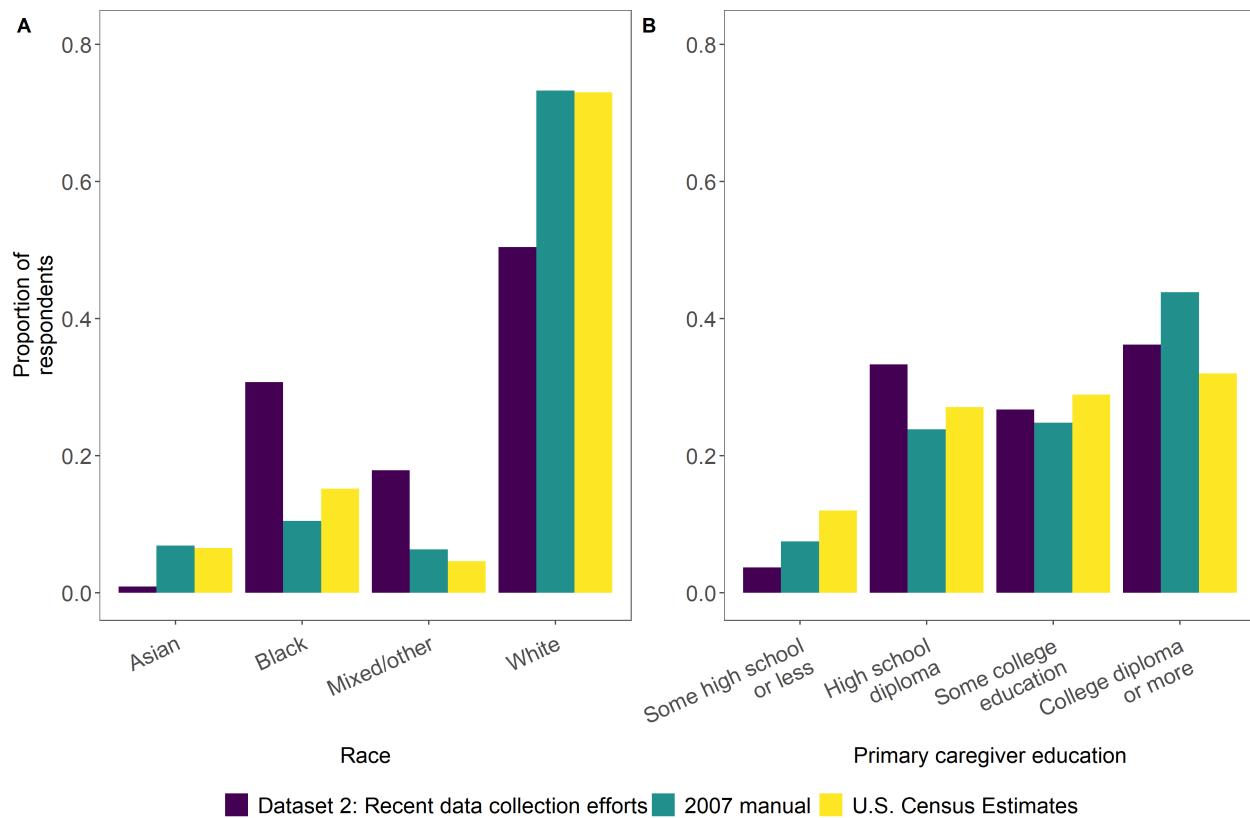


Figure 13. Proportion of respondents plotted by child race (A) and educational level of primary caregiver (B) from Dataset 2, recent data collection efforts aimed towards oversampling non-white, less highly-educated families ($N = 243$), compared with norming sample demographics from Fenson (2007). Latinx participants can be of any race and are thus not represented as a separate category here.

574 less highly-educated participants. After exclusions, Dataset 2 has a higher proportion of
 575 non-white participants than Dataset 1 (the overall Web-CDI sample) and the norms
 576 established by Fenson et al. (2007) (Figure 13). Black participants in particular showed a
 577 marked increase in representation, from 10.5% in the 2007 norms to 30.7% in Dataset 2,
 578 while the proportion of white participants decreased from 73.3% in the 2007 norms to
 579 50.5% in Dataset 2. Representation on the basis of families' reported primary caregiver
 580 education also improved (Figure 13). Participants with only a high school diploma

581 accounted for 33.3% of Dataset 2 as compared to 23.8% in the 2007 norms, and
582 representation of those with a college diploma or more education decreased from 43.8% in
583 the 2007 norms to 36.2% in Dataset 2. Notably, the distribution of Dataset 2 with regards
584 to primary caregiver education level is quite similar to Kristoffersen et al. (2013), who
585 collected a large, nationally-representative sample of CDI responses in Norway and
586 obtained a sample with 30%, 42%, and 24% for participants reporting 12, 14-16, and 16+
587 years of education, respectively.

588 **Discussion: Dataset 2.** The results from Dataset 2 indicate that Web-CDI could
589 be a promising platform to collect vocabulary development data in non-white populations
590 and communities with lower levels of education attainment when paired with online
591 recruitment methods that yield legitimate, representative participant samples. These data
592 do, however, convey clear limitations of our approach. Perhaps most conspicuously, more
593 than half of completed administrations in this sample had to be excluded, in many cases
594 because the information provided by participants appeared rushed or incomplete: over 40%
595 of administrations were completed in a shorter amount of time than that allowed by our
596 cutoff criteria (Tables 2 and 3), and of these quick completions, well over 90% were missing
597 demographic information that is rarely missing in other administrations of the form.
598 Determining the precise reasons for the high exclusion rate, and how (if at all) this
599 (self-)selection may bias data reflecting demographic trends in vocabulary development,
600 requires a more thorough assessment of who is submitting hastily-completed forms. Such
601 an assessment is beyond the scope of the current study. However, all respondents who got
602 to the end of the form were compensated regardless of how thoroughly they completed it,
603 creating the possibility that some participants who clicked the anonymous link may not
604 have been members of the population of interest, but rather were other individuals
605 motivated by compensation. To the extent that participants moved through the form
606 quickly because they found the length burdensome, a transition to short forms, including
607 computer adaptive ones (e.g., Chai, Lo, & Mayor, 2020; Kachergis et al., 2021; Makransky,

608 Dale, Havmose, & Bleses, 2016; Mayor & Mani, 2019), would potentially increase data
609 quality and completion rates substantially.

610 Additionally, the exclusion rates described previously only provide information on
611 those participants who did, at some point, submit a completed form, but many individuals
612 clicked the advertisement link and did not subsequently continue on to complete the form.
613 Without an in-depth exploration of who is clicking the link and why they might choose not
614 to continue, we cannot draw conclusions about the representativeness of the sample in
615 Dataset 2 with regards to the communities we would like to include in our research. As
616 such, a more thorough understanding of how users from different communities respond to
617 various recruitment and sampling methods is needed in future work in order to draw
618 conclusions about demographic trends above and beyond those already established in the
619 literature.

620 Participants in Dataset 2 were recruited through a targeted post on social media, a
621 technique that is considerably more anonymous than recruitment strategies which entail
622 face-to-face or extended contact between researchers and community members. Online
623 recruitment methods may not be suitable for all communities, especially when researchers
624 ask participants to report potentially sensitive information about the health, developmental
625 progress, ethnicity and geographic location of their children (even when such information is
626 stored anonymously). Our goal here was to assess whether general trends in past literature
627 could be recovered using such an online strategy, but future research should take into
628 account that other more personal methods of recruitment, such as direct community
629 outreach or liaison contacts, may improve participants' experiences and their willingness to
630 engage with the study.

631 Finally, a significant limitation of the data collection process in Dataset 2 is that
632 many people in the population of interest - particularly lower-income families - do not have
633 reliable internet access. Having participants complete the Web-CDI on a mobile device

634 may alleviate some of the issues caused by differential access to Wi-Fi, since the vast
635 majority of American adults own a smartphone (Pew Research Center, 2019). Accordingly,
636 improving Web-CDI's user experience on mobile platforms will be an important step
637 towards ensuring that caregivers across the socioeconomic spectrum can easily complete
638 the survey. For smartphone users on pay-as-you-go plans, who may be reluctant to use
639 phone data to complete a study, a possible solution could be compensating participants for
640 the amount of "internet time" they incurred completing the form.

641 **General Discussion and Conclusions**

642 In this paper, we presented Web-CDI, a comprehensive online interface for researchers
643 to measure children's vocabulary by administering the MacArthur-Bates Communicative
644 Development Inventories family of parent-report instruments. Web-CDI provides a
645 convenient researcher management interface, built-in data privacy protections, and a
646 variety of features designed to make both longitudinal and social-media sampling easy. To
647 date, over 3,500 valid administrations of the WG and WS forms have been collected on
648 Web-CDI from more than a dozen researchers in the United States after applying strict
649 exclusion criteria derived from previous norming studies (Fenson et al., 2007, 1994). Our
650 analysis of Dataset 1 shows that demographic trends from previous work using the
651 paper-and-pencil CDI form are replicated in data gleaned from Web-CDI, suggesting that
652 the Web-CDI is a valid alternative to the paper form and captures similar results.

653 Many research laboratories, not only in the United States but around the world,
654 collect vocabulary development data using the MacArthur-Bates CDI. With traditional
655 paper-based forms, combining insights from various research groups can prove challenging,
656 as each group may have slightly different ways of formatting and managing data from CDI
657 forms. By contrast, if all of these groups' data come to be stored in a single repository with
658 a consistent database structure, data from disparate sources can easily be collated and
659 analyzed in a uniform fashion. As such, a centralized repository such as Web-CDI provides

660 a streamlined data-aggregation pipeline that facilitates cross-lab collaborations, multisite
661 research projects and the curation of large datasets that provide more power to
662 characterize the vast individual differences present in children’s vocabulary development.

663 Beyond the goal of simply getting more data, we hope that Web-CDI can advance
664 efforts to expand the reach of vocabulary research past convenience samples into diverse
665 communities. A key question in the field of vocabulary development concerns the
666 mechanisms through which sociodemographic variables, such as race, ethnicity, income and
667 education are linked to group differences in vocabulary outcomes. Large,
668 population-representative samples of vocabulary development data are needed to
669 understand these mechanisms, but research to date (including the full sample of Web-CDI
670 administrations) has often oversampled non-Hispanic white participants and those with
671 advanced levels of education.

672 We explored the use of Web-CDI as part of a potential strategy to collect data from
673 non-white and less highly-educated communities in two phases (Dataset 2). Several overall
674 patterns emerged which we expected: vocabulary scores grew with age, providing a basic
675 validity check of the Web-CDI measure; females held a slight advantage in word learning
676 over males; and children of caregivers with a college education showed slightly higher
677 vocabulary scores. Nonetheless, the insights from these data, while aligned with past
678 norming studies, are necessarily constrained by several features of our method.

679 Limitations of our method notwithstanding, a transition to web-based data collection
680 streamlines the process by which historically underrepresented populations can be reached
681 in child language research. In particular, recruitment methods involving community
682 partners, such as parenting groups, childcare centers and early education providers, are
683 simplified substantially if leaders in these organizations can distribute a web survey to their
684 members that is easy to fill out, as compared with paper forms, which present more
685 logistical hurdles for distribution and collection. Additionally, we hope that Web-CDI can

686 serve as an accessible, free, and easy to use resource for researchers already doing extensive
687 work with underrepresented groups.

688 Web-based data collection can capture useful information about vocabulary
689 development from diverse communities, but future research will need to examine which
690 sampling methods can yield accurate, population-representative data that can advance our
691 understanding of the link between sociodemographic variation and variation in language
692 outcomes.

693 **Ethics statement**

694 Data collected in the United States for this project are anonymized according to
695 guidelines set forth by the United States Department of Health and Human Services. Data
696 collection at Stanford University was approved by the Stanford Institutional Review Board
697 (IRB), protocol 20398.

698 **Data, code and materials availability statement**

- 699 • Open data: All data analyzed in this work are available on the Open Science
700 Framework at <https://osf.io/nmdq4/>.
- 701 • Code: All code for this work is available on the Open Science Framework at
702 <https://osf.io/nmdq4/>.
- 703 • Materials: All code and materials for the Web-CDI are openly available at
704 <https://github.com/langcog/web-cdi>. If readers wish to view the Web-CDI interface
705 in full from the participants' or researchers' perspectives, they are encouraged to
706 contact webcdi-contact@stanford.edu.

707 **Author contributions**

- 708 • Conceptualization: Benjamin deMayo, Danielle Kellier, Mika Braginsky, Caroline
709 Rowland, Michael Frank and Virginia Marchman.

- 710 • Data Curation: Benjamin deMayo, Danielle Kellier and Virginia Marchman.
- 711 • Formal Analysis: Benjamin deMayo.
- 712 • Funding Acquisition: Caroline Rowland and Michael Frank.
- 713 • Investigation: Benjamin deMayo, Danielle Kellier and Virginia Marchman.
- 714 • Methodology: Benjamin deMayo, Danielle Kellier, Michael Frank and Virginia
715 Marchman.
- 716 • Project Administration: Caroline Rowland, Michael Frank and Virginia Marchman.
- 717 • Software: Danielle Kellier, Mika Braginsky, Christina Bergmann and Cielke Hendriks.
- 718 • Supervision: Caroline Rowland, Michael Frank and Virginia Marchman.
- 719 • Visualization: Benjamin deMayo.
- 720 • Writing - Original Draft Preparation: Benjamin deMayo, Michael Frank and Virginia
721 Marchman.
- 722 • Writing - Review & Editing: Benjamin deMayo, Danielle Kellier, Mika Braginsky,
723 Christina Bergmann, Cielke Hendriks, Caroline Rowland, Michael Frank and Virginia
724 Marchman.

725 Software used

726 R [Version 4.0.3; R Core Team (2020)] and the R-packages *broman* [Version 0.71.6;
727 *Broman* (2020)], *cowplot* [Version 1.1.0; Wilke (2020)], *dplyr* [Version 1.0.2; Wickham,
728 François, Henry, and Müller (2020)], *estimatr* [Version 0.26.0; Blair, Cooper, Coppock,
729 Humphreys, and Sonnet (2020)], *forcats* [Version 0.5.0; Wickham (2020a)], *fs* [Version 1.5.0;
730 Hester and Wickham (2020)], *ggplot2* [Version 3.3.2; Wickham (2016)], *here* [Version 0.1;
731 Müller (2017)], *kableExtra* [Version 1.3.1; Zhu (2020)], *papaja* [Version 0.1.0.9997; Aust and
732 Barth (2020)], *purrr* [Version 0.3.4; Henry and Wickham (2020)], *readr* [Version 1.4.0;
733 Wickham and Hester (2020)], *scales* [Version 1.1.1; Wickham and Seidel (2020)], *stringr*
734 [Version 1.4.0; Wickham (2019)], *tibble* [Version 3.0.4; Müller and Wickham (2020)], *tidyR*
735 [Version 1.1.2; Wickham (2020b)], *tidyverse* [Version 1.3.0; Wickham et al. (2019)],

⁷³⁶ *wordbankr* [Version 0.3.1; (**R-wordbankr?**)], and *xtable* [Version 1.8.4; Dahl, Scott,

⁷³⁷ Roosen, Magnusson, and Swinton (2019)]

References

- 738 Alcock, K., Meints, K., & Rowland, C. (2020). *The UK communicative development inventories: Words and gestures*. J&R Press.
- 739
- 740
- 741 Astivia, O. L. O., & Zumbo, B. D. (2019). Heteroskedasticity in multiple regression analysis: What it is, how to detect it and how to solve it with applications in r
- 742
- 743 and SPSS. *Practical Assessment, Research, and Evaluation*, 24(1), 1.
- 744 Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*.
- 745 *Journal of Open Source Software* (Vol. 4, p. 1686).
- 746 <https://doi.org/10.21105/joss.01686>
- 747 Bates, E., & Goodman, J. C. (2001). On the inseparability of grammar and the
- 748 lexicon: Evidence from acquisition.
- 749 Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., ... Hartung,
- 750 J. (1994). Developmental and stylistic variation in the composition of early
- 751 vocabulary. *J Child Lang*, 21(01), 85–123.
- 752 Blair, G., Cooper, J., Coppock, A., Humphreys, M., & Sonnet, L. (2020). *Estimatr: Fast estimators for design-based inference*. *Journal of Open Source Software* (Vol.
- 753 4, p. 1686). Springer-Verlag New York. <https://doi.org/10.21105/joss.01686>
- 754
- 755 Bornstein, M. H., & Putnick, D. L. (2012). Stability of language in childhood: A
- 756 multiage, multidomain, multimeasure, and multisource study. *Developmental*
- 757 *Psychology*, 48(2), 477.
- 758 Broman, K. W. (2020). *Broman: Karl broman's r code*. *Journal of Open Source*
- 759 *Software* (Vol. 4, p. 1686). Springer-Verlag New York.
- 760 <https://doi.org/10.21105/joss.01686>
- 761 Caselli, N. K., Lieberman, A. M., & Pyers, J. E. (2020). The ASL-CDI 2.0: An
- 762 updated, normed adaptation of the MacArthur bates communicative

- development inventory for american sign language. *Behavior Research Methods*, 1–14.
- 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.
- Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., & Swinton, J. (2019). *Xtable: Export tables to LaTeX or HTML*. Retrieved from <https://CRAN.R-project.org/package=xtable>
- Dale, P. S. (2015). Adaptations, Not Translations! Retrieved from <http://mb-cdi.stanford.edu/Translations2015.pdf>
- De Houwer, A. (2019). Equitable evaluation of bilingual children’s language knowledge using the CDI: It really matters who you ask. *Journal of Monolingual and Bilingual Speech*, 1(1), 32–54.
- Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement properties of the MacArthur communicative development inventories at ages one and two years. *Child Development*, 71(2), 310–322.
- Fenson, L., Bates, E., Dale, P. S., Marchman, V. A., Reznick, J. S., & Thal, D. J. (2007). *MacArthur-Bates Communicative Development Inventories*. Brookes Publishing Company.
- Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... Stiles, J. (1994). Variability in early communicative development. *Monogr Soc Res Child Dev*, 59(5).

- 787 Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & Reznick, J. S. (2000).
788 Short-form versions of the MacArthur communicative development inventories.
789 *Applied Psycholinguistics*, 21(1), 95–116.
- 790 Floccia, C., Sambrook, T. D., Delle Luche, C., Kwok, R., Goslin, J., White, L., ...
791 others. (2018). Vocabulary of 2-year-olds learning english and an additional
792 language: Norms and effects of linguistic distance. *Monographs of the Society for*
793 *Research in Child Development*, 83(1), 1–135.
- 794 Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ...
795 others. (2017). A collaborative approach to infant research: Promoting
796 reproducibility, best practices, and theory-building. *Infancy*, 22(4), 421–435.
- 797 Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability*
798 *and consistency in early language learning: The wordbank project*. MIT Press.
- 799 Henry, L., & Wickham, H. (2020). *Purrrr: Functional programming tools*. *Journal of*
800 *Open Source Software* (Vol. 4, p. 1686). <https://doi.org/10.21105/joss.01686>
- 801 Hester, J., & Wickham, H. (2020). *Fs: Cross-platform file system operations based*
802 *on 'libuv'*. *Journal of Open Source Software* (Vol. 4, p. 1686). Springer-Verlag
803 New York. <https://doi.org/10.21105/joss.01686>
- 804 Kachergis, G., Marchman, V., Dale, P., Mehta, H., Mankowitz, J., & Frank, M.
805 (2021). *An online computerized adaptive test (CAT) of children's vocabulary*
806 *development in english and mexican spanish*.
- 807 Kapalková, S., & Slanèová, D. (2007). Adaptation of CDI to the slovak language. In
808 *Proceedings from the first european network meeting on the communicative*
809 *development inventories: May 24-28 2006 dubrovnik croatia*. University of Gävle.
- 810 Kartushina, N., Mani, N., AKTAN-ERCIYES, A., Alaslani, K., Aldrich, N. J.,
811 Almohammadi, A., ... al., et. (2021). COVID-19 first lockdown as a unique

- 812 window into language acquisition: What you do (with your child) matters.
- 813 PsyArXiv. <https://doi.org/10.31234/osf.io/5ejwu>
- 814 Kristoffersen, K. E., Simonsen, H. G., Bleses, D., Wehberg, S., Jørgensen, R. N.,
- 815 Eiesland, E. A., & Henriksen, L. Y. (2013). The use of the internet in collecting
- 816 CDI data—an example from norway. *Journal of Child Language*, 40(03), 567–585.
- 817 Makransky, G., Dale, P. S., Havmose, P., & Bleses, D. (2016). An item response
- 818 theory-based, computerized adaptive testing version of the MacArthur–bates
- 819 communicative development inventory: Words & sentences (CDI: WS). *Journal*
- 820 *of Speech, Language, and Hearing Research*, 59(2), 281–289.
- 821 Mayor, J., & Mani, N. (2019). A short version of the MacArthur–bates
- 822 communicative development inventories with high validity. *Behavior Research*
- 823 *Methods*, 51(5), 2248–2255.
- 824 Müller, K. (2017). *Here: A simpler way to find your files*. *Journal of Open Source*
- 825 *Software* (Vol. 4, p. 1686). <https://doi.org/10.21105/joss.01686>
- 826 Müller, K., & Wickham, H. (2020). *Tibble: Simple data frames*. *Journal of Open*
- 827 *Source Software* (Vol. 4, p. 1686). <https://doi.org/10.21105/joss.01686>
- 828 Percentage of persons 18 to 24 years old and age 25 and over, by educational
- 829 attainment, race/ethnicity, and selected racial/ethnic subgroups: 2010 and 2017.
- 830 (2019). https://nces.ed.gov/programs/digest/d18/tables/dt18_104.40.asp?referer=raceindica.asp.
- 832 Pew research center mobile fact sheet. (2019). Retrieved from
- 833 <https://www.pewresearch.org/internet/fact-sheet/mobile/>
- 834 R Core Team. (2020). *R: A language and environment for statistical computing*.
- 835 *Journal of Open Source Software* (Vol. 4, p. 1686). Vienna, Austria: R
- 836 Foundation for Statistical Computing; Springer-Verlag New York.

- 837 <https://doi.org/10.21105/joss.01686>
- 838 Snyder, T. D., De Brey, C., & Dillow, S. A. (2019). Digest of education statistics
839 2017, NCES 2018-070. *National Center for Education Statistics*.
- 840 Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. *Journal of Open*
841 *Source Software* (Vol. 4, p. 1686). Springer-Verlag New York.
842 <https://doi.org/10.21105/joss.01686>
- 843 Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string*
844 *operations*. *Journal of Open Source Software* (Vol. 4, p. 1686).
845 <https://doi.org/10.21105/joss.01686>
- 846 Wickham, H. (2020a). *Forcats: Tools for working with categorical variables*
847 *(factors)*. *Journal of Open Source Software* (Vol. 4, p. 1686). Springer-Verlag
848 New York. <https://doi.org/10.21105/joss.01686>
- 849 Wickham, H. (2020b). *Tidyr: Tidy messy data*. *Journal of Open Source Software*
850 (Vol. 4, p. 1686). <https://doi.org/10.21105/joss.01686>
- 851 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ...
852 Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*,
853 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- 854 Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of*
855 *data manipulation*. *Journal of Open Source Software* (Vol. 4, p. 1686).
856 Springer-Verlag New York. <https://doi.org/10.21105/joss.01686>
- 857 Wickham, H., & Hester, J. (2020). *Readr: Read rectangular text data*. *Journal of*
858 *Open Source Software* (Vol. 4, p. 1686). <https://doi.org/10.21105/joss.01686>
- 859 Wickham, H., & Seidel, D. (2020). *Scales: Scale functions for visualization*. *Journal*
860 *of Open Source Software* (Vol. 4, p. 1686). <https://doi.org/10.21105/joss.01686>

- 861 Wilke, C. O. (2020). *Cowplot: Streamlined plot theme and plot annotations for*
862 *'ggplot2'*. *Journal of Open Source Software* (Vol. 4, p. 1686). Springer-Verlag
863 New York. <https://doi.org/10.21105/joss.01686>
- 864 Zhu, H. (2020). *kableExtra: Construct complex table with 'kable' and pipe syntax.*
865 *Journal of Open Source Software* (Vol. 4, p. 1686).
866 <https://doi.org/10.21105/joss.01686>

Table 3

Minimum times to completion, WG measure

Age in months	Minimum time to completion (minutes)
16	8.129
17	8.613
18	9.097
19	9.581
20	10.065
21	10.55
22	11.034
23	11.518
24	12.002
25	12.486
26	12.97
27	13.455
28	13.939
29	14.423
30	14.907

Table 4

Exclusions from Dataset 2: recent data collection using Facebook and Prolific.

Exclusion	WG	% of full	WS	% of full
	exclusions	WG sample excluded	exclusions	WS sample excluded
Not first administration	0	0.00%	0	0.00%
Premature or low birthweight	7	2.53%	1	0.33%
Multilingual exposure	18	6.50%	23	7.62%
Illnesses/Vision/Hearing	4	1.44%	4	1.32%
Out of age range	1	0.36%	26	8.61%
Completed survey too quickly	119	42.96%	133	44.04%
System error in word tabulation	0	0.00%	0	0.00%
Total exclusions	149	54%	187	62%

Appendix

Table A1

Settings customizable by researchers when creating new studies to be run on the Web-CDI platform.

Study setting	Default value	Notes
Study name	none	—
Instrument	none	—
Age range for study	none	Defaults based on instrument selected.
Number of days before study expiration	14	Must be between 1 and 28 days.
Measurement units for birth weight	Pounds and ounces	Weight can also be measured in kilograms (kg).
Minimum time (minutes) a parent must take to complete the study	6	—
Waiver of documentation	blank	Can be filled in by researchers to include a Waiver of Documentation for the participant to approve before proceeding to the experiment.
Pre-fill data for longitudinal participants?	No, do not populate any part of the form	Researchers can choose to pre-fill the background information and the vocabulary checklist.

Table A1

Settings customizable by researchers when creating new studies to be run on the Web-CDI platform. (continued)

Study setting	Default value	Notes
Would you like to pay subjects in the form of Amazon gift cards?	No	If checked, researchers can enter gift codes to distribute to participants once they have completed the survey.
Do you plan on collecting only anonymous data in this study? (e.g., posting ads on social media, mass emails, etc)	No	If checked, researchers can set a limit for the maximum number of participants, as well as select an option that asks participants to verify that the information entered is accurate.
Would you like to show participants graphs of their data after completion?	Yes	–
Would you like participants to be able to share their Web-CDI results via Facebook?	No	–
Would you like participants to answer the confirmation questions?	No	Asks redundant demographic questions to serve as attention checks.

Table A1

Settings customizable by researchers when creating new studies to be run on the Web-CDI platform. (continued)

Study setting	Default value	Notes
Provide redirect button at completion of study?	No	Used to redirect users to external site after form completion.
Capture the Prolific Id for the participant?	No	For integration with Prolific.
Allow participant to print their responses at end of Study?	No	—
End message	Standard end-of-study message	Can be changed to customize end-of-study message.

Table A2

Regression output for WG comprehension measure.

term	estimate	std.error	statistic	p.value	conf.low	conf.high	df
Intercept	122.274887	2.4269823	50.3814499	0.0000000	117.5145103	127.0352635	1610
Age	20.050405	0.7674302	26.1266830	0.0000000	18.5451379	21.5556721	1610
Caregiver education: Some college	17.444920	8.1788102	2.1329411	0.0330804	1.4026869	33.4871541	1610
Caregiver education: High school or less	21.862039	10.9352384	1.9992284	0.0457515	0.4132408	43.3108368	1610
Age * Caregiver education: Some college	-1.991340	2.2605963	-0.8808917	0.3785079	-6.4253612	2.4426803	1610
Age * Caregiver education: High school or less	-6.604347	3.1589628	-2.0906694	0.0367142	-12.8004580	-0.4082353	1610

Table A3

Regression output for WG production measure.

term	estimate	std.error	statistic	p.value	conf.low	conf.high	df
Intercept	29.7706296	1.3315210	22.3583630	0.0000000	27.158933	32.382326	1610
Age	7.5986384	0.4978017	15.2643875	0.0000000	6.622231	8.575046	1610
Caregiver education: Some college	5.6401603	4.9192167	1.1465566	0.2517353	-4.008581	15.288901	1610
Caregiver education: High school or less	20.4554991	7.6929658	2.6589874	0.0079153	5.366220	35.544779	1610
Age * Caregiver education: Some college	-1.3565558	1.3274944	-1.0218919	0.3069856	-3.960355	1.247243	1610
Age * Caregiver education: High school or less	-0.1206958	2.0946793	-0.0576202	0.9540583	-4.229280	3.987889	1610