How WEIRD-biased is CHILDES data on childrens linguistic input? Supplementary Materials 4

Contents

In this document, we provide code that allows readers to reproduce information provided in the main text that depend on calculations computed here.

Introduction

Code to reproduce Figure 1

Note that this Figure appears in a different page.

Code to reproduce information in the caption to Figure 1

We classified countries as Western and Democratic. For Western, we classified as Western all of the countries in Western Europe (according to EuroVoc: Andorra, Austria, Belgium, France, Germany, Ireland, Liechtenstein, Luxembourg, Monaco, Netherlands, Switzerland, United Kingdom) as well as a few others in Europe (Italy, Portugal, Spain, Norway, Sweden, Iceland, Denmark), most of North America (with the exception of Bermuda and Greenland), and two countries in Oceania (Australia, New Zealand). This resulted in 20 Western and 223 non-Western countries. For democratic, we simplified the OWID's classification of closed and electoral autocracies into autocracies, versus electoral and liberal democracies into democracies (Our World in data, 2022c). We had data on political regime for 178 countries (from Our World in Data, OWID, 2011); on the proportion of the population having completed lower secondary school 178 (from OWID, 2007-2014); on GDP for 189 countries (from WDI, 2011); on the proportion of the population that was urban for 212 countries (from WDI, 2011); on fertility for 199 countries (from OWID, 2011).

Code to reproduce Figure 2

Note that this Figure appears in a different page.

Code to reproduce text on p. 13

The answer to the question of who surrounds the child may also depend on how large the nuclear family is. There is considerable cross-population variation in terms of the average number of children a woman has in her reproductive lifetime, varying between a little over 1 (1.07, Taiwan) to over 6 (7.46, Niger).

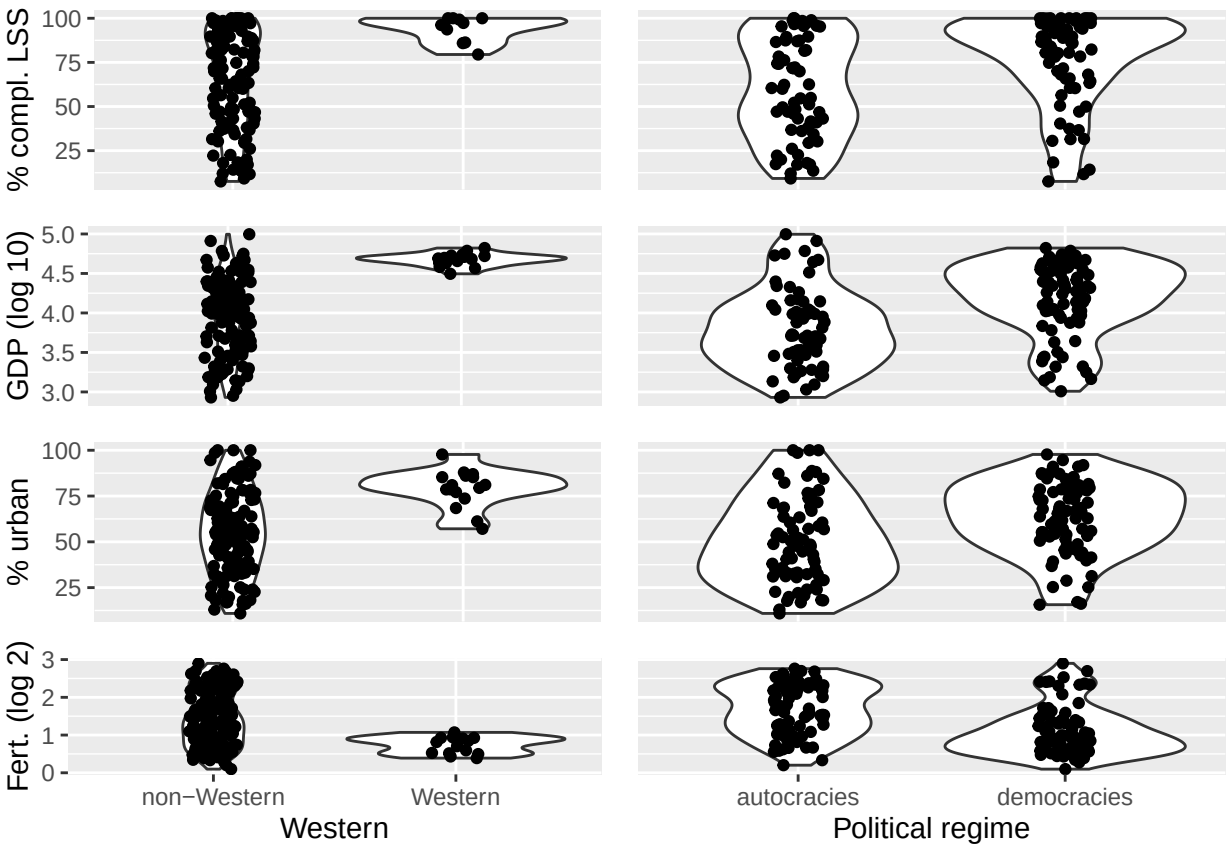

Figure 1: Evidence that values in the multidimensional WEIRD complex are only partially correlated (continued). Violin plots of continuous variables as a function of the two discrete variables: Western and type of political regime. Education is represented by proportion of the population completing lower secondary school; industrialization by percentage of the population living in urban (as opposed to rural) sites; richness by GDP per capita. In addition, we show women's average total fertility.

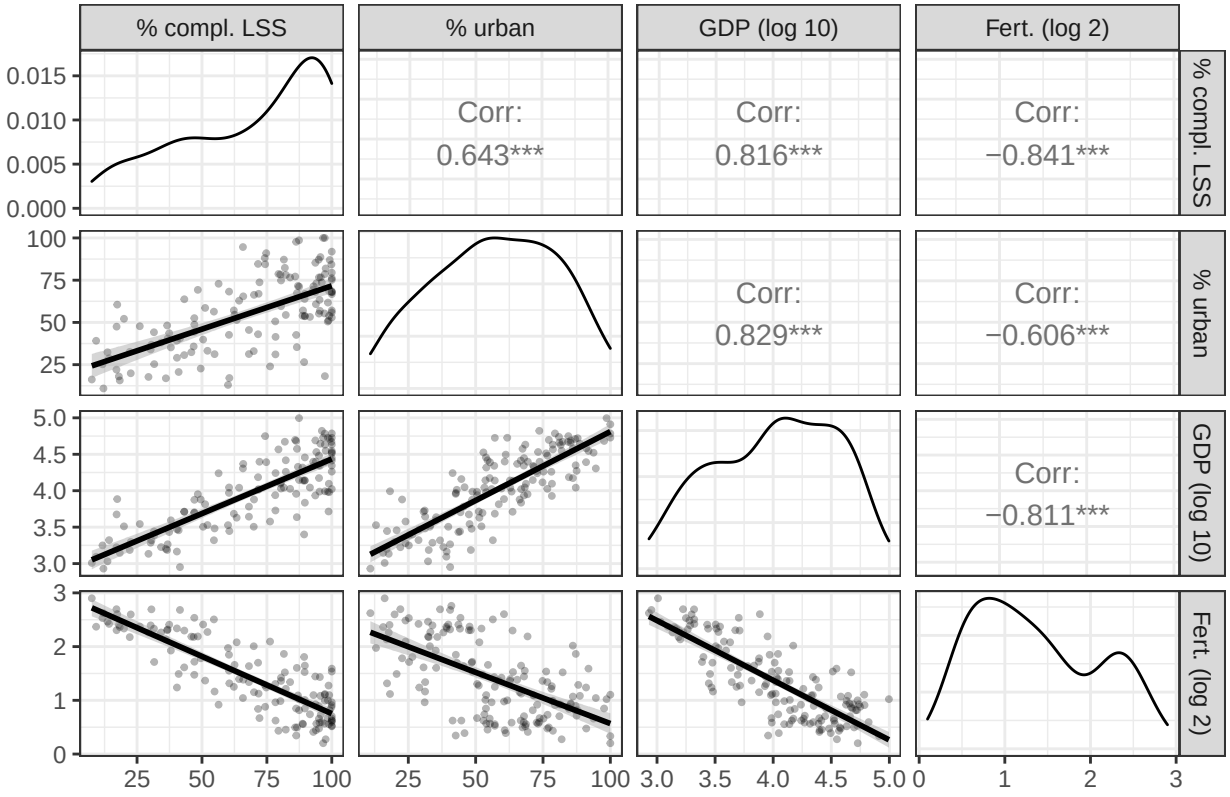

Figure 2: Evidence that values in the multidimensional WEIRD complex are only partially correlated, with a focus on continuous variables. The diagonal shows density of the distribution of each of the variables. Panels below the diagonal show the scatter plot for the two variables involved (e.g., GDP and proportion completed highschool for the second row, first column). Those above the diagonal show the Pearson correlation for the two variables involved. Education is represented by proportion of the population completing lower secondary school; industrialization by proportion of the population living in urban (as opposed to rural) sites; richness by GDP per capita. In addition, we show women's average total fertility.

Methods

Code to reproduce text on p. 16

A total of 321 corpora were initially considered. We excluded 10 because at least some of the children were not typically developing; 22 because data was collected in the lab or school; 9 because data was not available; 19 because only the child was transcribed; 11 because they were diary studies; 67 because speech was triggered by a task (elicitation, story-telling, etc.); 3 because the conversation involved exclusively unfamiliar adults. After these exclusions, 180 remained. The following analyses will continue only on these included corpora.

Code to reproduce information on Table 2

Code to reproduce information on Table 3

Results

Code to reproduce text on p. 20

The samples are varied in geographic terms, with corpora for every populated continent. Specifically, 3 corpora were collected in Africa; 32 in Asia; 73 in Western Europe, and a further 32 in Non-Western Europe; 34 in North America and 73 in Latin America. Only 1 was collected in Oceania.

Code to reproduce Figure 3

Note that this Figure appears in a different page.

Code to reproduce text on p. 20

Using country-level statistics, we were able to assess the extent to which the countries with data in CHILDES were a representative sample of countries in the world. Density plots are portrayed in Figure 3. The means for each of the variables differed for the countries in CHILDES versus overall in the world using unpaired samples t-tests without assuming equality of variance (Welch's t). Countries in CHILDES had a higher proportion of the population completing lower secondary school than the world wide sample (% compl. LSS, t(101.59)=-5.79, p = 0); they were more urban (% urban, t(79.77)=-3.51, p = 0); richer (log GDP per capita, t(103.29)=-6, p = 0) and had lower fertility rates (log fertility, t(117.09)=6.85, p = 0).

Code to reproduce text on p. 22

Education information was missing for more than half of the corpora (see Table 3); and 4 were described as diverse, without clarifying the range of education covered. Of the remaining 76 samples, 3 had at least some parents with primary-level education; 9 had parents with secondary school education as the lower bound of the education range, and a further 6 had some college as the lower bound. Thus, 76% of the samples (N = 58) portrayed children whose parents had at least a graduate, if not a postgraduate, degree. Samples were not representative of the countries they were collected, since in those same countries the proportion of the population with tertiary education was only 16%.

For socioeconomic status, there were 78 missing values (43%). Of the remaining 102 samples, 5 were described as having low SES; 16 were described as spanning both lower and middle or higher SES; and 81 were described as middle or higher SES exclusively. Given that most countries represented in CHILDES are in the Organization for Economic Cooperation and Development (OECD, 150 out of the 180 corpora), we can compare this proportion with the proportion of the population in these countries that are middle

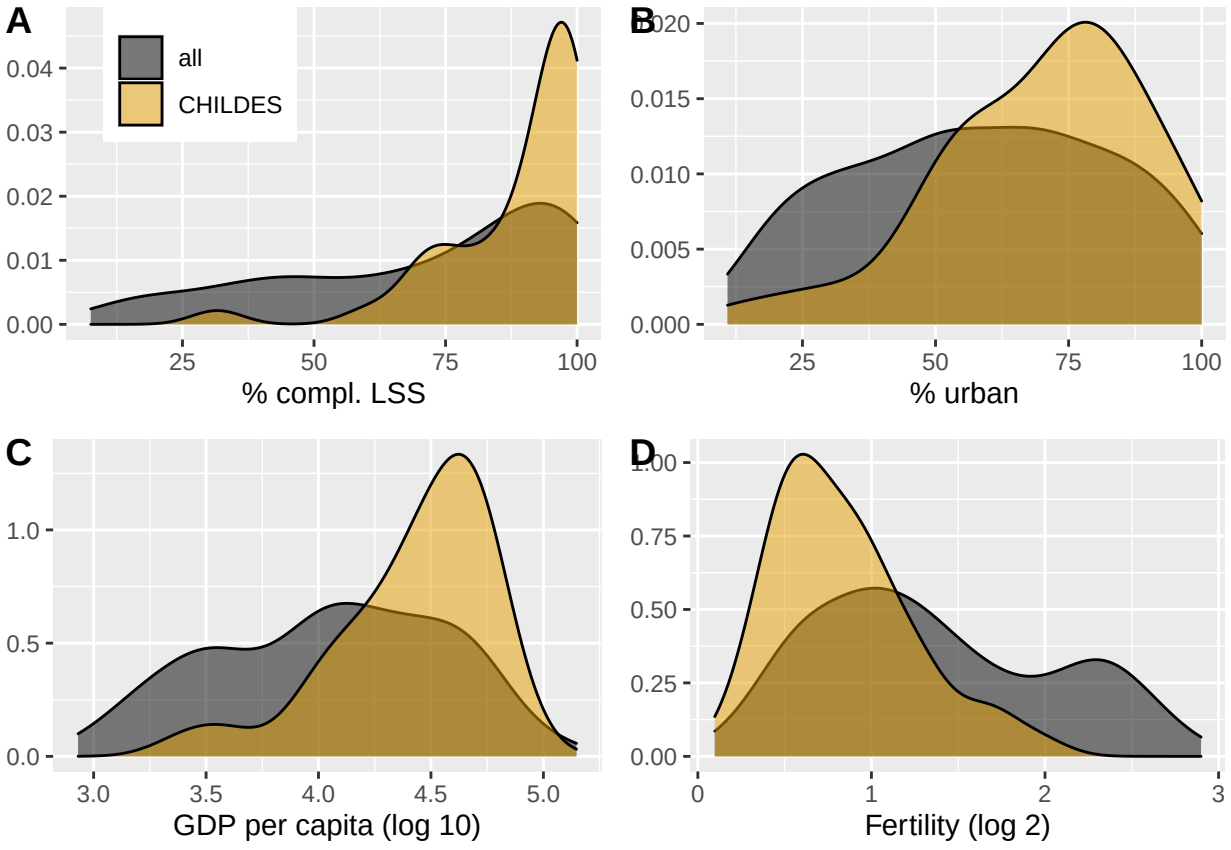

Figure 3: Figure 3: Density plots showing the distribution of country-level descriptors for all countries (dark gray) versus countries represented in CHILDES (orange). GDP stands for Gross Domestic Product. Percent completed lower secondary school is the percent of the country's population completing lower secondary school. Percent urban is the percent of the country's population residing in urban (as opposed to rural) locations. Total fertility rate is the average number of children a woman has over her whole reproductive period.

class. According to a 2016 report, "Almost two-thirds of people live in middle-income households in OECD countries", for whom "household net income [is] between 0.75 and 2 times the median". Thus, middle and higher class participants appear to be over-represented in CHILDES data, composing 81% of available data.

Information about parents' profession or activity was also missing for the majority of the corpora (see Table 3). Professions were overall varied, but it should be noted that 50% of the samples contained parents who were described as (Masters or PhD) students, professors, linguists, researchers, scientists, or academics. To give an idea of the extent to which this is not representative, consider the fact that in 2020, 6% of the American population would be included in that list of professions. Similar data is hard to find for all countries represented in CHILDES, but we suspect that the proportion of scientists, professors, Masters and PhD students found in most other countries will be the same or lower.

Only about a third of the samples had information about whether the community was rural or urban (see Table 3), and the remaining ones were very homogeneous. Setting aside 115 missing values (64%), and focusing on the remaining 65 samples, 58 were described as industrialized or urban, and an additional one as both rural and urban. Only 7 samples were described as farming or rural. In these same countries, the proportion of the population residing in urban settings was 76%, suggesting that samples were not representative of their countries in terms of rural versus urban settings either.

We then turned to the additional factors, starting with how varied language backgrounds were. A total of 62 different languages or language combinations (for bilingual and multilingual children) were reportedly spoken in the corpora. The samples in which only one language was reported spoke Afrikaans, Arabic (Egyptian or Kuwaiti), Basque, Cantonese, Catalan, Cree, Croatian, Czech, Danish, Dutch, English, Estonian, Farsi, French, German, Greek, Hebrew, Hungarian, Icelandic, Indonesian, Irish, Italian, Jamaican, Japanese, Korean, Mandarin, Norwegian, Nungon, Polish, Portuguese (Brazilian or European), Romanian, Russian, Serbian, Sesotho, Slovenian, Spanish, Swedish, Taiwanese, Tamil, Thai, Turkish, Welsh, and the samples in which multiple languages were reported spoke Catalan/Spanish, Dutch/English, Dutch/French, Dutch/Italian, English/Cantonese, English/Dutch, English/French, English/Hebrew, English/Japanese, English/Japanese/Danish, English/Mandarin, English/Mandarin/Cantonese, English/Russian, English/Spanish, French/Russian, German/Spanish, Hungarian/Catalan/Spanish, Hungarian/Farsi/English, Italian/German, Italian/Japanese, Portuguese/Swedish/English, Spanish/Catalan, Spanish/English, Spanish/Galician.

About a third (32%) of the included corpora that had available data for this variable (N = 110) were not monolingual. It is hard to find reliable estimates of the percentage of the population who are not monolingual in the world or in the countries represented in CHILDES, but for instance, in Europe in 2016, 65% of adults reported knowing multiple languages (Eurostat, 2022). According to such estimates, even if samples are linguistically diverse, it would appear that input data in CHILDES under-represents bilinguals and multilinguals.

As for family structure, 60 corpora (88% of those that had data for this variable) were based on nuclear families; in 7 extended families were portrayed; and in 1 one sample the structure was varied. We do not know of a country-level index that would allow us to check whether CHILDES corpora are representative of their countries for this variable.

A majority of corpora in CHILDES include children who have siblings. In fact, only 28% of samples (among the 93 corpora having information on siblings) were constituted exclusively by children with no siblings, and the remaining had at least one sibling, with the overall average being 0.8 siblings. Since 83% of countries in CHILDES are in the OECD, we draw a comparison point for such countries: 46% of children had no siblings in OECD countries according to 2015 data. In this sense, children with multiple siblings appear to be over-represented in CHILDES.

Discussion

Code to reproduce Table 4

Package and environment version

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] RColorBrewer_1.1-2 rjson_0.2.21       ggthemes_4.2.4     GGally_2.1.2
##  [5] plotly_4.10.0      ggpubr_0.4.0       kableExtra_1.3.4   stringr_1.4.0
##  [9] scales_1.1.1       purrr_0.3.4        tidyr_1.1.3        dplyr_1.0.7
## [13] ggplot2_3.3.5
##
## loaded via a namespace (and not attached):
##  [1] httr_1.4.2        splines_4.1.2     jsonlite_1.7.2    viridisLite_0.4.0
##  [5] carData_3.0-4     assertthat_0.2.1  highr_0.9         cellranger_1.1.0
##  [9] yaml_2.2.1        lattice_0.20-45   pillar_1.6.1      backports_1.2.1
## [13] glue_1.6.1        digest_0.6.27     ggsignif_0.6.3    rvest_1.0.0
## [17] colorspace_2.0-2  Matrix_1.3-4      cowplot_1.1.1     htmltools_0.5.2
## [21] plyr_1.8.6        pkgconfig_2.0.3   broom_0.7.8       haven_2.4.1
## [25] webshot_0.5.2     svglite_2.0.0     openxlsx_4.2.4    rio_0.5.27
## [29] tibble_3.1.2      mgcv_1.8-38       generics_0.1.0    farver_2.1.0
## [33] car_3.0-11        ellipsis_0.3.2    withr_2.4.2       lazyeval_0.2.2
## [37] magrittr_2.0.1    crayon_1.4.1      readxl_1.3.1      evaluate_0.14
## [41] fansi_0.5.0       nlme_3.1-153      rstatix_0.7.0     forcats_0.5.1
## [45] xml2_1.3.2        foreign_0.8-81    tools_4.1.2       data.table_1.14.0
## [49] hms_1.1.0         lifecycle_1.0.0   munsell_0.5.0     zip_2.2.0
## [53] compiler_4.1.2    systemfonts_1.0.2 rlang_0.4.11      grid_4.1.2
## [57] rstudioapi_0.13   htmlwidgets_1.5.4 labeling_0.4.2    rmarkdown_2.9
## [61] gtable_0.3.0      abind_1.4-5       DBI_1.1.1         reshape_0.8.9
## [65] curl_4.3.2        R6_2.5.0          knitr_1.33        fastmap_1.1.0
## [69] utf8_1.2.1        stringi_1.6.2     Rcpp_1.0.7        vctrs_0.3.8
## [73] tidyselect_1.1.1  xfun_0.24
```