

## Overview

The data and code in this replication package reproduce all tables and figures in Hengel (2021). Raw data are in the 0-data/fixed directory. Transformed data, estimation results and figures are saved in the 0-data/generated, 0-tex/generated and 0-images/generated directories, respectively. The replication code, described in detail below, will take 21–22 hours to run.

The data in this replication package are publicly available and licensed under a Creative Commons Attribution 4.0 International License. See [LICENSE.txt](#) for details.

## Data

### Main dataset: read.db

Almost all figures and tables in Hengel (2021) were generated using the raw data in 0-data/fixed/read.db. read.db is an SQLite database of biblio- and biographic data on articles and authors published in the *American Economic Review* (AER), *Econometrica*, *Journal of Political Economy* (JPE), *Quarterly Journal of Economics* (QJE) and the *Review of Economic Studies* (REStud). read.db contains 11 tables. Their contents and provenance are described below. Figure 1 displays read.db's entity-relationship diagram; Table 1 describes each column. Please see Hengel (2021), Section 2 and Appendices C and D for additional details on data provenance and variable construction.

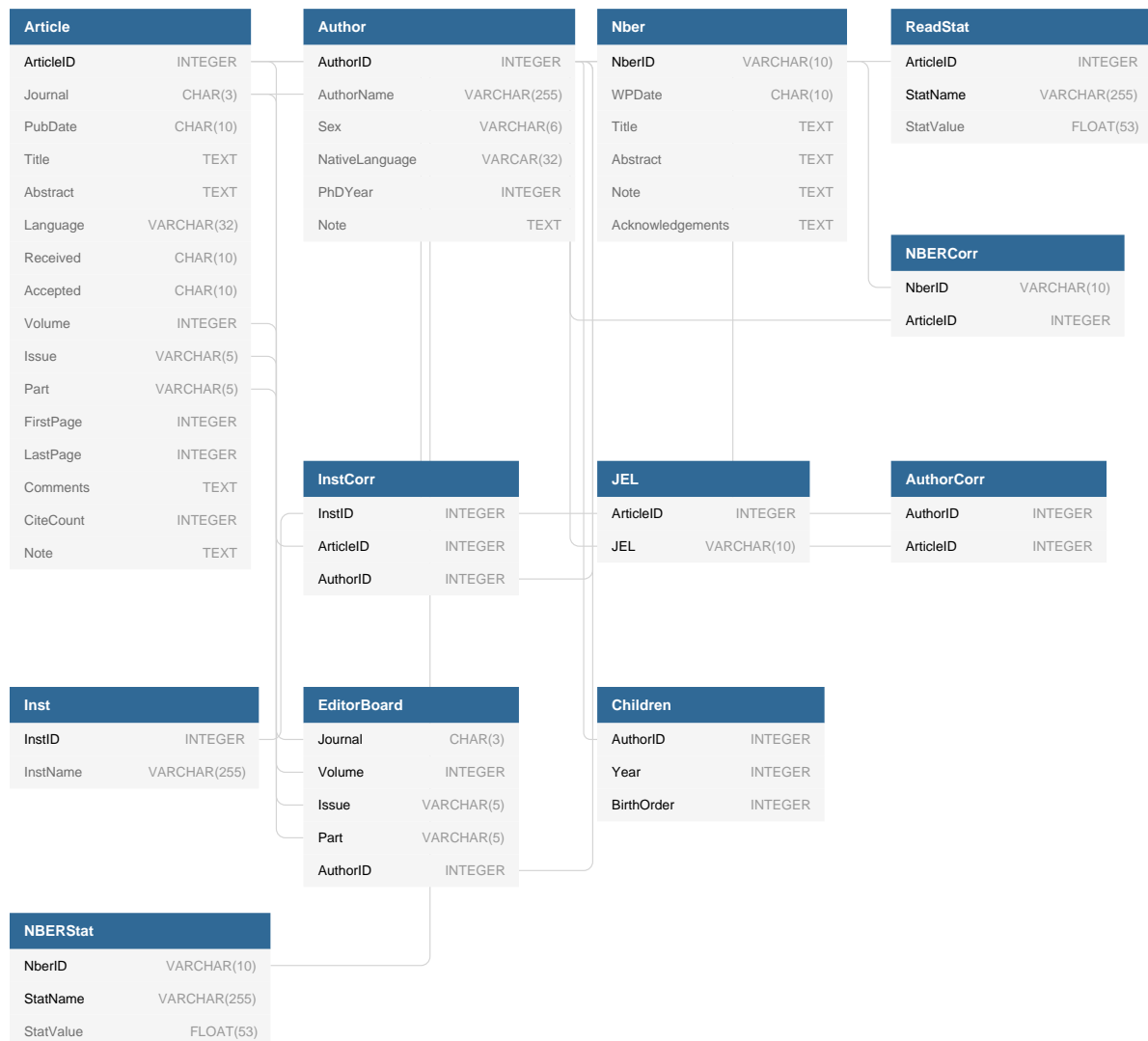


Figure 1: Entity-relationship diagram for read.db

- **Article.** The Article table contains bibliographic information from every English-language article published with an abstract in the *AER*, *Econometrica*, *JPE* and *QJE* between January 1950 and December 2015 (inclusive). It also contains data on *REStud* articles published with their submission and acceptance dates.

All data were collected from publicly available sources (*e.g.*, publishers' websites and JSTOR). The exception is citations which were obtained from [Web of Science](#) in September 2017 and January 2018. (*Citation data are proprietary to Web of Science and are included here for replication purposes only; please do not distribute these data or publish online.*) Data on submit-accept times and institutions were collected from journals' online archives, extracted from digitised articles using the open source command utility `pdftotext` or entered manually by me or a research assistant.

The column **Abstract** contains unicode textual data of articles' abstracts. When using the data, please ensure it is imported with the proper encoding (*e.g.*, `encoding("utf-8")` in Stata (14+) or `iconv(, from="macintosh", to="UTF-8")` in R).

- **Author.** The Author table contains biographic details on authors. Gender was initially assigned using [GenderChecker.com](#)'s database of male and female names. Three separate Mechanical Turk workers, a research assistant or I then manually verified them based on photos and other information found on faculty websites, Wikipedia articles, *etc.* In situations where the author could not be found, I emailed or telephoned colleagues and institutions associated with the author.

Authors were assumed to be native-English speakers if one or more of the following criteria were satisfied: (i) they were raised in an English-speaking country; (ii) they obtained all post-secondary education from English speaking institutions; or (iii) they spoke with no discernible non-native accent. This information was almost always found in authors' CVs, websites, Wikipedia articles, faculty bios or obituaries. In a small number of cases the criteria were ambiguously satisfied or not available; in these instances I asked friends and colleagues of the author or inferred English fluency from the author's first name, country of residence or surname (in that order). If one co-author on a paper was found to be a native English speaker, I did not necessarily check whether any of the other co-authors were also native English speakers.

- **AuthorCorr.** The AuthorCorr table maps **AuthorID** in Author to **ArticleID** in Article.
- **Children.** The Children table contains data on the year female authors with at least one exclusively female-authored paper published in *Econometrica* gave birth. This information was obtained from authors' published profiles, CVs, acknowledgements, Wikipedia, personal websites, Facebook pages, background checks and by consulting local school district/popular extra-curricular activity websites. Exact years were recorded whenever found; otherwise, they were approximated by subtracting a child's actual or estimated age from the date the source material was posted online. In several instances, I obtained or verified this information from acquaintances, friends and colleagues or by asking the woman directly. If an exhaustive search turned up no reference to children, I assumed the woman in question did not have any. Data only systematically collected for children potentially born during the time a woman had an exclusively female-authored paper under review at *Econometrica*.
- **EditorBoard.** The EditorBoard table contains the **AuthorID** for every editor and each issue of a journal. Editors were identified from issue mastheads.
- **Inst.** The Inst table maps each unique **InstID** to an institution name.
- **InstCorr.** The InstCorr table maps each (**ArticleID**, **AuthorID**) combination in AuthorCorr to at least one **InstID** in Inst.
- **JEL.** The JEL table maps *JEL* codes to each **ArticleID** in Article. The *JEL* system was significantly revised in the 1990s; because exact mapping from one system to another is not possible, I collected these data only for articles published post-reform. Codes were recorded whenever found in the text of an article or on the websites where bibliographic information was scraped. Remaining articles were classified using data from the American Economic Association's Econlit database.
- **NBER.** The NBER table contains basic bibliographic data on the NBER working papers that were eventually published in a top-four journal. Data were scraped from the [NBER website](#) or extracted from digitised working papers.

- **NBERCorr.** The NBERCorr table maps each `NberID` in NBER to at least one `ArticleID` in Article. (The mapping is not one-for-one because a small number of working papers were eventually published as multiple articles or combined into one.) Matches were identified using citation data from [RePEc](#) and by searching NBER’s database directly for unmatched papers authored by NBER family members. The column `Note` contains notes in situations where the matching process involved some degree of ambiguity (*e.g.*, because of title changes between draft and final versions of the paper).
- **ReadStat.** The ReadStat table contains readability statistics for every article with an abstract in Article. Readability scores were generated with the Python module `Textatistic` using the text in the `Abstract` column of the Article table. `Textatistic`’s code and documentation are available on [GitHub](#); a brief description is provided in Hengel (2021), Appendix D.3.
- **NberStat.** The NberStat table contains the readability statistics using `Abstract` text for every working paper in the NBER table. Statistics were generated using the Python module `Textatistic`.

Table 1: Description of variables in `read.db`

| Table       | Column name                 | Description   |
|-------------|-----------------------------|---|
| Article     | <code>ArticleID</code>      | Unique ID for each article  |
| Article     | <code>Journal</code>        | Journal   |
| Article     | <code>PubDate</code>        | Date of publication (YYYY-MM-DD)  |
| Article     | <code>Title</code>          | Title   |
| Article     | <code>Abstract</code>       | Abstract  |
| Article     | <code>Language</code>       | Language ( <i>e.g.</i> , English or French)                             |
| Article     | <code>Received</code>       | Date (YYYY-MM-01) submission received by the editorial office           |
| Article     | <code>Accepted</code>       | Date (YYYY-MM-01) manuscript accepted by the editorial office           |
| Article     | <code>Volume</code>         | Volume  |
| Article     | <code>Issue</code>          | Issue   |
| Article     | <code>Part</code>           | Part  |
| Article     | <code>FirstPage</code>      | First page  |
| Article     | <code>LastPage</code>       | Last page   |
| Article     | <code>CiteCount</code>      | Citation count (Web of Science)   |
| Author      | <code>AuthorID</code>       | Unique ID for each author   |
| Author      | <code>AuthorName</code>     | Author name   |
| Author      | <code>Sex</code>            | Gender  |
| Author      | <code>NativeLanguage</code> | English or non-English native speaker                                   |
| AuthorCorr  | <code>AuthorID</code>       | Unique ID for each author (maps to Author table)                        |
| AuthorCorr  | <code>ArticleID</code>      | Unique ID for each article (maps to Article table)                      |
| Children    | <code>AuthorID</code>       | Unique ID for each author (maps to Author table)                        |
| Children    | <code>Year</code>           | Year a child was born   |
| Children    | <code>BirthOrder</code>     | Order of birth for children born in the same year                       |
| EditorBoard | <code>AuthorID</code>       | Unique ID for each editor (maps to Author table)                        |
| EditorBoard | <code>Journal</code>        | Journal   |
| EditorBoard | <code>Part</code>           | Part  |
| EditorBoard | <code>Volume</code>         | Volume  |
| EditorBoard | <code>Issue</code>          | Issue   |
| Inst        | <code>InstID</code>         | Unique ID for each institution  |
| Inst        | <code>InstName</code>       | Institution name  |
| InstCorr    | <code>InstID</code>         | Unique ID for each institution (maps to Inst table)                     |
| InstCorr    | <code>ArticleID</code>      | Unique ID for each article (maps to Article table)                      |
| InstCorr    | <code>AuthorID</code>       | Unique ID for each editor (maps to Author table)                        |
| JEL         | <code>ArticleID</code>      | Unique ID for each article (maps to Article table)                      |
| JEL         | <code>JEL</code>            | Tertiary <i>JEL</i> code  |
| NBER        | <code>NberID</code>         | Unique ID for each NBER working paper                                   |
| NBER        | <code>WPDate</code>         | Date manuscript was released as a working paper (YYYY-MM-DD)            |
| NBER        | <code>Title</code>          | Title   |
| NBER        | <code>Abstract</code>       | Abstract  |
| NBER        | <code>Note</code>           | Note on observation and/or matching process with <code>ArticleID</code> |

| Table    | Column name | Description  |
|----------|-------------|--|
| NBERCorr | NberID      | Unique ID for each NBER working paper (maps to NBER table) |
| NBERCorr | ArticleID   | Unique ID for each article (maps to Article table)         |
| ReadStat | ArticleID   | Unique ID for each article (maps to Article table)         |
| ReadStat | StatName    | Name of statistic ( <i>e.g.</i> , flesch)                  |
| ReadStat | StatValue   | Value of statistic   |
| NBERStat | NberID      | Unique ID for each NBER working paper (maps to NBER table) |
| NBERStat | StatName    | Name of statistic  |
| NBERStat | StatValue   | Value of statistic   |

## Other datasets

A small number of figures and tables are generated from data contained in `introduction_text.txt`, `correlations.txt` and `JEL.csv`. Their contents and provenance are described below and in Table 2.

- **introduction\_text.txt**. This file contains the first paragraph of text to come after a heading explicitly titled “Introduction” in NBER working papers eventually published in top-four journals. Data are used to generate Figure D.2 in Appendix D.2 in Hengel (2021). Textual data kindly provided by Henrik Kleven and Data Scott.
- **correlations.txt**. This file contains coefficients of correlations between the five readability scores used in Hengel (2021) and alternative measures of text difficulty. These figures are from the studies listed in Appendix D.4. They are used to produce the top graphic of Figure D.1 in Appendix D.1.
- **JEL.csv**. The file `JEL.csv` categorises all tertiary *JEL* codes as either theory/methodology, empirical or other. Categorisation was done manually by me. Data in `JEL.csv` are used to generate Table C.1 and construct the theory/methodology, empirical and other dummies described in Appendix C.

Table 2: Description of variables in other datasets

| File name                          | Column name | Description  |
|------------------------------------|-------------|--|
| <code>introduction_text.txt</code> | NberID      | Unique ID for each NBER working paper                      |
| <code>introduction_text.txt</code> | Text        | First paragraph of text                                    |
| <code>correlations.txt</code>      | StatName    | Name of readability statistic                              |
| <code>correlations.txt</code>      | Correlation | Coefficient of correlation                                 |
| <code>correlations.txt</code>      | Test        | Name of alternative measure of text difficulty             |
| <code>correlations.txt</code>      | TestType    | Type of alternative measure                                |
| <code>correlations.txt</code>      | Source      | BibTeX label for source study ( <i>e.g.</i> , SurnameYYYY) |
| <code>correlations.txt</code>      | Note        | Notes on calculations, <i>etc.</i>                         |
| <code>JEL.csv</code>               | JEL         | Tertiary <i>JEL</i> code                                   |
| <code>JEL.csv</code>               | Description | Long name of <i>JEL</i> code                               |
| <code>JEL.csv</code>               | Type        | Classification (empirical, theory/methodology or other)    |

## Code

To generate all figures and tables in Hengel (2021), first download the replication package, expand it and navigate to project’s root directory. Then execute the following four steps.

1. Run `1-update-textatistic.py` in Python.
2. Run `2-update-readability.R` in R.
3. Run `3-master.do` in Stata.
4. Execute `Figure-3.nb` and `Figure-G.2.nb` (both in the `0-code/output` directory) in Mathematica.

Each step can be executed individually by following the steps outlined below. Alternatively, the Bash script `4-master.sh` completes all four steps automatically.

To run `4-master.sh`, install the latest version of [WolframScript](#) and follow the instructions under the `1-update-textatistic.py` and `3-master.do` headings for installing `Textatistic` and an SQLite driver. Then, navigate to the project’s root directory and issue the following command in a Bash shell:

```
sh 4-master.sh
```

4-master.sh was last run on X August 2021 on a 4-core Intel-based iMac running MacOS version 11.5. Computation took X hours, X minutes and X seconds to run.

### 1-update-textatistic.py

The Python script 1-update-textatistic.py calculates readability scores for every article and NBER working paper with an abstract in read.db and updates its ReadStat and NBERStat tables with the results. More details on the Textatistic program are available on [GitHub](#). Documentation on how it calculates readability scores are available at [erinhengel.com](#).

For 1-update-textatistic.py to work, you must first install the Python package Textatistic. If you're lucky, this can be done in a single step by issuing the following command in your terminal application:

```
pip install textatistic
```

But you're probably not lucky. The problem is the PyHyphen dependency; for reasons I do not fully understand, pip does not always properly install it before trying to install Textatistic.

If you encounter this error, you will need to install both PyHyphen and Textatistic from source. But don't panic; it's not that hard. Just navigate to the project's root directory and issue the following sequence of commands in your terminal application.

```
cd "0-code/programs/Textatistic/required_packages/PyHyphen-2.0.5/"
sudo python setup.py install
cd ../../
sudo python setup.py install
```

Once Textatistic has been properly installed, navigate back to the project's root directory and run the following command in the terminal application.

```
python 1-update-textatistic.py
```

You will be alerted when the ReadStat and NBERStat tables in read.db have been successfully updated with the newly calculated readability statistics.

1-update-textatistic.py was last run on 13 August 2021 on a 4-core Intel-based iMac running MacOS version 11.5. Computation took 1 minute and 59 seconds to run.

### 2-update-readability.R

The R script 2-update-readability.R (R version 4.1.0) calculates readability scores using the R [readability package](#). To run it, open R, set the current working directory as the project directory and issue the following command:

```
source("2-update-readability.R")
```

2-update-readability.R first installs the latest version of RSQLite (version 2.2.7), tidyverse (version 1.3.1), haven (version 2.4.3) and pacman (version 0.5.1) from CRAN and readability from [GitHub](#). It then connects to the read.db database, fetches published article and NBER abstracts, calculates readability scores and exports the results to readstat.dta and nberstat.dta in the 0-data/generated directory. Finally, it reads in introduction\_text.txt, calculates readability scores and exports the result to articlestat.dta, also in the 0-data/generated directory.

2-update-readability.R was last run on 13 August 2021 on a 4-core Intel-based iMac running MacOS version 11.5. Computation took 1 minute and 26 seconds to run.

### 3-master.do

The Stata script master.do (Stata version 15.1) generates all figures and tables in Hengel (2021) with the exception of Figure 3 and Figure G.2 (see below).

To run `master.do`, first install an SQLite driver—I use the open source driver from [Actual Technologies](#)—and update the file path in line 35 accordingly. Then, open a Stata terminal, navigate to the project’s root directory and issue the following command:

```
do 3-master.do
```

`3-master.do` first installs several third-party packages from SSC (`ftools`, `estout`, `psmatch2`, `xtabond2`, `listtex`, `reghdfe`, `binscatter`, `distinct`, `labutil` and `coefplot`) and `wordwrap` from [GitHub](#). It then copies the `ado`, `scheme`, `colors` and `estout` definition files in the `0-code/programs/stata` directory into your Stata personal `ado` directory. (Alternatively, manually load these files into Stata before running `3-master.do` and comment out lines 26–29.) It then transforms the raw data (results are saved in `0-data/generated`) and executes the Stata `do` files in the `0-code/output` directory. Estimation results are either saved as LaTeX output in the `0-tex/generated` directory or as image files in the `0-images/generated` directory. A log of all output is saved in the `0-log` directory as `YYYY-MM-DD-HH-MM-SS.smcl`.

`3-master.do` was last run on 15 August 2021 on a 4-core Intel-based iMac running MacOS version 11.5. Computation took 21 hours, 1 minute and 19 seconds to run.

### Create Mathematica graphs

Figures 3 and G.2 in Hengel (2021) were created using Mathematica (version 12.1). To generate them, follow the three steps below:

1. Navigate to the `0-code/output` directory and open the files `Figure-3.nb` and `Figure-G.2.nb` in Mathematica.
2. Select “Evaluate Notebook” from the Evaluation dropdown menu.
3. Click “Yes” to run the initialisation cells.

`Figure-3.nb` generates `Figure-3.png`; `Figure-G.2.nb` generates `Figure-G.2.png`. Both files are saved in the `0-images/generated` directory.

`Figure-3.nb` and `Figure-G.2.nb` were last run on 13 August 2021 on a 4-core Intel-based iMac running MacOS version 11.5. Combined computation took less than 4 seconds.

### References

Hengel 2021. “Publishing while female: Are women held to higher standards? Evidence from peer review.” Mimeo.