

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

Benjamin deMayo<sup>1</sup>, Danielle Kellier<sup>2</sup>, Mika Braginsky<sup>3</sup>, Christina Bergmann<sup>4</sup>, Cielke  
Hendriks<sup>4</sup>, Caroline Rowland<sup>4</sup>, Michael C. Frank<sup>5</sup>, and & Virginia Marchman<sup>5</sup>

<sup>1</sup> Princeton University

<sup>2</sup> University of Pennsylvania

<sup>3</sup> Massachusetts Institute of Technology

<sup>4</sup> Max Planck Institute for Psycholinguistics

<sup>5</sup> Stanford University

## Abstract

Understanding the mechanisms which drive variation in children’s language acquisition requires large, population-representative datasets of children’s word learning across development. Parent report measures such as the MacArthur-Bates Communicative Development Inventories (CDI) are commonly used to accrue such data, but the traditional paper-based forms make the curation of large datasets logistically challenging, and tend to rely on convenience samples located close in proximity to major research institutions. Here, we introduce Web-CDI, a web-based tool which allows researchers to collect CDI data online. Web-CDI contains functionality to collect and manage longitudinal data, share links, and download standardized vocabulary scores. To date, over 3,529 valid Web-CDI administrations have been completed. General trends found in past norming studies of the CDI are present in data collected from Web-CDI: scores of children’s productive vocabulary grow with age, female children show a slightly faster rate of vocabulary growth in early childhood, and participants with higher levels of education attainment report slightly higher vocabulary production scores. We also report results from an effort to oversample non-white, lower-SES participants ( $N = 241$ ), which showed similar demographic trends to the full sample but which had a high exclusion rate. Implications and challenges for the collection of large, population-representative datasets using Web-CDI in future research are discussed.

*Keywords:* vocabulary development, parent report, socioeconomic status

Word count: X

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

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

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

We begin by discussing parent report as a powerful method by which to address the challenge of measuring early language outcomes, as well as previous online parent report instruments. We then describe Web-CDI and its use from the perspective of participants and researchers. Finally, we report on our use of Web-CDI thus far and discuss the potential of Web-CDI to acquire vocabulary data from diverse, population-representative samples. We end with some challenges for future research.

## The Importance of Parent Report Data

Gaining empirical traction on variation in children’s early language requires reliable and valid methods for measuring language abilities, especially in early childhood (8 to 30 months). Parent report is a mainstay in this domain. Parent reports are based on their daily experiences with the child, which are much more extensive than a researcher or clinician can generally obtain. Moreover, they are less likely to be influenced by factors that may mask a child’s true ability in the laboratory or clinic (e.g., shyness; Frank et al. (2021)), and parents are remarkably accurate at reporting their children’s language, especially in the first two years of life (CITE). One widely used set of parent-report instruments is the MacArthur-Bates Communicative Development Inventories, originally designed for children learning American English (Fenson et al., 2007). The American English CDIs come in two versions, Words & Gestures for children 8 to 18 months, focusing on word comprehension and production, as well as gesture use, and Words & Sentences, for children 16 to 30 months, focusing on word production and sentence structure. Together, these instruments allow for a comprehensive picture of milestones that characterize language development in the first 2½ years of life.

A substantial body of evidence suggests that these instruments are both reliable and valid (e.g., Fenson et al., 1994, 2007) leading to their widespread use in thousands of research studies over the last few decades. Indeed, the popularity of the American English and Spanish CDI instruments has meant that many teams around the world have adapted the CDI format to the particular language and community (Dale, 2015). Importantly, these adaptations are not simply translations of the original form but rather incorporate the specific features of different languages and cultures, since linguistic variability exists even among cultures that share a native language (e.g., Cheerios are more common in American than British homes, so age of acquisition of this word may differ substantially). To date there are now more than 100 adaptations for languages around the globe.

Initial large-scale work to establish the normative datasets for the American English CDI not only provided key benchmarks for determining children’s progress, but also documented the extensive individual differences that characterize early language learning during this critical period of development (Bates et al., 1994; Fenson et al., 1994). Understanding the origins and consequences of this variability remains an important empirical and theoretical endeavor that has informed critical insights in the field (e.g., Bates & Goodman, 2001; Bornstein & Putnick, 2012; see also, Frank et al., 2021). The popularity of the instruments has remained strong over the years, leading to the development of extensions of the methodology to alternative formats, e.g., short forms (Fenson et al., 2000).

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

Here, we report on our recent efforts to create and distribute a web-based version of the MacArthur-Bates CDIs in order to address some of the limitations of the standard paper versions. Online administration of the CDI is not a novel innovation – a variety of research groups have 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 (CHECK CITE HERE). 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, and 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 standardized 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, far reducing the overhead efforts associated with bringing together data collected using paper forms by researchers across the world.

## Introducing Web-CDI

Web-CDI is our web-based platform for CDI administration and management. Parents are recruited by either receiving an individual URL directly from a researcher or by interacting with a targeted social media advertisement using general-purpose URLs. Web-CDI serves as a low-risk method that allows researchers to communicate with families electronically, facilitating access to families in areas distant from an academic institution and eliminating costly mailings or laboratory visits. It also permits other institutions to use Web-CDI as a resource, while still allowing members of the CDI Advisory Board to access and analyze the resulting data with the researchers' permissions. Since 2018, more than 3000 CDIs have been collected via 15 research groups throughout the US, demonstrating the potential for large-scale data collection and aggregation.

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

### *Participant interface*

Participants can complete the Web-CDI on a variety of devices, including personal computers and tablets; Web-CDI can be administered on a smartphone, although the experience is not as ideal for the user due to the length of the survey. When a participant clicks a link to the form, they are directed to a website displaying their own personal administration of the Web-CDI, regardless of whether the link was participant-specific or general-purpose. In some cases, they may be asked to read and accept a waiver of consent documentation, depending on whether the researcher has chosen to use that feature (see also Researcher Interface below).

*Demographics.* The parent is then asked to provide demographic information about their family and any health conditions that might impact their child's vocabulary development. The specific demographic questions asked of participants can be adjusted to vary between different versions of the form, allowing researchers to tailor the demographic questions to local norms<sup>1</sup>. Researchers can customize the presentation of these demographic questions in three ways. First, they can elect to show all of the demographics items on the landing page or they elect to present the majority of these questions at the end of the instrument. This choice is provided because some pilot work in the United Kingdom indicated that answering questions regarding personal health information at the beginning may deter participants from completing the instrument. Second, certain demographic questions can be asked at both the beginning and the end of the form to serve as validity checks, providing a check that can be used to screen for hasty or illegitimate completions. Third, researchers can tailor the questions to the societal and cultural context of their participants (e.g. country-specific education level descriptors and income categories).

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

---

<sup>1</sup> For example, the Dutch CDI omits questions about participant ethnicity, since census data in the Netherlands does not include ethnicity.

**Instructions:**

- This form can be filled any time before the due date.
- It can also be saved at any time and resumed later by using the same link ([create bookmark](#)).
- After the form is submitted, it cannot be altered.
- The form also cannot be altered after the due date.
- Please use the navigation buttons below. Do not use the "back" and "forward" buttons on your browser.
- You can use the tab button and arrow keys to quickly navigate and answer questions.

**Due date:** Feb 15, 2020 6:02 PM

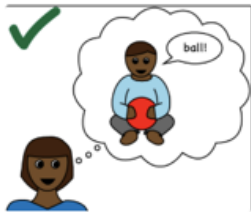

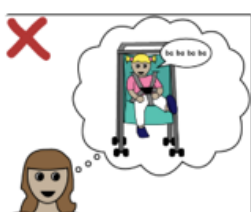
[Show waiver of documentation](#)  
Reach out to the Web-CDI Team!

[Save](#)

[Go back to Background Info](#)

[Table of Contents](#)

In this section, you will be asked about words that your child "understands and says." Your child "understands and says" a word on the list if they know what the word means AND they say it by themselves. Here are some examples. This assessment is for children of many ages. Your child may not be able to understand or say a lot of the words on the form. That is perfectly fine!

DO check the box if:

Your child says the word when trying to name an object or describe something that happened. You think s/he has a meaning for that word.

It's fine if your child can't say the whole word or says it his/her own "special" way. If you use another word in your family (e.g., Nana for Grandma), that is ok too!

DO check the box.

DON'T check the box if your child is just stringing sounds together. This is not a real word that means something.

**Figure 1**

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

directed to the instructions that are appropriate for either the Words & Gestures or Words & Sentences version (see Figure 1). At the top of the page are general instructions that inform participants that they should expect the study to take at least 30 minutes and that they should try to complete it in a quiet setting( e.g., while their child is sleeping). In addition, there are more detailed instructions for completing the vocabulary checklist. Unlike the traditional paper versions, instructions on how to properly choose responses are provided both in written and pictorial form. The pictorial instructions (Figure 1) aim to further increase caregivers' understanding of how to complete the checklist. For example, these instructions clarify that the child's understanding of a word requires them to have some understanding of the object that the word refers to or some aspect of the word's meaning. In addition, caregivers are reassured that "child like" forms (e.g., "raff" for



“giraffe”) or family- or dialect-specific forms (e.g., “nana” for “grandma” are acceptable). Lastly, caregivers are reminded that the child should be able to produce the words “on their own” and that imitations are not acceptable. These general “rules of thumb” for completing the form should be familiar to researchers who are distributing the forms to parents so they can field any questions that may arise. While this is not possible for certain use-cases (e.g., collecting data via Facebook), these instructions should ideally also be reviewed either in writing (e.g., via email) or verbally (e.g., over the phone), so that these pictured instructions serve merely as a reminder to caregivers when completing the form.

*Completing the instrument.* The majority of the participant’s time in the study is spent completing the main sections of the instruments. As shown in Figure 2, on the American English Words and Gestures form, the vocabulary checklist portion of the form (396 items) asks parents to indicate whether their child can “understand” or “understand and say” each word. Gesture communication and other early milestones are also assessed. In the American English Words and Sentences form, the vocabulary checklist (680 items) only asks parents to indicate which words their child “says”. Additional items assess children’s production of their three longest sentences, as well as morphological and syntactic development more broadly.

At the completion of the form, a graph is displayed illustrating the proportion of words from each semantic category that the child currently produces or understands. In addition, data from the norming studies are used to estimate the “hardest” (i.e., most advanced) word that the child currently understands or produces. This feedback to parents is intended to provide parents with a fun “thank you” and is intentionally not designed to provide specific feedback about their child’s progress. The closing page also reminds parents that their participation does not constitute a clinical evaluation and that they should contact their pediatrician or primary care physician if they have any concerns about their child’s development.

**Table 1**

Study setting
Study name
Instrument
Number of days before study expiration
Measurement units for birth weight
Minimum time (minutes) a parent must take to complete the study
Waiver of documentation
Pre-fill data for longitudinal participants?
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 e
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?

## *Researcher Interface*

One of the main goals of Web-CDI is to facilitate wide distribution of the platform to the child language research community. To that end, researchers are required to contact a member of the CDI Advisory Board to register an account on Web-CDI, from which they can create studies to distribute to participants. Note that we ask that researchers allow fully anonymised data to be shared with us, so that it can be added to Wordbank (<http://wordbank.stanford.edu/>). However, there is an opt-out option if researchers do not wish to share their data.

A study in the context of the Web-CDI system is a set of individual administrations created by a researcher that share certain specifications. Table 1 gives an overview of the

customizable features that are available at the study level in Web-CDI. These features are set when creating a study for the first time in Web-CDI using the “Create Study” tool, and most of the features can be updated continuously during data collection using the “Update Study” tool. While some of these features are only particularly relevant to specific use cases (e.g., longitudinal research and social media data collection, outline below), others are relevant to all researchers using Web-CDI.

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

Researchers can download participant data in two formats. Both formatting options output a comma-separated values file with one row per participant; the full data option includes participant-by-item responses, and allows researchers to explore item-level trends, while the summary data option omits item-level data and only provides summary scores (e.g., total number of words understood/produced, percentile scores by age and gender).

Below, we outline several possible use cases of Web-CDI, as well the features which may facilitate them from a researcher’s perspective.

*Individual recruitment.* One possible workflow using Web-CDI is to send unique study URLs to individual participants. Researchers do so by entering numerical participant IDs or by auto-generating a specified quantity of participant IDs, each with its own unique study URL, using the “Add Participants” tool in the researcher dashboard. New participants can be added on a continual basis so that researchers can adjust the sample size of their study during data collection. Unique links generated for individual

participants expire, by default, 14 days after creation, though the amount of days before link expiration is adjustable, which may be an important consideration for some researchers depending on their participant populations and specific project timelines. This workflow is most suitable for studies which pair the CDI with other measures, or when researchers contact specific participants from an existing database.

*Longitudinal studies.* Web-CDI also facilitates longitudinal study designs in which each participant completes multiple administrations. Researchers wishing to design longitudinal studies can do so by entering a list of meaningful participant IDs using the “Add Participants” tool in the researcher dashboard. If a certain participant ID is added multiple times, Web-CDI will create multiple unique study URLs in the study dashboard that have the same specified ID. In addition, when creating studies, researchers can select whether they would like the demographics information, vocabulary checklist, or no sections at all to be prefilled when a participant fills out a repeat administration of the instrument. Unless researchers are interested in cumulative vocabulary counts, it is strongly recommended that they do not use the option to pre-fill the vocabulary checklist portion of the instrument in longitudinal administrations as parents should complete the instrument at each time point independently.

*Social media and survey vendors.* Web-CDI contains several features designed to facilitate data collection from social media recruitment or through third-party crowd-sourcing applications and vendors (e.g., Amazon Mechanical Turk, Prolific). First, rather than creating unique survey links for each participant, researchers can also use a single, anonymous link. When a participant clicks the anonymous link, a new administration with a unique subject ID is created in the study dashboard. Additionally, Web-CDI studies have several customizable features that are geared towards anonymous online data collection. For example, researchers can adjust the minimum amount of time a participant must take to fill out the survey before they are able to submit; with a longer minimum time to completion, researchers can encourage a more thorough completion of

the survey. Researchers can also ask participants to verify that their information is accurate by checking a box at the end of the survey, and can opt to include certain demographic questions at both the beginning and end of the survey, using response consistency on these redundant items as a check of data quality.

*Paid participation.* If researchers choose to compensate participants directly through the Web-CDI interface, Web-CDI has built-in functionality to distribute redeemable gift codes when a participant reaches the end of the survey. Web-CDI contains several features to facilitate integration with third-party crowdsourcing applications and survey vendors 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>).

### ***Administrator Interface***

System administrators who oversee the development and usage of Web-CDI can log into a specific interface provided by Django, the open-source web framework used to develop Web-CDI. This Django interface gives access to the entire dataset collected by all participating researchers, as well as all of the instrument forms that can be distributed to participants. Administrators approve new researchers' requests to create accounts on Web-CDI, and give individual researchers access to specific instrument forms. As an example, if a researcher wants to distribute the Spanish Words and Gestures form to

participants, an administrator would need to give a researcher access to this instrument. Administrator privileges are limited to a handful of individuals, as most of the functionality accessible to administrators is not relevant to researchers using the survey for their own purposes.

### *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 by Amazon Web Services. Individual researchers can download data from their studies through the researcher interface, and Web-CDI admins have access to the entire aggregate set of data from all studies run with Web-CDI. Website code is available in a GitHub repository <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 from the ground up to be compliant with stringent human subjects privacy protections across the world. First, for US 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 dates are used to calculate age in months (with no decimal information) and never stored, and geographic zip codes are trimmed to the first 3 digits. Because of the architecture of the site, even though participants enter zip codes and dates of birth, these are never transmitted in full to the Web-CDI server. Since no identifying information is being collected by the Web-CDI system, this feature ensures that Web-CDI can be used by United States labs without a separate Institutional Review Board agreement between users labs and Web-CDI (though of course researchers using the site

will need Institutional Review Board approval of their own research projects)<sup>2</sup>.

In the European Union (EU), research data collection and storage is governed by the Generalized Data Protection Regulation (GDPR) and its local instantiation in the legal system of the member states. Some of the questions on the demographic form contain information that may be considered sensitive (e.g., information about children’s developmental disorders), and in some cases, the possibility of linking this sensitive information to participant IDs exists, particularly when researchers draw on local databases that contain full names and addresses for recruitment and contacting. As a result, issues regarding GDPR compliance arise when transferring data outside the EU, namely to Amazon Web Services servers housed in the United States. Following GDPR regulations, these issues would make a data sharing agreement between data collectors and Amazon Web Services necessary. In addition, all administrators who can access the collected data would have to enter such an agreement, which needs updating whenever personnel changes occur. To overcome these hurdles and in consultation with data protection officers, we opted to exploit the local technical expertise and infrastructure to set up a sister site housed on GDPR-compliant servers, currently available under [webcdi.mpi.nl](http://webcdi.mpi.nl). This site is updated synchronously with the main Web-CDI website to ensure a consistent user experience and access to the latest features and improvements. This site has been used in 135 successful administrations so far and is the main data collection tool for an ongoing norming study in the Netherlands. We are further actively advertising the option to use the European site to other labs who are following GDPR guidelines and are planning adaptations to multiple European languages, where copyright allows.

---

<sup>2</sup> 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.

330 **Current Usage and Data**

331 **Participants**

332 **Material**

333 **Procedure**

334 **Results**

335 **Discussion**



## References

- Bates, E., & Goodman, J. C. (2001). On the inseparability of grammar and the lexicon: Evidence from acquisition.
- Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., . . . Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *J Child Lang*, 21(01), 85–123.
- Bornstein, M. H., & Putnick, D. L. (2012). Stability of language in childhood: A multiage, multidomain, multimeasure, and multisource study. *Developmental Psychology*, 48(2), 477.
- Caselli, N. K., Lieberman, A. M., & Pyers, J. E. (2020). The asl-cdi 2.0: An updated, normed adaptation of the macarthur bates communicative development inventory for american sign language. *Behavior Research Methods*, 1–14.
- Dale, P. S. (2015). Adaptations, Not Translations! Retrieved from <http://mb-cdi.stanford.edu/Translations2015.pdf>
- 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).
- Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & Reznick, J. S. (2000). Short-form versions of the macarthur communicative development inventories. *Applied Psycholinguistics*, 21(1), 95–116.
- Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The wordbank project*. MIT Press.

- 361 Kristoffersen, K. E., Simonsen, H. G., Bleses, D., Wehberg, S., Jørgensen, R. N., Eiesland,  
362 E. A., & Henriksen, L. Y. (2013). The use of the internet in collecting cdi data—an  
363 example from norway. *Journal of Child Language*, 40(03), 567–585.