Return Prediction and Stock Selection from Unidentified Historical Data*

Doron Sonsino and Tal Shavit
School of Business
College of Management
December, 2008

Abstract: The experimental approach is applied to explore the value of unidentified historical information in stock-return prediction. Return sequences were randomly drawn cross section and time from historical S&P500 data. Subjects were requested to predict returns or select stocks from 12 preceding realizations. The hypothesis that predictions are randomly assigned to historical sequences is rejected in permutation tests and prediction-errors significantly decrease with expertise. The best-stock portfolios by experimental predictions significantly outperform worst-stock portfolios in joint examination of mean-return and volatility. Actual predictions are more effective than various statistical rules in separating the "best" stock from the "worst" in random 6-stock menus.

Keywords: Return-prediction, stock-selection, momentum and contrarian investment

JEL classification: D8, G1, C9

* 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 College of Management Business School for financial support. Corresponding author: Doron Sonsino. School of Business. The College of Management. 7 Rabin Blvd. P.O.B 9017. Rishon LeZion. Israel. 75190. Emails: sonsinod@colman.ac.il. ShavitT@colman.ac.il. Fax: 972-3-9634210.

1. Introduction

Empirical research on the dynamics of stock returns reveals surprising evidence on the existence of auto-correlations and cyclic trends. The early 1990 studies of Jegadeesh and Lehmann show that daily, weekly and monthly particular-stock returns exhibit negative first-order correlation. Jegadeesh and Titman renowned 1993 paper suggests that correlations become positive when longer periods and larger portfolios are examined. Momentum-based strategies that buy stocks that have performed well while shorting stocks that have performed poorly generate significant arbitrage payoffs in 3-12 months intervals. When the length of inspection is increased to several years, the reversal effect reappears. The famous studies of DeBondt and Thaler (1985, 1987) document a loser-winner effect where stocks with the worst (best) performance in the last 3-5 years become strong winners (losers) in subsequent 3-5 years horizon. The short-run momentum and long-run reversal in returns have been confirmed in dozens of subsequent studies (see, for example, Korajczyk and Sadka, 2004 and Chou e al., 2007).[1] Recent research moreover suggests that momentum profits may be enhanced by controlling for contrarian cycles and vice versa (Chan and Kot, 2006; Balvers and Wu, 2006).

The empirical studies on profitability of momentum and contrarian investment typically deal with very large or extreme portfolios.[2] Trend-riding and contrarian investment however are very popular among private investors and professionals (Menkhoff and Schmidt, 2005). The psychological judgment literature suggests that decision-makers detect patterns and search for regularities even in sequences of random independent draws (Kahneman et al., 1982). The popularity of momentum-investment and contrarian-

[1] Explanations to these empirical phenomena (behavioral, statistical or other) are topic of intense research and controversy which is beyond the scope of our experimental examination.

[2] Jegadeesh and Titman's arbitrage portfolios use the top and bottom deciles of the NYSE and AMEX stocks; De Bondt and Thaler's winners/losers portfolios include only the best/worst 35-50 stocks in the list.

trading has been attributed to investors' inclination to recognize patterns in return sequences. The belief that stocks that start increasing/decreasing will continue to rise/decline is frequently related to the hot-hand bias (Gilovich et al., 1985). On the opposite, believers in the gambler's fallacy (Tversky and Kahneman, 1982), expect that persistent trends must reverse in the future. Short-run trending and long-run reversals have been captured by many theoretical models with diverse behavioral assumptions (Barberis and Thaler, 2003). A regime-shift from "trending" to "contrarian" modes, in particular, is an essential of Barberis, Shleifer and Vishny's (1998) renowned model of under and over-reaction. Bloomfield and Hales (2002) find support to Barberis et al. regime-shifting hypothesis in experimental studies where the actual process of return is random. Subjects' reactions to fixed recent changes in prices get stronger when the changes are preceded by fewer reversals. Durham et al. (2005) contrarily show that bettors' expectations for price-reversals in the football wagering market are strengthen as streak-length grows. While the specific trigger for switching from trend-riding to contrarian modes and vice versa may warrant additional research, the inclination to detect trends and respond to patterns, in general, seems an essential factor in field financial prediction.[3]

With this motivation, we employ the experimental approach to examine the value of unidentified historical information on realized returns in investment-decision. Subjects with adequate background in finance are asked to examine "recent" return data for unidentified stocks. The field-data on realized returns may exhibit auto-correlations and periodical trends as documented in empirical studies. The investor does not receive any information on the identity of the stock, market conditions or economic fundamentals, but may thoroughly examine the historical data to detect trends or patterns. We wish to test the value of such technical non-formal forecasting. Is the historical information on unidentified stocks useful in return prediction or stock selection?

[3] Related issues are addressed – in broader context - in the judgmental forecasting literature (see the survey in Lawrence et al., 2006) and in the experimental literature on individual investment-decision (see, for example, Weber et al., 2005).

The experimental exercise is run on field empirical data. The return sequences for each prediction/selection task are randomly drawn for each subject from the S&P500 stock-return data for the last 40 years. Subjects are asked to predict returns or select stocks from 12 preceding realizations without receiving any information beyond the unidentified sequence of historical returns. The random design is intended to clear away sequence-specific effects and examine the significance of historical information in a most general setting. The drawing of sequences cross section and time, in addition, precludes inference about the identity of specific stocks or particular inspection periods and thereby confines information to the historical series.

We run 3 experiments on 163 MBA students with advanced background in finance. In Experiment I, subjects are asked to point-predict the 13-th period return from the historical sequences. Each subject is requested to fill-in 6 annual predictions (the ANN condition) and 6 monthly predictions (the MON condition). Experimental payoffs are determined by the accuracy of prediction in a randomly selected assignment. Non-surprisingly we find that the "anonymous" historical information is insufficient for close prediction of subsequent returns. The distance between predictions and actual returns is very large in absolute and relative terms and subjects are unable to predict the sign ("positive" or "negative") of hidden-returns from the historical data. Prediction-errors however strongly decrease with expertise and the hypothesis that subjects randomly assign predictions to tasks is rejected in permutation tests. The payoffs on "best" stocks, by subjects' predictions, are higher than the payoffs on "worst" stocks, and the differences cannot be explained by the risk-levels of the corresponding portfolios. Closer examinations suggest that subjects attempt to balance trend-based and contrarian forecasting. Predictions tend in general to increase with the average level of returns in the historical sequences, but decrease significantly after "heavy" streaks of positive returns. Actual predictions, moreover, significantly outperform various statistical prediction-rules in separating the best stock from the worst, in random 6-stock menus.

To further examine the results, we run two more experiments. Experiment II, where subjects select 2 stocks from random menus of 6 to fill-in their predictions and Experiment III where subjects select 2 stocks from random menus of 6 for investment. The results of the additional experiments confirm those of Experiment I. The mean payoff on selected stocks in Experiment III exceeds the mean payoff on "holding" all stocks, in both treatments. The difference is highly significant in MON where the payoff on the stocks selected by "high-skill" subjects is 3 times larger (on average) than the payoff on non-selected stocks, although the standard deviation of selected portfolios is not larger than random.

The random selection of return sequences from datasets that span over 40 years generates a very noisy experimental database. The standard deviations of returns and predictions are large and the empirical distributions exhibit high skeweness and kurtosis. All through the paper we avoid parametric tests and often use conservative sign-tests to check the significance of results. To increase statistical power, specialized randomization and permutation tests (Good, 2005) are applied whenever applicable. The bottom-line results of the experiments suggest that unidentified historical return data could be used for profitable stock selection and even generate significant experimental-arbitrage payoffs. The results are related to the literature on profitability of technical trading (Brock et al., 1992) and may complement recent AI studies on the predictability of returns from historical data (see concluding discussion).

Equilibrium asset-pricing implies that expected returns on financial assets increase with relative risks. In equilibrium, our subjects could use the empirical mean (say, the average realized return in the historical series) to predict the missing 13-th period return. Equilibrium-pricing however implies that the "best" stocks under such selection would also constitute the most risky investment. Our results interestingly reveal that while the mean payoff on the "best" stocks by subjects' predictions is significantly higher than the mean payoff on "worst" stocks, we cannot (in most comparisons) reject the hypothesis that the (ex-post) variability of "best" and "worst" payoffs is equal. In this sense, the

subjects are able to beat the (experimental) index without bearing additional risks. These results appear strongly for MON but are violated in ANN where additional returns are gained at the cost of higher ex-post variability.[4]

Using Harrison and List (2004) taxonomy of field-experimentation, the experiments run in this study can be classified as "framed-field experiments"; i.e., experiments with field context in subjects, commodities and information. The term "field experiment" in general refers to the case where experimentation takes place in a more natural setting than the "sterile" economic lab. While field experiments are gaining increased popularity in various fields of economic research, we are not aware of concrete applications regarding financial-investment decision. Field experiments could be used for collecting micro-level data on the strategies employed by investors, the type of information upon which investors react and other aspects of financial decision that are unobservable even in account-level empirical data.

2. Experiment I

2.1: Method and Subjects

Each questionnaire consisted of 12 prediction tasks: 6 annual-return prediction assignments (the ANN condition) and 6 monthly-return prediction tasks (the MON condition). In each task, subjects were requested to fill-in a point prediction for the 13-th period return from the series of 12 preceding (annual or monthly) realizations. An example to a typical prediction problem (for ANN) is provided in Figure 1.

< Insert Figure 1>

The historical return series were randomly drawn for each subject from historical records of the stocks that composed the S&P500 list at the end of April 2006. The instructions

[4] But randomizations (for ANN) suggest that larger differences in expected returns together with smaller differences in variability are observed in less than 5% of the cases under random selection of best and worst stocks (see Section 2.6)

explained that series of 13 successive returns were drawn at random from the S&P500 data for the last 40 years. The 13-th observation was concealed but kept in our records in order to validate predictions. No information was revealed on the identity of the underlying stock, 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 selection of stocks cross section and time for each subject was intended to decrease the chances that subjects would claim to identify the underlying stocks or inspection-periods. An alternative design where subjects receive cross-section historical data on 6 anonymous stocks (for the same 12-periods), for instance, could motivate the subjects into guessing the period of inspection and even identifying specific stocks. In such design, subjects might hinge their predictions on "private signals" rather than acting to the historical data provided in the questionnaires.[5] The random selection of distinct historical sequences for different subjects, on the other hand, was intended to cancel out sequence-specific effects and test the value of anonymous historical data, in a most general setting.

Experimental questionnaires were distributed to 65 (25 female; 40 male) advanced MBA students that have passed the basic core courses in finance. The average age of the participants was 30.2; 29% of the subjects (19 of 65) claimed having experienced investment-management in practice. Subjects were requested to cooperate and provide their "best" prediction in each task. In addition, we have used a performance-based payoff to motivate accurate prediction. The instructions explained that 1 of the 12 tasks ("the selected problem/prediction") will be randomly drawn for each subject to determine actual payouts. If the selected prediction was of the wrong sign (the subject has predicted negative return when actual return was positive or vice versa), the subject received a

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

minimal payoff of 20 N.I.S (about 5 U.S. dollars). If sign-prediction for the selected problem was correct, the minimal payoff increased by 50% to a level of 30 N.I.S.. Specifically, payoffs were determined by subtracting a "penalty for deviation" from a maximal balance of 80 N.I.S. The penalty in MON was 6 N.I.S for 1% (positive or negative) deviation, while the penalty in ANN was only 2 N.I.S for 1% absolute deviation. The decreased penalty for deviation in ANN was motivated by the larger volatility of annual returns.[6] The instructions clarified that if the penalty for deviation decreases the payoff below 30, the guaranteed minimum will be paid. The fairness of the experimental protocol was underlined in the instructions and subjects were invited to examine the data and payoff calculations at the end of the experiment. Subjects that did not wish to participate were exempted. The mean actual payout was about 40 N.I.S (about 10 U.S. dollars) with standard deviation 23.

The experiment was run in-class by distributing printed questionnaires to the students. The instructions were introduced on the blackboard and questions were privately answered. No time limits were imposed; the approximate mean participation time was 20-25 minutes. The instructions did not mention the possibility of using calculators, since this could have instructed subjects to calculate statistics. For similar reasons we did not use bar-charts or other graphic illustrations that might direct subjects into trend extrapolation.[7] Only few subjects used a calculator or calculated average-returns on their questionnaires (see the discussion of "statistical forecasting" in section 2.8). Each prediction task appeared on a separate page. The condition-type (ANN or MON) was bolded on the top of each page. The ANN and MON assignments appeared in distinct parts of the questionnaire; the order of treatments was randomly assigned. At the last

[6] The random selection of one task to determine actual payouts is a common practice in experimental economics (for a recent examination see Hey and Lee, 2005). The methods is intended to avoid wealth effects, diversification concerns and other "portfolio" type of considerations that may effect subjects' behavior if actual payouts are determined by several tasks concurrently. Since payouts decrease with prediction-errors subjects are incentivized to submit their most accurate point prediction in each problem.

[7] For a recent example to framing effects on return expectations see Galser et al. (2007).

page, subjects were asked to rank their knowledge in finance "theory" and their "familiarity" with the financial arena in 1-10 scales. In addition, subjects were encouraged to provide details on the "method" they used to predict missing returns and comment on the experiment. The variable "skill" henceforth denotes the sum of ranks "theory" + "familiarity".

2.2: Notation and Statistical Method

The following notation is used to simplify the discussion of results: PRED denotes experimental predictions while OBSi presents the i-th observation in the historical return-sequences. OBS12 thus denotes the last-observed return in each problem while OBS13 describes the hidden-return which subjects were requested to forecast. The product OBS13*100 is sometimes addressed as the "realized" payoff or "holding" payoff (on $100 investment). Subjects exhibit "under-prediction" when PRED<OBS13 while "over-prediction" refers to the case where PRED>OBS13. The absolute value metric is used to measure the accuracy of predictions: ERR=|PRED-OBS13| accordingly denotes the prediction-error. The symbols AVG12 and STD12 are reserved for the mean and standard deviation of returns in the 12-observations series presented to the subjects. AVG6/STD6 accordingly denote the mean/STD of returns in the series (OBS7, OBS8…. OBS12) and AVG3/STD3 are similarly defined. TRENDn finally denotes the slope of returns in the last n observations.[8]

We continue with a brief discussion of the statistical method of the paper. The statistical testing of experimental results is restricted by two obstacles. First note that we employed a multi-task design where each subject fills-in 6 predictions in each condition (ANN and MON). For independence of observations, statistical tests should be run on subject-level statistics. Prediction-errors (ERR), for example, must be averaged across the 6 tasks in each treatment for testing hypotheses regarding predictions.[9] The averaging of

[8] Trend was formally measured by linear regression of returns on observation-indices (1 to 12).

[9] We henceforth use "average" when talking about subject-level statistics and keep "mean" for the between-subject average.

observations across tasks however results in loss of information that obstructs statistical testing. Secondly, recall that the prediction/selection tasks were randomly drawn for each subject from empirical stock data. Each subject accordingly responded to a distinct collection of assignments. The distribution of individual-level statistics (average prediction-errors or average payoffs) is noisy, non-normal and skewed. The shape of distributions changes considerably when sub-samples of the experimental pool are examined. Because of the between-subject variability, noise, and non-normality of the data, we use non-parametric tests all through the paper. Sign-tests and permutation-tests (P-tests) for paired replicates (Siegel and Castellan, 1988) are applied for paired comparisons. To increase statistical power, specialized randomization or permutation tests are employed when applicable.[10] These tests exploit the multi-task design to check significance of payoffs and differences without imposing any assumptions on underlying distributions. We now illustrate the method of such "specialized" testing by describing in detail a randomization test for comparing the payoff on the "best" stock according to the experimental predictions to the payoff on the "worst" stock by these predictions.

Let "best1" denote the realized payoff (OBS13*100) on the stock that has received the highest prediction among the 6 predictions provided by the subject. Let "worst1" denote the payoff on the worst stock according to 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 directly determine if the difference is statistically significant. The null hypothesis is that best1=worst1, while the directed alternative stipulates that best1>worst1. To run the test we first subtract the mean worst1 payoff from the mean best1 payoff. Let D denote the difference in mean payoffs. The randomization test is used to examine if such large differences (as D) arise in random assignment of stocks into the best1/worst1 categories. In each randomization, 2 stocks are

---

[10] Good (2005) provides a detailed general discussion of permutation and randomization tests. SAS codes for some randomization/permutation tests are provided in http://www2.colman.ac.il/business/doron. Other programs will be provided by request.

randomly selected for each subject and arbitrarily designated as "randomized-best" and "randomized-worst". The process is repeated to create 1000 independent randomizations. The mean payoff on randomized-worst stocks is subtracted from the mean payoff on randomized-best stocks, in each randomization. Let D' denote the difference for an arbitrary randomization.[11] If D'≥D then the randomization suggests that larger differences in mean payoffs (than D) 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 relatively to the difference that arise from random selection. To conclude the test, we calculate the proportion of cases α (out of 1000 randomizations) where D'≥D. If α<0.1, the hypothesis that the best1=worst1 is rejected for the alternative best1>worst1; α is the (one-tail) significance level. Similar randomization tests are applied to determine if the average-payoff on the 2 (or 3) stocks that have received the highest predictions is higher than the average-payoff on the 2 (3) stocks that have received the lowest predictions. The only difference is that the randomized best/worst selections would now include 2 (or 3) stocks, accordingly.

The randomization tests exploit the nominal values of variables without transforming data into categories or ranks. A sign-test, on the contrary, would compare the proportion of subjects where best1>worst1 to the proportion where best1<worst1, but completely ignore the magnitude of differences in best1/worst1 payoffs. Sign-tests could not reject our null hypotheses in some cases where randomization tests (henceforth: R-tests) revealed statistical significance. When discussing such results we disclose the proportion of subjects with best1>worst1 payoffs but use the more powerful R-test to determine statistical significance.[12] The sample-size for each randomization-test was determined by

[11] Clearly, D' is close to 0 on average.

[12] Other tests for the equality of paired observations include the sign-rank test and the P-test for paired replicates (Siegel and Castellan, 1988). The sign-rank test assumes that distributions are symmetric, an assumption that is strongly violated in our data (Diebold and Lopez, 1996). The P-test ignores all data except for individual differences in best1 and worst1 payoffs. The R-test on the contrary directly examines the significance of differences relatively to the data from which the best1/worst1 stocks were selected.

trial and error to produce consistent results in independent repeated runs. In the following, we report the results for a sample of 1000 randomizations unless otherwise specified. We use a small α to denote one-tail significance and upper-case asterisks to characterize the corresponding levels of significance: 2 asterisks (**) denote significance at $\alpha<0.1$; single asterisk (*) is used for significance at $\alpha<0.05$ and a bolded asterisk (*) for $\alpha<0.025$.

<Insert Table I>

2.3: Preliminary

The mean values of PRED, OBS13, ERR and other statistics are disclosed in Table I. 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) although the frequency of underprediction is close to the frequency of overprediction in both treatments (see proportions on the table). Closer examination reveals a tendency to underpredict when the historical series ends-up with a negative trend. In the ANN tasks, for example, PRED is 9.7% lower than OBS13 (on average) in the series ending with OBS11>OBS12; while PRED is 7.7% higher than OBS13 when OBS11<OBS12. Smaller 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. Our subjects thus display "extreme" prediction patterns (De Bondt, 1991) submitting overly pessimistic predictions when latest trends are negative and too-optimistic forecasts when recent trends are positive.[13]

The distance between predictions and actual returns appears very large in absolute terms and relatively to OBS13. The median value of ERR in the monthly tasks is 6.4%. The

[13] The extreme prediction weakens/disappears when longer trends (TREND12, TREND6) 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 differences completely disappear in MON where PRED is about 1% lower than OBS13 in both sub-samples.

median ERR in ANN is more than 3 times larger: 21.4%.[14] Since the error data is skewed to the right, the mean ERR levels are even higher (see table). The error-size however appears more plausible when compared to measures of variability for the historical return-series (see the STD12 and STD6 columns on the table). Mean ERR is about 6% lower than mean STD12 and mean STD6 in ANN (sign-tests; $\alpha<0.06$ for each comparison). The differences diminish in MON where the mean prediction error is close to mean STD12 (STD6). The prediction errors of the subjects, in conclusion, are either significantly smaller (in ANN) or similar (in MON) to the STD of returns in the historical series that the subjects observed.

To further examine the relation between PRED and OBS13 while controlling for individual heterogeneity, we OLS-estimate the linear model $OBS13 = a + b * PRED$ on the N=6 observations collected from each subject in every treatment. 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 subjects with skill≥8, the mean values of *b* 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).[15] The regressions thus 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 predictions.

[14] The error-levels are similar to the ones observed by Törngren and Montgomery (2004) who find 10%-11% errors in the prediction of monthly returns on leading Swedish stocks by students and professionals.

[15] Similar levels of fit are obtained in (random) analysis where 1 of the 6 prediction tasks is drawn for each subject and regressions are run "between subjects" on N=65 observations. The mean R^2 levels obtained in 5000 random estimations are 0.3 for ANN and 0.09 for MON. The mean value of the coefficient b decreases to 0.06 in ANN and 0.05 in MON, possibly because heterogeneity across subjects weakens the relation between predictions and actual returns. The mean coefficients still increase (to 0.13 in ANN and 0.2 in MON) when the estimation is restricted to the 34 high-skill subjects.

Finally, we run a direct permutation test to examine if the historical information was useable for predicting missing returns. The null hypothesis of the test is that predictions are randomly assigned to prediction assignments. If this is the case, then the shuffling of forecasts across tasks should not increase (nor decrease) prediction-errors. If, on the other hand, the historical data provides clues for predicting future returns, then such permutations would decrease the accuracy of prediction and increase prediction errors. We use the absolute value metric 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 6 predictions of each subject are randomly shuffled across the 6 prediction tasks. The exercise is independently repeated 5,000 times to construct a sample of 5,000 permutations. We then calculate the proportion of permutations where the mean distance between shuffled predictions and actual returns is lower than 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. [16]

The shuffling of annual predictions has increased the distance between predictions and actual returns in about 96% of 5,000 permutations. The hypothesis of random-prediction is thus rejected at α=0.04 for ANN. The results for the monthly prediction tasks are slightly weaker but still evidently significant. The distance between shuffled predictions and actual returns was lower (on average) than actual ERR in only 6.22% of 5,000 permutations. The hypothesis of random prediction is accordingly rejected at α=0.06 for MON. The randomality of predictions with respect to historical data, in conclusion, is rejected for ANN and for MON.

[16] Alternatively, the tests could be run "between subject" by shuffling the 390 predictions across the 390 prediction tasks in each treatment. This procedure generates excess noise because of the heterogeneity across subjects. The "randomality" hypothesis is therefore rejected at α=0.04 for MON and α=0.02 for ANN.

2.4: Skill Effect on Prediction Errors

Recall that subjects were requested to rank their theoretical knowledge in finance and their familiarity with the field in 1-10 scales (see supplementary appendix B). The mean level of "theory" was about 4 compared to mean familiarity-rank of 4.1. The correlation between the two skill variables was 0.773. In this section, we briefly examine the effect of skill on prediction-errors. The results of such comparison must be qualified in advance. The subjective skill-ranks reported by subjects could approximate level of motivation, seriousness and other variables that may affect performance even if level-of-expertise per se does not change the quality of prediction. We are still interested in checking if the serious/motivated or knowledgeable subjects were able to exploit the historical information more effectively.

To examine the effect of skill on prediction-errors we run fixed-effect linear regressions where ERR is explained by variables that characterize the observed return-series (STD12, AVG12, TREND12 and others) and by the 2 skill variables: "theory" and "familiarity". The fixed effect specification allows for heterogeneity in individual intercepts while assuming common coefficients for all explanatory variables (Green, 2003). Since the number of variables in the fixed-effects model is very large, model-selection procedures are employed to eliminate insignificant coefficients. The regressions were run in many different versions with distinct collections of explanatory variables and various selection procedures in order to check the robustness of results. In all cases, we find a significant negative "familiarity" effect on prediction-errors in MON. Table II, for example, presents the results of the simplest estimation where ERR is explained by (an individual intercept,) STD12, and the 2 skill variables. As intuitively expected, the coefficients for STD12 are positive and significant, suggesting that errors increase with the volatility of historical series. More interestingly, the regressions suggest that "familiarity" significantly decreased prediction errors in MON. The coefficient of the familiarity index is negative -0.62, implying that prediction errors of subjects with maximal familiarity rank are about 5.5% lower than the errors of subjects with minimal familiarity. The difference is large relatively to the mean level of ERR in MON: 8.5%. When forcing the backward selection

procedure to estimate the effect of "familiarity" in the ANN tasks, the estimated coefficient is negative -0.47 but it is not statistically significant. Note that "theory" was removed in the model selection process for MON and ANN. Similar results are obtained in many alternative model specifications.

< Insert Table 2>

2.5: Predicting the Sign of OBS13

The rejection of randomality in 2.3 motivates further examination of the quality and economic significance of experimental prediction. This section examines if the historical information was useful for predicting the "trend" (positive or negative) of missing return. Specifically, we test if the frequency of positive OBS13 was larger in the cases where subjects submitted positive predictions. Slight differences are observed for ANN where the conditional frequency Pr(OBS13>0|PRED>0)=78.2% while Pr(OBS13>0|PRED<0)= 71.6%.[17] The difference is larger in the subsample of subjects with familiarity≥4 (N=34) where the conditional 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) suggest that the difference is marginally significant for the high-skill sample (χ^2=2.6; $\alpha\approx$0.1) but could not reject the hypothesis that sign(OBS13) is independent of sign(PRED) for the complete sample (χ^2= 1.73; N.S.).[18] The results of the comparison are even 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% in the complete sample and we cannot reject the hypothesis of

[17] "Pr" abbreviates empirical frequencies. The exact frequencies are provided in supplementary appendix C.

[18] The Chi-square test assumes that prediction-signs are independent across tasks. Comparison of proportions on an individual basis is misleading since negative predictions are relatively infrequent (especially in ANN). If some subject, for example, provides 5 positive predictions and only 1 negative prediction, and OBS13 is positive in all cases except one of the tasks where the subject has provided a positive prediction, then Pr(OBS13>0|PRED>0)=0.8 while Pr(OBS13>0|PRED<0)=1.

independence even for the high-skill subjects. Our first conclusion regarding the usefulness of historical-series for prediction is therefore strongly negative. Actual returns are positive in more than 60% of the cases where subjects submit negative predictions. With the exception of high-skill subjects' marginal ability to distinct trends in ANN, we cannot reject the independence of sign(PRED) and sign(OBS13).

2.6: Separating "Best" Stocks from "Worst"

This section examines subjects' ability to separate "good" and "bad" stocks on the basis of historical data. For this we compare the payoffs (OBS13*100) on the stocks that received subjects' highest predictions (the "best" stocks) to the payoffs on the stocks that obtained the lowest predictions (the "worst" stocks). The comparison is run in 3 levels. First, we compare the realized 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 2 stocks that obtained the highest predictions (Best2) to the average-payoff on the 2 stocks that have received the lowest predictions (Worst2). Finally, we split the "prediction-menu" of each subject into two groups of 3 stocks and compare the average-payoff on the 3 stocks with the highest predictions (Best3) to the average-payoff on the 3 other stocks (Worst3). The mean best/worst payoffs in the 3 levels of analysis are contrasted in Table III. The standard deviation of each payoff is disclosed in smaller brackets and asterisks are used to denote R-test significance (the shading is explained below).[19]

<Insert Table III>

[19] The bracketed "(22/19)" figures in the best2/worst2 column for example denote the standard deviations (rounded to integer numbers) of best2/worst2 payoffs across the N=65 subjects in the complete sample.

Consider the mean payoffs first. The mean payoff on "best stocks" is higher than the mean payoff on corresponding "worst stocks" in all 12 cells in the table.[20] The largest differences appear in comparison of Best1 and Worst1. The mean payoff on the best stock by subjects' predictions in ANN is almost twice larger than the mean payoff on the worst stock. Best1 is larger than Worst1 for 39 of 65 subjects in the complete sample and for 21 of 34 subjects in the sub-sample of high-skill subjects. The hypothesis that Best1=Worst1 in ANN is rejected in a direct randomization test (see Section 2.2) for the complete sample (α=0.02) and for the sub-sample of high-skill subjects (α=0.06).

Large differences between Best1 and Worst1 payoffs also appear 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 most of the subjects (34 of 65) and the equality of Best1 and Worst1 is not rejected, for the complete sample, in neither a sign-test nor a randomization-test (R-test; α=0.23). The mean Best1 payoff increases to 5 while the mean Worst1 payoff decreases to 1.5 when the comparison is restricted to the high-skill subjects. The number of subjects with Best1>Worst1 (19 of 34) is not large enough for sign-test significance but the randomizations suggest significance at α=0.047.

"Best" payoffs are still larger than "worst" payoffs, on average, when the portfolios are extended to include the 2 or 3 best/worst stocks in each questionnaire. The differences in means are not significant when the complete sample is examined but significance is obtained in 5 of 6 comparisons for the high-skill subjects (see R-test asterisks in Table III).

Significance moreover improves in joint examination of mean returns and variability. The bracketed STD data in Table III reveals that the STD of "best(n)" is higher than the STD of corresponding "worst(n)" in 5 of the 6 comparisons for ANN. The STD of "best" and

[20] Similar conclusions are derived in comparison of the average-payoff on the best n stocks (n=1,2,3,4,5) to the average-payoff on the other 6-n stocks.

"worst" payoffs however are similar in MON in spite of the larger mean return on the best stocks by subjects' predictions. To jointly compare the mean and STD of best/worst payoffs, we use the randomizations to calculate the proportion α' (out of 1000 randomizations) where randomized portfolio-selections lead to larger differences in mean-payoffs together with smaller differences in STD. The results of the test are summarized in Table III by shading the cells where the proportion $\alpha'<0.1$.[21] Light shading is used for $0.05\leq\alpha'<1$ and darker shading for cases where $\alpha'<0.05$. The equality of best/worst payoffs is now rejected in 11 of 12 comparisons. The randomizations for MON (complete sample), in particular, reveal that larger differences in mean best1/worst1 payoffs occur jointly with smaller differences in STD, in only 5% of 1000 randomizations. Recall that the randomization test for mean-payoffs alone revealed that larger differences in mean payoffs arise in about 23% of 1000 randomizations. The joint examination however proves that in random selections, large differences in means are accompanied – in most of the cases – by larger differences in variability. The best stocks by subjects' predictions, on the other hand, pays higher mean return than the worst stock without such increase in volatility. Similar conclusions are drawn in the other joint examinations.[22]

2.7: Trend-Based vs. Contrarian Prediction

At the last part of the questionnaire (see supplementary appendix B) subjects were requested to comment on the methods they employed to predict the missing return. The most common methods mentioned in the survey can be classified as variants of "momentum-based" or "contrarian" predictions. Following standard terminology we use the term "trend-chasing" or "momentum-based" prediction for cases where subjects provide positive/negative predictions for historical series that end with a sequence of

[21] Note that $\alpha'\leq\alpha$ by definition, so that the joint examination can only improve significance.

[22] Closer look into the data confirms that subjects decreased predictions (relatively to AVG12) when the volatility of historical returns increased. The mean Pearson coefficient of correlation between the difference (PRED-AVG12) and STD12 (where the correlation is calculated for each subject separately and correlations are averaged for the 65 subjects) are -0.61 for ANN and -0.20 for MON.

positive/negative returns, correspondingly. The term "contrarian prediction", on the opposite, is kept for cases where predictions are negative/positive when histories end with a sequence of positive/negative returns. In fact, many subjects mentioned both types of strategies saying they "switched" from trend-chasing to contrarian prediction after large streaks of positive returns which "increase the chance of price adjustment".

The interplay of momentum-based and contrarian forecasting is reflected in the data on Table IV. The Table summarizes subjects' response to streaks of positive returns. Say that the return sequence OBS1, OBS2… OBS12 ends with an up-streak if OBS12>0. The size of the up-streak is the sum of successive positive returns at the end of the sequence. In the table we median-split the series that ended up with an up-steak (N=278 for ANN and N=215 for MON) and calculate the proportion of contrarian (negative) predictions for each sub-group. The left column reveals the proportion of negative predictions in histories with streak-size smaller than median, while the right column presents the proportion for histories with streak-size larger than median. The proportion of negative predictions in the large streak-size group is higher than the proportion for the small streak-size group in both treatments, reflecting subjects increased expectations for price-adjustment following large streaks. The difference is more pronounced in MON where the proportion of contrarian predictions in the large streak-size group is 41% compared to 26% in the small size group (sign-test on individual-level proportions; $\alpha<0.04$). The differences are smaller in ANN where we cannot reject the hypothesis that contrarian predictions are as likely in the two streak-size groups ($\alpha=0.28$).

<Insert Table IV>

Similar results are obtained in comparing the experimental predictions to the historical mean returns. While predictions are significantly lower than AVG12 in the large streak-size group (mean PRED-AVG12=-6.6 in ANN, -0.8 in MON), the predictions are either closer or even higher than AVG12 in the small streak group (PRED-AVG12=-2.3 in ANN, +2 in MON). The non-monotonic relation between PRED and "empirical"

forecasts, like AVG12 or AVG6, also reflects in relatively low levels of correlation.[23] The next section compares the experimental predictions to various empirical rules in detail.

2.8 Empirical Prediction Rules

In this section we argue that actual predictions outperform "statistical forecasting" in separating the best stock from the worst in random 6-stock menus. Consider for example the mean level of return in the historical sequence presented to the subject: AVG12. The best/worst payoffs may be redefined using AVG12 (instead of PRED) to rank the 6 stocks in each questionnaire. "Best1", for AVG12, would be the realized payoff (OBS13*100) on the stock with (12-periods) highest historical mean. "Worst1" would similarly denote the payoff on the stock with lowest historical mean. Best2, Worst2, Best3 and Worst3 could be similarly redefined, for AVG12. The same exercise can be applied to other statistical rules: AVG6, AVG3, OBS12 and more.

In Table V we compare the payoffs implied from the experimental predictions to the payoffs implied from the empirical rules: AVG12, AVG6 and AVG3. For the analysis we examined a longer list of statistical rules (including OBS12, AVG2, TREND12 and more), but confined the discussion to the statistics that produced the highest best payoffs.[24] Randomization tests are employed to determine significance; asterisks denote

[23] Pearson correlations are calculated for each subject (N=6) and averaged across all subjects (N=65). The mean correlation between AVG12 and PRED is 0.276 in ANN and 0.198 in MON. The corresponding figures for AVG6 are 0.182 and 0.224. Closer examination reveals that none of the subjects consistently followed AVG12, AVG6 or AVG3 (exactly, or with a reasonable "band") in all 6 tasks.

[24] Note however that the statistical rules are closer than PRED to OBS13 on average. The mean error obtained when AVG12 is used to forecast OBS13 is 24.5% in ANN and 6.8% in MON (compared to actual mean ERR of 28.2% and 8.5%, correspondingly). The mean errors of alternative statistical rules (AVG6 and AVG3, in particular) are larger than the errors for AVG12 but still lower than ERR. Task-level comparisons however reveal that PRED is closer to OBS13 (than AVG12) in 45% (38%) of the tasks in ANN (MON) respectively.

cases where the payoff implied from statistical prediction is significantly different from the payoff implied from experimental predictions.[25] The noise in implied-payoff data is large and only 10 comparisons (out of 36) reveal statistical significance. Consistent significant differences between "PRED" and each of the 3 statistical rules appear in 2 levels of comparison, underlined by bolding the corresponding panels in Table V (the other significant differences are discussed below). In both cases, actual predictions significantly outperform each of the 3 statistical rules:

(I) The mean payoff on the stocks that received subjects' highest predictions (Best1) in ANN is about 80% larger from the mean payoff on the stocks that were ranked highest by each of the 3 statistical rules. The mean best1 payoff by PRED is 26.5 while the corresponding mean best1 payoff for AVG12 is only 17.7. Similar large differences appear for AVG6 and AVG3 (see table). Randomization tests suggest that the differences in payoffs are statistically significant at $\alpha<0.05$ although we cannot reject the hypothesis that standard deviations are equal (see the STD data in smaller brackets in the Table).

(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 the statistical rules. The mean payoff on the worst stock by AVG12-ranking for example is 4.5, about 250% larger than the mean Worst1 payoff by PRED (1.8). Similar large differences appear for AVG6 and AVG3 (see table). In all 3 comparisons, randomization tests suggest that subjects selected the worst stock more effectively than the empirical rules in MON (again, the STDs are not significantly different).

[25] The significance of the difference between Best1 payoff by PRED (mean: 26.5) and Best1 payoff by AVG12 (mean: 17.7), for example, is tested by drawing 1000 independent pairs of random 1-stock portfolios (independence is required since the highest prediction could be assigned to the stock with highest AVG12). For each randomized pair, let D' denote the difference in mean randomized best1 portfolios. The proportion of randomizations where the difference D' is larger than 26.5-17.7=8.8 constitutes the significance level of the test.

<Insert Table V>

To further examine these results, consider the difference between Best1 and Worst1 payoffs. As an interpretation, the difference (Best1-Worst1) may represent the payoff on hypothetical arbitrage where subjects buy (100 value of) the stock that obtained the highest prediction while short-selling an equal amount of the stock that received the lowest prediction. For convenience, let Arb1=Best1-Worst1 denote the payoff on this arbitrage. The mean Arb1 payoff, for each prediction rule, can be directly calculated from the data in Table V. When actual predictions are used to determine the best/worst stocks, the mean Arb1 payoff in ANN is 12. Randomization tests suggest that mean Arb1 payoff larger than 12 appears in only 1% of 1000 randomized selections of similar arbitrage strategies. The Arb1 payoff implied from actual prediction is therefore R-test significant at α=0.01. The mean Arb1 payoff however decreases to 4.8 when AVG12 is used to rank the 6 stocks in each questionnaire. The hypothesis that Arb1 (by AVG12) = 0 could not be rejected in the R-test (α=0.20). The Arb1 results for AVG6 and AVG3 are even weaker. The mean Arb1 payoff when AVG6 is used for ranking is 3.6 (R-test; α=0.28); the corresponding mean for AVG3 is negative -0.8 (α=0.55). Actual predictions thus strongly outperform the 3 empirical rules in separating Best1 from Worst1 in ANN.

Similar analysis reveals that actual predictions outperform the statistical rules in separating best1 from worst1 in MON as well, but the results are weaker (compared to ANN) and significance is obtained only for the high-skill sub-sample. Consider the complete sample first. The mean Arb1 payoff by PRED is 1.27 which is very close to the mean Arb1 payoff for AVG12: 1.1. The Arb1 payoff by PRED is positive for only 31 of 65 subjects and the hypothesis Arb1=0 could not be rejected in the R-test (α=0.22). The weaker results for MON are partially due to PRED's weaker ability to identify the best stock in MON (see Table V). The Best1 payoff implied from subjects actual predictions in MON (mean: 3) was significantly lower than the Best1 payoff implied by AVG12 (mean: 5.5) and from the Best1 payoff implied by AVG6 (mean: 5.8) (R-tests; α=0.06 for

AVG12 and α=0.04 for AVG6). The Arb1 payoff implied from PRED is still slightly higher, on average, from the Arb1 payoff by the statistical rules because of PRED's stronger success in assigning the lowest predictions to the worst stocks (point II above). The results for MON however improve when the examination is restricted to the sub-sample of 34 high-skill subjects. The mean Arb1 payoff by PRED increases to 3.4 (R-test; α=0.05) compared to a mean Arb1 payoff of 2.3 for AVG12 (α=0.15; N.S.) and 0.6 for AVG6 (α=0.39; N.S.). Actual predictions in this sense outperform the empirical rules in MON, when the high-skill sample is examined.[26]

3. Experiment II: Selecting Stocks for Prediction

To further examine the random-prediction hypothesis, we run another version of the experiment where subjects are asked to select 2 of 6 stocks in each treatment to fill-in their predictions. The possibility to select stocks for prediction should, on one hand, enable subjects to focus on cases of "stronger predictability" and decrease prediction errors relatively to the 6-tasks design. When subjects however fill-in only 2 predictions in each treatment, there is just one way to permute predictions across tasks. The unique permutation would reassign the first-prediction to the second-stock and the second-prediction to the first-stock. The testing of the random-prediction hypothesis is therefore more focused (strict) and exact in the modified design.

For comparability of results, we used questionnaires from Experiment I to construct the selection problems for Experiment II. The 6 prediction tasks in each treatment were replicated into 12X6 return-tables where each column refers to a different stock and the rows represent the sequence of historical returns for the corresponding stock (see example in supplementary appendix D). Subjects were asked to select exactly 2 of the 6

[26] Arbitrage payoffs can also be defined for larger portfolios. In particular let Arb2= Best2-Worst2 and Arb3=Best3-Worst3. The data on Table V implies that the Arb2 and Arb3 payoffs implied from actual predictions are lower on average than the corresponding payoffs for AVG12. None of the differences is statistically significant except for Arb3 in ANN: Arb3 based on actual predictions (mean: 0.4) is much lower than the corresponding payoff for AVG12 (mean: 7.05) (R-tests; α=0.05).

stocks/columns in each treatment and provide predictions only for the selected series. The instructions emphasized that the 6 return-series for each treatment were randomly selected from S&P500 historical data and that different columns may refer to completely different stocks and inspection periods. A "payment stock" was randomly selected from the 4 prediction tasks filled-in by each subject. The payoff scheme was otherwise identical to the method used in experiment I. The selection experiment was run on 37 advanced MBA students that did not participate in experiment I. The average age was about 33; the mean familiarity and theory ranks were about 6 and 6.2 correspondingly. The mean actual payout was about 40 NIS (10 U.S dollars).

The mean prediction-errors in the selection experiment are disclosed in the "actual" column of Table VI. To test the randomality hypothesis, we shuffle the 2 predictions of each subject, in each condition, across the 2 prediction-tasks. The "shuffled" column on Table VI presents the mean prediction errors for the shuffled series. The left panel presents the results for the complete sample (N=37) while the right panel deals with the sub-sample of subjects with familiarity≥6 (N=18).

<Insert Table VI>

Consider the results for the complete sample first. The mean ERR in ANN is 25.7 compared to a mean prediction-error of 28.4 in the permuted series. Task-level comparisons reveal that the shuffling has (strongly) increased prediction-errors in 43 cases while decreasing the errors in only 29 cases. The average prediction-error of 19 subjects increased while the average error of 10 subjects decreased as a result of the shuffling.[27] The hypothesis that prediction errors are as likely to increase or decrease as a

[27] Prediction errors did not change when subjects provided the same prediction twice (N=1 in ANN and N=2 in MON). Note also that when min{PRED1,PRED2}≥max{OBS1,OBS2} or min{OBS1,OBS2}>max{PRED1,PRED2}, the shuffling does not affect the average-error of the subject. The average-error of about 20%-33% of the subjects (depending on treatment and sample) therefore did not change as a result of the shuffling. This obstructs significance in individual-level testing.

result of the shuffling is rejected in a directed sign-test (α=0.07) and in a corresponding P-test for paired replicates (α=0.06) to conclude that the shuffling has increased prediction-errors in ANN. The complete sample results for MON are weaker. The mean prediction-error in the permuted series (7.5) is slightly lower than mean ERR (7.8). The number of tasks where the shuffling has increased the error is equal to the number of cases where prediction-error decreased (35 cases of each type) and we cannot reject the hypothesis that average-errors did not change because of the shuffling.

The results appear stronger when the sub-sample of 18 high-skill subjects is examined (see the right panel of Table VI). Task-level comparisons reveal that the shuffling of predictions across tasks increased prediction-errors in about 60% of the cases while decreasing the errors in only 35% of the tasks, in ANN and in MON. The averaging of errors for each subject however decreases statistical power and the hypothesis that average errors are as likely to increase or decrease is rejected for neither treatment. It is still interesting to note that the shuffling has increased both prediction errors for 7 subjects in ANN and 5 subjects in MON, while it decreased both errors for only 2 subjects in ANN and 1 subject in MON.

4. Experiment III: Selecting 2 Stocks of 6

The results of Experiment I revealed that subjects' intuitive predictions outperform the statistical rules in separating the best1 stock from the worst1 in random 6-stock menus. When the comparison was extended to the pair of best/worst stocks in each menu, the "best2" payoffs were still higher than "worst2" on average, but significance decreased (see Table III) and the payoffs implied from statistical forecasting were not significantly different from the payoffs implied from PRED (recall the "best2" data on Table V). Experiment III, was designed to test subjects' ability to choose the better 2 stocks in a random menus of 6, in a direct stock-selection design.

As in Experiment II, subjects observed 12X6 return tables (in each treatment) and were requested to select/mark 2 stocks in each menu. The instructions explained that one of the

4 stocks selected by the subject ("the payment-stock") will be used to determine the actual payout and advised subjects to "the most attractive stocks for investment". To equalize incentives in ANN and MON, the realized return was multiplied by 12 in cases where the payment stock was drawn from the MON menu. Actual payouts were determined by adding or subtracting the gain/loss on holding the payment-stock from an initial balance of 30 NIS. When the loss from investment was larger from the initial balance, subjects did not receive any payment. As in preceding experiments, subjects were invited to check and control our procedure and calculations at the end of the experiment. The experiment was run on 61 advanced MBA students (that did not participate in experiments I or II). The mean age of the participants was 33.9; the mean "theory" and "familiarity" ranks were 4.4 and 4.5 correspondingly. The mean actual payout was about 45 N.I.S with standard deviation 33.

The mean payoffs (OBS13*100) on selected and non-selected stocks are contrasted in Table VII. The left panel presents the results for the complete sample (N=61) while the right panel restricts the comparison to the subset of subjects with skill≥10 (N=31).[28] To test the significance of differences, we run randomization tests where 2 stocks are independently selected for each subject repeatedly to generate a sample of 1000 randomized selections. The proportion of cases (out of the 1000 randomizations) where the mean payoff on randomly selected stocks exceeds the mean payoff on actual selections constitutes the significance level α. When α<0.1, the hypothesis of random selection is rejected to conclude that unidentified historical-information has guided subjects into profitable stock-selection.

<Insert Table VII>

The mean payoff on selected-stocks is higher than the mean payoff on non-selected stocks in all 4 cells of Table VII. The differences are smaller in ANN where we cannot

[28] The sub-sample of subjects with skill≥10 (N=31) is almost identical to the sub-sample of subjects with familiarity≥5 (N=32). Similar results are obtained for the alternative split.

reject the hypothesis that payoffs on selected stocks are equal to the payoffs that arise from random-selection. The mean payoff on the stocks selected by the 31 high-skill subjects, for example, is 20.5 compared to a mean payoff of 16.7 on the stocks that were not selected. The randomization test reveals that randomized 2-of-6 stock selections yield mean payoff higher than 20.5 in about 22% of 1000 randomizations. We therefore cannot reject the hypothesis of "random selection" for ANN, even for the sub-sample of high-skill subjects.

The difference between payoffs on selected and non-selected stocks increases in MON. The mean payoff on selected stocks (3.7) is 75% higher than the mean payoff on non-selected stocks (2.1) in the complete sample. The randomizations suggest that larger mean-payoff (than 3.7) arise only in about 5% of 1000 randomizations. The hypothesis that selections are random is therefore rejected at α=0.05. The differences increase and significance improves when the examination is restricted to the high-skill sub-sample, where the mean-payoff on selected stocks (4.7) is more than 3 times larger than the mean payoff on non-selected stocks (1.3). The randomization test suggests significance at α=0.012.[29]

As in Section 2.6, we use STD to measure the variability of individual average-payoffs on selected stocks (parallel to STD of Best2). The STD data is provided, in smaller brackets, in the "selected" columns of Table VII. The mean STD values in 1000 randomized 2-stock selections are presented, for comparison, in the "non-selected" columns of the Table. The STD of actual payoffs is higher than the mean STD on randomized-selections in all 4 comparisons. Joint-examination of the proportion of cases (out of 1000 randomizations) where random-selection generates better portfolios, in terms of higher mean-payoff and lower STD, than actual selections, however, does not

[29] Sign-test significance is (again) weaker. The average-payoff on selected stocks is higher than the average-payoff on non-selected stocks for only 29/61 subjects in ANN and 32/61 subjects in MON when the complete sample is considered. The ratios increase to 17/31 (N.S.) and 20/31 (α=0.07) correspondingly in the sub-sample of high-skill subjects.

alter the conclusions of the R-tests on mean payoffs (alone). The proportion of cases where random 2-of-6 selections outperform actual selections (in both mean and STD dimensions) are α=0.37 for ANN and α=0.03 for MON when the complete sample is examined. Significance improves to α=0.18 for ANN and α=0.01 for MON when the examination is restricted to the 31 high-skill subjects. Again, the random-selection hypothesis is only rejected for MON.

Supplementary appendix E presents a detailed comparison of the payoffs on selected stocks to the payoffs that could be earned by using the empirical rules AVG12, AVG6 or AVG3 for selecting 2-stock portfolios. The comparison basically reconfirms the conclusions of the corresponding analysis for Experiment I (recall Table V). The payoffs implied from "statistical selection" are close to the actual payoffs and permutation tests suggest that almost all differences are not statistically significant.[30] The relative success of actual selection is stronger in MON where the mean payoff on the stocks selected by the high-skill subjects is higher than the mean payoff implied by each of the rules AVG12, AVG6 and AVG3, although none of the differences is statistically significant. It is still interesting to note that empirical selection rules like AVG12, AVG6 and AVG3 could be used to select 2-of-6 MON portfolios that significantly outperform random selection (recall Table VII), without increasing the variability of payoffs beyond the variability of random 2 stock portfolios.

5. Discussion

Brock, Lakonishok and LeBaron (1992) show that popular technical-trading heuristics like the "moving average" and the "trading-range break" obtain positive returns that cannot be explained by the stochastic process of the market. The returns for technically

[30] The only significant difference is obtained in comparison of actual payoff to the payoff implied from AVG12 in ANN. The payoff on the selections implied by AVG12 (mean: 22.2) is significantly higher than the payoff on actual selections (mean: 17.6) (R-test; p=0.04). Comparison of payoff variability however reveals that the higher payoff is obtained at the cost of higher STD (actual STD: 24.1 vs. STD for AVG12: 29.1).

generated “buy” signals are higher, and less volatile, than the returns for corresponding “sell” signals. Park and Irwin (2007) recent survey of the literature on technical-trading concludes that 56 out of 95 “modern” studies find positive results while only 20 studies reach negative conclusions regarding the profitability of technical-trading. Technical trading rules may provide additional channels for exploiting historical return-data in investment-decision. Several subjects mentioned the use of “technical rules” in their comments regarding prediction-methods. One subject mentioned a comparison of short-run trends vs. long-run trends in the spirit of the “moving average” rule. The intuitions underlying technical-trading heuristics (e.g., the expectation that a trend is initiated when the short-run average penetrates the long-run average by a significant “band”) might have directed the experimental 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, 2004; McNelis, 2005). Qi (1999) concludes that portfolios based on the recursive neural-network forecasts generate higher profits with lower risks than both the buy-and-hold market portfolio and portfolios based on linear recursive forecasts. Ince (2006) argues that neural network models may outperform ARCH/GARCH models in fitting empirical stock return data. The random selection of historical sequences in our experiments prohibits systematic analysis of the methods that lead subjects to successful performance. Our attempts to study the relation between predictions, deviations and various explanatory variables (empirical means, historical trends and other statistics) did not produce insightful results. The analysis is hindered by significant co-linearity in explanatory variables and large between-subject heterogeneity. Detailed analysis of the relation between prediction-methods and performance is therefore relegated to subsequent studies. One specific venue, for example, would be to directly compare the response of distinct subjects to similar investment menus in order to characterize the effective prediction/selection methods.

The experiments run in this study can be classified as "framed-field experiments"; i.e., experiments with field context in subjects, commodities and information.[31] Harrison and List (2004) propose six factors that can be used to characterize the field context of an experiment: the type of subject pool, the information that subjects bring into the experiment, the nature of commodity, the experimental task or trading-rules, the nature of stakes and the environment where subjects operate. The prediction/selection experiments run in this study have field context in most of these dimensions. The experiments were run on MBA students with knowledge and interest in finance. The data for the questionnaires was drawn from historical S&P500 records and payoffs were determined by the actual returns on corresponding stocks. While field experiments are gaining increased popularity in various fields of economic research, we are not aware of concrete applications regarding financial forecasting or investment decision. Field experiments could be used for collecting micro-level data on the strategies employed by investors, the type of information upon which investors react and other aspects of financial decision that are unobservable even at the level of specific investor accounts.

References

Balvers, R. J. and Wu, Y. (2006). "Momentum and mean reversion across national equity markets," *Journal of Empirical Finance*, 13(1), 24-48.

Barberis, N., Shleifer, A. and Vishny, R. (1998). "A model of investor sentiment," *Journal of Financial Economics*, 49, 307-343.

Barberis, N. and Thaler, R. (2003). "A survey of behavioral finance," in: G.M. Constantinides & M. Harris & R. M. Stulz (ed.), *Handbook of the Economics of Finance*, volume 1, chapter 18, pages 1053-1128 Elsevier.

[31] Recent examples for framed field experiments include Vollan (2008), Zhe Jon et al. (2008), List (2006) and Lusk et al. (2006).

Binner, J. M., G. Kendall, and S. H. Chen, eds. (2004) *Applications of Artificial Intelligence in Finance and Economics*. Advances in Econometrics, vol. 19. Elsevier

Bloomfield, R. and Hales, J. (2002). "Predicting the next step of a random walk: Experimental evidence of regime-shifting beliefs," *Journal of Financial Economics*, 65(3), 397-414.

Brock, W., J. Lakonishok and B. LeBaron (1992). "Simple technical trading rules and the stochastic properties of stock returns," *The Journal of Finance*, 47(5), 1731-1764.

Chan, K. and Kot H. W. (2006). "Can contrarian strategies improve momentum profits?," *Journal of Investment Management*, first quarter, 2006.

Chou, P. H., K. C. J. Wei, and H. Chung (2007). "Sources of contrarian profits in the Japanese stock market," *Journal of Empirical Finance,* 14(3), 261-286.

De Bondt, W. (1991). "What do economists know about the stock market," *Journal of Portfolio Management*, 17(2), 84-91.

De Bondt, W. and R. Thaler (1985). "Does the Stock market overreact?," *Journal of Finance,* 40, 793-805.

De Bondt, W. and R. Thaler (1987). "Further evidence on investor overreaction and stock market seasonality," *Journal of Finance,* 42, 557-581.

Diebold, F. X. and Lopez, J. A. (1996). "Forecast evaluation and combination," *Handbook of Statistics*, *Vol. 14.* G.S. Maddala and C. R. Rao, eds. Elsevier Science.

Durham, G. R., Hertzel, M. G., and Martin, J. S. (2005). "The market impact of trends and sequences in performance: New evidence," *The Journal of Finance*, 35(5), 2551-2569.

Gilovich, T., R. Vallone and A. Tversky (1985). "The hot hand in basketball: On the misperception of random sequences," *Cognitive Psychology*, 17, 295-314.

Glaser, M., Langer, T., Reynders, G. and Weber, M. (2007). "Framing Effects in Stock Market Forecasts: The Difference Between Asking for Prices and Asking for Returns," *Review of Finance,* 11(2), 325-357.

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

Green, W. H. (2003). *Econometric Analysis*. Prentice Hall.

Harrison, W.G and List, J. A. (2004). "Field Experiments," *Journal of Economic Literature*, **42**, 1009-1055.

Hey, J. and Lee, J. (2005). "Do Subjects Separate (or are They Sophisticated)?," *Experimental Economics,* 8(3), 233-265.

Ince, H. (2006). "Non-parametric regression methods," *Computational Management Science*, 3(2), 1619-1697

Jegadeesh, N. (1990). "Evidence of predictable behavior of security returns," *Journal of Finance*, 45(3), 881-898

Jegadeesh, N. and S. Titman (1993). "Returns for buying winners and selling losers: Implications for stock market efficiency," *Journal of Finance*, 48(1), 65-91.

Kahneman, A., P. Slovic and A. Tversky (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Korajczyk, R. A. and Sadka, R. (2004). "Are momentum profits robust to trading costs?," *The Journal of Finance*, 59(3), 1039–1082.

Lawrence, M., Goodwin, P., O'Connor, M. and Őnkal, D. (2006). "Judgmental Forecasting: A Review of Progress over the Last 25 Years," *International Journal of Forecasting*, 22(3), 493-518.

Lehmann, B. N. (1990). "Fads, martingles, and market efficiency," *Quarterly Journal od Economics*, 105, 1-28.

List, J. A. (2006). "Using Hicksian surplus measure to examine consistency of individual preferences: Evidence from a field experiment," *Scandinavian Journal of Economics*, 108(1), 115-134.

Lusk, J. L., Pruitt, J. R. and Bailey, N. (2006). "External validity of a framed field experiment," *Economic Letters*, 93(2), 285-290.

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

Menkhoff, L. and Schmidt, U. (2005). "The use of trading strategies by fund managers: Some first survey evidence," *Applied Economics*, 37, 1719-1730.

Park, C. H. and Irwin, S. H. (2007). "What do We know about the profitability of technical analysis?," *Journal of Economic Surveys*, 21(4), 786-826.

Qi M. (1999). "Nonlinear predictability of stock returns using financial and economic variables," *Journal of Business and Economic Statistics*, 17(4), 419-429.

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

Törngren, G. and Montgomery, H. (2004). "Worse than chance? Performance and confidence among professionals and laypeople in the stock market," *The Journal of Behavioral Finance*, 5(3), 148-153.

Tversky, A. and Kahneman, D. (1982), "Representativeness—belief in the Law of small numbers", in D. Kahneman, P. Slovic and A. Tversky (eds.), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge.

Vollan, B. (2008). "Socio-ecological Explanations for Crowding-out Effects from Economic Field Experiments in Southern Africa," *Ecological Economics*, 67(4), 560-573.

Weber, E. U., Siebenmorgen, N. and Weber, M. (2005). "Communicating Asset Risk: How Name Recognition and the Format of Historic Volatility Information Affect Risk Perception and Investment Decisions," *Risk Analysis*, 25(3), 597-609.

Zhe Jin, G., Kato, A., and List, J. A. (2008). "That's New to Me! Information Revelation in Professional Certification Markets," forthcoming, *Economic Inquiry*

Table I: Mean Predictions and Errors

The table presents mean statistics on predictions and observed returns. The number of observations is 390 (65 subjects that filled in 6 predictions in each treatment). The results for the ANN treatment are disclosed in the first row of the table and the results for MON in the second row. PRED denotes predictions. OBS13 is the corresponding hidden return. ERR is the absolute value of the difference between PRED and OBS13. "% Over-prediction" denotes the proportion of observations where PRED>OBS13. STD12 is the standard deviation of returns in the 12-observations series observed by the subjects. STD6 is the standard deviation of returns in the subsequence OBS7, OBS8… OBS12.

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%

Table II: Fixed Effect Linear Regression on ERR

The table presents the results of estimating the fixed effect model: $ERR_{ij} = \alpha_i + \beta \cdot STD12_{ij} + \gamma \cdot Familiarity_i + \delta \cdot Theory_i$. ERR_{ij} denotes the prediction error of subject *i* in problem *j*. α_i is an individual intercept for subject *i*. $STD12_{ij}$ is the standard deviation of returns in the 12-observations series presented to subject *i* in problem *j*. $Familiarity_i$ 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. A backward selection procedure is applied to delete coefficients that are insignificant at $\alpha \leq 0.05$. Theory is removed in both treatments.

	STD12	**Familiarity**	**R^2**
ANN (N=390)	0.31 (α<0.01)	N.S.	0.18
MON (N=390)	0.85 (α<0.01)	-0.62 (α<0.01)	0.23

Table III: Mean Payoff (Std) on Best vs. Worst Stocks

The table compares the payoff on the best stocks by subjects' predictions to the payoff on the worst stocks. The 6 predictions submitted by each subject in each treatment are sorted in descending order. The realized payoff (OBS13*100) on the stock that received the highest prediction is Best1. The average payoff on the 2 stocks that obtained the highest predictions is Best2 while Best3 is the average payoff on the 3 stocks with highest predictions. Worst1 similarly denotes the payoff on the last stock in the descending-prediction ranking. Worst2 is the average payoff on the last 2 stocks and Worst3 is the average payoff on the last 3 stocks in the ranking. Randomization tests are applied to test the significance of the differences between Best(n) and Worst(n) payoffs (in each of the 3 levels n=1,2,3). In each randomization the 6 stocks in each questionnaire are randomly ranked and the randomized best(n)/worst(n) payoffs are re-calculated for the shuffled series. The proportion (out of 1000 randomizations) where larger differences in mean randomized Best(n) and Worst(n) payoffs arise, constitutes the significance level of the test. The results of the test are summarized using the asterisk convention (a single asterisk denotes significance at $\alpha \leq 0.05$; 2 asterisks denote significance at $0.05 < \alpha \leq 0.1$). The randomizations are also applied to calculate the proportion α' of cases where larger difference in means arise with lower differences in standard deviations. The 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.

	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)

Table IV: Proportion of Negative Predictions in High/Low Streak-Size Group

The table demonstrates the increase in proportion of negative (contrarian) predictions following large streaks of positive returns. Observed return-series are classified as ending with an up-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 an up-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 sub-group. The left column of the table presents the proportion of negative predictions in the series ending with an up-streak smaller than the median. The right column presents the proportion for the series ending with an up-streak larger than the median. A sign-test (on individual proportions) is applied to determine the significance of differences.

	Streak size ≤ median	**Streak size > median**	**Sign Test**
ANN	14.5% (20/139)	24.5% (34/139)	N.S.
MON	26% (28/108)	41% (44/107)	α=0.04

Table V: Mean Payoff (Std) Implied by PRED, AVG12, AVG6 and AVG3

The payoffs implied by subjects' actual predictions are compared to the payoffs implied by statistical prediction. The Best(n) and Worst(n) payoffs for actual predictions are replicated from Table IV in the first row of each panel. The payoffs implied by AVG12 are calculated by ranking the 6 prediction tasks in each questionnaire in descending order by AVG12. Best1 for AVG12 is the payoff on the stock with highest AVG12; Best2 is the average payoff on the 2 stocks with highest AVG12 etc. The payoffs implied by AVG6 and AVG3 are similarly calculated. Randomization tests are applied to test the difference between the payoffs implied by subjects' actual predictions and the payoffs implied by each statistical prediction rule (D). In each randomization, the 6 prediction problems in each questionnaire are randomly sorted –twice- and the corresponding payoff is calculated for each of the random rankings. The proportion of cases (out of 1000 randomizations) where the difference in mean randomized payoffs is larger than D, constitutes the significance level of the test. A single asterisk is used to denote significance at $\alpha \leq 0.05$ while 2-asterisks denote significance at $0.05 < \alpha \leq 0.1$

		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	**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)
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)	**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)

Table VI: Mean ERR and Permutation-Tests for Selection Experiment

The table discloses the mean prediction errors in the selection experiment (Experiment II) and compares the actual prediction errors to the prediction errors obtained when predictions are shuffled across problems. The left panel presents the results for the complete sample (N=37) while the right panel presents the results for the sub-sample of high-skill subjects (N=18). The columns titled "Actual" present the actual mean prediction-errors. The columns titled "Shuffled" present the mean prediction error obtained when the first/second prediction of each subject is reassigned to second/first prediction problem selected by the subject. Sign-tests (on individual average errors) are applied to test if errors increase as a result of the shuffling.

	Complete Sample (N=37)			**High-Skill Sample (N=18)**		
	Actual	**Shuffled**	**Sign-test**	**Actual**	**Shuffled**	**Sign-test**
ANN	25.7	28.4	α=0.07	30.3	32.6	N.S.
MON	7.8	7.5	N.S.	7.3	8.0	N.S.

Table VII: Mean (Std) Payoff on Selected vs. Non-Selected Stocks

The average-payoff on the stocks selected by subjects in Experiment III is compared to the average-payoff on the 4 stocks that were not selected. The left panel of the table presents the results for the complete sample (N=61) while the right panel discloses the results for the subjects with skill≥10 (N=31). The columns titled "Selected" disclose the mean average payoff on selected stocks. The standard deviation of the corresponding average payoff is presented in smaller brackets. The columns titled "Non-Selected" present the mean payoff on the stocks that were not selected. To keep portfolio size fixed, we disclose the mean standard deviation in 1000 randomized 2 stock selections (of 6) in brackets. The randomizations are also used to determine the frequency α of cases (out of 1000 randomizations) where larger mean payoffs arise in random selection of 2 stocks for each subject.

	MBA (N=61)			**High-Skill MBA (N=31)**		
	Selected	**Non-selected**	**R-Test**	**Selected**	**Non-selected**	**R-Test**
ANN	17.6 (24.1)	17.4 (21.6)	N.S.	20.5 (26.1)	16.7 (21.5)	N.S.
MON	3.7 (6.2)	2.1 (5.9)	α=0.05*	**4.7** (6.8)	**1.3** (6.2)	**α=0.012***

Figure 1: Example to the Annual Return Prediction Task (Experiment I)

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 13-th year.
Please fill-in your prediction in the 13-th 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	