Overview

The data and code in this replication package reproduce all tables and figures in Hengel (2021). Raw data are contained in the `0-data/fixed` directory; transformed data are found in the `0-data/generated` directory. Estimation results are found in `0-tex/generated` and `0-images/generated`. The replication code, described in detail below, will take 8–12 hours to run.

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 bibliographic and author information for articles published in top-five economics journals. It contains 11 tables. Their contents and provenance are described below; Table 1 describes each column; Figure 1 displays `read.db`'s entity-relationship diagram. Please see Hengel (2021), Section 2 and Appendices C and D for additional information on data provenance and variable construction.

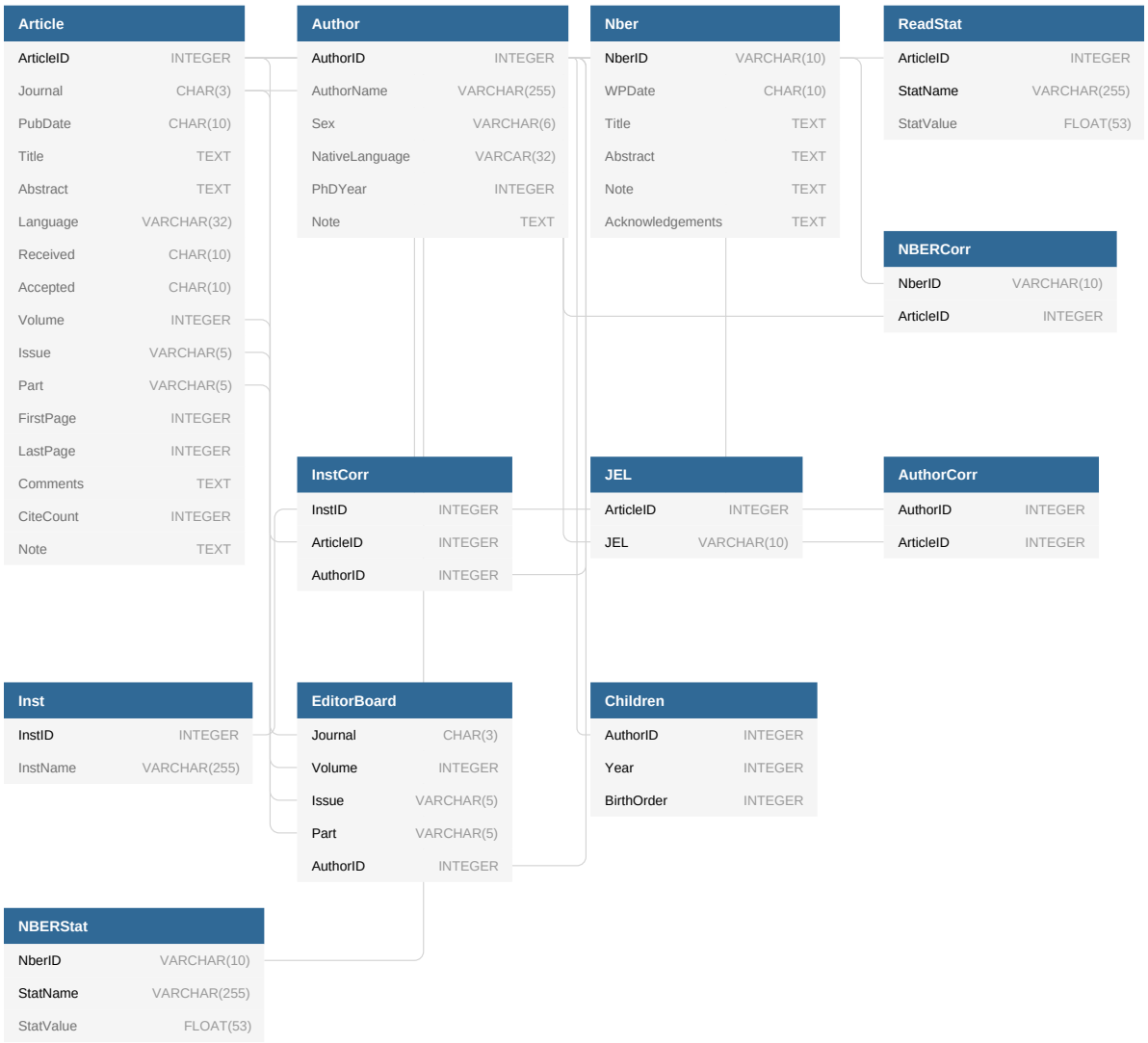

Figure 1: Entity-relationship diagram for `read.db`

- **Article**. The Article table contains bibliographic information from every English-language article published with an abstract in the *American Economic Review*, *Econometrica*, *Journal of Political*

Economy and the *Quarterly Journal of Economics* between January 1950 and December 2015 (inclusive) and *Review of Economic Studies* articles published with submit and accept 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. 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.

- **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 form 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` of each editor for every issue of a top-five journal. I recorded editor/editorial board member names 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. Matches were identified using citation data from RePEc and by searching NBER's database directly for unmatched papers authored by NBER family members.

- **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	`ArticleID`	Unique ID for each article
Article	`Journal`	Journal
Article	`PubDate`	Year of publication (YYYY-08-01)
Article	`Title`	Title
Article	`Abstract`	Abstract
Article	`Language`	Language (*e.g.*, English or French)
Article	`Received`	Date (YYYY-MM-01) submission received by the editorial office
Article	`Accepted`	Date (YYYY-MM-01) manuscript accepted by the editorial office
Article	`Volume`	Volume
Article	`Issue`	Issue
Article	`Part`	Part
Article	`FirstPage`	First page
Article	`LastPage`	Last page
Article	`CiteCount`	Citation count (Web of Science)
Article	`Note`	Note on observation
Author	`AuthorID`	Unique ID for each author
Author	`AuthorName`	Author name
Author	`Sex`	Gender
Author	`NativeLanguage`	English or non-English native speaker
AuthorCorr	`AuthorID`	Unique ID for each author (maps to Author table)
AuthorCorr	`ArticleID`	Unique ID for each article (maps to Article table)
Children	`AuthorID`	Unique ID for each author (maps to Author table)
Children	`Year`	Year a child was born
Children	`BirthOrder`	Order of birth for children born in the same year
EditorBoard	`AuthorID`	Unique ID for each editor (maps to Author table)
EditorBoard	`Journal`	Journal
EditorBoard	`Part`	Part
EditorBoard	`Volume`	Volume
EditorBoard	`Issue`	Issue
EditorBoard	`Part`	Part
Inst	`InstID`	Unique ID for each institution
Inst	`InstName`	Institution name
InstCorr	`InstID`	Unique ID for each institution (maps to Inst table)
InstCorr	`ArticleID`	Unique ID for each article (maps to Article table)
InstCorr	`AuthorID`	Unique ID for each editor (maps to Author table)
JEL	`ArticleID`	Unique ID for each article (maps to Article table)
JEL	`JEL`	*JEL* code
NBER	`NberID`	Unique ID for each NBER working paper
NBER	`WPDate`	Date manuscript was released as a working paper (YYYY-MM-DD)
NBER	`Title`	Title
NBER	`Abstract`	Abstract
NBER	`Note`	Note on observation
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 (*e.g.*, 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**. The file introduction_text.txt contains the first paragraph of text to come after a heading explicitly titled "Introduction" in NBER working papers eventually published in a top-four journal. Data are used to generate Figure D.2 in Appendix D.2 in Hengel (2021). Textual data were transcribed from pdfs by Henrik Kleven and Data Scott.
- **correlations.txt**. The file correlations.txt contains coefficients of correlations between the five readability scores used in Hengel (2021) and alternative measures of text difficulty. These figures were identified by me or a research assistant in the studies listed in Appendix D.4 and are used to produce the top graphic of Figure D.1 in Appendix D.1.
- **JEL.csv**. The file JEL.csv contains a list of 859 tertiary *JEL* codes manually classified by me as either theory/methodology, empirical or other. It is 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
introduction_text.txt	`NberID`	Unique ID for each NBER working paper
introduction_text.txt	`Text`	First paragraph of text
correlations.txt	`StatName`	Name of readability statistic
correlations.txt	`Correlation`	Coefficient of correlation
correlations.txt	`Test`	Name of alternative measure of text difficulty
correlations.txt	`TestType`	Type of alternative measure
correlations.txt	`Source`	Label of source study
correlations.txt	`Note`	Notes on calculations, etc.
JEL.csv	`JEL`	Tertiary *JEL* code
JEL.csv	`Description`	Long name of *JEL* code
JEL.csv	`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` will execute all four steps for you. To run it, install `Textatistic` and an SQLite Driver (see instructions under the `1-update-textatistic.py` and `3-master.do headings` below) as well as the latest version of WolframScript. 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 by issuing the following command in your terminal application:

```
pip install textatistic
```

But you probably won't be lucky. `Textatistic` requires the `PyHyphen` dependency, which, for reasons I do not understand, `pip` does not always properly download before installing `textatistic`. So if you can't get `textatistic` to install with `pip`, then you'll need to install both `PyHyphen` and `Textatistic` from source. To do this, 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.

`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, you can simply 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 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.

Create Mathematica graphs

Figures 3 and G.2 in Hengel (2021) were created using Mathematica (version 12.1.0.0). 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. Change the `fpath` variable in the grey box at the top of each notebook to point to the project's root directory.
3. Select "Evaluate Notebook" from the Evaluation dropdown menu. (Be sure to 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.