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 Rowland4,6, Michael C. Frank5, & Virginia A. Marchman5

5 1 Princeton University

6 2 University of Pennsylvania

7 3 Massachusetts Institute of Technology

8 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

112 example, Kristoffersen et al. (2013) collected a large normative sample of Norwegian CDIs

113 using a custom online platform. Similarly, the Slovak adaptation of the CDI uses an online

114 administration format (Kapalková & Slanèová, 2007). And many groups have used general

115 purpose survey software such as Qualtrics and Survey Monkey to administer CDIs and

116 variants online (e.g., Caselli, Lieberman, & Pyers, 2020). The innovation of Web-CDI is to

117 provide a comprehensive researcher management interface for the administration of a wide

118 range of CDI forms, allowing researchers to manage longitudinal administrations, download

119 scores, and share links easily, all while satisfying strong guarantees regarding privacy and

120 anonymity. Moreover, a key benefit of a unified data collection and storage system such as

121 Web-CDI is that data from disparate sources are combined into a single repository. This

122 substantially reduces the overhead efforts associated with bringing together data collected

123 by researchers across the world and allows for the analysis of large comparative datasets

124 with the power to detect general trends in vocabulary development that may emerge across

125 languages. Finally, due to an agreement between the CDI Advisory Board and Brookes

126 Publishing, the publisher of the print versions of the CDI suite, Web-CDI is free of charge

127 for those researchers who agree to contribute their data for the renorming of the long form

128 instruments.

# 129 Introducing Web-CDI

130 Web-CDI is a web-based platform for CDI administration and management.

131 Web-CDI allows researchers to communicate with families by sharing URLs (web links that

132 contain individual users’ own administration of the Web-CDI) via email or social media,

133 facilitating access to families in areas distant from an academic institution and eliminating

134 costly mailings and laboratory visits. Web-CDI also standardizes electronic administration

135 and scoring of CDI forms across labs and institutions, making possible the aggregation of

136 CDI data for later reuse and comparison across administrations by different labs. Indeed,

137 users of Web-CDI grant the CDI Advisory Board permission to access and analyze the

138 resulting data on an opt-out basis, providing a path towards continual improvement of CDI

139 instruments. Since 2018, more than 3,500 CDIs have been collected by 15 research groups

140 throughout the U.S. who are using Web-CDI, demonstrating the potential for large-scale

141 data collection and aggregation.

142 Below, we outline how Web-CDI is used. We begin by detailing the consent obtention

143 process and participant experience. Second, we describe the interface that researchers use

144 to collect data using Web-CDI, specifying a number of common use cases for the platform.

# 145 Participant interface

146 Participants can complete the Web-CDI on a variety of devices, including personal

147 computers and tablets. Web-CDI can be administered on a smartphone, although the

148 experience is not as ideal for the user due to the length of the survey. As Web-CDI moves

149 in the future to incorporate more short forms and computer adaptive tests (CATs) formats

150 (e.g., Chai, Lo, & Mayor, 2020; Makransky, Dale, Havmose, & Bleses, 2016; Mayor &

151 Mani, 2019), smartphone-responsive design will become a priority.

152 When a participant clicks a URL shared by a researcher, they are directed to a

153 website displaying their own personal administration of the Web-CDI. In some cases, they

154 may be asked to read and accept a waiver of consent documentation, depending on

155 whether the researcher has chosen to use that feature (see also Researcher Interface below).

156 *Instructions.* After completing the first demographics page, participants are provided

157 with detailed instructions that are appropriate for either the Words & Gestures or Words

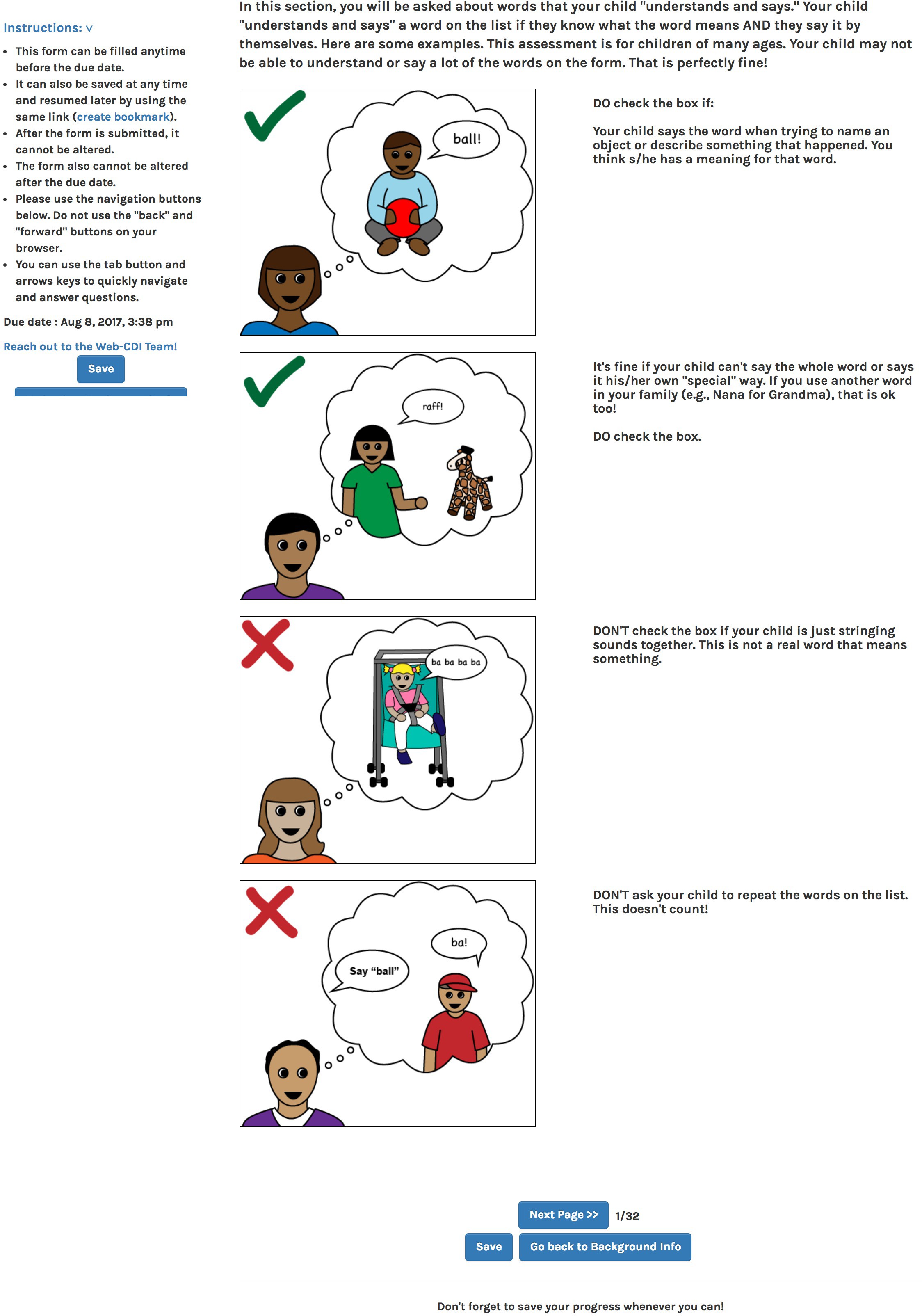
158 & Sentences version (see Figure [1](#_bookmark0)). In addition, there are more detailed instructions for

159 completing the vocabulary checklist. Unlike the traditional paper versions, instructions on

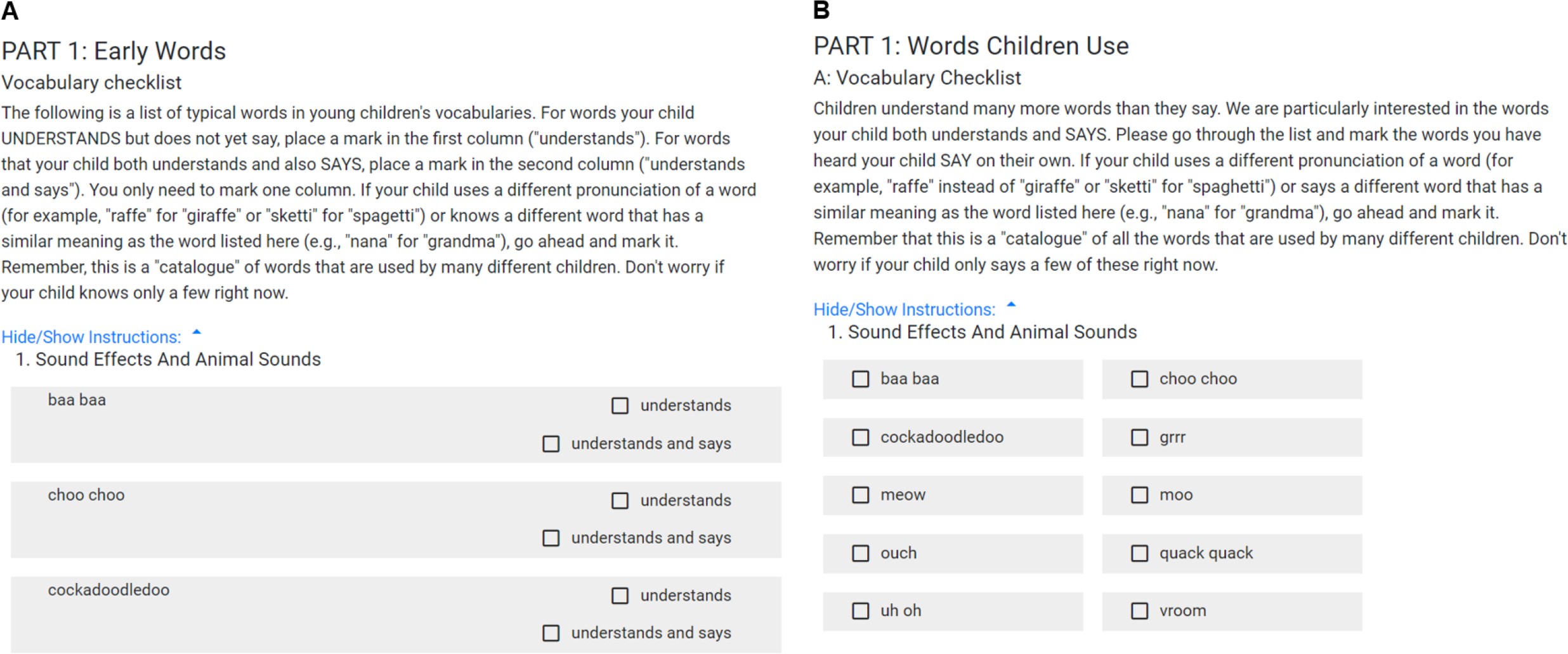
160 how to properly choose responses are provided both in written and pictorial form. The

161 pictorial instructions (Figure [1](#_bookmark0)) aim to further increase caregivers’ understanding of how to

162 complete the checklist. For example, these instructions clarify that the child’s



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



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

163 understanding of a word requires them to have some understanding of the object that the

164 word refers to or some aspect of the word’s meaning. In addition, caregivers are reassured

165 that “child-like” forms (e.g., “raff” for “giraffe”) or family- or dialect-specific forms (e.g.,

166 “nana” for “grandma”) are acceptable. Lastly, caregivers are reminded that the child

167 should be able to produce the words “on their own” and that imitations are not acceptable.

168 These general “rules of thumb” for completing the form should be familiar to researchers

169 who are distributing the forms to caregivers so they can field any questions that may arise.

170 While this is not possible for certain use-cases (e.g., social media recruitment), these

171 instructions should ideally also be reviewed either in writing (e.g., via email) or verbally

172 (e.g., over the phone), so that these pictured instructions serve merely as a reminder to

173 caregivers when completing the form. Pictured instructions are available for download on

174 the MB-CDI website at <http://mb-cdi.stanford.edu/about.html>.

175 *Completing the instrument.* The majority of the participant’s time is spent

176 completing the main sections of the instruments. As shown in Figure [2](#_bookmark1), on the American

177 English Words and Gestures form, the vocabulary checklist portion (396 items) asks

178 caregivers to indicate whether their child can “understand” or “understand and say” each

179 word; they can also indicate that their child neither understands nor says the word by

180 checking neither box. Additionally, gesture communication and other early milestones are

181 assessed. In the American English Words and Sentences form, the vocabulary checklist

182 (680 items) only asks caregivers to indicate which words their child “says.” Additional

183 items assess children’s production of their three longest sentences, as well as morphological

184 and syntactic development more broadly. All of these items are broken up across multiple

185 screens for easier navigation through the form.

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

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

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

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

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

191 (2021)) word that the child currently understands or produces. This feedback to caregivers

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

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

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

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

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

# 197 Researcher interface

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

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

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

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

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

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

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

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

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

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

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

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

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

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

212 ones we describe below in the *Dataset 1* section. Additionally, date of form completion will

213 be preserved when adding Web-CDI data into Wordbank, so that researchers can choose to

214 filter out data that may be affected by the particular point in time at which they were

215 collected (for example, the COVID-19 pandemic, Kartushina et al., 2021).

216 A study in the context of the Web-CDI system is a set of individual administrations

217 created by a researcher that share certain specifications. Table A1 in the Appendix gives

218 an overview of the customizable features that are available at the study level in Web-CDI.

219 These features are set when creating a study using the “Create Study” tool, and most of

220 the features can be updated continuously during data collection using the “Update Study”

221 tool. While some of these features are only particularly relevant to specific use cases (e.g.,

222 longitudinal research and social media data collection, described below), others are relevant

223 to all researchers using Web-CDI.

224 There are currently several CDI forms available for distribution on Web-CDI,

225 including multiple versions of the English WG and WS forms and forms in other languages

226 (see Cross-linguistic research, below). When creating a study, researchers choose one of the

227 forms that they would like to distribute to participants; only one can be used in a given

228 study. Researchers who wish to send multiple forms to participants simultaneously (e.g.,

229 those conducting multilingual research) should create multiple studies, each with a single

230 instrument associated with it.

231 Researchers can download participant data in two formats. Both formatting options

232 output a comma-separated values file with one row per participant; the full data option

233 includes participant-by-item responses, and allows researchers to explore item-level trends,

234 while the summary data option omits item-level data and only provides summary scores

235 and normative information, including total number of words understood/produced and

236 percentile scores by age in months and gender. Percentile scores based are calculated to a

237 single percentile resolution using norms from Fenson et al. (2007).

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

239 may facilitate them from a researcher’s perspective.

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

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

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

243 study URL, using the “Add Participants” tool in the researcher dashboard. New

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

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

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

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

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

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

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

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

252 each participant completes multiple administrations. Researchers wishing to design

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

254 “Add Participants” tool in the researcher dashboard. If a certain participant ID is added

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

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

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

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

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

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

261 the instrument in longitudinal administrations as caregivers should complete the

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

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

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

265 *Social media and survey vendors.* Web-CDI contains several features designed to

266 facilitate data collection from social media recruitment or through third-party

267 crowd-sourcing applications and vendors (e.g., Amazon Mechanical Turk, Prolific). First,

268 rather than creating unique survey links for each participant, researchers can also use a

269 single, anonymous link. When a participant clicks the anonymous link, a new

270 administration with a unique subject ID is created in the study dashboard. Additionally,

271 Web-CDI studies have several customizable features that are geared towards anonymous

272 online data collection. For example, researchers can adjust the minimum amount of time a

273 participant must take to fill out the survey before they are able to submit; with a longer

274 minimum time to completion, researchers can encourage a more thorough completion of the

275 survey. This feature is typically only relevant in research designs in which participants are

276 not vetted by the researcher or those in which there is no direct communication between

277 participants and researchers, as might be the case when recruiting respondents on social

278 media. Responses collected via personal communication with participants show low rates of

279 too-fast responding, mostly removing the need for the minimum time feature. Even in the

280 case of anonymous data collection, however, it is recommended that researchers not raise

281 the minimum completion time higher than 6 minutes, since some caregivers of very young

282 children may theoretically be able to proceed through the measure quickly if their child is

283 not yet verbal. Aside from the minimum time feature, researchers can ask participants to

284 verify that their information is accurate by checking a box at the end of the survey, and

285 can opt to include certain demographic questions at both the beginning and end of the

286 survey, using response consistency on these redundant items as a check of data quality.

287 *Paid participation.* If researchers choose to compensate participants directly through

288 the Web-CDI interface, Web-CDI has built-in functionality to distribute redeemable gift

289 codes when a participant reaches the end of the survey. Web-CDI contains several features

290 to facilitate integration with third-party crowdsourcing applications and survey vendors

291 should they choose to handle participant compensation through another platform. For

292 example, when creating studies, researchers can enter a URL to redirect participants to

293 when they reach the end of the survey. Researchers using the behavioral research platform

294 Prolific can configure their study to collect participants’ unique Prolific IDs and pre-fill

295 them in the survey.

296 *Cross-linguistic research.* Web-CDI forms are currently available in English (U.S.

297 American and Canadian), Spanish, French (Quebecois), Hebrew, Dutch and Korean. We

298 are looking to add more language forms to the tool, as the paper version of the forms has

299 been adapted into more than 100 different languages and dialects, and further ongoing

300 adaptations have been approved by the MB-CDI board

301 (<http://mb-cdi.stanford.edu/adaptations>).

302 **System Design**

303 Web-CDI is constructed using open-source software. All of the vocabulary data

304 collected in Web-CDI are stored in a standard MySQL relational database, managed using

305 Django and Python and hosted either by Amazon Web Services or by a European Union

306 (GDPR) compliant server (see below). Individual researchers can download data from their

307 studies through the researcher interface, and Web-CDI administrators have access to the

308 entire aggregate set of data from all studies run with Web-CDI. Website code is available in

309 a GitHub repository at <https://github.com/langcog/web-cdi>, where interested users can

310 browse, make contributions, and request technical fixes.

# 311 Data Privacy and GDPR Compliance

312 Web-CDI is designed to be compliant with stringent human subjects privacy

313 protections across the world. First, for U.S. users, we have designed Web-CDI based on the

314 United States Department of Health and Human Services “Safe Harbor” Standard for

315 collecting protected health information as defined by the Health Insurance Portability and

316 Accountability Act (HIPAA). In particular, participant names are never collected, birth

317 dates are used to calculate age in months (with no decimal information) but never stored,

318 and geographic zip codes are trimmed to the first 3 digits. Because of the architecture of

319 the site, even though participants enter zip codes and dates of birth, these are never

320 transmitted in full to the Web-CDI server. Since no identifying information is being

321 collected by the Web-CDI system, this feature ensures that Web-CDI can be used by

322 United States labs without a separate Institutional Review Board agreement between

323 users’ labs and Web-CDI (though of course researchers using the site will need Institutional

324 Review Board approval of their own research projects).[1](#_bookmark2)

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

325 In the European Union (EU), research data collection and storage is governed by the

326 Generalized Data Protection Regulation (GDPR) and its local instantiation in the legal

327 system of the member states. Some of the questions on the demographic form contain

328 information that may be considered sensitive (e.g., information about children’s

329 developmental disorders), and in some cases, the possibility of linking this sensitive

330 information to participant IDs exists, particularly when researchers draw on local databases

331 that contain full names and addresses for recruitment and contacting. As a result, issues

332 regarding GDPR compliance arise when transferring data outside the EU, namely to

333 Amazon Web Services servers housed in the United States. Following GDPR regulations,

334 these issues would make a data sharing agreement between data collectors and Amazon

335 Web Services necessary. In addition, all administrators who can access the collected data

336 would have to enter such an agreement, which needs updating whenever personnel changes

337 occur. To overcome these hurdles, and in consultation with data protection oﬀicers, we

338 opted to leverage the local technical expertise and infrastructure to set up a sister site

339 housed on GDPR-compliant servers, currently available at [http://webcdi.mpi.nl](http://webcdi.mpi.nl/). This site

340 is updated synchronously with the main Web-CDI website to ensure a consistent user

341 experience and access to the latest features and improvements. This site has been used in

342 135 successful administrations so far and is the main data collection tool for an ongoing

343 norming study in the Netherlands. We are further actively advertising the option to use

344 the European site to other labs who are following GDPR guidelines and are planning

345 adaptations to multiple European languages, where copyright allows.

# 346 Current data collection

347 We now turn to an overview of the data collected thus far using Web-CDI. First, we

348 examine the full sample of all of the Web-CDI administrations collected as of autumn 2020

re-identification in some cases and if potential users have questions, we encourage them to consult with an Institutional Review Board.

349 (Dataset 1); we then focus in on a specific subset of Dataset 1 which is comprised of data

350 from recent efforts to oversample non-white, less highly-educated U.S. participants

351 (Dataset 2). Across both datasets, we show that general trends from prior research on

352 vocabulary development are replicated using Web-CDI, and we discuss the potential for

353 using Web-CDI to collect vocabulary development data from diverse communities online.

# 354 Dataset 1: Full Current Web-CDI Usage

Table 1

*Exclusions from Dataset 1: full Web-CDI sample*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Exclusion** | **WG**  **exclusions** | **% of full WG sample excluded** | **WS**  **exclusions** | **% of full WS sample 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% |

355 In this section, we provide some preliminary analyses of Dataset 1, which consists of

356 the full sample of American English Web-CDI administrations collected before autumn

357 2020. At time of writing, researchers from 15 universities in the United States have

358 collected over 5,000 administrations of the American English CDI using Web-CDI since it

359 was launched in late 2017, with 2,868 administrations of the WG form before exclusions

360 and 3,594 administrations of the WS form before exclusions. We excluded participants from

361 the subsequent analyses based on a set of stringent criteria intended for the creation of

362 future normative datasets. We excluded participants if it was not their first administration

363 of the survey; if they were born prematurely or had a birthweight under 5.5 lbs (< 2.5 kg);

364 reported more than 16 hours of exposure to a language other than English per week on

365 average (amounting to > 10% exposure to English); had serious vision impairments,

366 hearing deficits or other developmental disorders or medical issues[2](#_bookmark3); were outside of the

367 correct age range for the survey; or spent less time on the survey than a pre-specified

368 timing cutoff. Timing cutoffs were determined by selecting two studies within Dataset 1

369 that, upon a visual inspection, appeared to contain high-quality responses (i.e., did not

370 contain a disproportionate number of extremely quick responders), and using these to

371 estimate the 5th percentile of completion time by the child’s age in months with a quantile

372 regression. Thus, for each age on the WG and WS measures, we obtained an estimate of

373 the 5th percentile of completion time and used this estimate as the shortest amount of time

374 participants could spend on the Web-CDI without being excluded from our analyses here.

375 The exclusion criteria we used were designed to be generally comparable with those

376 used in Fenson et al. (2007), who adopted stringent criteria to establish vocabulary norms

377 that reflect typically developing children’s vocabulary trajectories. A complete breakdown

378 of the number of participants excluded on each criterion is in Table 1. Of the completed

379 WG forms, 1,248 were excluded, leading to a final WG sample size of 1,620 administrations,

380 and 1,694 WS administrations were excluded, leading to a final WS sample size of 1,900.

381 **Demographic distribution and exclusions.** Figure [3](#_bookmark4) shows the distribution of

382 participant ethnicities in Dataset 1 as compared with previously reported numbers in a

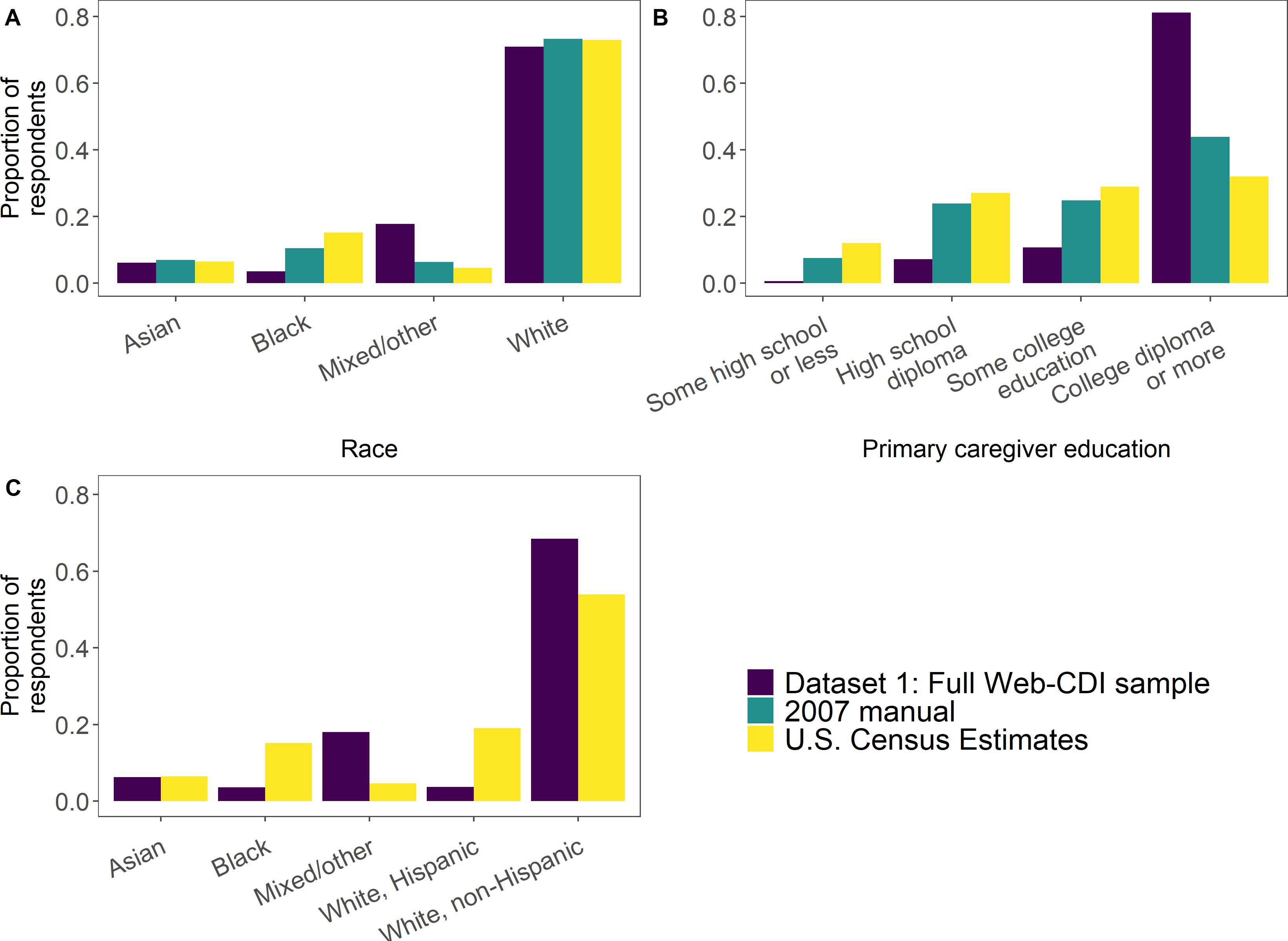
383 large scale norming study of the paper-based CDI form by Fenson et al. (2007). Several

384 issues pertaining to sample representativeness are appreciable. First, as shown in Figure

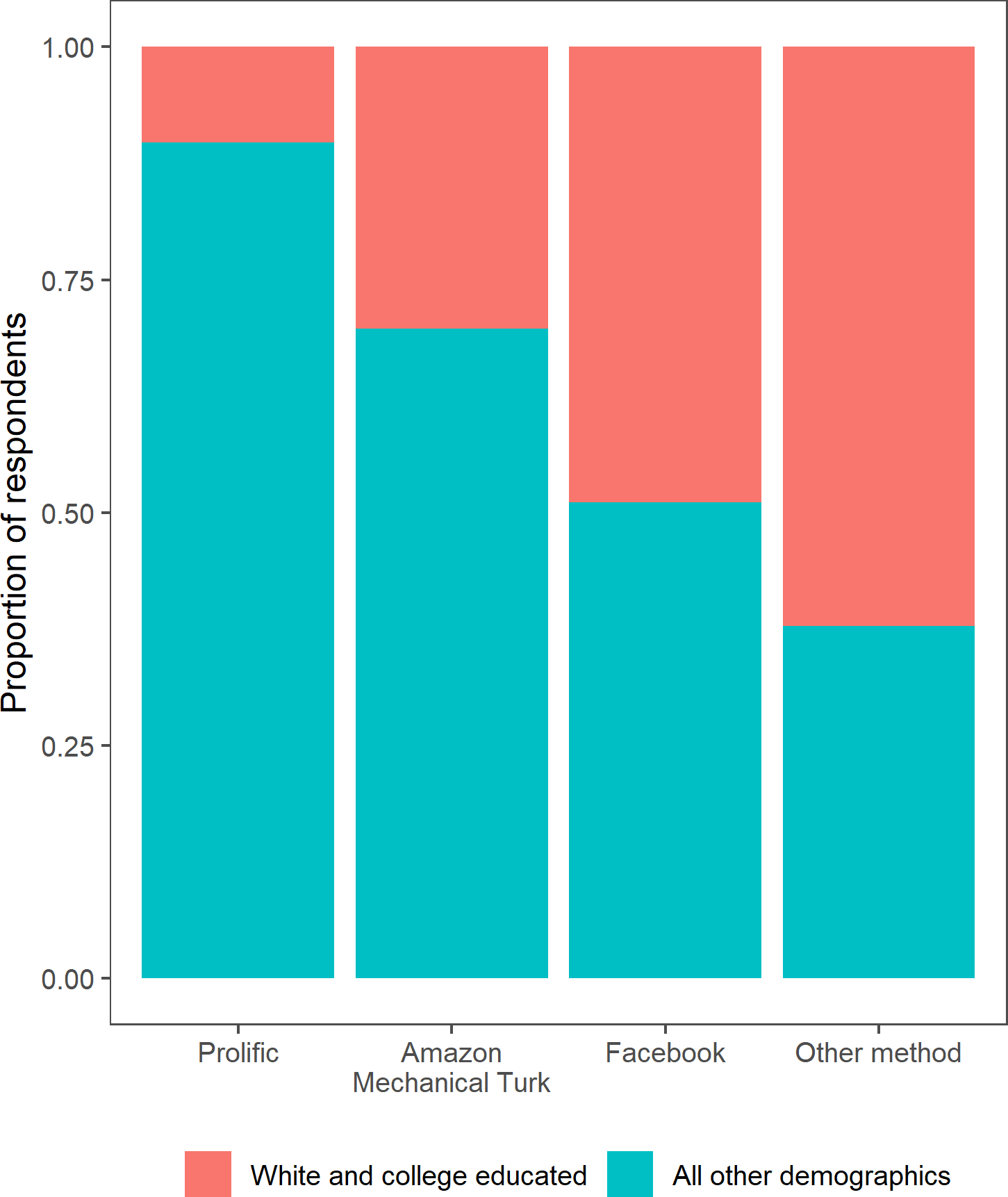
385 [3](#_bookmark4)A, white participants comprised nearly three quarters of Dataset 1, which is comparable

386 to U.S Census estimates in 2019 of U.S. residents between the ages of 15 and 34 in 2019;

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.



*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). Bot- tom 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.

387 however, Figure [3](#_bookmark4)C shows that, compared with U.S. Census estimates, many more white

388 participants in Dataset 1 were non-Hispanic than is true of the U.S. population in general,

389 indicating that Web-CDI is significantly oversampling white, non-Hispanic individuals (the

390 breakdown of white participants into Hispanic and non-Hispanic is not reported in the

391 2007 norms). Moreover, few participants identified as Hispanic/Latinx: 6.4% of WG

392 participants and 5.2% of WS participants reported Hispanic or Latinx heritage. The low

393 percentage of Hispanic/Latinx participants was due in part to our exclusion of children

394 with substantial exposure to languages other than English: before exclusions, 8.4% of WG

395 participants were Hispanic/Latinx, and 8.1% of WS participants were Hispanic/Latinx.

396 Finally, representation of Black participants is generally lower in Dataset 1 (3.5%) than in

397 the 2007 norms (10.5%), which is in turn lower U.S. Census estimates (15.2%). This

398 indicates that both Web-CDI data and existing norming samples tend to underrepresent

399 Black participants.

400 Participants’ educational attainment level, as measured by the primary caregiver’s

401 highest educational level reached[3](#_bookmark6), was similarly skewed. In Dataset 1, 81.2% of responses

402 came from families with college-educated primary caregivers compared to 43.8% from the

403 same group in the 2007 norms and 32.0% (Figure [3](#_bookmark4)). Furthermore, less than 1% of

404 participants report a primary caregiver education level less than a high school degree,

405 compared to 7% from the same group in the 2007 norms. The overrepresentation of white,

406 non-Hispanic Americans and those with high levels of education attainment points to a

407 general challenge encountered in vocabulary development research, which we return to

408 when we detail our efforts to recruit more diverse participants. Figure [4](#_bookmark5) shows that, of the

409 recruitment methods used in Dataset 1, the studies conducted using the platform Prolific

410 (which we detail in the *Dataset 2* section) contributed the least to the high proportion of

411 white, non-Hispanic, college educated participants. Respondents not known to be recruited

412 through an online channel or crowdsourcing platform (labeled “Other method” in Figure [4](#_bookmark5))

413 showed the most overrepresentation of white, college educated participants, suggesting that

414 reliance on university convenience samples may be driving the demographic skewness of

415 Dataset 1 most acutely.

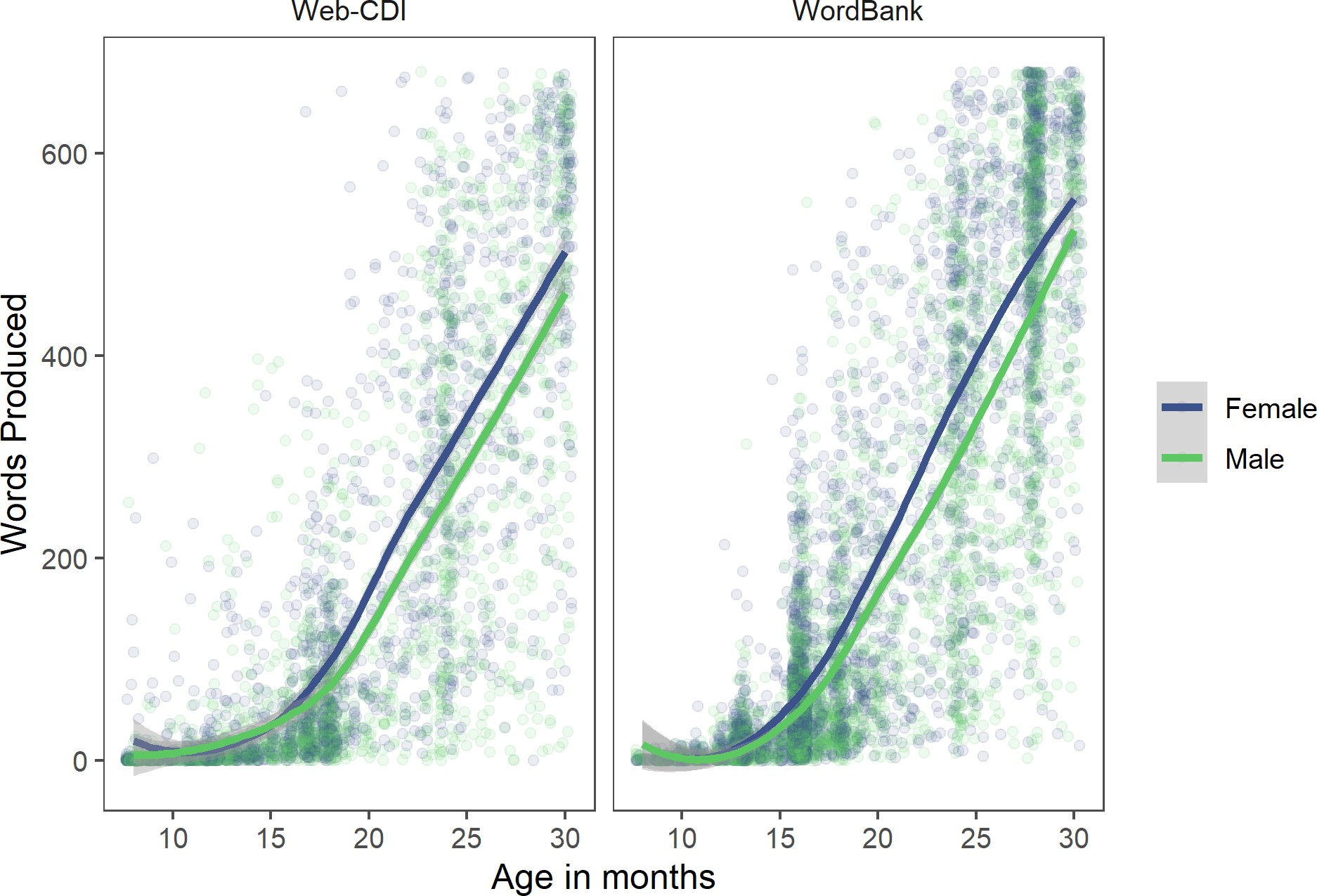
416 **Results: Dataset 1.** Although the CDI instruments include survey items intended

417 to measure constructs other than vocabulary size, such as gesture, sentence production and

418 grammar, we focus exclusively on the vocabulary measures here. We also visualize key

419 analyses from Dataset 1 alongside the analogous analyses on the American English CDI

3 Maternal education level is a common measure of family socioeconomic status; 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, but we expect that in the vast majority of cases this corresponds to 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.

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

administrations from the WordBank repository (Frank, Braginsky, Yurovsky, & Marchman, 2021) that include the relevant demographic information needed to provide a comparison dataset of traditional paper-and-pencil forms. Across both the WG and WS measures, Dataset 1 shows greater reported vocabulary comprehension and production for older children. Moreover, data from both the WG and WS measures in Dataset 1 replicate a subtle but reliable pattern such that female children tend to have slightly larger vocabulary scores than male children across the period of childhood assessed in the CDI forms (Frank, Braginsky, Yurovsky, & Marchman, 2021), though in these data this difference does not appear until around 18 months (Figure [5](#_bookmark7)).

On the WG form, respondents’ reports of children’s vocabulary comprehension and production both increased with children’s age (Figure [6](#_bookmark9)). We replicate overall patterns found by Feldman et al. (2000) in that, on both the “Words Understood” and “Words Produced” measures, vocabulary scores were slightly negatively correlated with primary caregivers’ education level, such that those caregivers without any college education reported higher vocabulary scores on both scales. A linear regression model with robust standard errors predicting comprehension scores with children’s age and primary caregivers’ education level (binned into categories of “High school diploma or less,” “Some college education” and “College diploma or more”[4](#_bookmark8)) as predictors shows main effects of both age

𝛽 𝑝 𝛽 𝑝

( = 20.05, < 0.001) and caregiver primary education ( ℎ𝑖𝑔ℎ𝑠𝑐ℎ𝑜𝑜𝑙 = 21.86, = 0.05).

439 Similarly, a linear regression model with robust standard errors predicting production

441 (𝛽 𝑝 < 0.001) and caregiver primary education (𝛽ℎ𝑖𝑔ℎ𝑠𝑐ℎ𝑜𝑜𝑙 = 20.46, 𝑝 = 0.008).

440 scores by children’s age and primary caregivers’ education level shows main effects of age

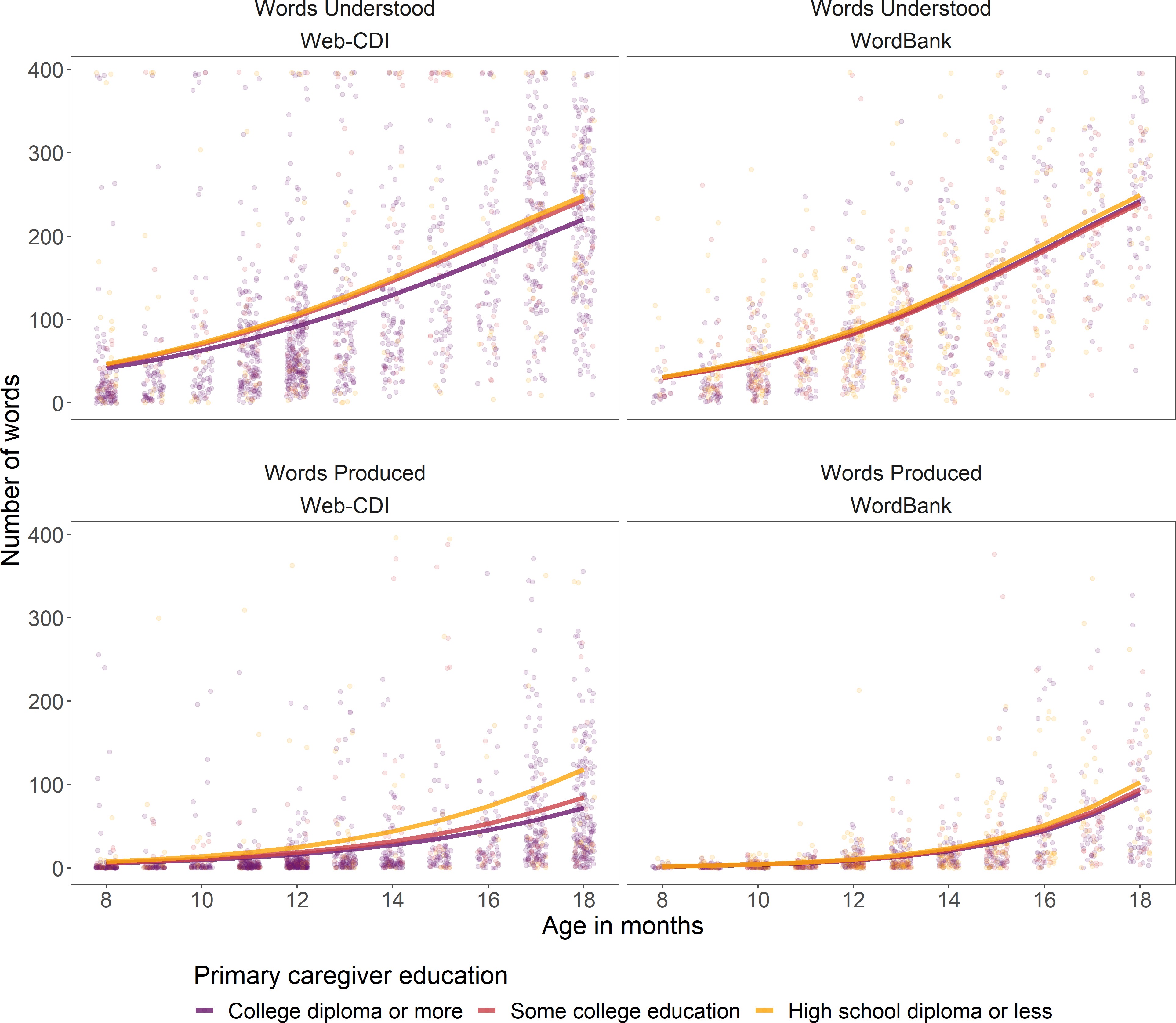
= 7.60,

442 These analyses were not preregistered, but generally follow the analytic strategy in Frank,

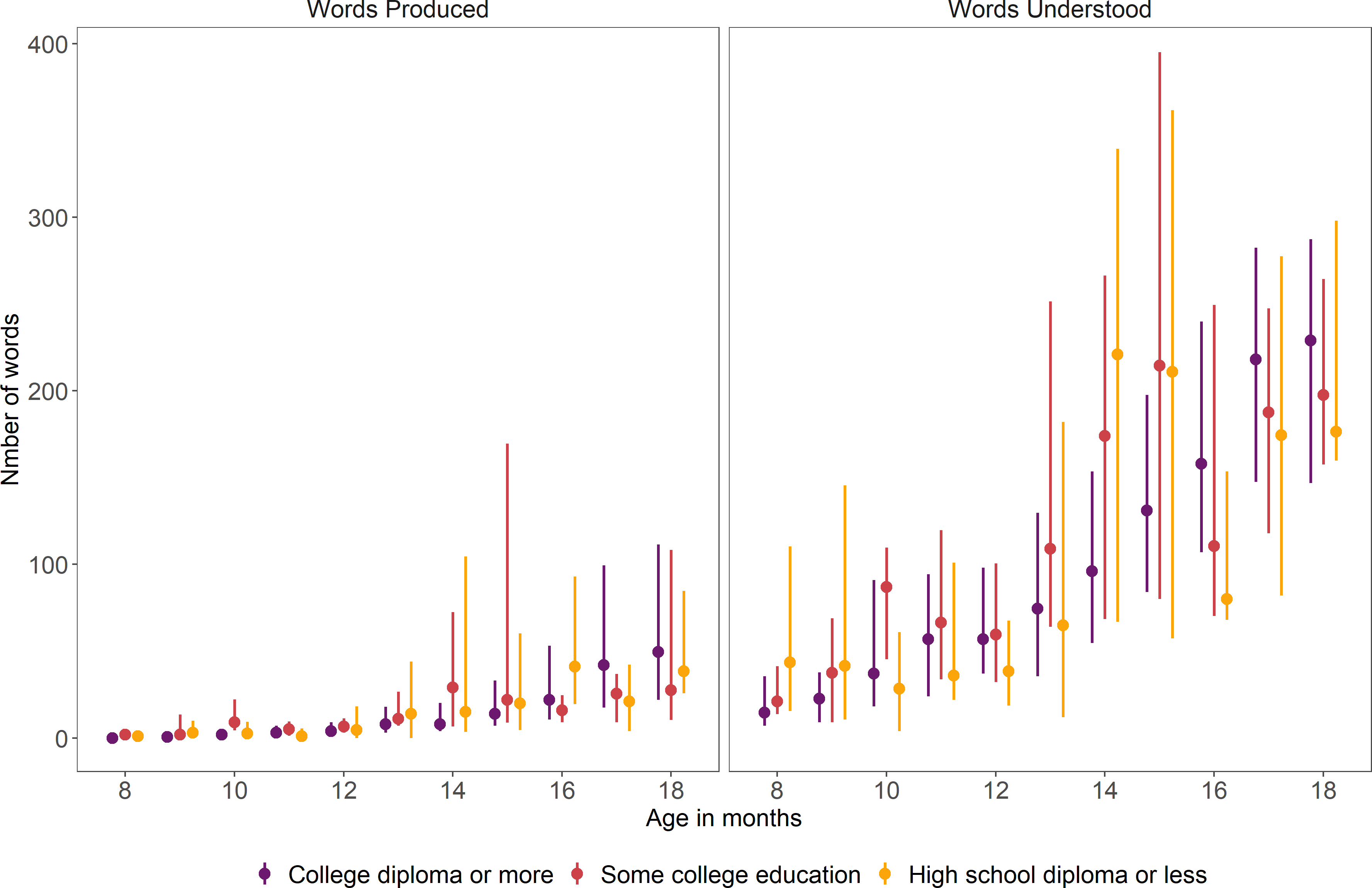
443 Braginsky, Yurovsky, and Marchman (2021); additionally, we fit linear models with robust

444 standard errors to account for heteroskedasticity in the data (Astivia & Zumbo, 2019).

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., 2021) 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 attain- ment on the WG form. Lines indicate span between first and third quartiles for each age.

445 Generalized linear model predictions for Web-CDI shown in Figure [6](#_bookmark9) differ somewhat from

446 those for WordBank; prediction curves for caregivers of different education attainment

447 levels diverge slightly more in the Web-CDI sample than in the WordBank sample.

448 The pattern of results seen in the WG subsample of Dataset 1 is consistent with prior

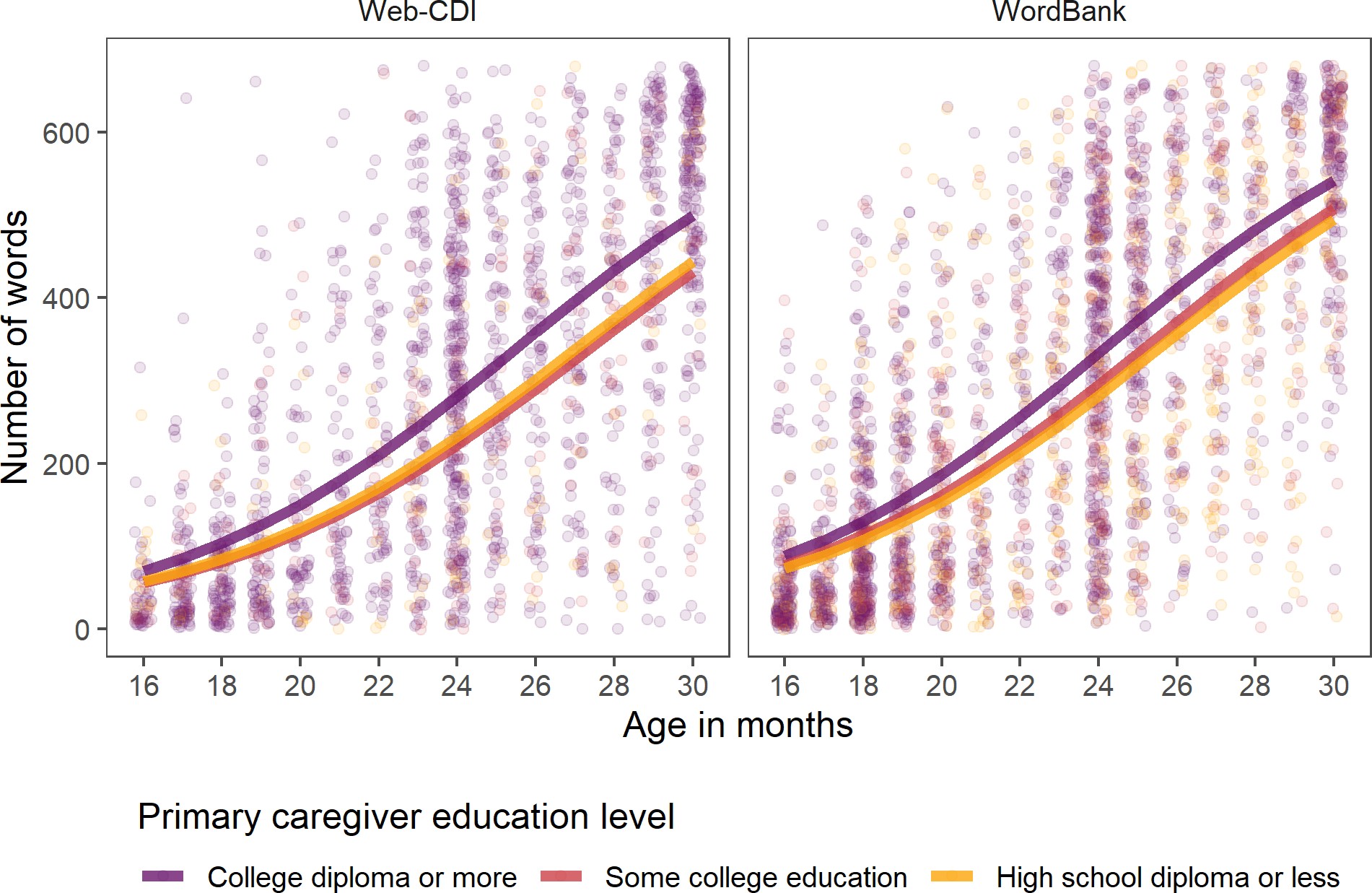
449 findings indicating that respondents with lower levels of education attainment report

450 higher vocabulary comprehension and production on the CDI-WG form (Feldman et al.,

451 2000; Fenson et al., 1994). Although caregivers with lower levels of education attainment

452 report higher mean levels of vocabulary production and comprehension, median vocabulary

453 scores (which are more robust to outliers) show no clear pattern of difference across



*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, left panel) and in the WordBank repository (N

= 2,776, right panel). Curves show generalized linear model fits.

454 primary caregiver education levels (Figure [7](#_bookmark10)). This discrepancy between the regression

455 effects and a group-median analysis suggests that the regression effects described

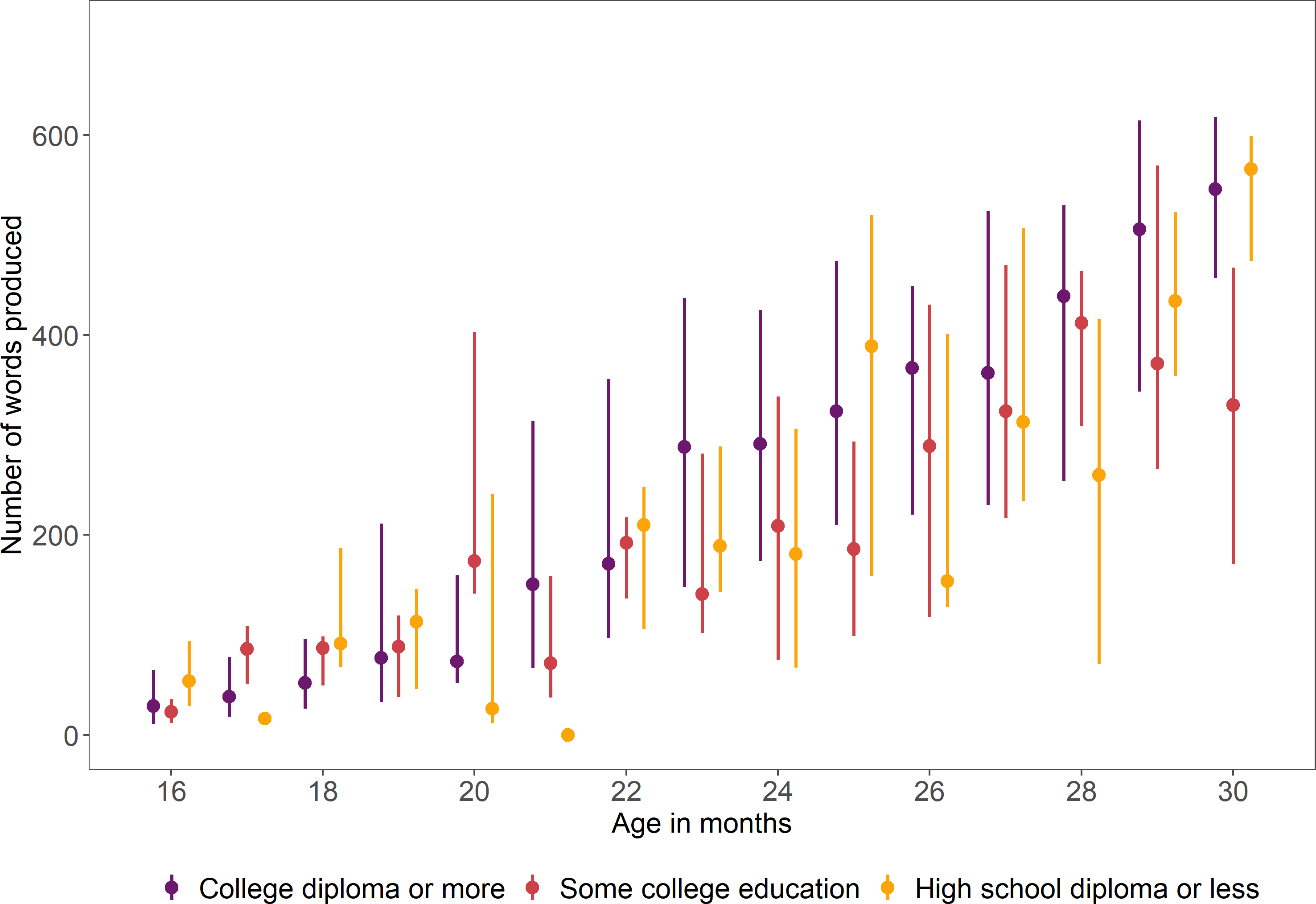
456 previously are driven in part by differential interpretation of the survey items, such that a

457 few caregivers with lower levels of education attainment are more liberal in reporting their

458 children’s production and comprehension vocabulary scores, especially for the youngest

459 children, driving up the mean scores for this demographic group.

460 Vocabulary production scores on the WS form show the expected pattern of increase



*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.

461 with children’s age in months; in addition, scores replicate the trend reported in Feldman

462 et al. (2000) and Frank, Braginsky, Yurovsky, and Marchman (2021) such that primary

463 caregiver education is positively associated with children’s reported vocabulary size (Figure

464 [8](#_bookmark11)). Because representation of caregivers without a high school diploma is scarce (N = 6 out

465 of a sample of 1,900), interpretation of the data from this group is constrained.

466 Nevertheless, as shown in Figure [8](#_bookmark11), a small but clear positive association between primary

467 caregiver education and vocabulary score exists such that college-educated caregivers

468 report higher vocabulary scores than those of any other education level. Notably, this

469 association is not the result of outliers and is still appreciable in median scores (Figure [9](#_bookmark12)),

470 unlike the data from the WG measure shown in Figure [7](#_bookmark10). The implications from these data

471 converge with previous findings which indicate that parental education levels, often used as

472 a metric of a family’s socioeconomic status, are related to children’s vocabulary size

473 through early childhood.

474 **Discussion: Dataset 1.** In general, the full sample of Web-CDI data after

475 exclusions (Dataset 1) replicates previous norming datasets used with the standard

476 paper-and-pencil form of the MB-CDI. We find that vocabulary scores grow with age and

477 that females hold a slight advantage over males in early vocabulary development.

478 Moreover, Dataset 1 replicates a previously documented relationship between primary

479 caregiver education level and vocabulary scores: on the WG form, primary caregiver

480 education shows a slight negative association with vocabulary scores, whereas the trend is

481 reversed in the WS form. Taken together, these data illustrate that Web-CDI and the

482 standard paper-and-pencil form of the CDI give similar results, and thus that Web-CDI

483 can be used as a valid alternative to the paper format.

484 The data discussed above have stemmed from efforts by many researchers across the

485 United States whose motivations for using the Web-CDI vary. As a result, they reproduce

486 many of the biases of standard U.S. convenience samples. In the next section, we describe

487 in more detail our recent efforts to use the Web-CDI to collect vocabulary development

488 data from traditionally underrepresented participant populations in the United States,

489 attempting to counteract these trends.

# 490 Dataset 2: Using Web-CDI to Collect Data from Diverse U.S.-based

491 **Communities**

492 Despite the large sample sizes we collected in the previous section, Dataset 1 is, if

493 anything, even more biased towards highly-educated and white families than previous

494 datasets collected using the paper-and-pencil form. How can we recruit more diverse

495 samples to remedy this issue? Here, we discuss and analyze Dataset 2, which consists of

496 those administrations from Dataset 1 which were part of recent data-collection efforts

497 (within the past year and a half) that were specifically aimed towards exploring the use of

498 online recruitment as a potential way to collect more diverse participant samples than are

499 typical in the literature. In other words, the following data from Dataset 2 were included in

500 the previous discussion and analysis of Dataset 1, but we examine them separately here to

501 give special attention to the issue of collecting diverse samples online.

502 While understanding that the performance of standard measurement tools like the

503 CDI among multilinguals is of immense import to the field of vocabulary development

504 research [Gonzalez et al., in prep; Floccia et al. (2018); De Houwer (2019)], we focused in

505 Dataset 2 only on vocabulary development in monolingual children, because collecting data

506 from multilingual populations introduces additional methodological considerations (e.g.,

507 how to measure exposures in each language) that are not the focus of our work here.

508 However, it will be imperative in future to collect large-scale datasets of vocabulary data in

509 bilingual children, both to better calibrate standard tools such as the CDI, as well as to

510 reduce the bias towards monolingual families in the existing literature on measuring

511 vocabulary development.

512 **Online data collection.** Online recruitment methods, such as finding participants

513 on platforms such as Amazon Mechanical Turk, Facebook and Prolific, represent one

514 possible route towards assembling a large, diverse sample to take the Web-CDI. These

515 methods allow researchers to depart from their typical geographical recruitment area much

516 more easily than with paper-and-pencil administration. Online recruitment strategies for

517 vocabulary development data collection have been used in the United Kingdom (Alcock,

518 Meints, & Rowland, 2020), but their usage in the U.S. context remains, to our knowledge,

519 rare. In a series of data collection efforts, we used Web-CDI as a tool to explore these

520 different channels of recruitment.

521 Dataset 2 consists of data that were collected in two phases. In the first phase, we

522 ran advertisements on Facebook which were aimed at non-white families based on users’



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

523 geographic locations (e.g., targeting users living in majority-Black cities) or other profile

524 features (e.g., ethnic identification, interest in parenthood-related topics). Advertisements

525 consisted of an image of a child and a caption informing Facebook users of an opportunity

526 to fill out a survey on their child’s language development and receive an Amazon gift card

527 (Figure [10](#_bookmark13)). Upon clicking the advertisement, participants were redirected to a unique

528 administration of the Web-CDI, and they received $5 upon completing the survey. This

529 open-ended approach to recruitment offered several advantages, namely that a wide variety

530 of potential participants from specific demographic backgrounds can be reached on

531 Facebook. However, we also received many incomplete or otherwise unusable survey

532 administrations, either from Facebook users who clicked the link and decide not to

533 participate, or those who completed the survey in an extremely short period of time (over

Table 2

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Exclusion** | **WG**  **exclusions** | **% of full WG sample excluded** | **WS**  **exclusions** | **% of full 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% |

534 half of all completed administrations, Table 2).

535 In the second phase, we used the crowdsourcing survey vendor Prolific

536 ([http://prolific.co](http://prolific.co/)) in the hopes that some of the challenges encountered with Facebook

537 recruitment would be addressed. Prolific allows researchers to create studies and post them

538 to individuals who are in the platform’s participant database, each of whom is assigned a

539 unique alphanumeric “Prolific ID.” Importantly, Prolific maintains detailed demographic

540 information about participants, allowing researchers to specify who they would like to

541 complete their studies. Prolific further has a built-in compensation infrastructure that

542 handles monetary payments to participants, eliminating the need to disburse gift cards

543 through Web-CDI.

544 In the particular case of Web-CDI, the demographic information needed to determine

545 whether an individual was eligible to complete our survey (e.g., has a child in the correct

546 age range, lives in a monolingual household, etc.) was more specific than the information

547 that Prolific collects about their participant base. We therefore used a brief pre-screening

548 questionnaire to generate a list of participants who were eligible to participate, and

549 subsequently advertised the Web-CDI survey to those participants. Given that we were

550 interested only in reaching participants in the United States who were not white or who

551 did not have a college diploma, our data collection efforts only yielded a sample that was

552 small (N = 68) but much more thoroughly screened than that which we could obtain on

553 Facebook.

554 Across both phases (Facebook and Prolific recruitment), we used the same exclusion

555 criteria as in the full Web-CDI sample to screen participants. A complete tally of all

556 excluded participants is shown in Table 2. In both the WG and WS surveys, exclusion

557 rates in Dataset 2 were high, amounting to 58% of participants who completed the survey.

558 The high exclusion rates were notably driven by an accumulation of survey administrations

559 which participants completed more quickly than our time cutoffs allow (Tables A4 and A5).

560 Many of the survey administrations excluded for fast completion had missing demographic

561 information reported: Among WG participants excluded for too-fast completions, 93% did

562 not report ethnicity, and among WS participants excluded for the same reason, 97% did

563 not report ethnicity. Absence of these data prevents us from drawing conclusions about the

564 origin or demographic profile of administrations that were excluded. After exclusions, full

565 sample size in Dataset 2 was N = 128 WG completions and N = 115 WS completions.

566 The results from Dataset 2 show overall similar patterns to the full Web-CDI sample

567 in several regards. Word production scores from both the WG and WS administrations

568 reflect growing productive vocabulary across the second and third years, with a very small

569 gender effect such that female children’s vocabularies are higher across age than males’

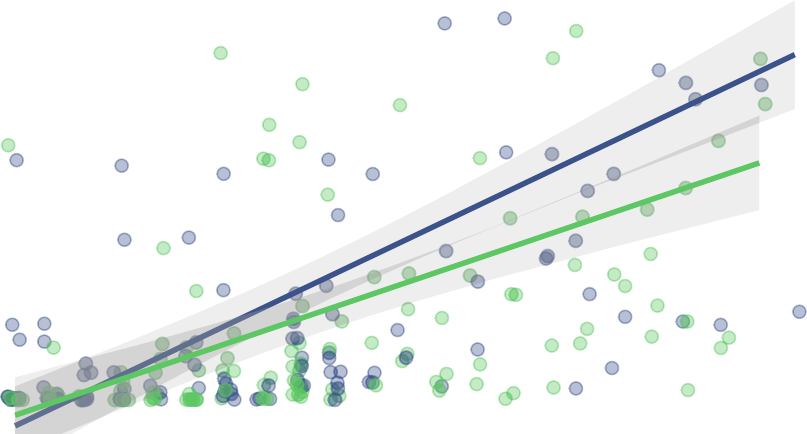
570 (Figure [11](#_bookmark14)). The relationship between caregivers’ reported levels of education and child’s

571 vocabulary score is not as clear as it is in the full Web-CDI sample (Figure [12](#_bookmark15)); however,

572 children of college-educated caregivers reported generally higher vocabulary scores across

573 age than did children of caregivers without any college degree. These patterns suggest that

600



400 Child gender

Total words produced

Female Male

200

0

10 15 20 25 30

Age in months

*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.

574 our data show similar general patterns to other CDI datasets with other populations

575 (Frank, Braginsky, Yurovsky, & Marchman, 2021).

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

577 less highly-educated participants. After exclusions, Dataset 2 has a higher proportion of

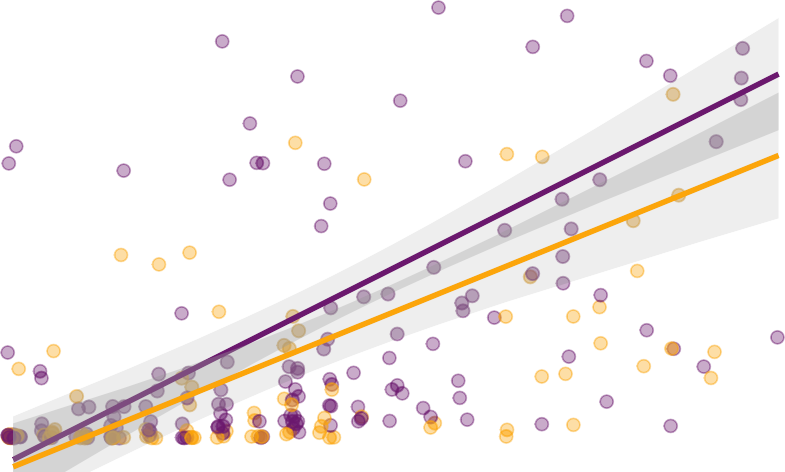
578 non-white participants than Dataset 1 (the overall Web-CDI sample) and the norms

579 established by Fenson et al. (2007) (Figure [13](#_bookmark16)). Black participants in particular showed a

580 marked increase in representation, from 10.5% in the 2007 norms to 30.7% in Dataset 2,

581 while the proportion of white participants decreased from 73.3% in the 2007 norms to

600



400

Words produced

College High school

200

0

10 15 20 25 30

Age in months

*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.

582 50.5% in Dataset 2. Representation on the basis of families’ reported primary caregiver

583 education also improved (Figure [13](#_bookmark16)). Participants with only a high school diploma

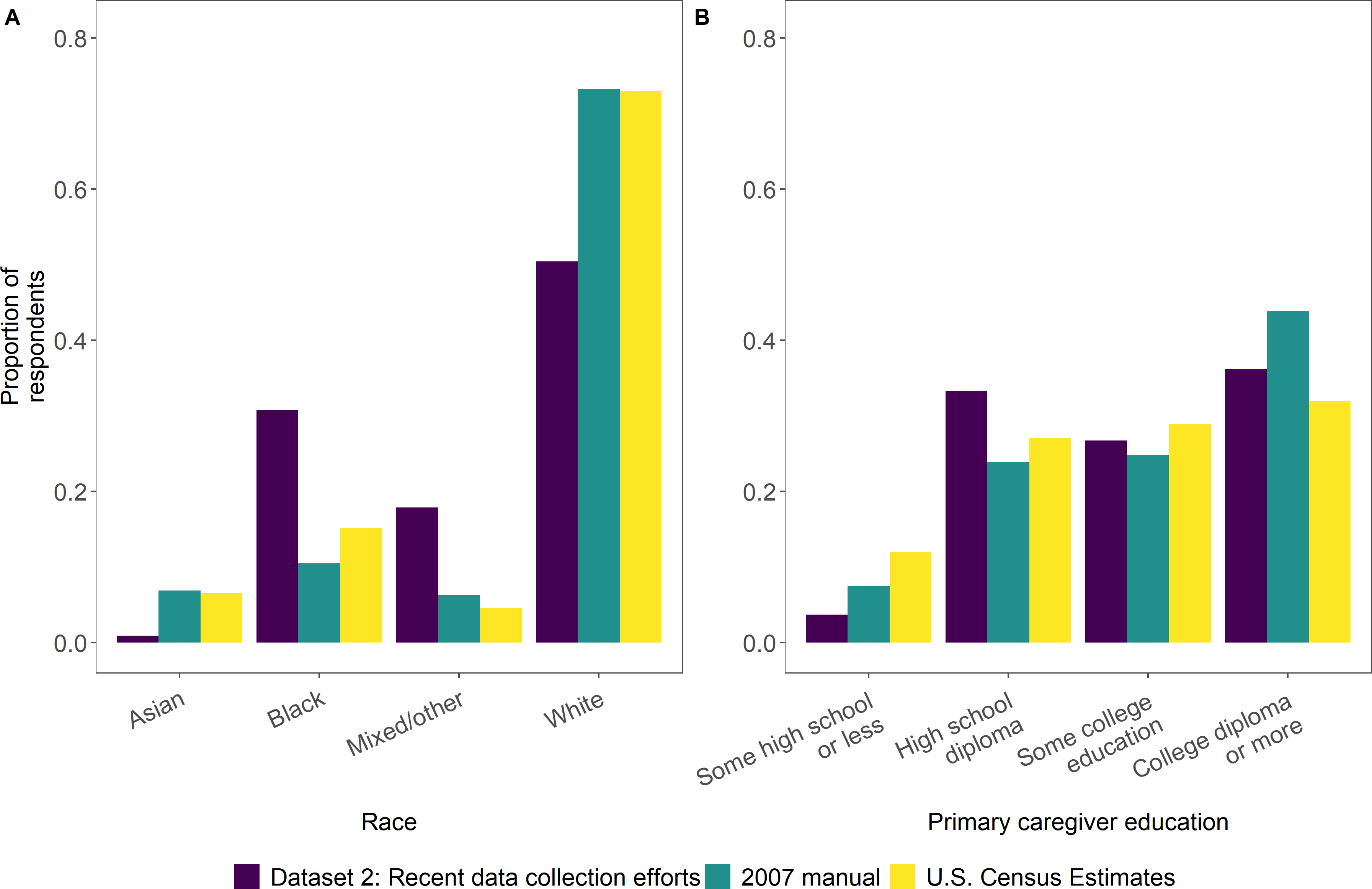
584 accounted for 33.3% of Dataset 2 as compared to 23.8% in the 2007 norms, and

585 representation of those with a college diploma or more education decreased from 43.8% in

586 the 2007 norms to 36.2% in Dataset 2. Notably, the distribution of Dataset 2 with regards

587 to primary caregiver education level is quite similar to Kristoffersen et al. (2013), who

588 collected a large, nationally-representative sample of CDI responses in Norway and



*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 oversam- pling 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.

589 obtained a sample with 30%, 42%, and 24% for participants reporting 12, 14-16, and 16+

590 years of education, respectively.

591 **Discussion: Dataset 2.** The results from Dataset 2 indicate that Web-CDI could

592 be a promising platform to collect vocabulary development data in non-white populations

593 and communities with lower levels of education attainment when paired with online

594 recruitment methods that yield legitimate, representative participant samples. These data

595 do, however, convey clear limitations of our approach. Perhaps most conspicuously, more

596 than half of completed administrations in this sample had to be excluded, in many cases

597 because the information provided by participants appeared rushed or incomplete: over 40%

598 of administrations were completed in a shorter amount of time than that allowed by our

599 cutoff criteria (Tables A4 and A5), and of these quick completions, well over 90% were

600 missing demographic information that is rarely missing in other administrations of the

601 form. Determining the precise reasons for the high exclusion rate, and how (if at all) this 602 (self-)selection may bias data reflecting demographic trends in vocabulary development, 603 requires a more thorough assessment of who is submitting hastily-completed forms. Such 604 an assessment is beyond the scope of the current study. However, all respondents who got 605 to the end of the form were compensated regardless of how thoroughly they completed it, 606 creating the possibility that some participants who clicked the anonymous link may not 607 have been members of the population of interest, but rather were other individuals

608 motivated by compensation. To the extent that participants moved through the form

609 quickly because they found the length burdensome, a transition to short forms, including 610 computer adaptive ones (e.g., Chai, Lo, & Mayor, 2020; Kachergis et al., 2021; Makransky, 611 Dale, Havmose, & Bleses, 2016; Mayor & Mani, 2019), would potentially increase data

612 quality and completion rates substantially.

613 Additionally, the exclusion rates described previously only provide information on

614 those participants who did, at some point, submit a completed form, but many individuals 615 clicked the advertisement link and did not subsequently continue on to complete the form. 616 Without an in-depth exploration of who is clicking the link and why they might choose not 617 to continue, we cannot draw conclusions about the representativeness of the sample in

618 Dataset 2 with regards to the communities we would like to include in our research. As 619 such, a more thorough understanding of how users from different communities respond to 620 various recruitment and sampling methods is needed in future work in order to draw

621 conclusions about demographic trends above and beyond those already established in the

622 literature.

623 Participants in Dataset 2 were recruited through a targeted post on social media, a 624 technique that is considerably more anonymous than recruitment strategies which entail 625 face-to-face or extended contact between researchers and community members. Online

626 recruitment methods may not be suitable for all communities, especially when researchers 627 ask participants to report potentially sensitive information about the health, developmental 628 progress, ethnicity and geographic location of their children (even when such information is 629 stored anonymously). Our goal here was to assess whether general trends in past literature 630 could be recovered using such an online strategy, but future research should take into

631 account that other more personal methods of recruitment, such as direct community

632 outreach or liaison contacts, may improve participants’ experiences and their willingness to

633 engage with the study.

634 Finally, a significant limitation of the data collection process in Dataset 2 is that

635 many people in the population of interest - particularly lower-income families - do not have

636 reliable internet access. Having participants complete the Web-CDI on a mobile device

637 may alleviate some of the issues caused by differential access to Wi-Fi, since the vast

638 majority of American adults own a smartphone (Pew Research Center, 2019). Accordingly,

639 improving Web-CDI’s user experience on mobile platforms will be an important step

640 towards ensuring that caregivers across the socioeconomic spectrum can easily complete

641 the survey. For smartphone users on pay-as-you-go plans, who may be reluctant to use

642 phone data to complete a study, a possible solution could be compensating participants for

643 the amount of “internet time” they incurred completing the form.

# 644 General Discussion and Conclusions

645 In this paper, we presented Web-CDI, a comprehensive online interface for researchers

646 to measure children’s vocabulary by administering the MacArthur-Bates Communicative

647 Development Inventories family of parent-report instruments. Web-CDI provides a

648 convenient researcher management interface, built-in data privacy protections, and a

649 variety of features designed to make both longitudinal and social-media sampling easy. To

650 date, over 3,500 valid administrations of the WG and WS forms have been collected on 651 Web-CDI from more than a dozen researchers in the United States after applying strict 652 exclusion criteria derived from previous norming studies (Fenson et al., 2007, 1994). Our 653 analysis of Dataset 1 shows that demographic trends from previous work using the

654 paper-and-pencil CDI form are replicated in data gleaned from Web-CDI, suggesting that

655 the Web-CDI is a valid alternative to the paper form and captures similar results.

656 Many research laboratories, not only in the United States but around the world,

657 collect vocabulary development data using the MacArthur-Bates CDI. With traditional

658 paper-based forms, combining insights from various research groups can prove challenging, 659 as each group may have slightly different ways of formatting and managing data from CDI 660 forms. By contrast, if all of these groups’ data come to be stored in a single repository with 661 a consistent database structure, data from disparate sources can easily be collated and

662 analyzed in a uniform fashion. As such, a centralized repository such as Web-CDI provides 663 a streamlined data-aggregation pipeline that facilitates cross-lab collaborations, multisite 664 research projects and the curation of large datasets that provide more power to

665 characterize the vast individual differences present in children’s vocabulary development.

666 Beyond the goal of simply getting more data, we hope that Web-CDI can advance 667 efforts to expand the reach of vocabulary research past convenience samples into diverse 668 communities. A key question in the field of vocabulary development concerns the

669 mechanisms through which sociodemographic variables, such as race, ethnicity, income and

670 education are linked to group differences in vocabulary outcomes. Large,

671 population-representative samples of vocabulary development data are needed to

672 understand these mechanisms, but research to date (including the full sample of Web-CDI

673 administrations) has often oversampled non-Hispanic white participants and those with

674 advanced levels of education.

675 We explored the use of Web-CDI as part of a potential strategy to collect data from 676 non-white and less highly-educated communities in two phases (Dataset 2). Several overall 677 patterns emerged which we expected: vocabulary scores grew with age, providing a basic 678 validity check of the Web-CDI measure; females held a slight advantage in word learning 679 over males; and children of caregivers with a college education showed slightly higher

680 vocabulary scores. Nonetheless, the insights from these data, while aligned with past

681 norming studies, are necessarily constrained by several features of our method.

682 Limitations of our method notwithstanding, a transition to web-based data collection 683 streamlines the process by which historically underrepresented populations can be reached 684 in child language research. In particular, recruitment methods involving community

685 partners, such as parenting groups, childcare centers and early education providers, are

686 simplified substantially if leaders in these organizations can distribute a web survey to their

687 members that is easy to fill out, as compared with paper forms, which present more

688 logistical hurdles for distribution and collection. Additionally, we hope that Web-CDI can 689 serve as an accessible, free, and easy to use resource for researchers already doing extensive 690 work with underrepresented groups.

691 Web-based data collection can capture useful information about vocabulary

692 development from diverse communities, but future research will need to examine which

693 sampling methods can yield accurate, population-representative data that can advance our

694 understanding of the link between sociodemographic variation and variation in language

695 outcomes.

# 696 Ethics statement

697 Data collected in the United States for this project are anonymized according to

698 guidelines set forth by the United States Department of Health and Human Services. Data

699 collection at Stanford University was approved by the Stanford Institutional Review Board

700 (IRB), protocol 20398.

# 701 Data, code and materials availability statement

702 • Open data: All data analyzed in this work are available on the Open Science

703 Framework at <https://osf.io/nmdq4/>.

704 • Code: All code for this work is avaiable on the Open Science Framework at

705 <https://osf.io/nmdq4/>.

706 • Materials: All code and materials for the Web-CDI are openly available at

707 <https://github.com/langcog/web-cdi>. If readers wish to view the Web-CDI interface

708 in full from the participants’ or researchers’ perspectives, they are encouraged to

709 contact [webcdi-contact@stanford.edu.](mailto:webcdi-contact@stanford.edu)

# 710 Author contributions

711 • Conceptualization: Benjamin deMayo, Danielle Kellier, Mika Braginsky, Christina

712 Bergmann, Caroline Rowland, Michael Frank and Virginia Marchman.

713 • Data Curation: Benjamin deMayo, Danielle Kellier and Virginia Marchman.

714 • Formal Analysis: Benjamin deMayo.

715 • Funding Acquisition: Caroline Rowland and Michael Frank.

716 • Investigation: Benjamin deMayo, Danielle Kellier and Virginia Marchman.

717 • Methodology: Benjamin deMayo, Danielle Kellier, Michael Frank and Virginia

718 Marchman.

719 • Project Administration: Caroline Rowland, Michael Frank and Virginia Marchman. 720 • Software: Danielle Kellier, Mika Braginsky, Christina Bergmann and Cielke Hendriks. 721 • Supervision: Christina Bergmann, Caroline Rowland, Michael Frank and Virginia

722 Marchman.

723 • Visualization: Benjamin deMayo.

724 • Writing - Original Draft Preparation: Benjamin deMayo, Michael Frank and Virginia

725 Marchman.

726 • Writing - Review & Editing: Benjamin deMayo, Danielle Kellier, Mika Braginsky,

727 Christina Bergmann, Cielke Hendriks, Caroline Rowland, Michael Frank and Virginia

728 Marchman.

729 **Software used**

730 R [Version 4.0.3; R Core Team (2020)] and the R-packages *broman* [Version 0.71.6;

731 Broman (2020)], *cowplot* [Version 1.1.0; Wilke (2020)], *dplyr* [Version 1.0.2; Wickham,

732 François, Henry, and Müller (2020)], *estimatr* [Version 0.26.0; Blair, Cooper, Coppock,

733 Humphreys, and Sonnet (2020)], *forcats* [Version 0.5.0; Wickham (2020a)], *fs* [Version 1.5.0;

734 Hester and Wickham (2020)], *ggplot2* [Version 3.3.2; Wickham (2016)], *here* [Version 0.1;

735 Müller (2017)], *kableExtra* [Version 1.3.1; Zhu (2020)], *papaja* [Version 0.1.0.9997; Aust and

736 Barth (2020)], *purrr* [Version 0.3.4; Henry and Wickham (2020)], *readr* [Version 1.4.0;

737 Wickham and Hester (2020)], *scales* [Version 1.1.1; Wickham and Seidel (2020)], *stringr*

738 [Version 1.4.0; Wickham (2019)], *tibble* [Version 3.0.4; Müller and Wickham (2020)], *tidyr*

739 [Version 1.1.2; Wickham (2020b)], *tidyverse* [Version 1.3.0; Wickham et al. (2019)], 740 *wordbankr* [Version 0.3.1; (**R-wordbankr?**)], and *xtable* [Version 1.8.4; Dahl, Scott, 741 Roosen, Magnusson, and Swinton (2019)]

742 **References**

743 Alcock, K., Meints, K., & Rowland, C. (2020). *The UK communicative development*

744 *inventories: Words and gestures*. J&R Press.

745 Astivia, O. L. O., & Zumbo, B. D. (2019). Heteroskedasticity in multiple regression 746 analysis: What it is, how to detect it and how to solve it with applications in r 747 and SPSS. *Practical Assessment, Research, and Evaluation*, *24*(1), 1.

748 Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*.

749 *Journal of Open Source Software* (Vol. 4, p. 1686).

750 <https://doi.org/10.21105/joss.01686>

751 Bates, E., & Goodman, J. C. (2001). On the inseparability of grammar and the

752 lexicon: Evidence from acquisition.

753 Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., … Hartung,

754 J. (1994). Developmental and stylistic variation in the composition of early

755 vocabulary. *J Child Lang*, *21*(01), 85–123.

756 Blair, G., Cooper, J., Coppock, A., Humphreys, M., & Sonnet, L. (2020). *Estimatr:* 757 *Fast estimators for design-based inference*. *Journal of Open Source Software* (Vol. 758 4, p. 1686). Springer-Verlag New York. <https://doi.org/10.21105/joss.01686>

759 Bornstein, M. H., & Putnick, D. L. (2012). Stability of language in childhood: A 760 multiage, multidomain, multimeasure, and multisource study. *Developmental* 761 *Psychology*, *48*(2), 477.

762 Broman, K. W. (2020). *Broman: Karl broman’s r code*. *Journal of Open Source*

763 *Software* (Vol. 4, p. 1686). Springer-Verlag New York.

764 <https://doi.org/10.21105/joss.01686>

765 Caselli, N. K., Lieberman, A. M., & Pyers, J. E. (2020). The ASL-CDI 2.0: An

766 updated, normed adaptation of the MacArthur bates communicative

767 development inventory for american sign language. *Behavior Research Methods*,

768 1–14.

769 Chai, J. H., Lo, C. H., & Mayor, J. (2020). A bayesian-inspired item response 770 theory–based framework to produce very short versions of MacArthur–bates 771 communicative development inventories. *Journal of Speech, Language, and* 772 *Hearing Research*, *63*(10), 3488–3500.

773 Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., & Swinton, J. (2019). *Xtable:*

774 *Export tables to LaTeX or HTML*. Retrieved from

775 [https://CRAN.R-project.org/package=xtable](https://CRAN.R-project.org/package%3Dxtable)

776 Dale, P. S. (2015). Adaptations, Not Translations! Retrieved from

777 <http://mb-cdi.stanford.edu/Translations2015.pdf>

778 De Houwer, A. (2019). Equitable evaluation of bilingual children’s language

779 knowledge using the CDI: It really matters who you ask. *Journal of Monolingual*

780 *and Bilingual Speech*, *1*(1), 32–54.

781 Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E.,

782 & Paradise, J. L. (2000). Measurement properties of the MacArthur

783 communicative development inventories at ages one and two years. *Child*

784 *Development*, *71*(2), 310–322.

785 Fenson, L., Bates, E., Dale, P. S., Marchman, V. A., Reznick, J. S., & Thal, D. J. 786 (2007). *MacArthur-Bates Communicative Development Inventories*. Brookes 787 Publishing Company.

788 Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., … Stiles,

789 J. (1994). Variability in early communicative development. *Monogr Soc Res*

790 *Child Dev*, *59*(5).

791 Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & Reznick, J. S. (2000). 792 Short-form versions of the MacArthur communicative development inventories. 793 *Applied Psycholinguistics*, *21*(1), 95–116.

794 Floccia, C., Sambrook, T. D., Delle Luche, C., Kwok, R., Goslin, J., White, L., …

795 others. (2018). Vocabulary of 2‐year‐olds learning english and an additional

796 language: Norms and effects of linguistic distance. *Monographs of the Society for*

797 *Research in Child Development*, *83*(1), 1–135.

798 Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., …

799 others. (2017). A collaborative approach to infant research: Promoting

800 reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435.

801 Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability*

802 *and consistency in early language learning: The wordbank project*. MIT Press.

803 Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools*. *Journal of*

804 *Open Source Software* (Vol. 4, p. 1686). <https://doi.org/10.21105/joss.01686>

805 Hester, J., & Wickham, H. (2020). *Fs: Cross-platform file system operations based* 806 *on ’libuv’*. *Journal of Open Source Software* (Vol. 4, p. 1686). Springer-Verlag 807 New York. <https://doi.org/10.21105/joss.01686>

808 Kachergis, G., Marchman, V., Dale, P., Mehta, H., Mankewitz, J., & Frank, M. 809 (2021). *An online computerized adaptive test (CAT) of children’s vocabulary* 810 *development in english and mexican spanish*.

811 Kapalková, S., & Slanèová, D. (2007). Adaptation of CDI to the slovak language. In

812 *Proceedings from the first european network meeting on the communicative*

813 *development inventories: May 24-28 2006 dubrovnik croatia*. University of Gävle.

814 Kartushina, N., Mani, N., AKTAN-ERCIYES, A., Alaslani, K., Aldrich, N. J.,

815 Almohammadi, A., … al., et. (2021). COVID-19 first lockdown as a unique

816 window into language acquisition: What you do (with your child) matters.

817 PsyArXiv. <https://doi.org/10.31234/osf.io/5ejwu>

818 Kristoffersen, K. E., Simonsen, H. G., Bleses, D., Wehberg, S., Jørgensen, R. N.,

819 Eiesland, E. A., & Henriksen, L. Y. (2013). The use of the internet in collecting

820 CDI data–an example from norway. *Journal of Child Language*, *40*(03), 567–585.

821 Makransky, G., Dale, P. S., Havmose, P., & Bleses, D. (2016). An item response

822 theory–based, computerized adaptive testing version of the MacArthur–bates 823 communicative development inventory: Words & sentences (CDI: WS). *Journal* 824 *of Speech, Language, and Hearing Research*, *59*(2), 281–289.

825 Mayor, J., & Mani, N. (2019). A short version of the MacArthur–bates

826 communicative development inventories with high validity. *Behavior Research*

827 *Methods*, *51*(5), 2248–2255.

828 Müller, K. (2017). *Here: A simpler way to find your files*. *Journal of Open Source*

829 *Software* (Vol. 4, p. 1686). <https://doi.org/10.21105/joss.01686>

830 Müller, K., & Wickham, H. (2020). *Tibble: Simple data frames*. *Journal of Open*

831 *Source Software* (Vol. 4, p. 1686). <https://doi.org/10.21105/joss.01686>

832 Percentage of persons 18 to 24 years old and age 25 and over, by educational

833 attainment, race/ethnicity, and selected racial/ethnic subgroups: 2010 and 2017.

[834](https://nces.ed.gov/programs/digest/d18/tables/dt18_104.40.asp?referer=raceindica.asp) (2019). [https://nces.ed.gov/programs/digest/d18/tables/dt18\_104.40.asp?](https://nces.ed.gov/programs/digest/d18/tables/dt18_104.40.asp?referer=raceindica.asp)

835 [referer=raceindica.asp](https://nces.ed.gov/programs/digest/d18/tables/dt18_104.40.asp?referer=raceindica.asp).

836 Pew research center mobile fact sheet. (2019). Retrieved from

837 <https://www.pewresearch.org/internet/fact-sheet/mobile/>

838 R Core Team. (2020). *R: A language and environment for statistical computing*.

839 *Journal of Open Source Software* (Vol. 4, p. 1686). Vienna, Austria: R

840 Foundation for Statistical Computing; Springer-Verlag New York.

841 <https://doi.org/10.21105/joss.01686>

842 Snyder, T. D., De Brey, C., & Dillow, S. A. (2019). Digest of education statistics

843 2017, NCES 2018-070. *National Center for Education Statistics*.

844 Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. *Journal of Open*

845 *Source Software* (Vol. 4, p. 1686). Springer-Verlag New York.

846 <https://doi.org/10.21105/joss.01686>

847 Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string*

848 *operations*. *Journal of Open Source Software* (Vol. 4, p. 1686).

849 <https://doi.org/10.21105/joss.01686>

850 Wickham, H. (2020a). *Forcats: Tools for working with categorical variables*

851 *(factors)*. *Journal of Open Source Software* (Vol. 4, p. 1686). Springer-Verlag

852 New York. <https://doi.org/10.21105/joss.01686>

853 Wickham, H. (2020b). *Tidyr: Tidy messy data*. *Journal of Open Source Software*

854 (Vol. 4, p. 1686). <https://doi.org/10.21105/joss.01686>

855 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … 856 Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 857 *4*(43), 1686. <https://doi.org/10.21105/joss.01686>

858 Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of*

859 *data manipulation*. *Journal of Open Source Software* (Vol. 4, p. 1686).

860 Springer-Verlag New York. <https://doi.org/10.21105/joss.01686>

861 Wickham, H., & Hester, J. (2020). *Readr: Read rectangular text data*. *Journal of*

862 *Open Source Software* (Vol. 4, p. 1686). <https://doi.org/10.21105/joss.01686>

863 Wickham, H., & Seidel, D. (2020). *Scales: Scale functions for visualization*. *Journal*

864 *of Open Source Software* (Vol. 4, p. 1686). <https://doi.org/10.21105/joss.01686>

865 Wilke, C. O. (2020). *Cowplot: Streamlined plot theme and plot annotations for* 866 *’ggplot2’*. *Journal of Open Source Software* (Vol. 4, p. 1686). Springer-Verlag 867 New York. <https://doi.org/10.21105/joss.01686>

868 Zhu, H. (2020). *kableExtra: Construct complex table with ’kable’ and pipe syntax*.

869 *Journal of Open Source Software* (Vol. 4, p. 1686).

870 <https://doi.org/10.21105/joss.01686>

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 6 – parent must take to complete

the study

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?

Do you plan on collecting only anonymous data in this study? (e.g., posting ads on social media, mass emails, etc)

Would you like to show participants graphs of their data after completion?

Would you like participants to be able to share their

Web-CDI results via Facebook?

Would you like participants to answer the confirmation questions?

No If checked, researchers can enter gift codes to distribute to participants once they have completed the survey.

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.

Yes –

No –

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?

Capture the Prolific Id for the participant?

Allow participant to print their responses at end of Study?

No Used to redirect users to external site after form completion.

No For integration with Prolific.

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 | 122.275 | 2.427 | 50.381 | 0.000 | 117.515 | 127.035 | 1610 |
| Age | 20.050 | 0.767 | 26.127 | 0.000 | 18.545 | 21.556 | 1610 |
| Caregiver education: Some college | 17.445 | 8.179 | 2.133 | 0.033 | 1.403 | 33.487 | 1610 |
| Caregiver education: High school or less | 21.862 | 10.935 | 1.999 | 0.046 | 0.413 | 43.311 | 1610 |
| Age \* Caregiver education: Some college | -1.991 | 2.261 | -0.881 | 0.379 | -6.425 | 2.443 | 1610 |
| Age \* Caregiver education: High school or less | -6.604 | 3.159 | -2.091 | 0.037 | -12.800 | -0.408 | 1610 |

Table A3

*Regression output for WG production measure.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value | conf.low | conf.high | df |
| Intercept | 29.771 | 1.332 | 22.358 | 0.000 | 27.159 | 32.382 | 1610 |
| Age | 7.599 | 0.498 | 15.264 | 0.000 | 6.622 | 8.575 | 1610 |
| Caregiver education: Some college | 5.640 | 4.919 | 1.147 | 0.252 | -4.009 | 15.289 | 1610 |
| Caregiver education: High school or less | 20.455 | 7.693 | 2.659 | 0.008 | 5.366 | 35.545 | 1610 |
| Age \* Caregiver education: Some college | -1.357 | 1.327 | -1.022 | 0.307 | -3.960 | 1.247 | 1610 |
| Age \* Caregiver education: High school or less | -0.121 | 2.095 | -0.058 | 0.954 | -4.229 | 3.988 | 1610 |

Table A4

*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 |

Table A5

*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 |