

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 Refinitiv (<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 Reflection Test score (raw) |
| crt | Cognitive Reflection Test score (standardized) |
| apm_raw | Advanced Progressive Matrices score (raw) |
| apm | Advanced Progressive Matrices score |
| 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, $N(0, \sigma_y)$, random noise, with $\sigma_y = 3$ used for Tenure, $\sigma_y = 150$ for AUM, and $\sigma_y = 7$ 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 |
|------------------------------|---|
| <code>FundId</code> | Fund identifier |
| <code>pid</code> | Manager identifier |
| <code>YM</code> | year-month |
| <code>TeamSize</code> | Number of additional managers in team |
| <code>Team</code> | =1 if team managed in that month |
| <code>ret_gross</code> | Gross return |
| <code>ret_gross_1</code> | Gross return in previous month (t-1) |
| <code>ret_gross_1_neg</code> | $\min(\text{ret_gross_1}, 0)$ |
| <code>ret_gross_1_pos</code> | $\max(\text{ret_gross_1}, 0)$ |
| <code>ret_net</code> | Net return |
| <code>ret_abn</code> | Abnormal return |
| <code>ret_abn_global4</code> | Abnormal return (from global 4-factor model) |
| <code>ret_abn_FF_EU</code> | Abnormal return (from Fama-French Europe model) |
| <code>V</code> | Value added |
| <code>SR</code> | Sharpe ratio |
| <code>RV</code> | Relative Volatility |
| <code>TE</code> | Tracking Error |
| <code>TE_global4</code> | Tracking Error (from global 4-factor model) |
| <code>TE_FF_EU</code> | Tracking Error (from Fama-French Europe model) |
| <code>ER</code> | Fund level expense ratio |
| <code>ER_source</code> | Source used for ER data |
| <code>AUM</code> | Fund level total AUM |
| <code>AUM_source</code> | Source used for AUM data |
| <code>AUM_1</code> | AUM from previous month (t-1) |
| <code>lAUM_1</code> | $\log(\text{AUM_1})$ |
| <code>cat_fi</code> | =1 if fixed income fund |
| <code>cat_eq</code> | =1 if equity fund |
| <code>cat_all</code> | =1 if allocation fund |
| <code>cat_rest</code> | =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 | $\min(\text{ret_gross_1}, 0)$ |
| ret_gross_1_pos | $\max(\text{ret_gross_1}, 0)$ |
| SR | Sharpe ratio |
| RSV | Relative Semi-Volatility |
| RV | Relative Volatility |
| TE | Tracking Error |
| AUM | Fund level total AUM |
| AUM_1 | AUM from previous month (t-1) |
| lAUM_1 | $\log(\text{AUM_1})$ |
| cat_fi | =1 if fixed income fund |
| cat_eq | =1 if equity 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 y_{it} denote an observation of one of these variables for fund i in month t . We create $\tilde{y}_{it} = y_{it} + c\epsilon_{it}$ with $\epsilon_{it} \sim N(0, \sigma_{yi}^2)$,

where σ_{yi} 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 $c = 0.62$ because this leads to $\text{Corr}(\tilde{y}_{it}, y_{it}) = 0.85$.

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`, `1AUM_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

- 0_main.do first installs packages if needed and then executes the rest of the do-files from the replication package. These do files do not necessarily have to be executed in order.
- 1_summary_experimental.do creates all the tables and figures that rely solely on the experimental data. These are: Table 2, Table S1, and Figure S3.
- 2_summary_funds2017.do creates Table 3.
- 3_summary_sample.do provides summary statistics of the sample. It creates Table 4, Table S2, Table S3, and Table S4.
- 4_regression_monthly.do creates all results that rely on the monthly fund data. It creates Table 5, Table 6, and Tables S5 to S10.
- 5_regression_halfyearly.do creates the results that rely on the half-yearly fund data. It creates Table S11.

List of tables and programs

The provided code reproduces all tables and figures in the paper

| | Data File ¹ | 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 |

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

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.

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.