Supplementary to How WEIRD is what we know about children's linguistic input from CHILDES? Appendix 1: Knitted analyses and supplementary analyses with a different implementation of Education

Contents

Load annotations

```
## [1] "Faroe Islands"
```

Introduction

Figures 1 & 2

The following graphs are based on 20 Western and 223 countries. We had data on political regime for 178 countries (from 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 was available for 212 countries (from WDI, 2011); on fertility for 199 countries (from OWID, 2011).

```
## pdf
##   2
```

```
## pdf
##   2
```

Table 2

```
##                                         country
##                                             147
##                                 Education.Naomi
##                                              70
##                                       SES.Naomi
##                                              85
##                             Parental.profession
##                                              72
##      Language.or.Languages.spoken.in.recordings
##                                             147
##                                        Minority
##                                             147
## Type.of.community.at.the.time.of.the.recordings
```

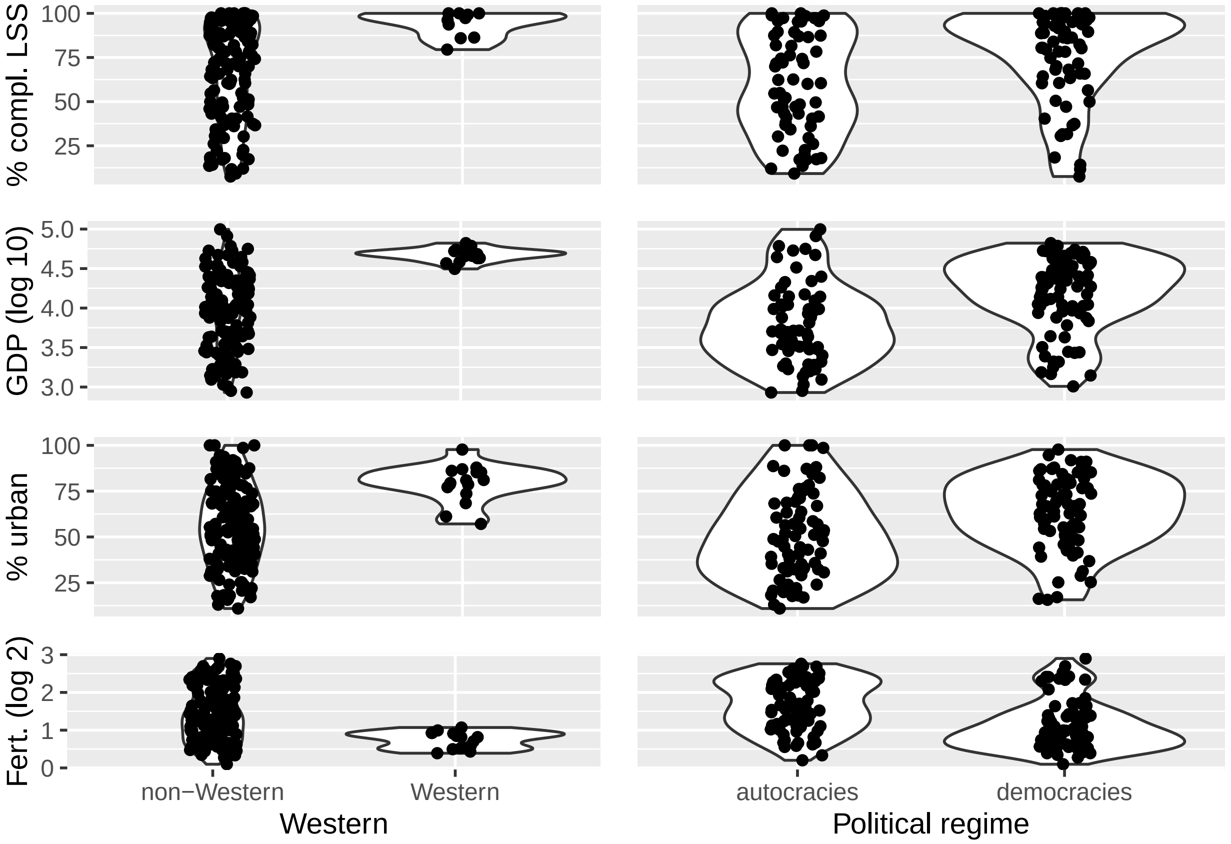

Figure 1: Evidence that values in the multidimensional WEIRD complex are only partially correlated (continued). Violin plots of continuous variables as a function of the two discrete variables: Western and type of political regime. Education is represented by proportion of the population completing lower 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.

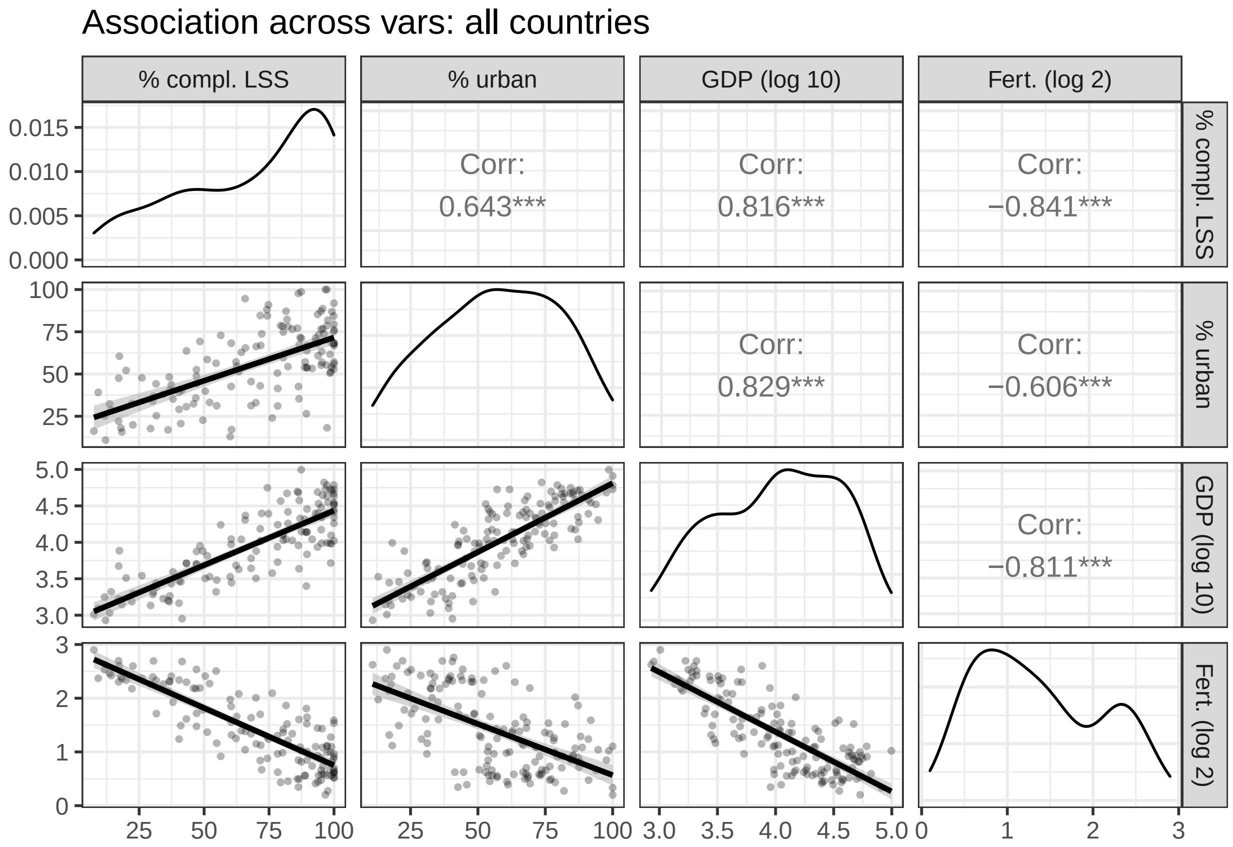

Figure 2: Evidence that values in the multidimensional WEIRD complex are only partially correlated, with a focus on continuous variables. The diagonal shows density of the distribution of each of the variables. Panels below the diagonal show the scatter plot for the two variables involved (e.g., GDP 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.

```
##                                            52
##                           Household.structure
##                                            59
##                      Proportion.With.Siblings
##                                            86
##                    Average.number.of.siblings
##                                            79
```

Text

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 (Taiwan, 1.07) to over 6 (Niger, 7.46).

Methods info

Results

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; 18 because only the child was transcribed; 11 because they were diary studies; 66 because speech was triggered by a task (elicitation, story-telling, etc.); 3 because the conversation involved exclusively unfamiliar adults. After these exclusions, 147 remained. The following analyses will continue only on these included corpora.

Comparison between corpora and world data in terms of education, income, urbanity and family size

The samples are varied in geographic terms, with corpora for every populated continent. Specifically, 3 corpora were collected in Africa; 30 in Asia; 66 in Western Europe, and a further 31 in Non-Western Europe; 11 in North America and 66 in Latin America. Only 1 was collected in Oceania.

The samples, however, over-represent richer countries, with relatively educated populations, living mostly in urban settings, and having relatively small families (Figure 3).

```
## pdf
##   2
```

Description of included families

We then investigated the extent to which samples themselves captured diverse families. We started with how varied language backgrounds were. All corpora could be coded in terms of which language or languages were spoken in the recordings. 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/French, Dutch/Italian, English/Cantonese, English/Dutch, English/French, English/Hebrew, English/Japanese, English/Mandarin, English/Mandarin/Cantonese, English/Russian, English/Spanish, French/Russian, German/Spanish, Hungarian/Catalan/Spanish, Hungarian/Farsi/English, Italian/German, Portuguese/Swedish/English, Spanish/Catalan, Spanish/English, Spanish/Galician. That said, for 50 corpora, we could not establish based on the description whether the corpora and/or the community from which the data were collected were mostly monolingual or not. Among the 97 corpora for which this could be established, 31% were not monolingual. It is hard to find reliable estimates of the percentage of the population who are multilingual in the world (which some estimate at 57%) or the countries represented in

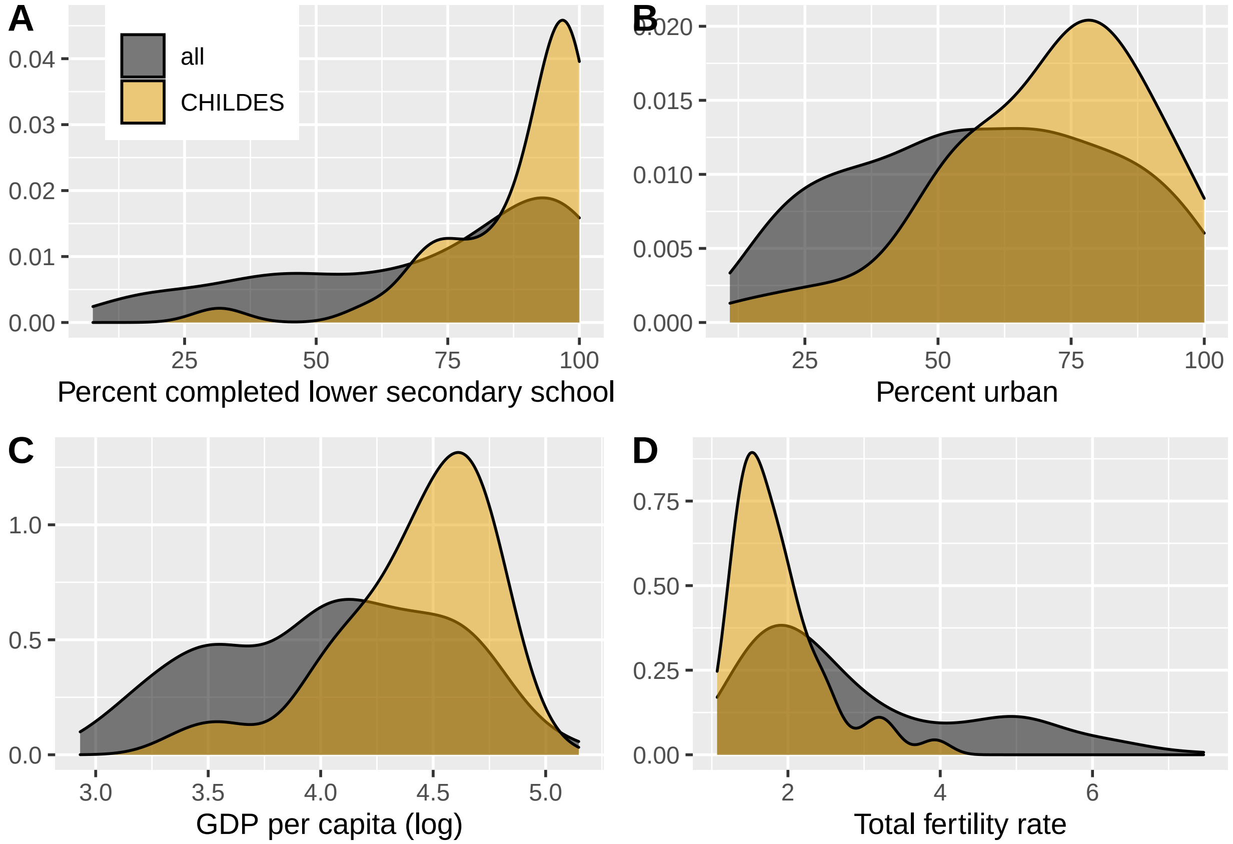

Figure 3: Figure 3: Density plots showing the distribution of country-level descriptors for all countries (dark gray) versus countries represented in CHILDES (orange). GDP stands for Gross Domestic Product. Percent completed lower secondary school is the percent of the country's population completing 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.

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.

Education information was missing for more than half of the corpora (see Table 2); and 3 were described as diverse, without clarifying the range of education covered. Of the remaining 67 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 3 had some college as the lower bound. Thus, the majority of the samples (N = 52, 78%) 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 15% in 2010.

For socioeconomic status, there were 62 missing values (42%). Of the remaining 85 samples, 3 were described as having low SES; 13 were described as spanning both lower and middle or higher SES; and 69 were described as middle or higher SES exclusively. Given that most countries represented in CHILDES are in the OECD (120 out of the 147 corpora), we can compare this proportion with the proportion of the population in these countries that are middle 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 a majority of available data.

Information about parents' profession or activity was also missing for the majority of the corpora (see Table 2). Professions were overall varied, but it should be noted that 49% 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, 47,000 people were scientists in the USA (including both in academia and the private sector), 135,000 were professors, 20 million were Masters students, and 3 million were PhD students. Given that the USA's population is 392 million, 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 lower.

Only about a third of the samples had information about whether the community was rural or urban (see Table 2), and the remaining ones were very homogeneous. Setting aside 95 missing values (65%), and focusing on the remaining 52 samples, 47 (90% of the samples for which this variable was available) were described as industrialized or urban, and one additional one as both rural and urban. Only 5 samples were described as farming or rural. In those same countries, the proportion of the population residing in urban settings was 76%) in 2011, suggesting that samples were not representative of their countries in terms of rural versus urban settings either.

As for household structure, there was a great deal of missing data, with only 59 (40%) of the samples specified. Among these, 51 (86%) were nuclear; in 7 (12%) were extended; and in 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.

We had information about whether the children had siblings or not for more than half of the samples (see Table 2). Out of the 86 samples with this information, most contained data from multi-child families. In fact, only 21 samples (24% of samples 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 0.91327 siblings. As a comparison point, 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. Note that this could emerge from a bias in reporting, with corpora creators more often reporting that there are siblings and their number when there are some than when there are none.

Discussion

Table 3. Low variability makes it difficult to tease apart empirically dimensions like family size and parental education: It may be reasonable to compare the input afforded to children who have at least one sibling (1+)

versus none for samples consisting of parents with a completed college education, but families with lower levels of education are too under-represented to allow a meaningful study. NA indicates missing values for this variable.

```
##
##                      None 1+ NA
##   Some primary          0  3  0
##   Some secondary        1  8  0
##   Some college          1  2  0
##   College and above    15 26 11
##   NA                    4 26 50
```

Additional analyses

Figures on education using WDI's proportion of the population completing high school (rather than Our world in data's proportion of the population completing lower secondary school).

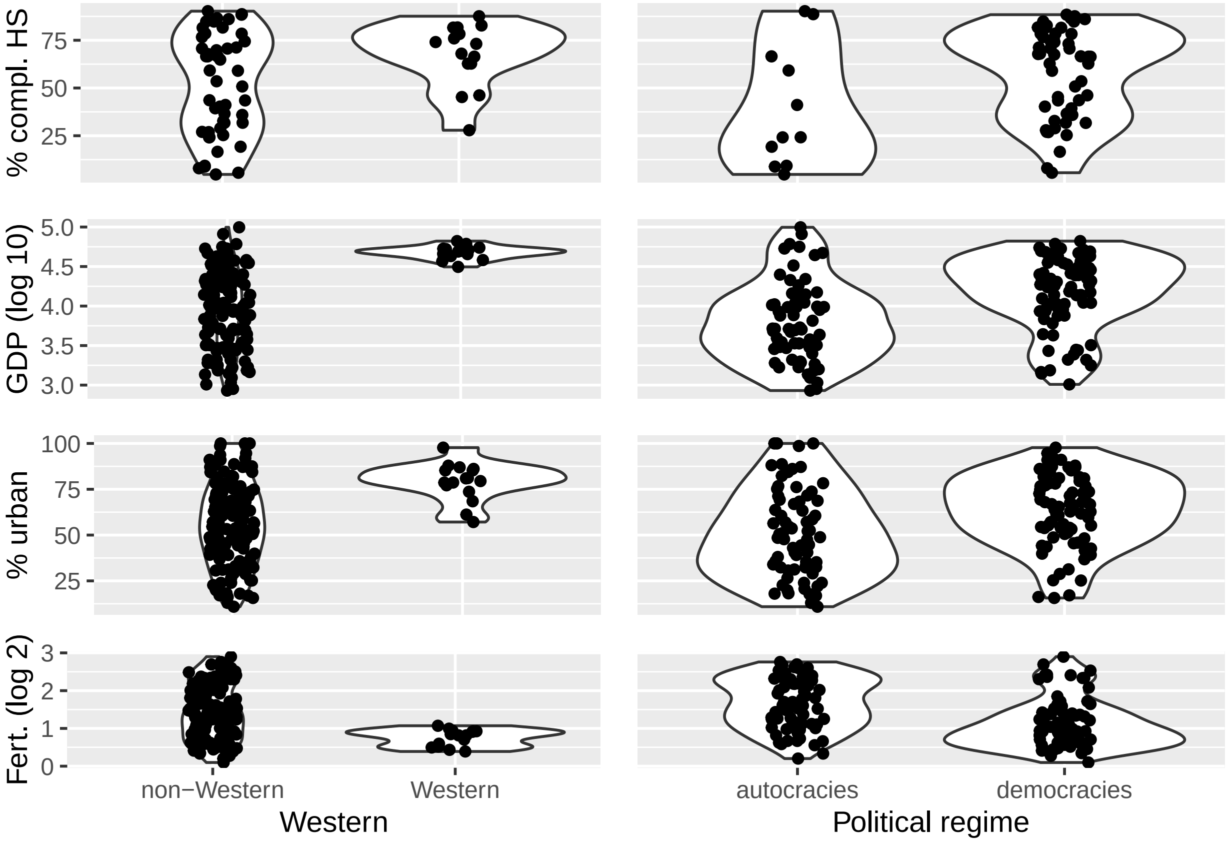

Figure 4: Equivalent to Fig 1, only the education variable has changed.

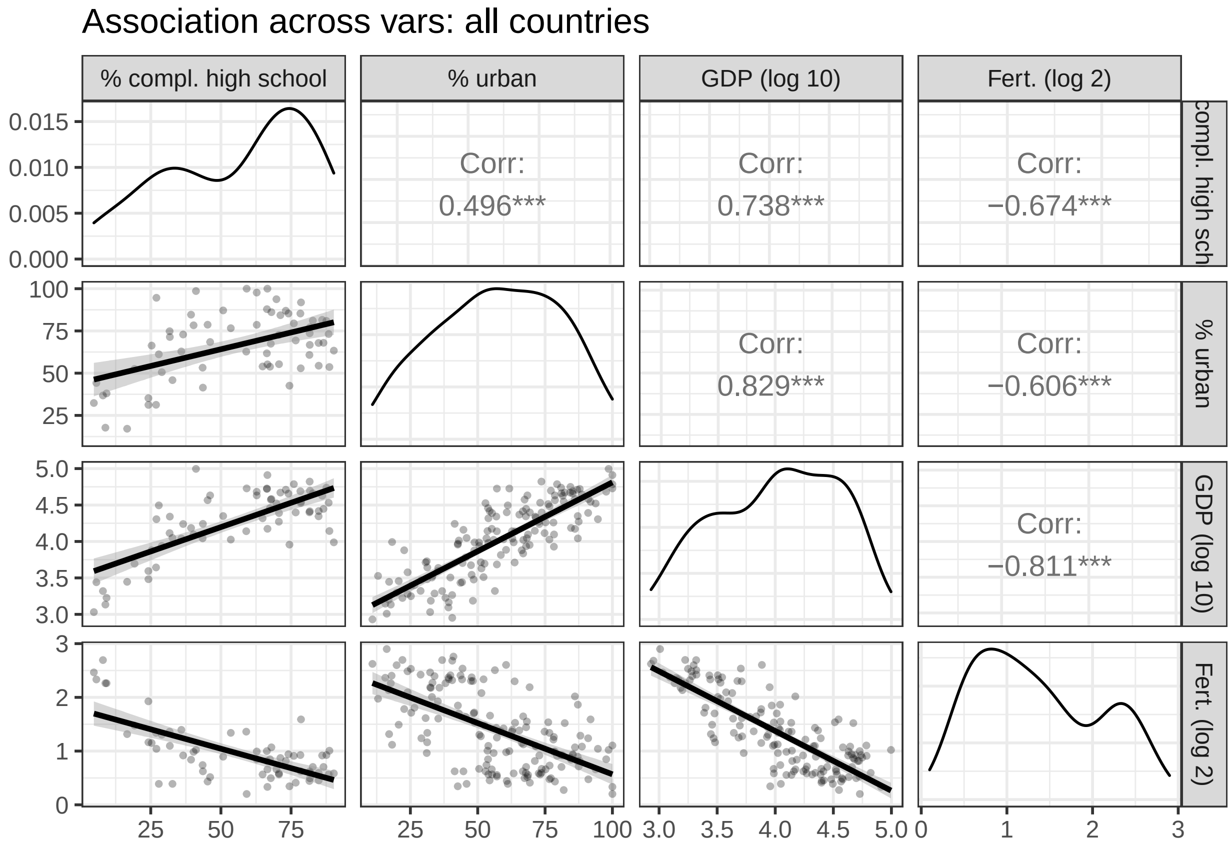

Figure 5: Equivalent to Fig 2, only the education variable has changed.

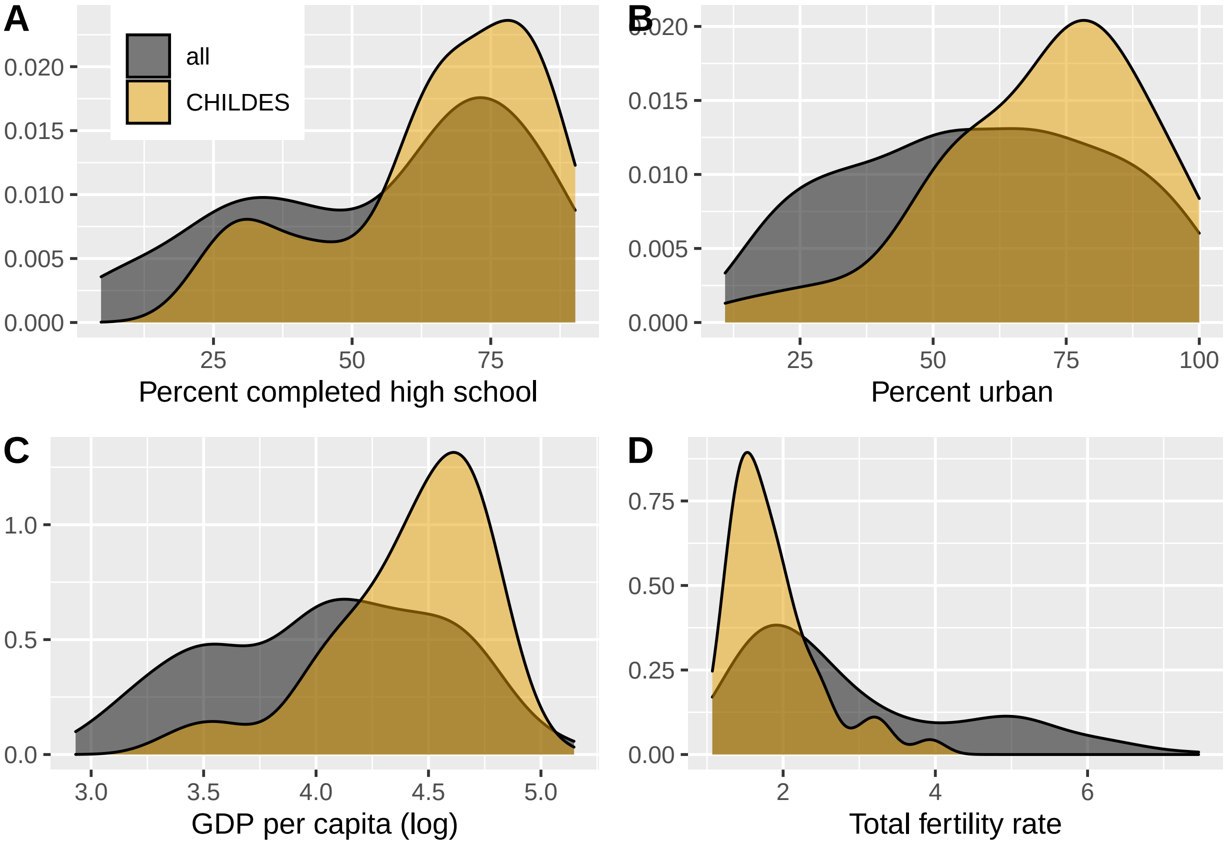

Figure 6: Equivalent to Fig 3, only the education variable has changed.