Replication package to “Cognitive Skills and Economic Preferences in the Fund Industry”

The code in this replication package constructs all the figures and tables of the paper using Stata. Some of the data in this package exactly corresponds to the data used in the paper, and hence the corresponding tables and figures are exactly reproduced. Another part of the data in this package is synthetic, and thus the results in the corresponding tables are not exactly the same as those in the paper. See the details later in this document.

## Data Sources

Three data sources are used in the paper. First, measures of cognitive skills and economic preferences of mutual fund managers were obtained via online experiments conducted by the authors. This source is referred to as the *experimental data*. The second and third sources of data are *Morningstar Direct* and *Lipper*. These sources provide empirical data on the funds managed by the participants of the online experiments. The data from Morningstar Direct and Lipper cannot be made publicly available. There are two reasons for this. First, these sources are subscription-based (fee-paying) services. Second, and more importantly, the participants of the experiments were ensured that their identity, or the identity of the funds they manage would not be traceable or disclosed publicly through any publications that rely on the data collected during the experiment. The direct publication of the fund-related data would make the funds in the sample identifiable. We provide synthetic data for the variables that are derived from the empirical sources (see the details under section “Construction of synthetic data”).

The following three subsections provide details for the data sources used.

### Experimental data

The experimental data were collected by the authors. The experiment has been conducted online using oTree (Chen et al., 2016). The software, including all instructions as used for the data collection, is available for download as a zipped oTree project at <https://osf.io/dq3t8/> and as a live demo version via <https://fea-2018-en.herokuapp.com>. Details on the experimental tasks are available in section A.2. of the Online Appendix. In case of this data source, the replication package contains the original data used in the paper.

Datafile:

* data\_experimental.dta

### Morningstar data

The main source of empirical data on mutual funds is *Morningstar Direct*. Morningstar Direct is a subscription based (fee-paying) service (<https://www.morningstar.com/products/direct>), and therefore it is not freely available to the public. The replication package provides synthetic data on variables derived from this data source.

Datafiles:

* data\_funds2017.dta
* data\_monthly.dta
* data\_halfyearly.dta

### Lipper data

The empirical data on mutual funds is complemented with data from *Lipper*. Lipper is a subscription based (fee-paying) service of Refinitv (<https://www.refinitiv.com/en/financial-data/fund-data/lipper-fund-data>), and therefore it is not freely available to the public. The replication package provides synthetic data on variables derived from this data source.

Datafiles:

* data\_monthly.dta
* data\_halfyearly.dta

## 

## Dataset list

The following table lists the data files provided as part of this replication package.

|  |  |  |
| --- | --- | --- |
| Data file | Source | Notes |
| data\_experimental.dta | Authors’ collection (experiment) |  |
| data\_funds2017.dta | Morningstar | Synthetic data |
| data\_monthly.dta | Morningstar and Lipper | Synthetic data |
| data\_halfyearly.dta | Morningstar and Lipper | Synthetic data |

The file data\_experimental.dta contains the experimental data. The data contain the original values used in the paper. The variables are listed in the following table.

Variables in the dataset data\_experimental.dta

|  |  |
| --- | --- |
| Variable name | Variable label |
| pid | Manager identifier |
| exper | Experience in industry (years) |
| crt\_raw | Cognitive Refelection Test score (raw) |
| crt | Cognitive Refelection Test score (standardized) |
| apm\_raw | Advanced Progressive Matrices score (raw) |
| apm | Advanced Progressive Matrices score (standardized) |
| tom\_raw | Theory of Mind score (raw) |
| tom | Theory of Mind score (standardized) |
| competitive\_raw | Competitiveness score (raw) |
| competitive | Competitiveness score (standardized) |
| risk | Risk tolerance score |
| loss | Loss tolerance score |
| ambiguity | Ambiguity tolerance score |
| patience | Time preference score |

The file data\_funds2017.dta contains data describing the funds domiciled in the four sample countries at the end of 2017. The variables are listed in the following table. This file contains synthetic data, created by adding random noise to the original variables. In particular, the variables Tenure, Ret, and AUM are transformed by adding normally distributed, , random noise, with used for Tenure, for AUM, and for Ret. The above transformed variables are also winsorized.

Variables in the dataset data\_funds2017.dta

|  |  |
| --- | --- |
| Variable name | Variable label |
| Category | Fund investment category |
| MSRating | Fund's Morningstar rating |
| Tenure | Manager's tenure at the fund on Dec 2017 |
| AUM | Assets under management on Dec 2017 (m EUR) |
| Ret | Fund's return in 2017 (%) |
| Team | =1 if team managed fund |
| Sample | =1 if the fund enters our sample |

The files data\_monthly.dta and data\_halfyearly.dta contain fund-related variables that are derived from raw data obtained from Morningstar and Lipper. The frequency of the former dataset is monthly, while that of the latter is half-yearly. The variable construction is described in section B of the Online Appendix. These files contain synthetic data; see the details of constructing the synthetic data under section “Construction of synthetic data”. The variables are listed in the following two tables.

Variables in the dataset data\_monthly.dta

|  |  |
| --- | --- |
| Variable name | Variable label |
| FundId | Fund identifier |
| pid | Manager identifier |
| YM | year-month |
| TeamSize | Number of additional managers in team |
| Team | =1 if team managed in that month |
| ret\_gross | Gross return |
| ret\_gross\_1 | Gross return in previous month (t-1) |
| ret\_gross\_1\_neg | min(ret\_gross\_1,0) |
| ret\_gross\_1\_pos | max(ret\_gross\_1,0) |
| ret\_net | Net return |
| ret\_abn | Abnormal return |
| ret\_abn\_global4 | Abnormal return (from global 4-factor model) |
| ret\_abn\_FF\_EU | Abnormal return (from Fama-French Europe model) |
| V | Value added |
| SR | Sharpe ratio |
| RV | Relative Volatility |
| TE | Tracking Error |
| TE\_global4 | Tracking Error (from global 4-factor model) |
| TE\_FF\_EU | Tracking Error (from Fama-French Europe model) |
| ER | Fund level expense ratio |
| ER\_source | Source used for ER data |
| AUM | Fund level totat AUM |
| AUM\_source | Source used for AUM data |
| AUM\_1 | AUM from previous month (t-1) |
| lAUM\_1 | log(AUM\_1) |
| cat\_fi | =1 if fixed income fund |
| cat\_eq | =1 if equtiy fund |
| cat\_all | =1 if allocation fund |
| cat\_rest | =1 if 'rest' fund |

Variables in the dataset data\_halfyearly.dta

|  |  |
| --- | --- |
| Variable name | Variable label |
| FundId | Fund identifier |
| pid | Manager identifier |
| YH | Year-half |
| Team | =1 if team managed in that month |
| ret\_gross | Gross return |
| ret\_gross\_1 | Gross return in previous month (t-1) |
| ret\_gross\_1\_neg | min(ret\_gross\_1,0) |
| ret\_gross\_1\_pos | max(ret\_gross\_1,0) |
| SR | Sharpe ratio |
| RSV | Relative Semi-Volatility |
| RV | Relative Volatility |
| TE | Tracking Error |
| AUM | Fund level totat AUM |
| AUM\_1 | AUM from previous month (t-1) |
| lAUM\_1 | log(AUM\_1) |
| cat\_fi | =1 if fixed income fund |
| cat\_eq | =1 if equtiy fund |
| cat\_all | =1 if allocation fund |
| cat\_rest | =1 if 'rest' fund |

### Construction of synthetic data

The following steps are used to create the synthetic data that appears in the dataset data\_monthly.do.

**Step 1**: We drop some observations. First, some participants manage many funds in our sample. In order for these participants not to be identifiable by the number of managed funds in the Morningstar database, for every participant who managed more than eight different funds throughout our sample period, we drop zero to two funds. The number of funds to drop (0, 1, or 2) and the actual funds to be dropped are randomly chosen for each eligible participant. Second, to ensure that participants in the experiment are not identifiable using the dates of fund-changes, for each manager-fund observation we drop monthly observations from the beginning and end of the manager’s tenure at the specific fund. We drop observations from the beginning if (i) the manager’s tenure is longer than 12 months during our sample period and (ii) the first month of tenure is not January 2008. Similarly, we drop observations from the end if (i) the manager’s tenure is longer than 12 months during our sample period and (ii) the last month of tenure is not December 2019. In all cases the number of monthly observations to drop is a randomly chosen number between 0 and 4.

**Step 2**: We add randomly generated noise to fund performance related variables, so that the funds themselves are not identifiable. In particular, we add noise to the following variables: ret\_gross, ret\_net, ret\_abn, ret\_abn\_global4, ret\_abn\_FF\_EU, ER, SR, RV, TE, TE\_global4, TE\_FF\_EU, and AUM. Let denote an observation of one of these variables for fund *i* in month *t*. We create

where is the standard deviation of the original variable *y* for fund *i* in the sample. That is, the variance of the generated noise is fund-specific. We use fund-specific variance for the generated noise because there are very different funds in the sample (e.g., fixed income vs. equity funds), and using a common variance would change the structure of the original data too much. Finally, we use because this leads to .

**Step 3**: We winsorize the transformed variables from step 2 so that funds with extreme values on any of these variables (e.g., very large funds, or funds with extreme monthly returns) are not identifiable. Generally, we winsorize from below using the 1st percentile, and winsorize from above using the 99th percentile. There are two exceptions from this rule. First, for AUM we use the 10th and 90th percentiles instead, because the possibility of identifying funds in the tails of the size distribution is a particular concern of ours. Second we use the 3rd percentile when winsorizing the TE, TE\_global4, TE\_FF\_EU variables from below, since the transformed versions of these variables (including the random noise term) would have negative values otherwise.

**Step 4**: The rest of the fund performance variables that appear in the replication dataset (ret\_gross\_1, ret\_gross\_1\_neg, ret\_gross\_1\_pos, V, AUM\_1, lAUM\_1) are generated from the above variables after the previous steps have been carried out.

To create the synthetic data contained in the file data\_halfyearly.dta, the same steps are carried out as above with the difference that the variables transformed in step 2 are ret\_gross, SR, RV, RSV, TE, and AUM.

## Computational requirements

### Software Requirements

* Stata (code was last run with version SE 16.1). Packages used:
  + outreg2
  + tab2docx

The program “0\_main.do” will start by checking if the above packages are installed, and will install them if needed.

### Memory and Runtime Requirements

#### Summary

Approximate time needed to reproduce the analyses on a standard 2021 desktop machine is less than 10 minutes.

#### Details

The code was last run on a 4-core Intel-based (i7) desktop computer with 32GB RAM and Windows 10 operating system.

## Description of programs/code

INSTRUCTIONS: Give a high-level overview of the program files and their purpose. Remove redundant/ obsolete files from the Replication archive.

* Programs in programs/01\_dataprep will extract and reformat all datasets referenced above. The file programs/01\_dataprep/main.do will run them all.
* Programs in programs/02\_analysis generate all tables and figures in the main body of the article. The program programs/02\_analysis/main.do will run them all. Each program called from main.do identifies the table or figure it creates (e.g., 05\_table5.do). Output files are called appropriate names (table5.tex, figure12.png) and should be easy to correlate with the manuscript.
* Programs in programs/03\_appendix will generate all tables and figures in the online appendix. The program programs/03\_appendix/main-appendix.do will run them all.
* Ado files have been stored in programs/ado and the main.do files set the ADO directories appropriately.
* The program programs/00\_setup.do will populate the programs/ado directory with updated ado packages, but for purposes of exact reproduction, this is not needed. The file programs/00\_setup.log identifies the versions as they were last updated.
* The program programs/config.do contains parameters used by all programs, including a random seed. Note that the random seed is set once for each of the two sequences (in 02\_analysis and 03\_appendix). If running in any order other than the one outlined below, your results may differ.

### Details

* programs/00\_setup.do: will create all output directories, install needed ado packages.
  + If wishing to update the ado packages used by this archive, change the parameter update\_ado to yes. However, this is not needed to successfully reproduce the manuscript tables.
* programs/01\_dataprep:
  + These programs were last run at various times in 2018.
  + Order does not matter, all programs can be run in parallel, if needed.
  + A programs/01\_dataprep/main.do will run them all in sequence, which should take about 2 hours.
* programs/02\_analysis/main.do.
  + If running programs individually, note that ORDER IS IMPORTANT.
  + The programs were last run top to bottom on July 4, 2019.
* programs/03\_appendix/main-appendix.do. The programs were last run top to bottom on July 4, 2019.
* Figure 1: The figure can be reproduced using the data provided in the folder “2\_data/data\_map”, and ArcGIS Desktop (Version 10.7.1) by following these (manual) instructions:
  + Create a new map document in ArcGIS ArcMap, browse to the folder “2\_data/data\_map” in the “Catalog”, with files “provinceborders.shp”, “lakes.shp”, and “cities.shp”.
  + Drop the files listed above onto the new map, creating three separate layers. Order them with “lakes” in the top layer and “cities” in the bottom layer.
  + Right-click on the cities file, in properties choose the variable “health”… (more details)

## List of tables and programs

The provided code reproduces all tables and figures in the paper

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Data File1 | Program | Lines | Output file |
| Table 1 |  | n.a. (no data) |  |  |
| Table 2 | experiment | 1\_summary\_experimental.do | 7-11 | Table2.doc |
| Table 3 | funds\* | 2\_summary\_funds2017.do | 7-149 | Table3.xlsx |
| Table 4 | monthly\* | 3\_summary\_sample.do | 63-146 | Table4\_\*.doc |
| Table 5 | monthly\* | 4\_regression\_monthly.do | 11-45 | Table5.doc |
| Table 6 | monthly\* | 4\_regression\_monthly.do | 49-83 | Table6.doc |
| Table S1 | experiment | 1\_summary\_experimental.do | 15-55 | TableS1.xlsx |
| Table S2 | monthly\* | 3\_summary\_sample.do | 7-48 | TableS2.xlsx |
| Table S3 | monthly\* | 3\_summary\_sample.do | 53-54 | TableS3.docx |
| Table S4 | monthly\* | 3\_summary\_sample.do | 58-59 | TableS4.docx |
| Table S5 | monthly\* | 4\_regression\_monthly.do | 87-184 | TableS5.xlsx |
| Table S6 | monthly\* | 4\_regression\_monthly.do | 509-636 | TableS6.doc |
| Table S7 | monthly\* | 4\_regression\_monthly.do | 187-211 | TableS7.doc |
| Table S8 | monthly\* | 4\_regression\_monthly.do | 215-430 | TableS8.xlsx |
| Table S9 | monthly\* | 4\_regression\_monthly.do | 434-468 | TableS9.doc |
| Table S10 | monthly\* | 4\_regression\_monthly.do | 472-505 | TableS10.doc |
| Table S11 | halfyearly\* | 5\_regression\_halfyearly.do | 11-37 | TableS11.doc |
| Figure S1 |  | n.a. (no data) |  |  |
| Figure S2 |  | n.a. (no data) |  |  |
| Figure S3 | experiment | 1\_summary\_experimental.do | 59-92 | FigureS3\_\*.png |

1 Abbreviations used in the Data File column: “experiment” refers to data\_experimental.dta, “funds” refers to data\_funds2017.dta, “monthly” refers to data\_monthly.dta, and “halfyearly” refers to data\_halfyearly.dta.

\* indicates that the data is synthetic, i.e., the values generated via the replication package do not exactly match those in the paper.

## References

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree: An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9, 88–97.