Replication Report for “Demand Forecasting Behavior: System Neglect and Change Detection”

By: Mirko Kremer, Brent Moritz and Enno Siemsen

Primary Team: Elena Katok and Kyle Hyndman
University of Texas at Dallas

Secondary Team: Samantha Keppler and Stephen Leider
University of Michigan

Kremer et al. (2011) study how individuals make forecasting decisions based on time-series data. They consider the effect of two kinds of random errors: temporary shocks and permanent shocks. Their experiment varies the magnitude of temporary shock at two levels ($n = 10$ and $n = 40$) and the magnitude of permanent shocks at three levels ($c \in \{0, 10, 40\}$) for the total of six treatments. The main finding is that forecasters over-react to forecast error in more stable environments and under-react to them in less stable environments.

Hypothesis to replicate:

Hypothesis 1 (system neglect). Individuals show relatively more overreaction for low values of $W = {}^{c^2}\!/{}_{n^2}$, and relatively more underreaction for high values of W.

Power Analysis

According to the normative benchmark, a forecast (F) in period $t+1$ is:

$$F_{t+1} = F_t + \alpha^*(W)(D_t - F_t)$$

where F_t is the last period forecast, $D_t - F_t$ is observed forecast error, and $\alpha^*(W)$ is the weight placed on the error, where the optimal weight is:

$$\alpha^*(W) = \frac{2}{(1 + \sqrt{1 + 4/W})}$$

and $W = {}^{c^2}\!/{}_{n^2}$. So the key metric for system neglect is the comparison between $\alpha(W)$ estimated from observed decisions and $\alpha^*(W)$. The system neglect hypothesis predicts that in treatments with low W, $\alpha(W) > \alpha^*(W)$ and in treatments with high W, $\alpha(W) < \alpha^*(W)$.

In the original study, Kremer et al. (2011) estimated and compared four different models of forecasting and found that the model that fits the data best has four parameters (Equation 10), with α being one of them. When α thus estimated is compared to α^*, the data is consistent with system neglect hypothesis. The original paper uses a full factorial design, with 6 treatments. Our aim is to select two treatments to test the main result (system neglect). The most promising design is to keep n constant at 10 and vary c at 0 and 40. This would correspond to comparing Condition 1 ($c = 0, n = 10$) where $\alpha^* = 0$ and $\alpha = 0.39$ and Condition 5 ($c = 40, n = 10$) where $\alpha^* = 0.94$ and $\alpha = 0.7$. In both of those conditions the null hypothesis $\alpha = \alpha^*$ is rejected at $p \leq 0.01$.

In order to achieve 90% power to replicate the original result, we needed 16 subjects[1]. As 16 is smaller than the original sample size of 86 subjects (43 subjects in each treatment), the original sample size is binding. We targeted at least 43 subjects in each treatment for this replication, for a total of 86 subjects.

Sample

Participants in the original study were students, probably from Penn State, incentivized for forecast accuracy, defined at (1-MAPE) (MAPE = Mean Absolute Percentage Error). Conditions 1 and 5 had 43 subjects each. The sample for the primary replication will consist of at least 86 (divided equally for each treatment) University of Texas at Dallas students. The sample for the secondary replication will consist of at least 86 subjects from the University of Michigan. Due to in-person laboratory interruptions from Covid-19, each replication will first be conducted online. Subsequently, if the p-value associated with the primary hypothesis is greater than .05, that location will repeat the study in-person. In both cases, students will be recruited from the general laboratory populations.

Materials

The instructions and treatment materials were provided by the authors. The original experiment was conducted using zTree. Due to the online nature of the replication, we recoded the experiment in SoPHIE software while ensuring that the task, decision support, and interface was similar to the original experiment. The time series were the same as in the original paper, and were graciously provided to us by the original authors.

Procedure

We followed the protocols outlined in section "4.4.4 Experimental design" on pages 1331–1832 with some minor deviations, detailed in a later section. The pre-registration report for the experiment is available at https://aspredicted.org/7r4i2.pdf.

Analysis

We fit "Model 4" from Appendix C.2 of the original paper and test the null hypothesis $\alpha = \alpha^*$. We fit the model using the method described in section 4.2.3 "Estimation" pp. 1835-1836, and we use the original Stata program used by the authors (footnote 7 p. 1835). One of the authors also vetted our experimental data and the estimation to make sure that it is consistent with the original.

For Condition 1, the relevant hypothesis to test is that the coefficient on $\mu(E_t)$ (i.e., α) is equal to 0, while for Condition 5, the relevant hypothesis to test is that the coefficient on $\mu(E_t)$ (i.e., α) is equal to 0.94.[2]

Differences from Original Study

The differences with respect to the original study are as follows:

1. We ran the experiments at the University of Texas at Dallas and the University of Michigan.
2. We conducted the replications online and asynchronously using SoPHIE software.
3. A "Stage 1" that consisted of individual response questions prior to the main experiment, but which was not included in the data analysis of the original study, was not conducted. The original authors agreed to this modification.

[1] We preregistered a higher needed N based on the p-value threshold of $p \leq 0.01$ reported in the paper (N=134). However, once we obtained the original data we observed the true p-value ($p = 4.5 \times 10^{-14}$). Using the true p-value, we calculate a needed sample size of 16 subjects to achieve 90% power.

[2] See Table 1 for the estimated empirical models.

Replication Results

In total 135 students participated at UT Dallas, while 139 students participated at the University of Michigan. Thus, in both samples, we recruited enough subjects to achieve the desired power. The results from estimating Model 4 from Appendix C.2 for the replication samples are contained in Table 1. For comparison purposes, we also include the results from the original work.

In all cases, the estimated coefficients in the replication sample are somewhat closer to the predictions than in the original sample. However, the directional deviation is the same as the original paper. Furthermore, for both replication samples and for both hypothesis tests at each location we reject the hypotheses at $p \ll 0.01$.[3] Therefore, the results replicate at both locations.

Unplanned Protocol Deviations

There were no unplanned protocol deviations.

[3] Specifically, the relevant p-values are: UTD Condition 1: 2.316×10^{-13}, UTD Condition 5: 3.377×10^{-5}, Michigan Condition 1: 3.268×10^{-25}, and Michigan Condition 5: 1.881×10^{-6}.

Table 1 Replicating Table C.2 (Model 4) for Conditions 1 and 5

(a) Condition 1

	UTD		Michigan		Original	
$\mu(E_t)$	0.367^{***}	(0.050)	0.286^{***}	(0.028)	0.39^{***}	(0.04)
F_t	-0.469^{***}	(0.052)	-0.551^{***}	(0.021)	-0.30^{***}	(0.03)
$\mu(\Delta D_t)$	0.002	(0.029)	-0.033^{***}	(0.012)	0.11^{***}	(0.03)
ΔD_{t-1}	0.079^{**}	(0.032)	0.047^{***}	(0.012)	0.01	(0.02)
ΔF_t	-0.034^{***}	(0.012)	0.000	(0.002)	-0.08^{***}	(0.02)
ΔF_{t-1}	0.020^{**}	(0.009)	0.000	(0.001)	-0.05^{***}	(0.02)
$\mu(con.)$	236.409^{***}	(26.279)	276.318^{***}	(10.488)	151	(15)
$\sigma_S(E_t)$					0.00	(0.00)
$\sigma_S(\Delta D_t)$					0.00	(0.00)
$\sigma_S(con.)$					0.65	(0.29)
$\sigma_i(E_t)$	0.011^{***}	(0.005)	0.031^{***}	(0.006)	0.17^{***}	(0.03)
$\sigma_i(\Delta D_t)$	0.000	(0.000)	0.002^{**}	(0.001)	0.13^{***}	(0.02)
$\sigma_i(con.)$	56.573^{***}	(12.095)	1.667^{***}	(0.426)	0.59	(0.23)
N (Subjects)	3030 (65)		3240 (69)		2021 (43)	

(b) Condition 5

	UTD		Michigan		Original	
$\mu(E_t)$	0.790^{***}	(0.036)	0.786^{***}	(0.032)	0.70^{***}	(0.03)
F_t	-0.004	(0.004)	-0.011^{***}	(0.004)	0.00	(0.00)
$\mu(\Delta D_t)$	0.012	(0.026)	-0.011	(0.022)	0.31^{***}	(0.07)
ΔD_{t-1}	0.144^{***}	(0.021)	0.087^{***}	(0.020)	0.10^{***}	(0.02)
ΔF_t	0.009	(0.010)	0.013	(0.010)	-0.14^{***}	(0.02)
ΔF_{t-1}	-0.017^{*}	(0.009)	-0.002	(0.009)	-0.05^{***}	(0.01)
$\mu(con.)$	5.339^{**}	(2.514)	6.922^{***}	(2.662)	1.2	(2.2)
$\sigma_S(E_t)$	0.003	(0.004)	0.001	(0.002)	0.00	(0.00)
$\sigma_S(\Delta D_t)$	0.002^{*}	(0.002)	0.001^{***}	(0.001)	0.11^{***}	(0.05)
$\sigma_S(con.)$	4.302	(5.985)	7.927^{*}	(8.637)	1.7	(1.1)
$\sigma_i(E_t)$	0.032^{***}	(0.007)	0.031^{***}	(0.006)	0.10^{***}	(0.02)
$\sigma_i(\Delta D_t)$	0.009^{***}	(0.003)	0.000	(0.000)	0.10^{***}	(0.03)
$\sigma_i(con.)$	44.313^{***}	(11.180)	24.360^{***}	(6.841)	3.1^{***}	(1.1)
N (Subjects)	3288 (70)		3168 (70)		2018 (43)	

Note: Standard errors are in parentheses. *, ** and *** denote significance at the 10, 5 and 1% levels respctively.

References

Kremer, Mirko, Brent Moritz, Enno Siemsen. 2011. Demand forecasting behavior: System neglect and change detection. *Management Science* **57**(10) 1827–1843.