Quantitative Finance

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rquf20

Return prediction and stock selection from unidentified historical data

Doron Sonsino [a] & Tal Shavit [a]

[a] School of Business, College of Management, 7 Rabin Blvd, POB 9017, Rishon LeZion, Israel 75190

Version of record first published: 13 Sep 2012.

To cite this article: Doron Sonsino & Tal Shavit (2012): Return prediction and stock selection from unidentified historical data, Quantitative Finance, DOI:10.1080/14697688.2012.712210

To link to this article: http://dx.doi.org/10.1080/14697688.2012.712210

Routledge
Taylor & Francis Group

Return prediction and stock selection from unidentified historical data

DORON SONSINO* and TAL SHAVIT

School of Business, College of Management, 7 Rabin Blvd, POB 9017,
Rishon LeZion, Israel 75190

(*Received 25 January 2010; in final form 11 July 2012*)

The experimental approach was applied to test the value of historical return series in technical prediction. Return sequences were randomly drawn cross-sectionally and over time from S&P500 records and participants were asked to predict the 13th realization from 12 preceding returns. The hypothesis that predictions (nominal or real) are randomly assigned to historical sequences is rejected in permutation tests, and the best-stock portfolios by experimental predictions significantly outperform the worst-stock portfolios in joint examination of mean return and volatility. The participants dynamically adjust their predictions to the observed series and switch from momentum riding to contrarian extrapolation when recent trends get extreme. The implicit tuning of predictions to specific series captures variabilities that could not be inferred by schematic statistical forecasting.

Keywords: Judgmental return forecasting; Predictability; Shift in prediction regime

JEL Classification: C9, D8, G1

1. Introduction

The extensive research on judgmental return forecasting has been motivated by the need to understand the belief formation process of possibly biased investors (Barberis and Thaler 2003), disentangle the axioms laid by models in behavioural finance (Bloomfield and Hales 2002, Durham *et al*. 2005), and examine methodological issues regarding the elicitation of beliefs (Glaser *et al*. 2007). Experimental studies on financial prediction typically utilize artificial sequences, especially tailored to explore the hypotheses under examination (see the comprehensive survey in Lawrence *et al*. 2006). Some recent studies (e.g. Du and Budescu 2007) alternatively adopt an empirical method, selecting specific samples of empirical field data for testing underlying theories. The current study extends the empirical approach, devising a 'random-sequence design' in which historical series are independently drawn for each subject and task from empirical records. Subjects are asked to predict subsequent monthly or annual returns from realized series; experimental payouts are determined by the accuracy of predictions.

The random drawing of historical sequences cross-section and time is primarily intended to preclude inference about the identity of specific stocks or particular inspection periods, and confine information to the historical data. It therefore opens the possibility for testing purely technical forecasting on an ecologically valid platform (Griffin and Brenner 2004). The drawing of series from a large empirical dataset, moreover, cleans the experiment from sequence-specific bias and tests the significance of historical technical data on a very general platform. Testing the economic value of historical return series in judgmental forecasting is another major interest of the study.

The return series for the prediction tasks were randomly drawn from the 40 year records of the stocks composing the S&P500, and advanced MBA students were asked to point-predict the 13th return from 12 preceding realizations.† Each participant ($N = 65$) made

*Corresponding author. Email: sonsinod@colman.ac.il

†By Harrison and List (2004) typology of economic field experiments, our experiment classifies into the framed-field category, as it employs field-based tasks on a selective subject pool of MBA students with extensive background in finance.

Quantitative Finance
ISSN 1469–7688 print/ISSN 1469–7696 online © 2012 Taylor & Francis
http://www.tandfonline.com
http://dx.doi.org/10.1080/14697688.2012.712210

six annual predictions (henceforth: the ANN condition) and six monthly forecasts (the MON condition) and experimental payouts were determined by the absolute value of the prediction error in a randomly selected assignment. The next paragraphs discuss the main goals of the experiment in light of recent literature and outline some of the major results.

Our first major goal is to explore the patterns of purely technical return prediction on a platform of empirical data. This line of interest directly relates to topics explored in the literature on return forecasting and expectations formation. Several recent studies propose that expectations—in financial domains—may switch from trend-chasing to contrarian when streaks become extreme. MacDonald's (2000) survey of financial forecasting concludes that short-run financial market expectations exhibit 'bandwagon effects', while long-run expectations tend to be regressive and stabilizing. Durham *et al.* (2005) show that spreads in the football wagering market move in favour of teams on short winning streaks (three games or less) while moving against teams experiencing long positive streaks. Switch in regime of prediction is consistent with de Long *et al.*'s (1990) classic model where noise traders and chartists boost prices in the short-run while smart-money investors correct overreactions in the long run. At the level of individual decision, momentum-based expectations are frequently related to the hot hand bias (Gilovich *et al.* 1985) while expectations for mean-reversion are commonly associated with the gambler's fallacy (Tversky and Kahneman 1982). Our experimental results indeed demonstrate that subjects attempt to balance trend-chasing and contrarian-forecasting when making experimental predictions. Predictions tend in general to increase with historical mean returns, but decrease significantly following heavy streaks of positive returns. As a result, the correlation between mean historical returns and current predictions is positive but low. The hypothesis of consistent statistical forecasting is robustly rejected in the analysis.†

A related goal of the study is to examine the value of unidentified historical return series in empirical prediction. This complementary goal communicates with parts of the vast empirical literature on the predictability of stock returns (Kaul 1996). Interestingly we show that predictions based on unidentified historical series explain, on average, more than 25% of the variance in hidden annual returns. The payoffs on the best stocks, according to experimental predictions, significantly exceed the payoffs on the worst stocks, and the differences cannot be attributed to larger volatility of the best-stocks portfolios. Actual predictions, moreover, significantly outperform six plausible statistical rules in separating the best stock from the worst, in random six-stock menus. These surprising results relate to the literature on the profitability of technical trading (Brock *et al.* 1992) and may complement recent AI studies on nonlinear neural network models of stock returns (see the concluding discussion).

Finally, we seek to use the random-series platform to test the effect of skills on prediction qualities. The extensive research on expertise effects in financial decision-making provides mixed evidence regarding the relative performance of experts versus unprofessional or less professional subjects.‡ The unambiguous prediction tasks composing the current study and the random platform of historical series, however, define an interesting interface for testing the interaction of skills with prediction abilities. Analysis of individual prediction errors along our random questionnaires indeed suggests that prediction errors robustly decrease with proclaimed familiarity with the financial arena. The familiarity effect is significant and large in magnitude in the monthly prediction tasks where regressions on absolute prediction errors reveal a negative significant FAMILIARITY coefficient, even when the volatility of specific series and individual heterogeneity are accounted. Participants with more than median FAMILIARITY, moreover, exhibit stronger ability to separate the best stocks from the worst in random six-stock menus, and are more successful than others in predicting the sign of missing annual observations.

We conclude the introduction with three general comments regarding the scope and method of the paper:

First, we should emphasize that anonymous series are obviously insufficient for precise prediction of eventual returns. Prediction errors are large and, in fact, larger than the errors that arise when the empirical mean is used as stationary forecasting rule. We still demonstrate that, by dynamically adjusting their predictions to the observed histories, subjects are able to capture implicit trends, unrecognized by standard statistical rules. The judgmental predictions, in particular, significantly outperform six plausible statistical rules in identifying the best stock for annual investment in random six-stock menus.

The second short comment concerns the equilibrium tradeoffs between risk and return. In stationary equilibrium, expected returns increase with relative risks. Investors can use the historical mean to select the most profitable stock, but the best stock under such selection

†The tendency for momentum-riding versus contrarian-trading may vary with individual traits and depend on the framing of the prediction task. Several studies suggest that experienced professionals chase trends less frequently, being more inclined to follow contrarian strategies (e.g. De Bondt 1991 versus 1993, Thomson *et al.* 2003, Menkhoff *et al.* 2006). Glaser *et al.* (2007) demonstrate that expectations for trend reversal are more common when prediction tasks are formulated with price data, while trend extrapolation is prevalent when series are presented in terms of returns. The current study ignores these issues, testing the shift in regime hypothesis on series of historical returns with a relatively homogenous sample of MBA students.

‡The literature is too large and diverse for comprehensive review. We therefore cite a non-representative sample briefly: the prediction/information processing abilities of financial experts are superior by the studies of Önkal *et al.* (2003), Elliott *et al.* (2008) and Andersson *et al.* (2009).The evidence is negative in Thomson *et al.* (2003), Törngren and Montgomery (2004), Glaser *et al.* (2005) and Haigh and List (2005).

would also constitute the most risky investment. Since systematic risk (or other factor-related risk) could not be extracted from the anonymous series, we use ex-post standard deviations to compare risk between portfolios. The experimental results interestingly reveal that while the best-stocks by individual predictions significantly outperform the worst-stocks by the same predictions, the variability of best and worst payoffs is similar or relatively close. In this (restricted) sense, subjects are able to select the best stocks for investment without bearing excessive risk.†

Finally, we briefly comment on the statistical method of the paper. The statistical testing of experimental results is restricted by two obstacles. Since we employed a multi-task design where subjects made six predictions in each condition, tests must be run on subject-level statistics (for independence). Tests of prediction errors, for example, should refer to the average error of each participant in each condition.‡ The use of individual average data results in loss of information that obstructs standard statistical testing. In addition, the acknowledged heterogeneity in individual forecasting patterns (Dominitz and Manski 2011) together with the random-sequence design generate an extremely noisy dataset. The distributions of individual-level statistics are non-normal, skewed and sometimes change considerably when sub-samples of the pool are inspected. We therefore avoid parametric tests throughout the paper and apply specialized randomization or permutation tests where applicable.§ The randomization/permutation tests exploit the multi-task design to check the significance of payoffs and differences without imposing assumptions on underlying distributions. The exact methodology is illustrated in section 3 after a detailed description of the experiment in section 2. Section 4 briefly summarizes the main experimental results while section 5 characterizes the individual prediction patterns along the questionnaires. Section 6 focuses on the separation of BEST stocks from WORST by experimental predictions while section 7 contrasts the judgmental predictions with the statistical prediction models. Section 8 is a concluding discussion.

2. Method and subjects

The typical prediction problem is illustrated in figure A1 in the appendix. A sequence of 12 historical returns is presented in tabulated form, and the subject is requested to point predict the 13th observation. Each questionnaire consisted of 12 prediction tasks of this type: 6 annual (ANN) and 6 monthly (MON) technical prediction assignments.

The historical return series (adjusted for dividends and splits in line with CRSP standards) were randomly drawn for each participant and each prediction task from historical records of the stocks composing the S&P500 list at the end of April 2006. The instructions explained that series of 13 successive returns were drawn at random from the S&P500 data for the last 40 years. The 13th observation was concealed, but kept in our records to validate predictions. No information was revealed about the identity of stocks, the exact inspection period, or economic fundamentals. The instructions (see supplementary appendix A) emphasized that different problems may refer to distinct stocks and diverse inspection intervals.¶ The random-sequence design was adopted to decrease the chances that subjects claim to identify stocks or inspection periods. An alternative design in which, for instance, participants receive cross-sectional historical data on 6 anonymous stocks (for the same 12 periods), could motivate subjects into guessing the dates of inspection and the identity of stocks. In such a design, subjects might hinge their predictions on private signals rather than acting on the historical data.⊥ In addition, the random design rules out sequence-specific bias and tests the value of anonymous historical series on a very general platform.

The ANN and MON assignments were separated in each questionnaire, and the order of conditions was randomly assigned. Each prediction task appeared in a separate page and the forecast horizon (ANN or MON) was marked in bold type at the top of the page. On the last page of the questionnaire (see supplementary appendix B), participants were asked to rank their knowledge of finance THEORY and their FAMILIARITY with the financial arena in 1–10 scale and characterize their prediction METHOD.

Experimental questionnaires were distributed in class to 65 advanced MBA students (25 female; 40 male) who had passed the basic core courses in finance. The average age was 30.2; 29% had experienced investment-management (personal portfolio or others) in practice. Following standard practice in experimental economics, we used a performance based payoff to motivate accurate prediction. The instructions explained that one of the 12 tasks ('the selected problem/prediction') would be randomly

†In the annual prediction assignments, the best stocks show higher volatility compared to the worst stocks but randomizations suggest that the difference in volatilities is small relative to the difference in means. The stocks that obtained the highest predictions could still be riskier in terms of systematic risk (higher betas) or other factor-related risk (e.g. size), but we do not believe that subjects could consistently recognize such stocks from the unidentified random series.

‡We henceforth use 'average' for subject-level statistics and keep 'mean' for the between subject average.

§See Good (2005) for detailed discussion of permutation and randomization tests. SAS codes will be provided on request; see http:// www2.colman.ac.il/business/ doron for examples.

¶Supplementary appendices are available at http://www2.colman.ac.il/business/doron/

⊥ Subjects could also use such cross-section samples to predict the market conditions for the 13th period; predictions may then be affected by relative-risk considerations. The cross-section and time design alleviates these concerns as well.

drawn to determine individual payouts.† If the selected prediction was of the wrong sign (the subject had predicted a negative return when the actual return was positive, or vice versa), the final payout would be 20 New Israeli Shekels (₪, about US$5). If sign-prediction for the selected problem was correct, the guaranteed payout increased to 30₪ and the exact amount was determined by adding a bonus for accurate prediction to the minimum. The bonus formula was $80 - P \times |\text{predicted return} - \text{actual return}|$, where the penalty ($P$) for 1% absolute deviation between predicted and actual return was 6 in MON and 2 in ANN. The decreased penalty for errors in ANN was motivated by the fact that annual returns are larger, and the instructions clarified that if the bonus turns negative, the guaranteed minimum will be paid. The organizers guaranteed the fairness of experimental procedures and subjects were invited to verify calculations when payouts were announced. The mean payout was 40₪ (about US$10) with standard deviation 23. The mean participation time was 20–25 minutes.‡

3. The randomization tests

First, we introduce some notation to simplify subsequent discussion. PRED will henceforth denote the experimental predictions while OBS*i* will present the *i*th observation in the historical return-sequences. OBS12 thus denotes the last observed return while OBS13 describes the target return that subjects were asked to forecast. The product OBS13 × 100 is sometimes addressed as the *realized payoff* on a $100 investment. Subjects exhibit *under-prediction* when PRED < OBS13 while *over-prediction* refers to cases where PRED > OBS13. The absolute value metric is used to measure prediction accuracy; ERR = |PRED − OBS13| accordingly denotes the prediction error. The symbols AVG12 and STD12 are reserved for the mean and standard deviation of the observed returns OBS1, OBS2…OBS12 (excluding OBS13). AVG6/STD6 accordingly denote the mean/STD of returns in the series OBS7–OBS12 and AVG3/STD3 are similarly defined. TREND*n* denotes the slope of returns in the last *n* observations.

We now illustrate the randomization tests employed throughout the paper by describing in detail a randomization test for comparing the eventual performance of the best stock by individual predictions to the payoff on the worst stock by these predictions. The term BEST1 denotes the realized payoff (OBS13 × 100) on the stock that received the highest prediction among the 6 predictions provided by each subject. WORST1 denotes the payoff on the worst stock by the same ranking. The experimental evidence reveals that BEST1 payoffs are substantially higher than WORST1 payoffs on average. The specialized randomization test is used to determine directly if the difference is statistically significant. The null hypothesis is BEST1 = WORST1, while the directed alternative stipulates that BEST1 > WORST1. To run the test we subtract the mean WORST1 payoff from the mean BEST1 payoff, using D to denote the difference. The randomizations are used to test if differences larger than D arise in random assignment of stocks into the BEST1/WORST1 categories. In each randomization, two distinct stocks are randomly designated as randomized-best and randomized-worst in each six-stock menu. The process is repeated to generate 1000 independent randomizations. The mean payoff on randomized-worst stocks is subtracted from the mean payoff on randomized-best stocks, using d' to denote the corresponding difference for some arbitrary randomization.§ If $d' \geq D$ then the randomization suggests that differences larger than D in mean BEST1−WORST1 payoffs may arise in random selection of stocks. If, on the other hand, $d' < D$ then the randomization reveals that the actual difference D is large relative to the difference that arises from random selection. To conclude the test, we calculate the proportion of cases α (out of 1000 randomizations) where $d' \geq D$. If $\alpha < 0.1$, the null hypothesis that BEST1 = WORST1 is rejected for the alternative BEST1 > WORST1; α is the one-tail significance level.

In some parts of the analysis (especially in ANN) the BEST1 portfolios pay higher returns than the WORST1 portfolios, but at the same time exhibit much higher standard deviation. To test the differences in means and standard deviations jointly, we use the 1000 randomizations to calculate the percentage of cases α' where larger difference in mean returns emerge together with smaller difference in standard deviations. When $\alpha' < 0.1$, the randomizations show that the larger volatility of BEST1 compared to WORST1 cannot explain the difference in means. The hypothesis BEST1 = WORST1 is rejected in the joint randomization test at the α' significance level. Since $\alpha' \leq \alpha$ by definition, the joint test can only improve significance compared to the test on mean returns alone.

Similar randomization tests (separate or joint) are applied to determine if the average payoff on the two (or three) stocks that attracted the highest predictions is higher than the average payoff on the two (three) stocks that obtained the lowest predictions. The only difference is that the randomized best/worst selections now include two (or three) stocks, accordingly. The sample size for the

†The random selection of a single task to determine payouts is a common practice in experimental economics; e.g. Hey and Lee 2005. The method is intended to avoid wealth effects, diversification concerns and other considerations that might otherwise interfere with task level behaviour.

‡The instructions did not mention the possibility of using calculators since this could direct subjects into statistical/schematic prediction. For similar reasons we did not use bar-charts or other graphic illustrations that might implicitly direct subjects into trend extrapolation. Only few subjects used calculators or showed signs of formally calculating return statistics.

§Clearly, d' is close to zero on average.

Table 1. Mean predictions and errors ($N = 390$).

	PRED	OBS13	ERR	Over-prediction	STD12	STD6
ANN	15.3%	17.5%	28.2%	50.8%	34.7%	33.5%
MON	1.6%	2.6%	8.5%	49.7%	8.3%	8.1%

The table presents the mean values of experimental predictions and return-series statistics. The number of observations is 390 (65 subjects × 6 prediction tasks in each condition). The results for the annual series (ANN) are displayed in the first row and the results for the monthly series (MON) in the second. PRED denotes experimental predictions. OBS13 is the corresponding hidden return. ERR is the absolute value of the error |PRED − OBS3|. 'Over-prediction' denotes the percentage of observations where PRED > OBS13. STD12 is the standard deviation of historical returns (OBS1–OBS12); STD6 is the standard deviation in the second half of the series.

randomizations was determined by trial and error but the results typically converged within 1000 randomizations. Small α denotes one-tail significance levels throughout the paper. Upper-case asterisks are used to characterize levels of significance: a single asterisk (*) denotes marginal significance at $0.05 < \alpha < 0.1$; two asterisks (**) are used for significance at $\alpha < 0.05$; the respective figures are marked in bold type when $\alpha < 0.01$. The abbreviations *R*-test and *P*-test are henceforth applied for randomization/permutation tests, respectively.†

4. Preliminary analysis

The mean PRED, OBS13, ERR and other pertinent statistics are disclosed in table 1. The mean predictions, 15.3% in ANN and 1.6% in MON, are lower than the mean values of OBS13 (17.5% and 2.6% correspondingly), but the frequency of under-prediction is close to the frequency of over-prediction in both conditions (see the table).‡ The mean absolute error levels (mean ERR 28.2% in ANN versus 8.5% in MON) are similar to those reported in preceding studies; e.g. Törngren and Montgomery (2004) find 10–11% errors in the prediction of monthly returns on leading Swedish stocks by students and professionals. Errors also appear plausible when compared to measures of variability for the historical sequences (see, for instance, the mean STD12 and STD6 figures in the table). The next paragraphs briefly summarize an exhaustive analysis of the data. To smooth the exposition, we separate the discussion into subsections using short titles to underline the main points.

4.1. *Extremity of predictions (De Bondt 1991)*

Trying to understand the gap between the average PRED and OBS13, we find a strong tendency of the participants to under-predict when the historical series end with a negative trend. In the ANN tasks, for example, PRED is 9.7% lower than OBS13 (on average) in the series ending with OBS12 < OBS11, while PRED is 7.7% higher than OBS13 when OBS12 > OBS11. Smaller, but qualitatively similar, differences are observed in MON where PRED is 2.9% lower than OBS13 (on average) when TREND2 < 0 while PRED is 1% higher than OBS13 when TREND2 > 0. Predictions are therefore 'extreme' in the sense of De Bondt (1991), displaying over-pessimism when latest trends are negative and exaggerated optimism when recent trends are positive. The extremity may be ascribed to biased trend extrapolation: if subjects respond to TREND2 when forecasting future returns but actual series are mean-reverting, then OBS13 might be less extreme than predicted.§ The extremity of predictions, however, weakens and even disappears when longer trends (TREND6, TREND12) are examined. In ANN, for example, PRED is 3.6% lower than OBS13 when TREND12 < 0, while PRED and OBS13 are approximately equal when TREND12 > 0. The difference completely disappears in MON where PRED is about 1% lower than OBS13 in both samples.

4.2. *PRED explains 27% of the variation in OBS13, in ANN*

To further examine the predictive power of PRED for OBS13 while controlling for possible heterogeneity in prediction skills, we OLS estimate the linear model $\text{OBS13} = a + b \times \text{PRED}$ on the $N = 6$ observations collected from each participant in ANN and MON. The mean value of the coefficient b is 0.196 in ANN and 0.159 in MON ($N = 65$). The mean R^2 values are 0.275 for ANN and 0.084 for MON. When the estimation is restricted to the subset of $N = 34$ participants with

†Standard tests for the equality of paired observations include the signed-rank test, the permutation test for paired replicates and the conservative sign-test (Siegel and Castellan 1988). The signed-rank test assumes symmetric distributions (Diebold and Lopez 1996), an assumption that is strongly violated in our data. The signed-rank test, permutation test, and sign-test moreover ignore the other problems in the six problems set, focusing, for instance, on the difference between BEST1 and WORST1. The *R*-tests, on the contrary, directly examine the significance of BEST1 − WORST1 relative to the data from which the stocks were selected. Sign-tests could not reject the null in some cases where the randomizations revealed significance. When discussing such results we disclose the proportion of subjects with BEST1 > WORST1 but use the more powerful *R*-test to determine significance.

‡In annualized terms, the mean monthly OBS13 accumulates to 36% compared to the 17.5% mean OBS13 in ANN. The large difference is attributed to sampling error: the median OBS13 figures (1.5% in MON versus 15.1% in ANN) are more compatible. Similar large differences between mean and median statistics emerge for ERR. The median ERR values are 21.4% in ANN versus 6.4% in MON. The skewness and kurtosis measures for PRED, ERR and other variables are provided in supplementary table C.

§Haruvy *et al.* (2007) alternatively demonstrate that prices in the lab converge more rapidly than expectations, as subjects try to sell stocks before anticipated peaks. This may provide a frictional type of explanation to the consistent gaps between predictions and returns.

FAMILIARITY ≥ 4, the mean b coefficients increase to 0.199 (in ANN) and 0.362 (in MON). The mean R^2 values do not change significantly (0.281 in ANN and 0.082 in MON).† The regressions thus robustly reveal a positive relation between predictions and actual returns and roughly suggest that more than 25% of the variation in ANN returns and about 8% of the variation in MON returns are captured by the judgmental forecasts. While the distance between the current experimental study and the empirical predictability literature is far too large for concrete comparisons, it is anecdotally interesting to note that our mean levels of fit are close to the fit levels observed in long-run predictability studies where the list of predictive variables include dividend-price ratios, earning-price ratios and more. Fama and French (1988), for a classic reference, show that dividend yields may explain 3% of the variation in monthly returns and up to 25% of the variation in annual returns. Eleswarapu and Reinganum (2004), more recently, obtain 25% fit when regressing annual access market returns on past glamour stock returns and various macroeconomic variables. Since the experimental regressions, however, were run on small samples of nominal returns, the relatively high levels of fit could technically follow from the variability in nominal rates across assignments. To control implicitly for underlying levels of inflation, we normalize OBS13 and PRED with respect to AVG12 and rerun the adjusted model $(\mathrm{OBS13}-\mathrm{AVG12}) = a + b \times (\mathrm{PRED}-\mathrm{AVG12})$.‡ The results, however, are reinforced and even improve slightly for ANN. The mean b coefficients are 0.35 for ANN and 0.12 for MON and levels of fit are 0.3 and 0.08, respectively.

4.3. *'Randomality' rejected in permutation tests*

To complement the regressions, we run direct permutation tests to examine the link between experimental predictions and missing returns. The null hypothesis of the tests is that predictions are randomly assigned to forecasting assignments. If this is the case, then shuffling of predictions across tasks should not increase (nor decrease) prediction errors. If, on the other hand, the historical data provides useful cues for predicting future returns, then such permutations would decrease accuracy and increase prediction errors. We use the absolute value metric ERR to measure the distance between predictions and OBS13 in the actual and shuffled series and run the permutation test separately for ANN and MON. In each permutation, the six predictions of each subject are randomly shuffled across the six prediction assignments. The exercise is independently repeated 5000 times to construct a sample of 5000 permutations. We then calculate the proportion of permutations where the mean distance between shuffled predictions and actual returns is lower than the mean actual ERR. This proportion constitutes the significance level of the test. If the proportion is smaller than 0.1, the random prediction hypothesis is rejected to conclude that the historical data was useful for prediction.§ The permutations for ANN reveal that the shuffling of predictions across tasks increases the mean prediction error in about 96% of 5000 permutations. The hypothesis of random-prediction is thus rejected at $\alpha = 0.04$. The results for MON are slightly weaker but still evidently significant. The distance between shuffled predictions and actual returns is lower (on average) than actual ERR in only 6.22% of 5000 permutations. The hypothesis of random prediction is accordingly rejected at $\alpha = 0.06$. The randomality of predictions with respect to historical data is therefore rejected for both experimental conditions. Significance moreover improves (to $\alpha = 0.001$ for ANN and $\alpha = 0.03$ for MON) when returns (PRED and OBS13) are normalized with respect to AVG12 before running the permutations.

4.4. *Is it possible to recognize sign(OBS13) from the sequence?*

To check if subjects are able to recognize the sign of OBS13 from the sequence, we compare the frequency of OBS13 > 0 in cases where subjects submitted positive versus negative predictions. Slight differences are observed in ANN where the conditional frequency $\Pr(\mathrm{OBS13}>0 \mid \mathrm{PRED}>0) = 78.2\%$ while $\Pr(\mathrm{OBS13}>0 \mid \mathrm{PRED}<0) = 71.6\%$.¶ The differences moreover increase when the comparison is confined to the sample with FAMILIARITY ≥ 4 ($N = 34$), where the frequency of positive returns in cases of positive predictions is 81.3% compared to 70.2% in cases of negative predictions. Chi-square tests for independence (Siegel and Castellan 1988) marginally reject the independence of sign(PRED) and sign(OBS13) for the high-skill sample $\left(\chi^2 = 2.6; \alpha \approx 0.1\right)$, but cannot reject independence for the complete sample $\left(\chi^2 = 1.73; \mathrm{N.S.}\right)$.⊥ The results of the comparison,

†Similar levels of fit are obtained in (random) analysis where one of the six prediction tasks is drawn for each subject and regressions are run cross-sample on $N = 65$ observations. The mean R^2 levels obtained in 5000 random estimations are 0.3 for ANN and 0.09 for MON.

‡We thank a referee for suggesting this line of analysis.

§Alternatively, the tests could be run between-subjects by shuffling the 390 predictions across the 390 tasks in each treatment. This procedure generates excess noise because of the heterogeneity across participants. The randomality hypothesis is rejected at $\alpha = 0.04$ for MON and $\alpha = 0.02$ for ANN.

¶'Pr' abbreviates empirical frequencies. The exact figures are provided in supplementary appendix D.

⊥The Chi-square test assumes that individual sign-predictions are independent across tasks. Comparison of proportions on an individual basis, however, is misleading since negative predictions are relatively infrequent, especially in ANN. If some subject, for example, provides five positive predictions and only one negative prediction, and OBS13 is positive in all cases except one of the tasks where the subject has provided a positive prediction, then $\Pr(\mathrm{OBS13}>0 \mid \mathrm{PRED}>0) = 0.8$ while $\Pr(\mathrm{OBS13}>0 \mid \mathrm{PRED}<0) = 1$.

Table 2. Fixed effects linear regression on ERR ($N = 390$).

	STD12	FAMILIARITY	THEORY	R^2
ANN	**0.31**** ($\alpha < 0.01$)	−0.21 (N.S.)	−0.76 (N.S.)	0.17
MON	0.87** ($\alpha = 0.02$)	**−0.62**** ($\alpha < 0.01$)	+0.28 (N.S.)	0.21

The table presents the results of estimating the fixed effects model:
$\mathrm{ERR}_{ij} = \alpha_i + \beta \cdot \mathrm{STD12}_{ij} + \gamma \cdot \mathrm{FAMILIARITY}_i + \delta \cdot \mathrm{THEORY}_i$.
ERR_{ij} denotes the prediction error of subject i in problem j. α_i is an individual intercept for subject i. $\mathrm{STD12}_{ij}$ is the STD12 in the series presented to subject i in problem j. $\mathrm{FAMILIARITY}_i$ is the familiarity index provided by subject i and THEORY_i is the theory rank provided by the subject. The model is separately estimated for ANN and MON. (Individual intercepts are not reported for obvious reasons.)

however, are weaker for MON where the conditional frequency Pr(OBS13 > 0 | PRED > 0) = 60.3% is slightly lower than Pr(OBS13 > 0 | PRED < 0) = 62.4% and independence could not be rejected even for the high-skill sample.

4.5. *Prediction errors increase with STD12 but decrease with FAMILIARITY*

To extract the factors that affect prediction errors across the sample, we run fixed effect linear regressions where ERR is explained by characteristics of the observed return series (STD12, AVG12, TREND12 and others) and individual attributes including the two skill measures: FAMILIARITY and THEORY. Our fixed effects specification allows for heterogeneity in individual intercepts while assuming common coefficients for all explanatory variables. Since the number of possibly relevant explanatory variables is very large, model selection procedures (Green 2003) are employed to eliminate insignificant coefficients. The regressions were run in different versions using distinct collections of explanatory variables and alternative model selection methods in order to verify robustness. The estimations consistently suggest that prediction errors increased with historical volatility (STD12), but interestingly reveal a significant negative FAMILIARITY effect on prediction errors in MON. Table 2 discloses the results of the simplest estimation where ERR is explained by an individual intercept, STD12, and the two skill variables. The coefficients of STD12 are positive and significant, confirming that the participants could not compensate for larger volatility in their judgmental predictions. More interestingly, the FAMILIARITY coefficient, for MON, is negative −0.6 ($p < 0.01$), implying that the prediction errors of subjects with maximal familiarity are about 5.5% lower than the errors of subjects with minimal familiarity, when volatility is controlled.† The difference is large in magnitude relative to the mean ERR in MON: 8.5%. A negative FAMILIARITY coefficient of −0.21 also emerges in ANN, but the noise (standard deviation 0.96) is far too large for significance. The effect of THEORY on prediction errors is insignificant in our sample.

5. Predicting PRED

5.1. *Regressions on individual PREDs*

To characterize individual prediction patterns along the questionnaire, we regress PRED on a variety of statistics including AVG12, STD12, various TREND proxies and more. As the pertinent statistics may strongly vary with the prediction horizon, we separately run the regressions for each condition, ANN and MON. Since prediction patterns, moreover, may strongly differ among individuals, we estimate PRED at the individual level, testing if the six predictions elicited from each participant show consistent response to specific statistics across assignments. If some subject, for instance, used the difference OBS12 − AVG12 (a simplistic moving average rule) to extrapolate OBS13 in all six MON tasks, then the regressions should expose the method exactly. If other participants intuitively adapted PRED across series, the regressions might still disclose the statistics that played a major role in forming individual predictions. Since the sample of PREDs for each participant consists of only six observations, we ran preliminary model-selections to choose the statistics that explained PRED most effectively, in terms of highest average $R2$ across the sample.‡ The selected model for table 3, $\mathrm{PRED} = \beta_0 + \beta_{\mathrm{AVG12}} \times \mathrm{AVG12} + \beta_{\mathrm{OBS12}} \times \mathrm{OBS12} + \beta_{\mathrm{AVG3}} \times \mathrm{AVG3}$, assumes that individual PREDs respond to the average level of return in the historical sequence (AVG12), the most recent observation (OBS12) and the intermediate quarterly trend (AVG3). The mean estimated coefficients and the corresponding standard deviations are displayed in table 3, together with the proportion of positive estimates.§ To summarize the marginal effect of each statistic, we additionally run

†Eight participants reported minimal FAMILIARITY = 1; three participants reported FAMILIARITY = 9 while only one subject proclaimed maximal familiarity. The Spearman correlation between individual THEORY and FAMILIARITY ranks is 0.77.

‡Because of the small number of observations per subject (six) we refrain from statistical inference and model selection at the individual level. The subject-level OLS estimations are used to approximate individual predictions, but tests and model selections are only applied on the complete sample.

§The hypothesis $\beta_0 + \beta_{\mathrm{AVG12}} + \beta_{\mathrm{OBS12}} + \beta_{\mathrm{AVG3}} = 1$ could not be rejected for MON but was marginally rejected for ANN at $\alpha = 0.07$. Our sample size appears to be too small (relatively to the highly volatile random historical sequences) to demand PRED to take the form of a weighted average of sequence statistics.

Table 3. Regression of PRED on sample statistics.

	β_0	β_{AVG12}	β_{OBS12}	β_{AVG3}	R^2
ANN					
Mean	−0.02	**0.79∗∗**	**0.25∗∗**	**−0.19∗**	0.72
(Standard deviation)	(0.32)	**(1.75)**	**(0.78)**	**(0.98)**	(0.24)
% positive coeffs	54%	**75%∗∗**	62%∗∗	47%	—
Fixed effects regressions	−0.01	**0.66∗∗**	**0.17∗∗**	−0.13∗	0.69–0.75
MON					
Mean	0.01	**0.55∗∗**	0.25∗∗	−0.02	0.75
(Standard deviation)	(0.08)	**(4.04)**	(1.35)	(2.69)	(0.24)
% positive coeffs	55%	**66%∗∗**	**65%∗∗**	47%	—
Fixed effects regressions	−0.00	**1.07∗∗**	**0.15∗∗**	−0.13	0.68–0.72

The equation PRED $= \beta_0 + \beta_{\text{AVG12}} \times \text{AVG12} + \beta_{\text{OBS12}} \times \text{OBS12} + \beta_{\text{AVG3}} \times \text{AVG3}$ is OLS estimated for each participant and each condition. The results for ANN/MON are separately summarized in the upper/lower panels of the table. The upmost line in each panel displays the mean and standard deviation of estimated coefficients ($N = 65$). The hypothesis $\beta_x = 0$ (throughout the sample) is tested using the Wilcoxon signed-rank test and the asterisk convention is used to represent significance. The intermediate line in each panel presents the proportion of positive coefficients, using asterisks to represent significance by sign-test. The line at the bottom of each panel shows the results of fixed effects estimations where all coefficients may vary at the individual level, except for the specific coefficient tested in the corresponding estimation; e.g. the fixed effects estimations for AVG12 allow for individual variation in β_0, β_{OBS12}, β_{STD12} but assume a single β_{AVG12} coefficient to summarize the response to AVG12 across the sample.

fixed effects regressions, allowing for individual heterogeneity in intercepts and slopes, except for the specific coefficient tested in each estimation; e.g., the fixed effects estimations for AVG12 allow for individual variation in β_0, β_{OBS12} and β_{STD12} but assume a single β_{AVG12} coefficient to summarize the response to AVG12 across the sample.

Unsurprisingly, the estimations reveal positive significant coefficients for both AVG12 and OBS12, in both conditions (see the table). Individual predictions therefore appear to increase with past return levels, while the most recent observation plays an additional significant role that could not be observed for the preceding realizations. When OBS11 is used to approximate recent trends instead of OBS12, the fixed effects regressions reveal an insignificant coefficient $\beta_{\text{OBS11}} = 0.04$ in ANN and a negative coefficient $\beta_{\text{OBS11}} = -0.15$ ($p < 0.01$) in MON (the AVG3 coefficients become positive to substitute for the loss of OBS12). The heterogeneity in individual response to AVG12 and OBS12 still reflects in negative β_{AVG12} (β_{OBS12}) coefficients for 25%–35% of the participants. The (unconditional) correlation between PRED and AVG12 is accordingly negative for 32% of the participants in ANN and 28% of the subjects in MON.

In addition, the estimations suggest that—when AVG12 and OBS12 are accounted—predictions tend to decrease with AVG3, in both conditions. The effect is consistent, across the sample, in ANN where the fixed effect regressions expose a marginally significant negative coefficient $\beta_{\text{AVG3}} = -0.13$ ($\alpha = 0.09$), but fail to reach significance in MON (see the table). The negative role of AVG3 may represent a natural tendency to reduce predictions when recent trends are extreme (direct comparisons follow in section 5.2). In the annual tasks, moreover, the negative response may relate to the 3–5 years contrarian cycles documented by De Bondt and Thaler (1985, 1987). If subjects recall the empirical evidence, then they could expect AVG3 to affect returns adversely starting at period 13, at least for extreme AVG3 levels.†

When other statistics that could explain predictions intuitively, like various TREND measures, are appended to the prediction model of table 3, the individual-level estimates are mixed in sign and the fixed effects coefficients are insignificant. In particular, the model selections suggest that the volatility of historical returns did not affect predictions consistently across the sample. When STD12, for example, is appended to the list of explanatory statistics, the coefficients β_{STD12} are mixed in sign and the fixed effects regressions confirm that subjects did not account historical volatility consistently while forming their predictions ($\beta_{\text{STD12}} = +0.09$ in ANN and $\beta_{\text{STD12}} = -0.09$ in MON; $\alpha = 0.3$ N.S. in both conditions).

5.2. Switch in prediction regimes, from trend-chasing to contrarian

In response to the METHOD query on the last page of the questionnaire, more than 20% of the participants mentioned attempts to 'recognize patterns' in the historical sequences. Several subjects were more specific, claiming to switch from trend-chasing to contrarian forecasting following large positive streaks which 'increase the chances for price adjustment'. To illustrate the concurrent effects on experimental predictions, let the terms *trend-chasing* or *momentum-based prediction* denote cases where subjects provide positive/negative PREDs for

†Other proxies for recent trends like AVG5 generate similar but weaker results.

Table 4. Proportion of negative predictions in high/low streak-size group.

	Streak size $\leq$ median	Streak size $>$ median	Sign test
ANN	14.5% (20/139)	24.5% (34/139)	N.S.
MON	26% (28/108)	41% (44/107)	$\alpha = 0.04$

Return series are classified as ending with a positive streak if OBS12 > 0. The size of the streak is the sum of consecutive positive returns at the end of the series. The sample of series ending with a positive streak ($N = 278$ for ANN and $N = 215$ for MON) are median-split by the size of the streaks. The proportion of negative (contrarian) predictions is calculated for each subgroup. The left column presents the proportion of negative predictions in the series ending with a positive streak smaller than the median. The right column presents the proportion for the series ending with a positive streak larger than the median. A sign-test (on individual proportions) is applied to determine significance.

historical series that end with a sequence of positive/negative returns, respectively. The term *contrarian prediction*, on the other hand, is kept for cases where predictions are negative/positive when histories end with a sequence of positive/negative returns. The interplay of momentum-based and contrarian forecasting is illustrated in table 4. The table summarizes the judgmental response to streaks of positive returns. Say that the return sequence OBS1, OBS2, ..., OBS12 ends with a positive streak if OBS12 > 0. The size of the streak is the sum of successive positive returns at the end of the sequence. The table median splits the series ending with a positive steak ($N = 278$ for ANN and $N = 215$ for MON) by the size of the streak, and calculates the proportion of contrarian (negative) predictions for each subgroup. The left column reveals the proportion of PRED < 0 in histories with streak-size smaller than median, while the right column presents the corresponding proportion for histories with streak-size larger than median. The proportion of negative predictions is higher in the large streak group in both conditions, reflecting increased expectation for trend reversal following heavy streaks. The difference is more pronounced in MON where the proportion of contrarian predictions following large streaks is 41% compared to 26% for the small streak group (sign-test on individual proportions; $\alpha < 0.04$). The differences are smaller for the ANN series where we cannot reject the hypothesis that negative PREDs are as likely in the two groups ($\alpha = 0.28$).†

Similar conclusions emerge in alternative forms of analysis: while predictions, for instance, are significantly lower than AVG12 in the large streak-size group (mean PRED − AVG12 = −6.6 in ANN, −0.8 in MON), PRED is either closer or even higher than AVG12 in the small streak-size group (PRED − AVG12 = −2.3 in ANN, +2 in MON).

6. Separating BEST stocks from WORST

The comparison of BEST versus WORST stocks by individual predictions is run at three levels. First, we compare the payoff on the stock that received the highest prediction in each questionnaire (BEST1) to the payoff on the stock that got the lowest prediction (WORST1). We then expand the portfolios of best and worst stocks and compare the average payoff on the two stocks with highest predictions (BEST2) to the average payoff on the two stocks that have received the lowest predictions (WORST2). Finally, we split the six-stock menu of each participant into two groups of three stocks and compare the average payoff on the three stocks with highest predictions (BEST3) to the average payoff on the remaining three stocks (WORST3). The mean best and worst payoffs are contrasted in table 5. The upper panel presents the results for the complete sample while the lower panel restricts the comparisons to the 34 participants with FAMILIARITY $\geq$ 4. Standard deviations are disclosed in smaller brackets and asterisks denote cases where the hypothesis BESTn = WORSTn is rejected in randomization tests as illustrated in section 3 (the shading is explained below).‡

Consider the mean payoffs first. The BESTn payoff is larger than the corresponding WORST n payoff in all 12 cells of table 5. The largest differences appear in comparison of BEST1 and WORST1. The mean payoff on the stocks that attracted the highest predictions in ANN is almost twice larger than the mean payoff on the worst stocks. BEST1 is larger than WORST1 for 39 of 65 participants, but the proportion increases to 21 of 34 subjects in the high-skill group. The hypothesis BEST1 = WORST1 is rejected in randomizations for the complete sample ($\alpha = 0.02$) and for the high-skill subsample ($\alpha = 0.07$).

Large differences between BEST1 and WORST1 payoffs also emerge in MON. The mean BEST1 payoff (3.0) is 2/3 higher than the mean WORST1 payoff (1.8) when the complete sample is examined. BEST1, however, is lower than WORST1 for 34 of the 65 subjects, and the randomizations reveal that larger differences in mean payoffs emerge in 22% of the cases when stocks are randomly assigned to the BEST1/WORST1 categories. The mean difference BEST1 − WORST1, however, increases to 3.5 (representing 3.5% monthly return) when the comparison is restricted to the high-skill subjects. The number of high-skill subjects with BEST1 > WORST1 (19 of 34) is still not large enough for sign-test significance but the randomizations suggest that larger differences in mean payoffs arise in less than 6% of the cases under random selection of BEST1/WORST1 stocks for the restricted sample.

†The weaker results for ANN compared to MON seem surprising given the established empirical evidence on momentum and contrarian cycles in annual return series. Our random sample of historical series, however, (apart from individual heterogeneity) may not be large enough to pick the typical extent of predictable return variation. A more powerful test could be constructed by selectively picking a sample of historical series where momentum versus contrarian predictions could be separated more easily.

‡The bracketed '(22/19)' at the BEST2 / WORST2 column, for example, denotes the standard deviations (rounded to integers) of the BEST2 / WORST2 payoffs across the sample ($N = 65$).

Table 5. Mean payoff (standard deviation) on best versus worst stocks.

	Best1/Worst1	Best2/Worst2	Best3/Worst3
Complete sample ($N = 65$)			
ANN	**26.5/14.5**$**$ (42/27)	19.7/15.3 (22/19)	17.7/17.3 (18/16)
MON	3/1.8 (10/10)	3.4/2.5 (6/6)	3.0/2.2 (4/5)
High-skill subjects ($N = 34$)			
ANN	25.1/13.2$*$ (41/26)	19.5/12.5$*$ (23/18)	19.0/16.2 (18/18)
MON	5.0/1.5$*$ (10/10)	4.2/2.2$*$ (5/6)	3.5/1.5$*$ (4/4)

The table compares the payoffs on the best stocks by individual predictions to the payoffs on the worst stocks. The realized payoff (OBS13 × 100) on the stock that received the highest prediction is BEST1. The average payoff on the two stocks that obtained the highest predictions is BEST2, while BEST3 is the average payoff on the three stocks with highest PRED. WORST1, WORST2 and WORST3 are similarly defined. Randomization tests are applied to test the significance of the differences between BESTn and WORSTn payoffs (for $n = 1,2,3$). In each randomization, the six stocks in each questionnaire are ranked randomly and the randomized BESTn/WORSTn payoffs are re-calculated for the shuffled series. The proportion (out of 1000 randomizations) where larger differences in mean randomized BESTn and WORSTn payoffs arise constitutes the significance level of the test. The results of the test are summarized using the asterisk convention. The randomizations are also used to calculate the proportion of cases α' where larger differences in mean payoffs arise with smaller differences in standard deviations. Cases where $\alpha' \leq 0.05$ are marked by darker shading while cases where $0.05 < \alpha' \leq 0.1$ are marked with lighter shading. The lower panel restricts the comparison to the subjects with FAMILIARITY ≥ 4.

The best-stock portfolios (by actual predictions) still show higher average payoffs compared to worst-stock portfolios, when the comparison is extended to include the two or three best/worst stocks in each questionnaire. The differences in mean payoffs are not significant when the complete sample is examined, but significance emerges in five of six comparisons for the high-skill subjects (see the asterisks in table 5).†

Significance, moreover, improves in joint testing of mean return and volatility. The bracketed STD data in table 5 reveals higher standard deviations on the BESTn portfolios, compared to the WORSTn portfolios, in five of the six comparisons for ANN. The standard deviations, however, are similar in MON, despite the larger mean payoffs on the best stock portfolios. To compare jointly the mean and standard deviation of best/worst payoffs, we use the randomizations to calculate the proportion α', where randomized portfolio selections show larger differences in mean payoffs together with smaller differences in STD. The results are summarized in table 5 by shading the cells where the proportion $\alpha' < 0.1$. Light shading is used for $0.05 \leq \alpha' < 1$ and darker shading for cases where $\alpha' < 0.05$. The equality of best and worst payoffs is now rejected in 11 of 12 comparisons. The randomizations for MON (complete sample), in particular, reveal that less than 5% of the randomized BEST1/WORST1 portfolios show larger differences in mean return jointly with smaller differences in STD. The hypothesis BEST1 = WORST1 is therefore rejected at $\alpha = 0.05$ in the joint test. Recall that when standard deviations are ignored, the randomizations show larger BEST1 − WORST1 differences in 22% of the cases. The joint test thus reveals that equivalently large differences in mean payoffs are achieved—in most cases—jointly with larger differences in volatility. The BEST1 stocks by experimental predictions, on the other hand, pay higher mean returns than the WORST1 stocks without such increases in STD. Similar conclusions are drawn in the other joint examinations.

7. Statistical forecasting

To conclude the analysis, we compare PRED to several intuitively plausible statistical rules. The analysis for AVG12, for example, compares the stationary prediction rule PRED′ ≡ AVG12 to the experimental PRED, in terms of prediction errors (|PRED′ − OBS13|), the ability to predict sign(OBS13), and the separation of BEST stocks from WORST. The analysis for other statistics repeats the same methodology, assuming the designated statistic is used to predict OBS3 across the sample, and comparing the performance of actual PRED to the statistical PRED′ in these various dimensions.

The comparison of prediction errors is summarized in table 6, where we contrast PRED with the empirical rules: AVG12, AVG6, AVG3 and OBS12. In the groundwork, we examined a longer comprehensive list of possibly relevant statistics, but confined the table to the statistics with lowest mean absolute error. The comparison does

†Note, however, that the payoff on the second-best stock (alone) in ANN (averaging at 13) is lower than the payoff on the second-worst stock (16). Similarly, the third-best stock pays less than the third-worst stock. The relative strength of BEST2 and BEST3 portfolios in ANN therefore follows from the stronger performance of BEST1 compared to WORST1. The results are more robust in MON where the stocks that received the second- (third-)highest PRED outperform the stocks that received the second- (third-)worst PRED, respectively. Table E in the supplement presents the single-stock version of table 5.

Table 6. Statistical prediction.

	ANN			MON		
	Mean	ERR	*P*-test	Mean	ERR	*P*-test
OBS13	17.5%	Na	Na	2.6%	Na	Na
PRED	15.3%	28.2	$\alpha = 0.039$	1.6%	8.5	$\alpha = 0.06$
AVG12	22.9%	24.5	$\alpha = 0.044$	1.7%	6.8	$\alpha = 0.17$
AVG6	20.5%	24.9	$\alpha = 0.07$	1.6%	7.0	$\alpha = 0.11$
AVG3	17.2%	26.2	$\alpha = 0.09$	1.5%	7.5	$\alpha = 0.08$
OBS12	15.6%	31.5	$\alpha = 0.28$	1.5%	10.0	$\alpha = 0.42$

The experimental predictions (PRED) are compared with the statistical rules: AVG12, AVG6, AVG3 and OBS12. The mean predictions by each model are disclosed in the 'Mean' columns. The mean absolute prediction errors are presented in the columns headed ERR. The value of ERR for AVG12, for instance, is the mean value of the distances $|\text{AVG12} - \text{OBS13}|$. The columns headed *P*-test present the results of a permutation test, as discussed in section 3, on the corresponding prediction model.

not provide clear insights. PRED is closer to OBS13, compared to most statistical rules, in ANN, but the absolute prediction errors of the statistical forecasts are lower in both conditions (except for OBS12, which is discussed separately below).† The permutation tests for each statistic are summarized in the respective columns of table 6. The *P*-test for AVG12, for instance, repeats the permutation procedure described in 4.3, assuming $\text{PRED}' \equiv \text{AVG12}$ throughout the sample. The permutations suggest that the correspondence between PRED and OBS13 is stronger than the link between each sample statistic and subsequent returns. The shuffling of AVG12 across series, for instance, decreases the mean absolute prediction error in 4.4% of the permutations for ANN and 17% of the permutations for MON (compared to the 3.9% and 6% *P*-test significance levels for PRED).

The statistical prediction rules (with the exception of OBS12), moreover, fail completely in crude prediction of sign(OBS13). In the ANN assignments, for example, AVG12, AVG6 and AVG3 are positive in 99%, 96% and 83% of the cases, while $\text{OBS13} > 0$ in only 76% of the series. The proportion of experimental $\text{PRED} > 0$, on the other hand, is 75%, which almost coincides with the proportion of positive OBS13. Similar conclusions emerge for MON where $\text{PRED} > 0$ in 62% of the cases; $\text{OBS13} > 0$ for 60% of the series; while AVG12/AVG6/AVG3 are positive in 81%, 73%, 66% of the cases. The results for OBS12, in this respect, are different, as the proportion of $\text{OBS12} > 0$ is naturally close to the proportion of positive OBS13 in the random series. OBS12, however, exhibits the weakest performance in *P*-tests: the random shuffling of OBS12 across historical series decreases the mean ERR in 28% of the permutations for ANN and 42% of the permutations for MON.

The strongest result for the experimental predictions emerges regarding the separation of BEST1 from WORST1 in the random six-stock menus. To compare PRED to the statistics in this respect, we recalculate BESTn and WORSTn using the designated statistics to sort the six stocks in each menu. BEST1, for AVG12, for example, represents the eventual payoff on the stock with highest historical mean. WORST1, for AVG12, similarly denotes the payoff on the stock with lowest mean historical return. BEST2, WORST2, BEST3 and WORST3 are similarly defined, for AVG12, and the paired differences BESTn − WORSTn are used to test AVG12's ability to separate BEST stocks from WORST. The best and worst payoffs for other statistics are calculated similarly, using each statistic in turn to rank the six stocks comprising the individual prediction menus. Table 7 compares the best and worst payoffs by experimental PRED to the corresponding payoffs for AVG12, AVG6, AVG3 and OBS12, using the asterisks and shading conventions of table 5 to mark cases where the hypothesis (payoff by experimental PRED) = (payoff by statistical rule) is rejected in randomization tests.‡ To test if historical volatility could screen the best stocks from the worst, we also sort the menus by STD12. Finally, we include an extrapolated return measure, TREND, to test if the best stocks could be separated from the worst by extrapolating OBS13 from the series.§ Consistent significant differences between PRED and the statistical rules appear in two levels of comparison, underlined by marking in bold type the corresponding columns in table 7.¶ In both columns, actual predictions significantly outperform the statistical rules.

(I) The mean payoff on the stocks that received the highest predictions in ANN (BEST1) is about

†Prediction errors could get closer to those implied by empirical rules like AVG12 in a different experimental design where each subject is requested to forecast many more series, say 100 series, and payouts are derived from average prediction errors along the 100 tasks. Our experiment is different, as subjects specifically attempt to decipher the code of each series. The attempts are surprisingly successful in terms of separating BEST1 from WORST1, but prediction errors are larger.

‡The only difference is that the randomizations now independently select the two portfolios (allowing for the possibility that the same portfolio would be selected twice), since the best or worst portfolio by PRED may coincide with the corresponding portfolio for the statistical rule.

§Our specific TREND formula is disclosed in table 7.

¶The differences are almost consistent in the BEST1 column for MON where the mean payoff by PRED is significantly lower than the mean payoff by AVG12, AVG6, STD12 and TREND. PRED, however, beats PRED3 in terms of identifying BEST1 in MON. These results are discussed later in the text.

Table 7. Comparison of the payoffs implied by PRED to the payoffs implied by statistical ranking.

		Best1	Best2	Best3	Worst1	Worst2	Worst3
ANN	PRED	26.5 (42)	19.7 (22)	17.7 (18)	14.5 (27)	15.3 (19)	17.3 (16)
	AVG12	17.7^{**} (36)	22.3 (28)	21.0^{*} (20)	12.9 (21)	13.4 (17)	14.0^{*} (14)
	AVG6	$\mathbf{16.5}^{**}$ (37)	16.9 (24)	17.8 (19)	13.2 (25)	15.5 (18)	17.2 (15)
	AVG3	17.3^{**} (39)	18.4 (27)	18.6 (19)	18.1 (32)	16.8 (23)	16.4 (16)
	STD12	20.3^{*} (44)	23.8 (29)	22.7^{**} (21)	14.4 (22)	12.1 (16)	12.3^{**} (13)
	OBS12	14.1^{**} (37)	16.3 (26)	16.3 (18)	15.2 (33)	19.8^{*} (25)	18.7 (19)
	TREND	12.4^{**} (30)	15.4^{*} (22)	14.8 (18)	10.1 (25)	15.1 (18)	20.2 (15)
MON	PRED	3.0 (10)	3.4 (6)	3.0 (4)	1.8 (10)	2.5 (6)	2.2 (5)
	AVG12	5.5^{*} (12)	3.9 (6)	3.3 (5)	4.5^{**} (10)	2.7 (7)	1.9 (5)
	AVG6	5.8^{**} (11)	3.5 (7)	3.0 (5)	$\mathbf{4.9}^{**}$ (11)	2.7 (6)	2.2 (4)
	AVG3	2.8 (12)	3.4 (6)	3.1 (4)	4.0^{*} (9)	2.3 (6)	2.1 (5)
	STD12	5.4^{*} (13)	4.6 (8)	3.9^{*} (6)	1.6 (8)	1.2^{*} (4)	1.3^{*} (4)
	OBS12	2.0 (8)	2.5 (6)	2.4 (4)	3.3 (11)	3.0 (6)	2.8 (5)
	TREND	4.6 (9)	4.5 (7)	3.2 (4)	1.6 (10)	1.8 (7)	2.0 (5)

The payoffs on the best/worst stocks by experimental predictions are compared to the payoffs on best/worst stocks by statistical rankings. BEST*n* and WORST*n* for PRED are replicated from table 5. BEST1, for AVG12, is the payoff on the stock with highest AVG12; Best2 is the average payoff on the two stocks with highest AVG12, etc. TREND is approximated multiplying AVG6 by the ratio of (mean return in the second half of the series: OBS7 – OBS12) to (mean return in the first half: OBS1 – OBS6).†Randomization tests are applied to test the hypotheses (payoff by PRED = payoff by the statistical model) for the various payoffs. In each randomization, the six prediction problems in each questionnaire are randomly sorted, twice, and the corresponding BEST*n* or WORST*n* payoffs (for each level of comparison $n = 1,2,3$) are calculated for each series. The proportion of randomizations where the difference in mean randomized payoffs is larger than |mean payoff by PRED – mean payoff by the statistical model| constitutes the significance level of the test and the asterisks convention is applied to mark significant differences. As in table 5, the randomizations are also used to calculate the proportion of cases α' where larger differences in mean payoffs arise with smaller differences in standard deviations. Cases where $\alpha' < 0.05$ are marked by darker shading while cases where $0.05 \leq \alpha' \leq 0.1$ are marked with lighter shading.

60% larger from the mean payoff on the stocks that were ranked highest by the empirical rules AVG12, AVG6 and AVG3. The mean BEST1 payoff by PRED is 26.5, while the corresponding payoff for AVG12 is only 17.7. Similar large differences emerge for AVG6 and AVG3 (see the table). The randomizations suggest that the differences in mean payoffs are statistically significant at $\alpha < 0.05$, although we cannot reject equality of standard deviations. The most recent annual return (OBS12) surprisingly shows very weak results in terms of BEST1, with an average payoff of 14.1, almost 50% lower than the corresponding PRED-based payoff. Momentum strategies, by way of interpretation, appear ineffective for the cross-section and time samples utilized in the experiments. The STD12 criterion for tracing the BEST stock approaches PRED with mean BEST1 payoff of 20.3, but the randomizations still suggest that PRED defeats STD12 significantly, in particular when the lower volatility of BEST1 payoffs by PRED (standard deviation = 42) compared to the volatility of BEST1 payoffs by STD12 (standard deviation = 44) is taken into consideration (*R*-test significance; $\alpha = 0.01$).‡

†The mean absolute prediction errors of TREND were extremely large: 62.8 in ANN and 15.8 in MON. The results for other extrapolation models were qualitatively similar but TREND did best in terms of identifying WORST1 stock in MON.

‡If this deserves clarification, significance by the joint test would be weaker (3.3% compared to 1.3%) if the standard deviations were reversed, with BEST1 by PRED showing the higher standard deviation (44) and BEST1 by AVG12 showing the lower (42) volatility.

(II) The mean payoff on the stocks that received the lowest predictions (WORST1) in MON is significantly lower than the mean payoff on the stocks that were ranked lowest by AVG12, AVG6, AVG3 and OBS12. The payoff on the worst stock by AVG12, for example, averaged at 4.5, compared to the 1.8 WORST1 payoff by PRED. Similar large differences appear for AVG6, AVG3 and OBS12, while STD12 and TREND deliver WORST1 payoffs similar to PRED (see the table). The randomizations confirm that subjects selected the worst stock more effectively than each of the four empirical-mean rules: AVG12, AVG6, AVG3 and OBS12, but cannot reject the equality of WORST1 payoffs for PRED, STD12 and TREND.

To further discuss results **(I)** and **(II)**, let ARB1 ≡ BEST1 − WORST1 represent the payoff on hypothetical 'arbitrage' where subjects buy the stock that is expected to pay the highest return while short-selling an equal amount of the stock that appears least attractive. When the best and worst stocks are determined by PRED, the mean ARB1 payoff in ANN is about 12 (representing 12% annual return on the arbitrage volume) with standard deviation 49. Randomizations of similar one-stock arbitrage portfolios suggest that more efficient portfolios, in terms of higher mean return and lower standard deviation, emerge in less than 2% of the randomizations, confirming the significance of ARB1 by experimental predictions. The ARB1 payoffs for each of the six statistics examined in table 7, however, are more than 50% lower, on average. The strongest results (among the statistical rules) emerge for STD12, with mean ARB1 payoff 5.9 and standard deviation 52. The randomizations suggest that the highly volatile STD12-based ARB1 portfolio is inefficient, as more than 13% of the randomized ARB1 portfolios show higher mean return with lower standard deviation. The ARB1 payoffs for AVG12, on the other hand, are efficient (mean return 4.8, standard deviation 41; *R*-test significance $\alpha = 0.06$), but pay 60% less, on average, than the corresponding PRED-based portfolio. The ARB1 results for the remaining statistics are even weaker, with negative mean returns for AVG3 and OBS12 and modest 2.3–3.4% returns on TREND and AVG6.† Actual predictions, in conclusion, strongly outperform the six empirical rules (and other rules examined in the groundwork), in terms of separating BEST1 from WORST1 in ANN.‡

The ARB1 results for MON however are much weaker. The mean ARB1 payoff by PRED (1.27) is close to the mean ARB1 payoff by AVG12 (1.10) and the two portfolios show similar standard deviations (around 15). Table 7 clarifies that the advantage of PRED in terms of identifying WORST1 (point **II** above) is almost cancelled out by AVG12's superior ability in picking BEST1. The randomizations reveal that neither experimental PRED nor AVG12 could direct the subjects into efficient stock-selection: the proportion of randomized ARB1 portfolios with higher mean return and lower standard deviation is about 20% for PRED and 23% for AVG12. The ARB1 results for AVG6 are even weaker (mean 0.9, standard deviation 16; *R*-test significance 0.30) while AVG3 and OBS12 show negative mean ARB1 payoff (about −1.3) with standard deviation 14. The hypotheses (ARB1 by PRED) = (ARB1 by AVG3) and (ARB1 by PRED) = (ARB1 by OBS12) are rejected at $\alpha < 0.09$ in randomization tests. PRED therefore exhibits similar or stronger ARB1-performance compared to the prediction rules based on empirical means, including OBS12. PRED, however, is strongly defeated by STD12 and TREND, which outperform the experimental predictions in terms of identifying the BEST1 stock while showing similar performance in terms of WORST1 payoffs. The mean ARB1 payoff by STD12, 3.9, is about three times larger than the corresponding payoff for PRED, with standard deviation 16 and *R*-test significance $\alpha < 0.01$. The respective statistics for TREND are 2.9, 14, $\alpha = 0.03$. More generally, the STD12 rows in table 7 suggest that except for PRED's ability to identify BEST1 more effectively in ANN, STD12 outperforms PRED consistently. As mentioned in section 5, STD12 does not emerge as a significant factor in the analysis of prediction patterns across the sample. It is interesting to note in conclusion that historical volatility exhibits relatively strong performance in terms of separating the best stocks from the worst in our cross-section and time samples, while the participants neglect this criterion in their judgmental predictions.

8. Discussion

The rich literature on judgmental return forecasting demonstrates that prediction patterns may vary drastically based on subtle properties of the underlying series (Czaczkes and Ganzach 1996) and respond to a variety of salient statistics (Mussweiler and Schneller 2003). It is also acknowledged that prediction patterns are heterogeneous and strongly differ between individuals (Dominitz and Manski 2011). Lawrence *et al*. (2006) contend that judgmental adjustment of statistical models may improve the quality of predictions significantly, especially when the forecaster holds domain knowledge that the rigid statistical rule may miss. Our experiment accordingly

†The ARB1 results for AVG6 (in ANN) are: mean return 3.4, standard deviation 46, *R*-test $\alpha = 0.17$. The corresponding results for AVG3/OBS12/TREND are: mean return −0.8/−1.1/2.3, standard deviation 49/45/36, and *R*-test significance 0.47/0.61/0.02. The ARB1 portfolios by TREND are therefore highly efficient, but the mean return is about 80% lower than PRED.

‡The relative strength of PRED partially disappears when the arbitrage is extended to include the two or three best and worst stocks in each menu. The mean ARB2 payoff for PRED, for instance is 4.5 (standard deviation 30), compared to mean ARB2 payoff of 8.9 with STD = 35 for AVG12 ($p = 0.08$ by joint *R*-test).

demonstrates that MBA students with a keen interest in finance are able to surpass the most intuitive empirical rules in certain aspects of prediction (e.g. identifying BEST1 in ANN and WORST1 in MON). The strong results are intriguing in light of our random sequence procedure that kept the identity of underlying stocks and specific market conditions strictly concealed.

Our results also relate to the literature on technical trading and recent AI attempts to 'decipher the code' of stock market returns. The path breaking study of Brock *et al.* (1992) showed that popular technical-trading heuristics earn positive returns that cannot be explained by the stochastic process of the market. The payoffs on technically generated 'buy' signals are higher, and less volatile, than the payoffs on corresponding 'sell' signals. Park and Irwin (2007) more recently conclude that 56 out of 95 modern studies find positive results, while only 20 studies reach negative conclusions, regarding the effectiveness of technical trading. Technical rules provide additional channels for exploiting historical data in judgmental forecasting. Several participants indeed mentioned the use of technical heuristics in their comments regarding prediction methods.† The intuitions underlying technical trading heuristics; e.g. the belief that trend is initiated when the short-run average penetrates the long-run average by a significant band, could direct predictions when appropriate. The predictability of returns from historical data is also at the focus of many recent AI studies (see for examples the surveys in Binner *et al.* 2005 and McNelis 2005). Qi (1999) concludes that portfolios based on recursive neural network forecasts generate higher profits with lower risks than the buy-and-hold market portfolio. Ince (2006) argues that neural network models outperform ARCH/GARCH in fitting empirical stock return data. Neural network techniques could, in principle, shed more light on the dynamic tuning of experimental predictions to specific properties of the underlying series, but the current study was not designed for this purpose and the samples of only six observations per subject preclude serious exploration of individual prediction patterns.

Acknowledgements

We have benefited from comments and conversations with Uri Benzion, Daniel Citrin, Asaf Heftol, Anat Hinezon, Ido Kallir, Guy Levy, Shiki Levy, Marco Monti, Hillel Mogel, Carsten Schmidt, Ziv Segal, Dale Stahl and participants at the FURXII conference in Barcelona, the 2007 meetings of the French Economic Association and the 2008 'Psychology and Investment' meetings at the College of Management. We thank the research authority at the Business School, College of Management, for financial support and two anonymous referees for constructive suggestions.

References

Andersson, P., Memmert, D. and Popowicz, E., Forecasting outcomes of the World Cup 2006 in football: Performance and confidence of betters and laypeople. *Psychol. Sport & Exercise*, 2009, **10**, 116–123.

Barberis, N. and Thaler, R., A survey of behavioral finance. In *Handbook of the Economics of Finance*, edited by G.M. Constantinides, M. Harris and R.M. Stulz, Vol. 1, pp. 1053–1128, 2003 (Elsevier: The Netherlands), Chap. 18.

Binner, J.M., Kendall, G. and Chen, S.H., (Eds.), *Applications of Artificial Intelligence in Finance and Economics*, Advances in Econometrics Vol. 19, 2005 (Emerald Group Publishing Limited).

Bloomfield, R. and Hales, J., Predicting the next step of a random walk: Experimental evidence of regime-shifting beliefs. *J.Finan. Econ.*, 2002, **65**, 397–414.

Brock, W., Lakonishok, J. and LeBaron, B., Simple technical trading rules and the stochastic properties of stock returns. *J. Finan.*, 1992, **47**, 1731–1764.

Czaczkes, B. and Ganzach, Y., The natural selection of prediction heuristics: Anchoring and adjustment versus representativeness. *J. Behav. Decis. Mak.*, 1996, **9**, 125–139.

De Bondt, W. and Thaler, R., Does the stock market overreact? *J. Finan.*, 1985, **40**, 793–805.

De Bondt, W. and Thaler, R., Further evidence on investor overreaction and stock market seasonality. *J. Finan.*, 1987, **42**, 557–581.

De Bondt, W.F., What do economists know about the stock market? *J. Portfolio Mgmt*, 1991, **17**, 84–91.

De Bondt, W.F., Betting on trends: Intuitive forecasts of financial risk and return. *Int. J. Forecasting*, 1993, **9**, 355–371.

de Long, B.J., Shleifer, A., Summers, L.H. and Waldmann, R.J., Noise trader risk in financial markets. *J. Polit. Econ.*, 1990, **98**, 703–738.

Diebold, F.X. and Lopez, J.A., Forecast evaluation and combination. In *Handbook of Statistics*, edited by G.S. Maddala and C.R. Rao, Vol. 14, 1996 (Elsevier Science: The Netherlands).

Dominitz, J. and Manski, C.F., Measuring and interpreting expectations of equity returns. *J. Appl. Econometr.*, 2011, **26**, 352–370.

Du, N. and Budescu, D.V., Does past volatility affect investors' price forecasts. *Int. J. Forecasting*, 2007, **23**, 497–511.

Durham, G.R., Hertzel, M.G. and Martin, J.S., The market impact of trends and sequences in performance: New evidence. *J. Finan.*, 2005, **35**, 2551–2569.

Eleswarapu, V. and Reinganum, M.R., The predictability of aggregate stock market returns: Evidence based on glamour stocks. *J. Bus.*, 2004, **77**, 275–294.

Elliott, W.B., Hodge, F. and Jackson, K.E., The association between nonprofessional investors' information choices and their portfolio returns: The importance of investing experience. *Contemp. Account. Res.*, 2008, **25**, 473–498.

Fama, E.F. and French, K., Dividend yield and expected stock returns. *J. Finan. Econ.*, 1988, **22**, 3–25.

Gilovich, T., Vallone, R. and Tversky, A., The hot hand in basketball: On the misperception of random sequences. *Cogn. Psychol.*, 1985, **17**, 295–314.

Glaser, M., Weber, M. and Langer, T., Overconfidence of professionals and lay men: Individual differences within and between tasks? Working Paper, Mannheim University, 2005.

Glaser, M., Langer, T., Reynders, G. and Weber, M., Framing effects in stock market forecasts: The difference between asking for prices and asking for returns. *Rev. Finan.*, 2007, **11**, 325–357.

†One subject mentioned a comparison of short-run trends versus long-run trends in the spirit of the 'moving average' rule.

Good, P., *Permutation, Parametric and Bootstrap Tests of Hypotheses*, 2005 (Springer: New York).

Green, W.H., *Econometric Analysis*, 2003 (Prentice Hall: Upper Saddle River, NJ).

Griffin, D. and Brenner, L., Perspectives on probability judgment calibration. In *Blackwell Handbook of Judgment and Decision Making*, edited by D.J. Koehler and N. Harvey, pp. 177–199, 2004 (Blackwell: Oxford, UK).

Haigh, M.S. and List, J.A., Do professional traders exhibit myopic loss aversion? An experimental analysis. *J. Finan.*, 2005, **60**, 523–534.

Harrison, W.G. and List, J.A., Field experiments. *J. Econ. Lit.*, 2004, **42**, 1009–1055.

Haruvy, E., Lahav, Y. and Noussair, C.N., Traders' expectations in asset markets: Experimental evidence. *Amer. Econ. Rev.*, 2007, **97**, 1901–1920.

Hey, J. and Lee, J., Do subjects separate (or are they sophisticated)? *Exp. Econ.*, 2005, **8**, 233–265.

Ince, H., Non-parametric regression methods. *Comput. Mgmt Sci.*, 2006, **3**, 1619–1697.

Kaul, G., Predictable components in stock returns. In *Handbook of Statistics*, edited by G.S. Maddala and C.R. Rao, Vol. 14, 1996 (Elsevier Science: The Netherlands).

Lawrence, M., Goodwin, P., O'Connor, M. and Őnkal, D., Judgmental forecasting: A review of progress over the last 25 years. *Int. J. Forecasting*, 2006, **22**, 493–518.

MacDonald, R., Expectations formation and risk in three financial markets: Surveying what the surveys say. *J. Econ. Surv.*, 2000, **14**, 69–100.

McNelis, P.D., *Neural Networks in Finance: Gaining Predictive Edge in the Market*, 2005 (Elsevier Academic Press: Burlington, MA).

Menkhoff, L., Schmidt, U. and Brozynski, T., The impact of experience on risk taking, overconfidence, and herding of fund-managers: Complementary survey evidence. *Eur. Econ. Rev.*, 2006, **50**, 1753–1766.

Mussweiler, T. and Schneller, K., What goes up must come down—how charts influence decisions to buy and sell stocks. *J. Behav. Finan.*, 2003, **4**, 121–130.

Önkal, D., Yates, J.F., Simga-Muganm, C. and Öztin, Ş., Professional vs. amateur judgment accuracy: The case of foreign exchange rates. *Organiz. Behav. & Hum. Decis. Proc.*, 2003, **91**, 169–185.

Park, C.H. and Irwin, S.H., What do we know about the profitability of technical analysis? *J. Econ. Surv.*, 2007, **21**, 786–826.

Qi, M., Nonlinear predictability of stock returns using financial and economic variables. *J. Bus. & Econ. Statist.*, 1999, **17**, 419–429.

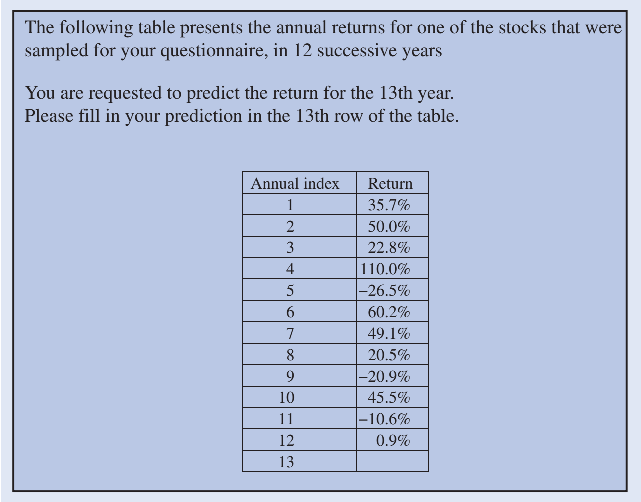
Siegel, S. and Castellan, N.J., *Nonparametric Statistics*, 1988 (McGraw-Hill: New York).

Thomson, M.E., Önkal-Atay, D., Pollock, A.C. and Macaulay, A., The influence of trend strength on directional probabilistic currency predictions. *Int. J. Forecasting*, 2003, **19**, 241–256.

Törngren, G. and Montgomery, H., Worse than chance? Performance and confidence among professionals and laypeople in the stock market. *J. Behav. Finan.*, 2004, **5**, 148–153.

Tversky, A. and Kahneman, D., Representativeness—belief in the law of small numbers. In *Judgment Under Uncertainty: Heuristics and Biases*, edited by D. Kahneman, P. Slovic and A. Tversky, 1982 (Cambridge University Press: Cambridge, UK).

Appendix A



The following table presents the annual returns for one of the stocks that were sampled for your questionnaire, in 12 successive years

You are requested to predict the return for the 13th year.
Please fill in your prediction in the 13th row of the table.

Annual index	Return
1	35.7%
2	50.0%
3	22.8%
4	110.0%
5	–26.5%
6	60.2%
7	49.1%
8	20.5%
9	–20.9%
10	45.5%
11	–10.6%
12	0.9%
13	

Figure A1. Example of the annual return prediction task (experiment 1).