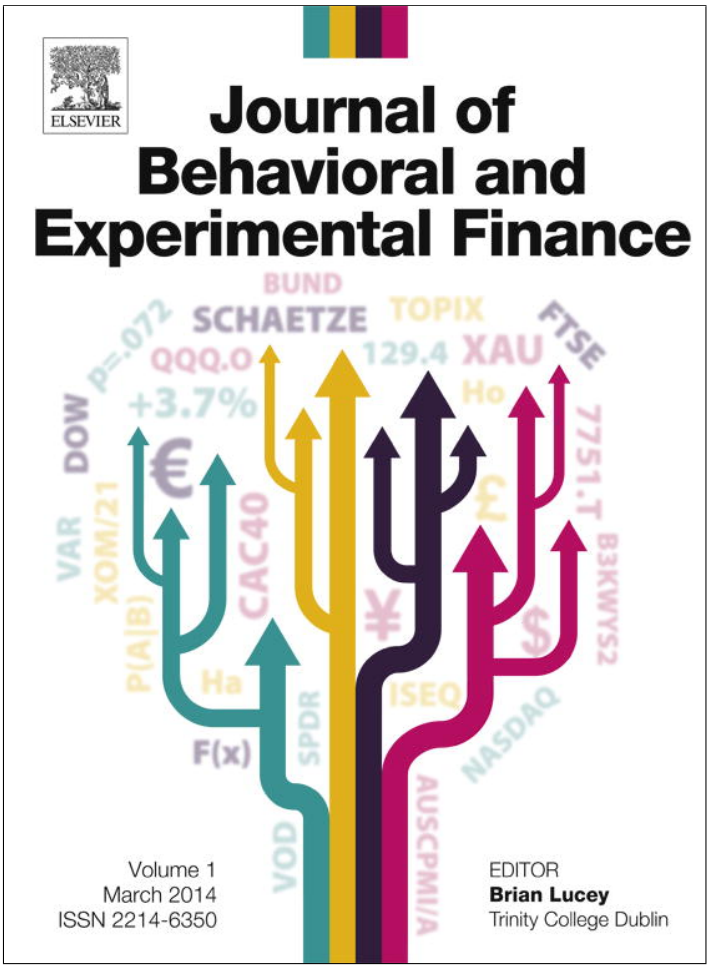

Journal of Behavioral and Experimental Finance 1 (2014) 61–73

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Finance

journal homepage: www.elsevier.com/locate/jbef

Full length article

Optimistic, but selling riskier stocks—An arbitrage experiment in crisis market☆

Doron Sonsino*, Tal Shavit

School of Business, COMAS, Rishon LeZion, Israel

ARTICLE INFO

Article history:
Received 11 November 2013
Received in revised form 29 January 2014
Accepted 29 January 2014
Available online 7 February 2014

ABSTRACT

The field-based experimental approach was utilized to collect zero-investment portfolios from more than 100 competent investors at the peak of the financial crisis. The average annual return on 117 arbitrage portfolios was 5.2% with 55% profitability rate, but prior self-confidence strongly correlates with eventual performance with yearly returns reaching 26% for the highest confidence quartile. The stocks selected for short-sale were riskier than the stocks selected for purchase and time-series estimations show that the unbalanced positions diminished profitability while markets recuperated. As most participants anticipated the recovery at the time of decision, the selling of riskier stocks suggests that "misperception of financial risk" (Shefrin, 1999) impaired performance.

1. Introduction

The increase in market transparency and swift reduction in transaction costs have boosted the volume and scope of arbitrage trading in recent years (Pole, 2007). Investment firms put vast resources into careful devising of arbitrage strategies (Beunza and Stark, 2004), while even private investors actively seek paired mispricing opportunities (Whistler, 2004). Empirical studies of arbitrage transactions, however, remain rather scarce, as arbitrage records are rarely exposed and micro-level data is hardly available. This study alternatively employs the field-based experimental approach to collect zero-investment portfolios from competent investors. The experimental arbitrages are closely analyzed to test if professionals could pre-detect under or overvalued stocks, against the efficient market hypothesis.

More than 130 participants took arbitrage positions on the Israeli stock market at the second half of 2008, in the midst of the sub-prime crisis. The task was introduced as "arbitrage in expectation" (e.g. Gatev et al., 2006) and participants were instructed to select few stocks for purchase and short-sale, assuming positions would be closed after one calendar year and payouts would be derived from market returns. The paper studies the stock selection and performance of the experimental arbitrageurs, controlling for the unique crisis conditions.

The analysis of zero-investment portfolios would be pointless in a perfectly efficient market (Fama, 1970). If prices instantly accommodate all relevant information, then it is impossible to separate underpriced or overpriced securities for profitable arbitrage. Our experiment therefore builds on the prevalence of market inefficiencies

☆ We thank participants at the Erasmus–Technion workshop on Decision and Prediction January 2012; ICABEEP/IAREP/SABE conference on Behavioral Economics and Economic Psychology in Exeter, July 2011; the ESA international meetings in Copenhagen July 2010, FUR XIV conference in Newcastle June 2010, and the 4th annual conference on psychology and investments at COMAS, College of Management Academic Studies, Israel 2012 for comments and suggestions. We also benefited from communications with Uri BenZion, Jan Bronfman, Jason Shachat, Eyal Ert, Nave Eshkenazi, Ido Erev, Shirly Farkas, Shmuel Hauser, Falahati Kazem, Eran Regev, Amnon Rapoport, Mosi Rosenboim, Eyal Shalom and Yossi Shvimer. We thank the research authority at COMAS for financial support and few anonymous referees for productive comments.

* Correspondence to: School of Business. The College of Management. 7 Rabin Blvd. P.O.B 9017. Rishon LeZion. 75190, Israel. Tel.: +972 544 996773.

E-mail address: sonsinod@colman.ac.il (D. Sonsino).

http://dx.doi.org/10.1016/j.jbef.2014.01.004
2214-6350/

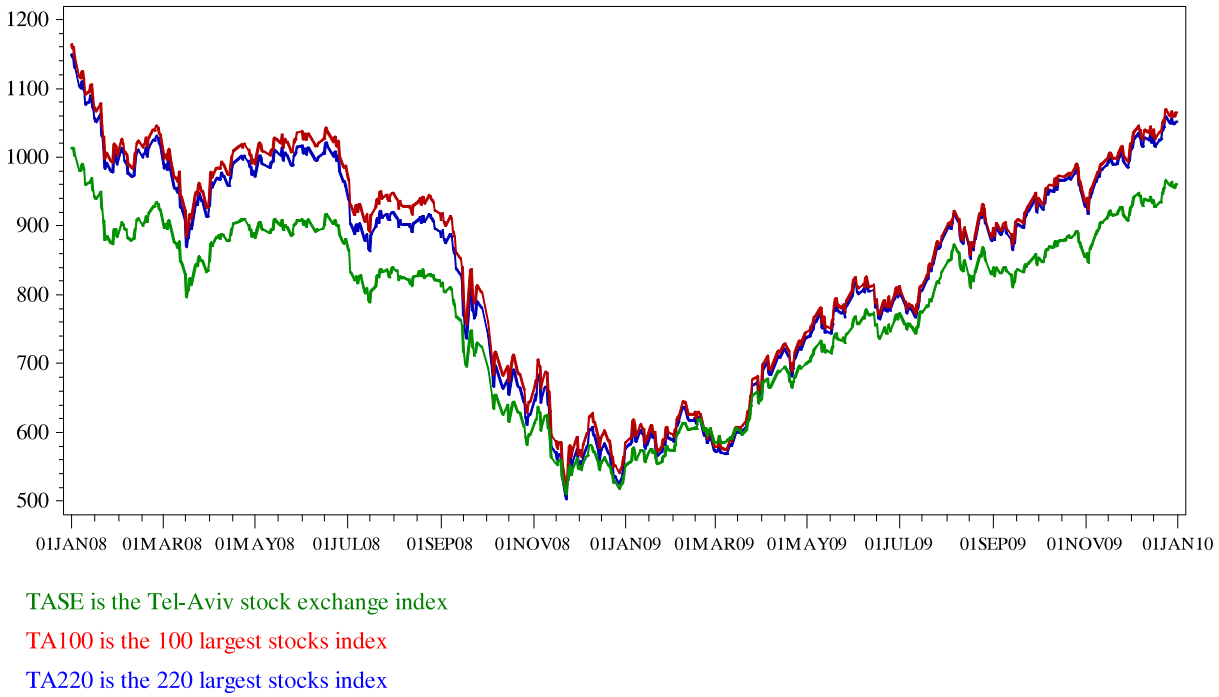

Fig. 1. Market trends. TASE indices from January 2008 to December 2009.

(Schleifer, 2003), testing the possibility of identifying mispricing opportunities for predetermined 12 months horizon. A specific virtue of our schematic few-stocks arbitrage task is the direct exploration of investors' ability to point at overpriced stocks. Most selling decisions in reality are portfolio dependent, and an alternative experiment where investors pick stocks for buy-and-hold would only test purchase decisions, relinquishing the dual task of tracing the worst performers.

The experimental arbitrageurs were carefully recruited (see Section 2) to avoid casual participation of unmotivated subjects that "throw darts" at the stock list for 50% profitability chance (Törngren and Montgomery, 2004).[1] While the purchase and sell commands were not actually implemented, the experiment was carefully designed to elicit meaningful portfolios that competent investors could pursue for speculative profit. Stock selection was restricted to the 220 largest stocks in Tel-Aviv Stock Exchange (TASE) to prohibit illiquid items, and the Web-based program was linked to major financial portals where updated information on stocks and market conditions could be freely inspected. The script has introduced the concept of mispricing arbitrage in detail, emphasizing that skillful stock selection should show profits independently of realized market conditions. The examples still demonstrated that strong losses may accrue if prices keep diverging.[2] Each participant received a personal password and could logoff the site and reenter repeatedly while deliberating choices. To borrow friction into the experiment, we introduced random short lags to the open and close dates and subtracted preannounced fees from eventual payouts.

The experiment was launched in mid-2008 and the first portfolio was received on May 17th where the leading TA100 index stood at 1026.5 points. The economic crisis had struck the market with full force at the 4th quarter of the year, pushing TA100 down to a 5-year slump of 518 points in late November (Fig. 1). Prices started recovering towards the end of the year and trends were mostly positive in 2009. Anecdotally, the index was back at the start level of 1027 points on December 19th 2009 where our latest yearly portfolio was liquidated. Our final sample consisted of 133 portfolios, but 16 arbitrages were later removed for choosing stocks that stopped trading along the yearly arbitrage. The participants appear highly professional with 66% reporting more than 15 years of formal education and 49% proclaiming investment-industry experience (typically, short experience of less than 5 years). On average, the permitted-stocks index (TA220) increased by 10.9% along the 117 eligible yearly arbitrages, but variability is high with the index decreasing in 48% of the cases. Since the portfolios were submitted at separate dates in times of extreme volatility, we control for specific yearly conditions throughout the analysis.

The bottom-line results of the test are positive but statistically marginal. The mean annual return (for real dates; before fees) was 5.2%, with median 4.8%, and the proportion of positive results was 55% ($p = 0.16$). The results are very robust and comparable statistics emerge when extreme performers are ignored and when portfolios that borrowed or deposited funds are dropped. The 117 portfolios show large discrepancy: 82 distinct stocks were selected for purchase and 97 stocks were chosen for short-sale, although each arbitrageur could select up to 3 items on each side. The crisis sensitive CONSTRUCTION

[1] If portfolios are selected at random from a given stock list, then the likelihood of (buying portfolio A and selling B) equals the likelihood of (buying B and selling A). Since this holds for all possible A–B combinations, profitability chances are 50%.

[2] The translated script is available at http://www2.colman.ac.il/business/doron.

and FINANCE stocks captured about 30% of long-side investments, but concurrently accounted for 29.6% of the short-side volume.

Beyond the key investigation of professional stock-selection ability, the experiment aimed at running exploratory analysis of the determinants of successful arbitrage performance. The Internet program included a detailed personal questionnaire and few subsidiary tasks, attempting to characterize participants in terms of risk-preferences, market familiarity, TASE expectations, confidence, academic backgrounds and more. The supplementary tasks were minimized to limit experimental load, and for obvious reasons, we cannot discuss the many hypotheses underlying the (exploratory) design. The significant findings are discussed directly in the following sections. The remaining of the introduction introduces 2 intriguing contributions of the paper: showing that crude, intuitive self-confidence scores may strongly predict performance in complicated tasks such as annual arbitrage choices, and documenting a puzzling gap between personal expectations and arbitrage styles in line with Shefrin and Statman's (1999) "misperception of risk".

The motivation for collecting prior confidence scores from the experimental arbitrageurs arrived from the core psychology judgment literature. In particular, we drew on familiar studies where subjects mark their selections in multi-choice knowledge problems, while stating their percentile level of confidence in choosing the correct option (Lichtenstein et al., 1982). Dozens of experiments reveal that the correct choice rates typically increase with subjective confidence, although subjects exhibit significant overconfidence in terms of exaggerated expectations (cf. Fig. I in Klayman et al., 1999 where the correct choice rate when subjective confidence lies at 70%–79% is 65%, compared to correct choice rate of about 75% when confidence is at 90%–99%). The exact patterns of calibration and overconfidence, however, vary with experimental design, and recent studies demonstrate that the correlation between self-confidence and realized performance may deteriorate in non standard assignments (Clark and Friesen, 2009) and even vanish in challenging binary stock-selection tasks (Sonsino and Regev, 2013). Our crisis-market arbitrage tasks therefore appear as an interesting application for confronting intuitive confidence and eventual performance.

Subjective confidence was elicited at the final screen of the program, where each participant point estimated her percentile chances for positive return.[3] The mean confidence level was 57% with median 60%. Given the complexity of the zero-investment task, our null hypothesis was that confidence would not show predictive power for personal performance. The intuitive conjecture, however, was strongly rejected with arbitrage results turning quite impressive when the analysis is restricted to participants that showed strong self-confidence a priori. The annual return of participants with higher than median confidence was almost 19%, while the arbitrageurs with CONF $>$ 75% earned more than 26%. A comprehensive regression analysis suggested that CONF is the most significant predictor of yearly payoffs, beyond direct competency or motivational controls such as performance in market-familiarity quiz, years of industry experience, and time spent on various screens along the experiment. In Section 5 of the paper we employ time-series estimations in the spirit of Fama and French (1993) 3-factor model and Carhart (1997) 4-factor generalization to test if the ex-ante confident arbitrageurs benefited from building more risky portfolios or otherwise show common arbitrage styles that explain their superior results. The estimations however could not generally explain the success of confident types. While some high-CONF traders successfully loaded on familiar return-generating factors such as momentum or SIZE, the strong performance of others mostly reflects in risk-adjusted ALPHA. The one-year lagged confidence scores interestingly emerge as an applicable instrument for separating the successful arbitrageurs in advance.[4]

Our second intriguing observation concerns the arbitrage styles of participants. The emerging literature on the role of emotions in financial decision proposes that anxiety may decrease willingness to take investment risk (e.g. Kuhnen and Knutson, 2011). Since the arbitrage portfolios were collected in the time of unprecedented crisis, we hypothesized that participants would tend to build crisis-prone portfolios, investing in relatively safer stocks while short-selling riskier companies. Comparison of the stocks selected for purchase to the stocks selected for short-sale indeed revealed that the experimental arbitrageurs sold riskier stocks in terms of standard risk factors (Fama and French, 1992) such as SIZE, book-to-market (BTM) and multiplier-related ratios. The stocks selected for sale, in particular, were 50% smaller (in market capitalization) than the stocks selected for purchase, sell-side BTM ratios averaged at 0.75 compared to 0.59 on the buy-side, and negative earnings-per-share stocks were traded twice more often on the sell-side. To approximate the impact of unbalanced positions on eventual returns, we employ again the time-series factor estimations. The estimations were separately run for each portfolio using the daily return series for the calendar year starting at the DATE where the arbitrage was submitted, and risk premia calculations were slightly adapted to take into account the experimental constraints on stock selection (e.g., the restricted list of permitted stocks). The most consistent results were obtained for the SIZE factor. The adapted SIZE premia averaged at 30.9%, suggesting that the stocks with lowest market capitalization outperformed the largest SIZE stocks significantly along the 117 yearly arbitrages. Since the arbitrageurs however sold relatively smaller size stocks, the estimations unsurprisingly exposed negative β_{SIZE} coefficients (representing the reversed loading on SIZE premia) averaging at -0.12. The negative loadings on positive premia represent the loss that the arbitrageurs suffered from

[3] Only 16 arbitrageurs were female. Female results were not significantly different.

[4] Prior-confidence similarly showed strong predictive power for eventual performance in short-run (quarterly) arbitrage tasks. The quarterly arbitrages are discussed in Section 5.

Stock Purchase or Deposit			Stock Sale or Loan	
Stock Name	Amount Purchased		Stock Name	Amount Sold
▼			▼	
▼			▼	
▼			▼	
Total	1000		Total	1000

Fig. 2. The arbitrage table.

selling riskier stocks while prices climbed back to pre-crisis levels. Interestingly, the participants delivered optimistic expectations regarding the market trends for their annual arbitrage, while picking riskier stocks on the sell-side. About 73% of the participants expected an increase in TA100 along their arbitrage year and 38% expected an increase of at least 10%. While the selling of riskier stocks could be rationalized if the crisis was expected to linger, the unbalanced positions seem puzzling in light of the expected recovery. We attribute the apparent contradiction to misperception of financial risk (Shefrin and Statman, 1999; Shefrin, 1999). Our crisis-affected investors confuse the stereotype of riskier companies with the quality of investment, choosing to sell small size riskier stocks although prices hit 5-year low records and recovery was anticipated for the arbitrage interval.

The paper proceeds as follows: Section 2 provides more details on the method of the experiment. The bottom-line results are summarized in Section 3. Section 4 studies stock selection more closely while Section 5 discusses the time-series estimations. Section 6 concludes.

2. Method and participants

The experiment was carefully designed to encourage serious deliberation of the arbitrage. Participants could disconnect from the site at any point and login later to proceed from the same point where they logged-off. The instructions could be reexamined repeatedly but answers or choices could not be revised after submission. The program measured the number of repeated LOGINs, the total login time and the time spent on selected pages. On average, the participants took 2.5 entries to complete the questionnaire; only 23% completed all tasks in a single login. The mean login time was about 47 min, but 56% (65 of 117) took less than 15 min on site.[5] An automatic control-mail was sent to our address whenever a portfolio was delivered. A confirmation note with details on payout procedures was sent to the participant.

Fig. 2 replicates the table through which the portfolios were delivered. In addition to selecting stocks from TA220, respondents could either deposit or borrow funds in fixed annual interest rate of 8%.[6] The "DEPOSIT" and "LOAN" options were appended to TA220 and the joint list was accessible, as a drop-down menu, in each row of the table. Participants could select up to 3 stocks for purchase and 3 distinct stocks for sale. The arbitrage volume was arbitrarily set at 1000 New Israeli Shekel (NIS) and the software validated that the volume of stock-purchases (including deposits) equals the volume of short-sales (plus loans). Three examples illustrated the possibilities to construct minimal arbitrage (buying one stock and selling another), borrow funds within the arbitrage, or deposit some of the amount. The instructions were linked to 3 financial portals, including the formal TASE site, with vast information on available stocks and underlying companies. William Goetzmann's Internet chapter on "arbitrage in expectations" was connected for those seeking additional background.[7]

Since the purchase and sale commands were not actually executed, we introduced artificial noise to capture the obstacles that may arise in implementing such orders in practice. The instructions emphasized that arbitrage positions would be opened 1–5 days after submission and closed 355–365 days after opening. The exact delay and duration were randomly assigned, and 20 NIS were deducted for "fees and commissions".[8] The ensuing analysis however addresses the actual returns before fees, for the exact dates.

Attempting to exclude unmotivated respondents, we downplayed monetary incentives, emphasizing the interest that investors might find in the experiment and its results. Preliminary calls for registration were distributed in professional portals and alumni lists. Usernames and password were awarded only to registrants that proclaimed adequate academic background and familiarity with the local market. The preliminary announcement explained that 100 randomly selected participants would receive checks in amounts that increase with their arbitrage result. The exact payout formula was provided as an optional link towards the end of the instructions. The "final balance" of each participant was determined by adding or subtracting the arbitrage payoff from an initial balance of 200 NIS. Payout checks were set at 50% of the balance.[9]

[5] Since we cannot control for the actual engagement of participants, the login times are used as proxies for the time spent on various screens.

[6] TA220 was tentatively constructed by joining the stocks in TA100 (the largest 100 stocks) and YETER-120 (the next 120 largest companies). The stocks in TA220 accounted for 90%–93% of the total stock market capitalization of TASE along the relevant periods.

[7] http://viking.som.yale.edu/will/finman540/classnotes/class6.html.

[8] As the opening (closing) of substantial arbitrage positions might affect market prices, positions may be gradually acquired (closed) in different rates. The random lags were implemented to import such price uncertainty into the experiment. In general, the lags and fees make the experimental task more challenging to increase external validity.

[9] In devising the payout scheme we did not anticipate the extreme turbulence. The initial balance of 200 NIS could only cover 20% annual loss, while losses were steeper for 28 portfolios.

Table 1
Illustrative regressions.

Equation	Model	CONF	T(ARB)	LOSS_AVERSION	LOAN	Δ(TA220)	R^2
(1)	Linear	0.28** (0.15)	−0.014* (0.009)	10.55** (5.88)	28.7** (13.6)	25.1* (17.7)	10.3%
(2)	Linear	0.32** (0.15)	−0.015* (0.009)	10.6** (5.87)	27.8** (13.2)	–	8.6%
(3)	Logistic	**0.015**** (**0.0055**)	−0.0005 (0.0004)	0.36* (0.21)	−0.17 (0.48)	0.30 (0.62)	–

The dependent variable in the linear regressions (1)–(2) is the percentile arbitrage return R. An indicator for profitable arbitrage ($R > 0$) is used in the logistic specification (3).

To encourage participation of industry professionals, we guaranteed confidentiality explaining that feedbacks would be distributed anonymously using ids. Many registrants still logged off the site without submitting their portfolio, and in early January 2009 we stopped recruitment with (smaller than planned) sample of 133 arbitrages. The mean age was 32 and formal education years averaged at 16.5, which implies 1–2 years of graduate studies. Only 24 participants held investment industry jobs, but almost 50% proclaimed past experience. About half of the sample thus consists of young industry veterans that did not face the barriers that could intimidate present professionals.

The supplementary appendix (see Appendix A) presents the list of personal attributes collected along the questionnaire. Subjects ranked their FAMILIARITY with the local market and their academic background in finance (THEORY) in 1–7 scales, filled-in a standard risk-preference task, and took a short familiarity quiz. The risk-preference assignment consisted of 6 standard binary choice problems, between two-outcome lotteries and risk-free payoffs. The first 3 problems roughly measured the inclination to take risk with gains, while the last 3 problems similarly tested for risk attitudes where losses are possible. The 0–3 variables RISK AVERSION and LOSS AVERSION henceforth denote the proportion of safe choices in each type of problems.[10] The Web-tailored quiz program randomly selected 3 multi-choice problems, from 3 pools of 25 similarly-challenging questions, for each participant. Quiz time was limited to 180 s and the program automatically proceeded to the next page when time was done. QUIZ performance was measured by subtracting the number of mistakes from the number of correct choices.

To close the introductions we briefly outline the statistical conventions of the paper. We use non-parametric sign tests and signed rank Wilcoxon tests, depending on context, to test the significance of returns, risk premia and related variables, keeping the Pitman permutation test for between-sample comparisons. We report 1-tail significance levels, using bolded $\mathbf{p}^{**}$ in cases where $p \leq 0.01$; plain p^{**} for $0.01 < p \leq 0.05$, and p^{*} for marginal $0.05 < p < 0.1$. Arbitrage returns (if this deserves clarification) were calculated by dividing the net payoff upon closing the arbitrage by the 1000 NIS volume. If the arbitrageur, for example, purchased equal amounts of 2 stocks that increased by 10% and 40% respectively, while short-selling another stock that increased by 15%, then the payoff on the arbitrage is 100 and the return R is 10% (−15% negative return on the sell-side vs. positive 25% return "long").[11] To account the specific market conditions for each portfolio, we calculate the change in TA220 along the arbitrage year, using Δ(TA220) to represent the variable.

3. Results

The mean yearly return on the 117 eligible portfolios was 5.2% with median 4.8%. Only 64 arbitrages (54.7%) closed in gain and the hypothesis that payoffs are centered at zero could not be rejected (Wilcoxon signed-rank; $p = 0.16$). Volatility is large with returns ranging between −217% and +277% (see the Web supplement Appendix A for a sketch of distribution), but the results are robust to removal of extreme performers. The participants with $|R| < 50\%$ gained 4.9% with profitability rate 57% ($N = 91$; $p < 0.05$). When the 20 portfolios that purchased the "most popular" stock on the buy-side are ignored, the mean return slightly decreases to 3.9% but the $N = 44$ participants with CONF > 60% still deliver 18.2%. Removal of the 17 portfolios that selected the most traded stock "short" similarly leaves the mean return at 5.7%. Alternatively, the results improve when the portfolios that selected stocks that stopped trading are reconsidered. Assuming that positions in nonviable stocks were closed at the last trading day, the mean return on these 16 portfolios amounts to 25% with median 9%.

Closer examination of the eligible sample reveals 11.4% gain on long-side investments, offset by 6.1% loss on sell-side positions. The stronger performance on the buy-side clearly follows from the mostly positive trends along the arbitrage year (Fig. 1). Arbitrage returns show mild positive correlation with Δ(TA220) ($\rho = 0.18$; $p < 0.05$). Marginally significant Δ(TA220) coefficient also emerges in line (1) of Table 1, where we account other significant determinants of performance (discussed below). Similar results, however, emerge in line (2), where the control for market conditions is removed.

[10] The RISK AVERSION lottery paid +200 or +50 with equal probabilities. Subjects chose between the lottery and risk-free payoffs of 165, 125 and 85. The LOSS AVERSION problems similarly used +200 or −50 lottery with risk-free payoffs of 105, 75 and 45.

[11] In general, returns were determined using dividend-adjusted prices.

Table 2
Returns by the level of confidence.

CONF range	*N*	Mean *R*	Median *R*	% ($R > 0$)	Mean CONF
CONF < 50	31	−2.8	−6.8	42%	22%
$50 \leq$ CONF ≤ 60	35	−7.3	0.6	51%	54%
$60 <$ CONF ≤ 75	22	8.7	4.5	55%	70%
CONF > 75	29	26.2	21.8	72%	86%
Complete sample	117	5.2	4.8	55%	57%

To extract the variables that affected performance across the sample, we employed a regression with model-selection procedure that iteratively removes variables for insignificance and repeatedly tests the expunged factors again as the iterations proceed. The analysis tested different collections of explanatory variables (a partial list is available in the supplement, see Appendix A), alternative model specifications (linear, logistic) and various model selection methods to extract the mutually significant bottom-line effects. The next paragraphs discuss the main findings directly, using Table 1 to illustrate intensity.

Prior self-confidence regarding profitability emerges as the strongest predictor of annual returns. Confidence was elicited at the final screen of the program where participants point estimated their chances for positive payoff in 0%–100% scale. The median CONF was 60%, but more than 25% (31 of 117) estimated their profitability chances at less than 50%. Since gains and losses are equally probable under random selection, CONF $< 50\%$ represents irrational pessimism or under-confidence. The pessimistic expectations however materialized in our sample. The mean return of the pessimists was negative −2.8% with arbitrage positions closing at loss in 58% of the cases. On the opposite extreme, the 29 participants that assigned probability $> 75\%$ to eventual profitability, earned 26% with profitability rate 72%. Statistics for the intermediate CONF groups are provided in Table 2.

The strong link between confidence and performance could be natural if confidence increased with competency. Individual confidence levels, however, negatively correlated with years of industry experience ($\rho = -0.15$) and quiz performance ($\rho = -0.10$), showing close to zero correlation with proclaimed market familiarity ($\rho = 0.05$) and academic background ($\rho = -0.03$). The skills related variables were constantly removed in model selections, suggesting that performance did not improve with standard competency scores. The 2 columns at the left of Table 2 however demonstrate that the confident participants overestimated their profitability chances. The mean CONF of the 51 participants with CONF $> 60\%$, for example, was 79% while the actual profitability rate for the respective portfolios was 65%. The literature on psychological bias in financial decision mostly suggests that over-confidence, in various manifestations, undermines performance. Over-confident investors trust their private signals excessively, trade too often, and accordingly close transactions in unfavorable prices (for comprehensive survey discussions see Subrahmanyam, 2008; Skala, 2008). The current results still propose that intuitive confidence may powerfully predict trading performance, even when the highly confident fall into the over-confidence trap. The arbitrage styles of the confident participants are closely explored in Section 5 to test if their successful performance may be ascribed to particular strategies such as purchase of relatively aggressive stocks or successful use of momentum strategies.

Table 3
Returns by CONF and T(ARB).

	Low CONF	High CONF	All portfolios
T(ARB) < 119	−4.7 −0.2 48%	+28.6 +17.7 69%	+11.9 +13.6 59%
T(ARB) ≥ 119	−3.2 −3.5 47%	+0.7 +6.3 55%	−1.3 +3.2 51%
All portfolios	−4.0 −0.2 47%	+14.6 +12.5 62%	+5.2 +4.8 55%

The numbers in each cell represent the average *R* (top), median *R* (middle), and the proportion of $R > 0$ (bottom) for each group. Sample sizes starting at the up-left cell and rotating clockwise are 29/29/29/30. In cases of tie in CONF, participants with higher T(ARB) were assigned to the low CONF category.

The second variable showing significance by the illustrative regressions is the login time to the screen with the arbitrage table (Fig. 2); henceforth denoted T(ARB). On average, the participants spent less than 3 min on the arbitrage screen, but variability is large with T(ARB) ranging between 21 s and more than 100 min. The rightmost column of Table 3 runs a median split of the sample by T(ARB). The average return on the 59 portfolios with T(ARB) higher than median was negative −1.3% compared to mean return of +11.9% on the remaining 58 portfolios ($p = 0.12$). The left columns of the table split each sub-sample further by CONF, demonstrating the contrasting effects of confidence and login time on performance.[12] The strongest results, mean return 28.6% with 69% profitability rate, emerge for the high-confidence participants that spent less than 119 s on the arbitrage screen. The respondents in the low-CONF high-T(ARB) category, on the opposite, earned −3.2% with 47% $R > 0$ rate. T(ARB) negatively correlated with skills related measures such as years of education ($\rho = -0.14$), academic background ($\rho = -0.09$), and industry experience ($\rho = -0.09$), suggesting that the negative coefficient indirectly represents the weaker performance of generally less competent participants.

[12] The Pearson coefficient of correlation between CONF and T(ARB) was 0.13. The correlation between CONF and TA100 expectations was insignificant 0.09.

Table 4
Comparison of stock-selection long vs. short.

	Between-stock comparison			Within-portfolio comparison		
	Long	Short	Significance	Long	Short	Significance
Market BETA	1.24	1.26	$p = 0.31$	1.12	1.01	$p = 0.12$
SIZE	5628	2515	$p < 0.01$	6335	3230	$p < 0.01$
BTM	0.60	0.76	$p < 0.01$	0.59	0.75	$p < 0.01$
PREV_3MON	−15.0%	−20.6%	$p < 0.01$	−12.7%	−18.5%	$p < 0.01$
PREV_6MON	−22.0%	−29.1%	$p < 0.01$	−19.4%	−26.7%	$p = 0.01$
Trading sector						
Construction	17.0%	22.5%	$p = 0.07$	15.8%	16.8%	$p = 0.33$
Financial	15.4%	14.1%	$p = 0.35$	14.5%	12.9%	$p = 0.35$
Investments	16.2%	8.5%	$p < 0.01$	14.2%	6.0%	$p < 0.01$
Industry	37.2%	31.5%	$p = 0.10$	33.4%	25.4%	$p = 0.03$
Trading	13.0%	21.6%	$p < 0.01$	11.2%	17.6%	$p = 0.06$

The BETA of each stock was calculated from the daily return series for the calendar year preceding the arbitrage; SIZE is the NIS market capitalization of the stock at the arbitrage date converted to million USD; BTM is the book-equity to market-capitalization ratio for the latest quarter preceding the arbitrage by (at least) 3 months; PREV_3MON represents the return on each stock in the 3 months preceding the arbitrage date; PREV_6MON represents the historical returns for the 6 months preceding the arbitrage. The split into sectors follows the formal categorization of TASE (ENERGY stocks are omitted). Sample sizes for between-stock comparisons are: 247 (long) vs. 213 (short) for BETA, SIZE, PREV_3MON, PREV_6MON and sectoral-affiliations; 246 vs. 208 for BTM (a few cases where BTM < 0 are ignored). Deposits and loans are ignored in the weighting of SIZE/BTM/PREV_3MON/PREV_6MON at the portfolio level (see the Web appendix for details). Sample sizes for within-portfolio comparisons are $N = 117$ for BETA and sectoral-affiliations (the sectoral-affiliation proportions sum up to 100% when loans, deposits and energy are appended); $N = 89$ for SIZE, PREV_3MON, PREV_6MON (participants that borrowed or deposited the complete arbitrage volume are ignored) and $N = 87$ for BTM (ignoring cases of BTM < 0). The Pitman test is employed for between-stock comparisons while the Wilcoxon signed rank test is used for within-portfolio comparisons.

Interestingly, the analysis also revealed that arbitrage returns and profitability-likelihood increased with LOSS AVERSION, while RISK AVERSION did not affect eventual performance. The correlation between the 2 discrete 0–3 risk-preference measures was significant but far from perfect $\rho = 0.31$. The mean return on the portfolios of the 73 relatively loss-averse participants (those that preferred the risk-free payoff at least in 2 of 3 respective choice problems) was 6.8% compared to 2.6% return on other portfolios, and the regressions suggest that LOSS AVERSION shows significance at $p < 0.05$ when other determinants of performance are controlled (equations (1)–(2) in Table 1). We return to LOSS AVERSION in the concluding discussion, noting that the crude 0–3 measure interestingly interacts with the arbitrage styles of participants.

4. Stock-selection long vs. short

This section examines more closely the stock selections of the 117 participants. First, we test if the combined selections could be utilized to successfully detect the best stocks for purchase or sale. We start by sorting the 131 stocks that were selected for purchase or sale into 3 distinct categories. The 48 stocks traded on both sides of the arbitrage (bought and sold at least once within different portfolios) compose the BUY-and-SELL list; the 34 stocks that were purchased but never sold constitute the BUY-ONLY list; while the 49 stocks that were sold and never purchased similarly sort into the SELL-ONLY category. About 60% of the arbitrage volume (67% of long-side volume and 52% of short-side volume) was run with BUY-and-SELL stocks. The BUY-ONLY stocks attracted 24% of the long-side volume while the SELL-ONLY items drew 28% of the short-side volume (the proportions add up to 100% with deposits and loans). Comparison of the dates where each of the 48 BUY-and-SELL stocks was selected for purchase to the dates where the same stock was selected for sale did not reveal consistent trends. Particular stocks were traded in opposite directions within few weeks and the hypothesis that DATEs of purchase equal DATEs of sale was meaningfully rejected for only 3 stocks.[13] With 60% of the volume involving stocks that were bought and sold concurrently by distinct traders, the experimental portfolios cannot be used to effectively separate TA220 into disjoint BUY and SELL lists. The Pearson coefficient of correlation between (total volume of sale) and (volume-weighted sell-side return) for the 97 stocks selected for sale, however, was positive 0.17, suggesting that performance improved with volume of trading on the sell-side. The results are weaker on the buy-side where the correlation between (volume of purchase) and (buy-side weighted-return) was close to 0 ($\rho = -0.04$; $N = 82$).[14]

Another interesting angle for closer inspection deals with the balancing of risks long vs. short. Table 4 compares the stocks selected for purchase to the stocks selected for sale in terms of historical risk-factors (BETA, SIZE and BTM), recent performance, and sectoral-affiliation (see precise definitions below the table). The left panel runs the comparisons between-stock; e.g., comparing the 247 stock-purchases along the experiment to the 213 short-sales in terms of historical BETA. The right panel alternatively runs the comparisons within-portfolio; e.g., calculating the volume-weighted BETA for each side of the arbitrage (with BETA $= 0$ for deposits and loans) and using the 117 paired differences to test equality at the portfolio level.[15]

[13] 28 stocks in the BUY-and-SELL list were purchased or sold only once along the experiment; we ignore these cases when testing the hypothesis that buy-side dates equal sell-side dates.

[14] The weighted-return for buying (selling) each stock is calculated as a volume-weighted average of the returns earned by subjects that bought/sold the stock. Stock-separation improves when the analysis is restricted to the 57 participants with $R > 5\%$. The proportion of trading with SELL-only (BUY-only) stocks doubles to 46% (52%). The correlation between volume and return turns 0.30 on the sell-side.

[15] Comparisons are run both ways as each method compensates for limitations of the other. The between-stock comparisons ignore the

While the experimental portfolios are almost balanced in terms of market-risk (average volume-weighted BETA 1.12 long vs. 1.01 short; $p = 0.12$), arbitrage positions appear extremely unbalanced in terms of SIZE and BTM. The stocks selected for purchase were about twice larger, on average, from the stocks selected for sale ($p < 0.01$, in both levels of comparison) and the average BTM ratio on the sell-side was about 0.75 compared to 0.60 on the buy-side ($p < 0.01$). The stocks selected for sale thus appear significantly riskier than the stocks selected for purchase in terms of Fama and French (1992) historical risk factors. About 54% of the portfolios (that did not borrow or invest the 1000 NIS volume; $N = 89$) sold smaller-SIZE and higher-BTM stocks compared to the stocks purchased; 83% satisfied at least one of the conditions.

The PREV_3MON line in the table additionally suggests that the participants tended to purchase stocks that decreased less steeply (or climbed more rapidly) in the 3 months preceding the delivery of the arbitrage. The stocks selected for purchase decreased on average by −15% in the quarter preceding the arbitrage, while the stocks selected for sale plunged by −20.6% ($p < 0.01$, in both levels of comparison). The inclination to sell weaker stocks reemerges when recent performance is measured for the 6 months preceding the arbitrage date (see the PREV_6MON line in the table). Only 37% of the portfolios (33 of 89) adopted "contrarian positions" in the sense of buying stocks with weaker 3 or 6 months performance compared to the stocks sold.

The comparison of sectoral-affiliation finally reveals that the CONSTRUCTION and FINANCIAL sectors (that plunged most drastically along the early stages of the crisis) were actively traded on both sides, attracting together 30.3% of the long-side volume and 29.7% of the short-side trade. The mixed parallel trading represents again the strong uncertainty conditions under which the portfolios were collected. The next section employs the factor estimation methodology to test if participants could still separate stocks effectively to produce risk-adjusted ALPHA.

5. Time-series estimations

5.1. Method

Applications of factor estimations in the spirit of Fama and French (1993) and Carhart (1997) are numerous, but our paper is the first utilizing the approach to closely investigate the decisions of experimental traders.[16] Since loadings on momentum show significance for 52% of the portfolios, we discuss the 4-factor version, noting that similar insights emerge in 3-factor estimations.

Since the experimental portfolios were collected on distinct dates and arbitrage styles could strongly differ amongst participants, the estimations are separately run for each portfolio. As the portfolios were constructed for a relatively short horizon of one calendar year, the estimations are run on daily returns. The daily returns on each portfolio were regressed on corresponding daily MARKET, SIZE, BTM and MOMENTUM premia (detailed definitions follow). The number of observations in the daily regressions lied between 261 and 271 depending on the specific calendar dates. Following common practice in the literature, estimations were run using GMM with the Newey–West (1987) method to correct errors for violation of standard assumptions. The 4-factor model is represented by the equation

$$R_{j,t} = \alpha_j + \beta_{\mathrm{MARKET},j} * (\mathrm{MARKET}_{j,t}) + \beta_{\mathrm{SIZE},j} * (\mathrm{SIZE}_{j,t}) + \beta_{\mathrm{BTM},j} * (\mathrm{BTM}_{j,t}) + \beta_{\mathrm{MOM},j} * (\mathrm{MOMENTUM}_{j,t}),$$

where $R_{j,t}$ is the return on portfolio j at day t and $\mathrm{MARKET}_{j,t}$, $\mathrm{SIZE}_{j,t}$, $\mathrm{BTM}_{j,t}$, and $\mathrm{MOMENTUM}_{j,t}$ denote the risk premia for portfolio j at day t.[17] The coefficients $\beta_{i,j}$ then represent the loadings of portfolio j on premia i while the intercept α_j represents the excess daily return beyond loading on the 4 factors. The daily returns and daily premia were measured in percentile form. An estimated alpha of 0.02 would therefore represent 0.02% (2 basis points) daily adjusted return on the arbitrage.

In calculating the daily premia on risk, we adopt an applicative approach that takes into account the experimental constraints on stock selection and directly approximates the premia that participants could collect by exposing to common risk factors.[18] Consider the MARKET-risk first. The most risky portfolio in terms of exposure to market-risk could be constructed by ranking the 220 available stocks by their BETA for the arbitrage DATE; buying the 3 stocks with highest BETA and selling the 3 stocks with smallest BETA (or vice versa). Let MARKET1 denote the return on the corresponding portfolio assuming the specific amounts of purchase or sale were derived from the relative market capitalizations (SIZE) for the sorting date.[19] If MARKET1 is used to measure the daily market premia and some participant followed the respective strategy, buying the 3 most aggressive and selling the 3 most defensive stocks in TA220, then the time-series regression of daily arbitrage

volumes of purchases and sales. The within-portfolio comparisons of SIZE, BTM, historical returns and sectoral-affiliation, on the other hand, must completely ignore the 28 participants with 1000 NIS LOAN or DEPOSIT. The two comparison methods are further motivated/ illustrated in the Web supplement (see Appendix A).

16 Recent examples of 4-factor estimations on empirical arbitrage portfolios include Doukas et al. (2010), Clark and Kassimatis (2012), Von Lilienfeld-Toal and Ruenzi (forthcoming). For recent example where the model is run on daily return series see Amihud and Goyenko (2013).

17 The premia is indexed by j and t since our risk-riding portfolios may change with the DATE where each arbitrage was delivered; the point is clarified next.

18 Carhart (1997, p. 61) notes that the 4-factor model may either represent an equilibrium model with 4 risk factors or a "performance attribution" model that could be used to explore profitability sources. In line with the latter interpretation, our adapted version makes some necessary adjustments to approximate the premia that participants could collect within the experiment.

19 To skip unnecessary notation we use "MARKET1" universally in discussing the arbitrage portfolio, the daily return series and the annual premia. The market premium is typically calculated using the difference between market return and risk-free interest, but we preferred using BETA-sorted portfolios since these match the "performance attribution" purpose more closely. Similar results, but 1/3 lower fit levels, emerge when the premia is approximated using the daily TA220 return (see the Web appendix).

Table 5
Annual premia (percentile form).

	Market	Size	BTM	Momentum
Mean	5.2	**30.9****	**18.2****	10.5
Median	−3.8	**21.9**	**21.2**	0.23
(Standard deviation)	(35)	(**24.6**)	(**17.3**)	(41.9)
Minimum/Maximum	−39/118	1/ 91	−14/49	−42/145
Proportion positive (of $N = 117$)	44%	100%	87%	50%
Sign-test significance	N.S.	$p < 0.01$	$p < 0.01$	N.S

The risk premia were separately calculated for each portfolio. The upmost row presents the average, median, and standard deviation of the annual premia. The asterisks mark significance by Wilcoxon signed rank test ($N = 117$). The second row presents the maximum and minimum premia. The third row discloses the proportion of portfolios with positive premia, while the 4th line discloses sign-test significance.

returns on daily market premia would generate positive coefficient 1, revealing the extreme position adopted with respect to market-risk. If, on the other extreme, the portfolio buys the 3 least aggressive stocks while selling the 3 most aggressive, then the regression should reveal negative coefficient $\beta_{\text{MARKET}} = -1$, suggesting that the portfolio rides market-risk reversely, possibly in anticipation of further market decline. Since MARKET1 however is an extreme benchmark and the participants could expose to market-risk less intensely, we remove the 3 smallest and largest stocks from the sorted TA220 list and repeat the construction again with the truncated list of 114 stocks. MARKET2 is used to represent (the returns on) the portfolio that buys the 4–6 most aggressive stocks and sells the 4–6 most defensive stocks in TA220. The exercise is repeated 8 additional times to construct the portfolios MARKET3–MARKET10. The simple average of the daily returns on MARKET1–MARKET10 is used to approximate the returns that could be earned by exposing to market-risk in the restricted settings of the experiment. The term "MARKET" (with no adjunct indexing) represents the average. Since the sorting of available stocks by historical BETA is run at the specific DATE where each arbitrage was submitted, the sorted portfolios represent simple strategies that the arbitrageurs could employ in an attempt to exploit market-risk for profitable arbitrage. The β_{MARKET} coefficients summarize the loadings on this premia.

Similar methods were applied to calculate the SIZE, BTM and MOMENTUM (henceforth abbreviated to MOM) premia. To calculate the daily SIZE premia we sort TA220 by SIZE (market capitalization for the arbitrage DATE) in descending order and construct 10 disjoint portfolios SIZE1–SIZE10 that represent different degrees of exposure to SIZE risk. The average daily return on the 10 portfolios represents the daily premia on SIZE. The daily premia on BTM (MOM) were similarly calculated by sorting TA220 (in ascending order) by BTM (PREV_3MON) as defined in Table 4 and constructing 10 disjoint arbitrage portfolios BTM1–BTM10 (MOM1–MOM10) that represent different levels of exposure to the corresponding factor. The average daily return on the 10 portfolios is used again to represent the BTM (MOM) premia in the particular settings of each arbitrage. The use of exactly 10 portfolios to calculate each premia is clearly arbitrary, but similar conclusions emerge when the number of portfolios is increased or decreased, when momentum portfolios are determined from 180 days histories (instead of 90), when MARKET returns are directly approximated from TA220, and in other robustness analyses (see the Web supplement Appendix A).

5.2. Annual premia

Before running the estimations, we briefly examine the behavior of MARKET, SIZE, BTM and MOMENTUM along the experiment. The yearly premia were separately calculated for each arbitrage using the 10 portfolios approach introduced above. Since the calculations take into account the experimental constraints on stock selection (the restricted TA220 list and the maximal number of stocks on each side), the annual figures approximate the returns that arbitrageurs could collect by exposing to familiar return-generating factors. Table 5 summarizes the annual premia distributions; the data is presented in percentile form as basis points are inconvenient for large annual figures.[20]

The most consistent results are observed for SIZE. The mean annual SIZE premium was about 31% and the median was 22% ($p < 0.01$). The annual SIZE premium, moreover, was positive for all 117 portfolios, reflecting stronger performance of relatively small-size stocks along the 2009 recovery.[21] Post-hoc, SIZE risk appears as a consistent source of profitability throughout the experiment. Since the stocks purchased within the arbitrages however were twice larger than the stocks sold, the steeper recovery of smaller stocks could damage profitability. The hypothesis is tested directly in the estimations.

Essentially positive premia also emerge for BTM. The mean annual BTM premium was positive 18.2% ($p < 0.01$) and the median (21.2%) was close to the median SIZE. BTM premia ranged from a minimum of −14% to a maximum of 49% with about 87% of the arbitrages running in times of positive BTM premium. BTM, similarly to SIZE, emerges as a solid source of profitability along the experiment, but again we suspect that the crisis-prone strategies of the participants damaged performance.[22]

[20] The correlations in annual risk premia are disclosed in the Web supplement (see Appendix A), where we also analyze the correlations in daily risk premia.

[21] The stronger performance of smaller stocks also shows at the level of stock indices: TA100 decreased by 26.8% from 17MAY2008 to 17MAY2009 (the earliest arbitrage) while YETER-120 decreased by 20.0% at the same interval; TA100, on the other hand, increased by 70.4% from 19DEC2008 to 19DEC2009 (the latest arbitrage) while YETER increased by 124.9%.

[22] Alternatively, we calculated BTM using the ratios for 31DEC2007, independently of DATE. The annual premia became positive for all portfolios but the results stayed similar (see the Web supplement Appendix A).

Table 6
Estimation results.

	ALPHA	β_{MARKET}	β_{SIZE}	β_{BTM}	β_{MOM}	R^2
Mean	0.02**	0.03	**−0.12**[a]	−0.04	−0.002	0.27
Median	0.04	0.04	**−0.13**	−0.04	−0.007	0.21
(Standard deviation)	(0.16)	(0.53)	(**0.40**)	(0.39)	(0.43)	(0.21)
Proportion positive	59%	56%	40%	47%	48%	–
Sign-test significance	$p = 0.03$	$p = 0.10$	$p = 0.02$	N.S.	N.S	–
Coefficient positive and 10% significant	5%	40%	24%	19%	26%	–
Coefficient negative and 10% significant	1%	31%	49%	28%	26%	–

The table summarizes the time-series estimation of Carhart (1997) 4-factor model on the $N = 117$ portfolios. Alpha is denominated in basis points. The upmost line of the table presents the mean, median and standard deviation of the estimated coefficients. The second line presents the proportion of positive estimates, with sign-test significance at the third line. The 2 lines at the bottom present the proportion of statistically significant positive and negative estimates.

[a] Wilcoxon signed rank test significance.

The results for MARKET and MOMENTUM are essentially different. The average annual MARKET premium was positive 5.2%, but the median was negative −3.8% ($p = 0.18$). The mean MOMENTUM premium was 10.5%, with median close to zero ($p = 0.18$). The portfolio-level inspections reveal that annual MARKET was positive in 44% of the cases, while the proportion of positive annual MOM was about 50%. Arbitrage strategies that sort the available stocks by market-sensitivity (BETA) or recent historical returns (PREV_3MON) therefore showed inconsistent results along the turbulent period of the experiment.

5.3. Estimation results

The estimation results are summarized in Table 6 where we disclose the mean, median and standard deviation of the estimated coefficients and the proportion of significant estimates. Again, the most consistent results are observed for SIZE. The mean β_{SIZE} coefficient is −0.12 and the proportion of portfolios with $\beta_{SIZE} < 0$ is 60% (sign test; $p < 0.02$). β_{SIZE} is negative and significant for 49% of the portfolios while it is positive and significant for only 24%. When the β_{SIZE} coefficients are multiplied by corresponding annual SIZE premia, the average product is negative −4.2%, indicating that the selling of relatively smaller stocks damaged performance significantly. The average product is even lower (−6.2%; $p < 0.01$) for the $N = 85$ optimists that expected a market recovery for their arbitrage year.

The time-series results for the other factors are much less consistent. The mean β_{BTM} (−0.04; $p = 0.44$) is close to 0, with 47% of the portfolios showing positive $\beta_{BTM} > 0$. When the portfolio-specific β_{BTM} estimates are multiplied by matching annual BTM premia, the average product (0.14%) is small and statistically insignificant. Loadings on BTM ($\beta_{BTM} *$ annual BTM premium) still ranged from −50% to +49%, suggesting that BTM could contribute or cut performance considerably depending on specific market conditions and individual stock selection. Similar, mixed patterns of loadings with diverse effects on eventual payoffs — are observed for MARKET and MOMENTUM.[23] Across the sample, $\beta_{MARKET,j} *$ (annual MARKET premium$_j$) + $\beta_{SIZE,j} *$ (annual SIZE premium$_j$) + $\beta_{BTM,j} *$ (annual BTM premium$_j$) + $\beta_{MOM,j} *$ (annual MOMENTUM premium$_j$), however averages at −3.1% with median −2.5% ($p < 0.05$). The sum of loadings is negative for 61% of the portfolios, illustrating that the selling of riskier stocks significantly damaged profitability.

Most portfolios (69 of 117), however, reveal positive intercept ALPHA. The average ALPHA is 0.0194, representing daily return of 1.94 basis points (bps) on the arbitrage. The median ALPHA is twice larger, at 4.25 bps. When the mean alpha is accumulated for 264 days (median number of daily observations), the product, 5.25%, is close to the average gross return on the 117 portfolios. When the median alpha is annualized, the yearly return (11.8%) is twice larger, demonstrating again the negative effect of unbalanced risk-exposure on eventual performance. As in many empirical studies, ALPHA is statistically significant in less than 10% of the cases (6 cases of positive and 1 case of negative significant intercept).[24] The Pearson correlation between ALPHA and gross return, however, is positive 0.79, suggesting that stronger performance cannot be solely attributed to differences in arbitrage styles but also represents higher risk-adjusted return. The supplementary appendix (see Appendix A) illustrates the robustness of results; e.g., the mean (median) ALPHA in 3-factor estimations are 1.7 (4.9) bps while the mean R^2 decreases slightly from 0.27 to 0.25.

5.4. Cross-sample comparisons

Finally, we use the factor estimations to reexamine the results of Section 3; e.g., the positive correlation between prior confidence and returns. We should note at the outset that such analysis would show meaningful results only when the tested variables (CONF) consistently affect loading styles or ALPHA. It is possible, for instance, that the more confident participants administered diverse, generally successful, arbitrage strategies, riding distinct factors within different portfolios. If this is the case, the complementary analysis would be unproductive as it is impossible to generally characterize the success factors. Unfortunately, the additional analysis indeed produced

[23] The product $\beta_{MARKET} *$ annual MARKET premium averaged at 0.69% ($p = 0.31$) while $\beta_{MOM} *$annual MOMENTUM premium averaged at 0.24% ($p = 0.28$), but variability is large with $\beta_{MARKET} *$annual MARKET ranging from −72% to 78% and $\beta_{MOM} *$ annual MOMENTUM ranging from −126% to 128%.

[24] Fama and French (2010), for example, suggest that only 2.3% of US equity mutual funds deliver annual alpha > 2.5% (before fees).

only few insights. The next paragraphs briefly summarize the contributions.[25]

Consider self-confidence first. The analysis could not expose a consistent confidence effect on loading styles or ALPHA, except for stronger tendency of confident arbitrageurs to load on momentum. The Pearson coefficient of correlation between β_{MOM} and CONF is 0.24, with mean $\beta_{MOM} + 0.19$ for the portfolios with CONF $> 75\%$ compared to negative -0.15 for the participants with CONF $< 50\%$ ($p < 0.01$). The average product $\beta_{MOM} *$ annual MOM premia is 7.9% for the highly confident, compared to -5.4% for the least confident ($p < 0.05$). Momentum therefore emerges as a significant source of profitability for the most confident, but the premia collected by momentum-riding roughly covers only 30% of the gross return for this group (26.2%, Table 2).

The CONF effect on ALPHA turns out positive but too noisy for cross-sample significance (ρ(CONF, ALPHA) $= 0.09$). The average 4-factor adjusted return for the 51 portfolios with CONF $> 60\%$ is 3 bps ($p < 0.05$), compared to insignificant 1 bps for the portfolios with CONF $\leq 60\%$. The results still appear quite impressive for the participants in the most confident quartile: mean daily alpha 6.3 bps (more than 18% in annualized terms) with 72% ALPHA > 0 rate ($p < 0.01$).

Another major mediator of performance, by the preliminary analysis, is the time spent on the arbitrage screen. A median split of the sample by T(ARB) indeed reveals a positive significant ALPHA (3 bps; $p < 0.05$) for the 58 participants that spent least time delivering their arbitrage, compared to insignificant 1 bps ($p = 0.23$) for others, but the correlation ρ(T(ARB), ALPHA) $= -0.12$ is too weak for significance. The analysis of loading patterns could not expose consistent T(ARB) effects on arbitrage styles, except for stronger negative loading of the slow participants on SIZE. The median split, for instance, reveals negative $\beta_{SIZE} = -0.19$ ($p < 0.01$) for the participants with higher T(ARB), compared to insignificant $\beta_{SIZE} = -0.06$ for the relatively rapid. The mean ($\beta_{SIZE} *$ annual SIZE premium) for the slower participants was negative -5.4% ($p < 0.01$), confirming that the selling of smaller stocks damaged performance. Interestingly, the slower arbitrageurs were also significantly more optimistic regarding the market trends for their arbitrage year (e.g., proportion of pessimists that expect further decline of the market: 19% for the slower compared to 38% for the rapid), therefore exhibiting stronger inconsistency between expectations (positive) and stock selection (crisis-prone).

While the interaction between self-confidence (T(ARB)) and performance seems weaker when returns are measured in risk-adjusted terms, LOSS AVERSION still appears to boost performance significantly when loadings are accounted. The average daily ALPHA of the 73 relatively loss-averse participants was 2.3 bps ($p = 0.025$) compared to insignificant mean ALPHA of 0.5 bps ($p = 0.32$) for other respondents. Regressions on ALPHA suggest that risk-adjusted returns increased by 3.4 bps with each choice of the risk-free alternative over the 50–50 gain or loss lottery (see the Web supplement Appendix A). In the concluding discussion we briefly demonstrate that our crude 0–3 LOSS AVERSION measure shows interesting interaction with the stock-picking patterns of participants in both annual and 3-monthly arbitrages.

6. Discussion: annual vs. 3-months arbitrage

In parallel to the annual arbitrage, the participants selected short-run arbitrage portfolios for 3 months. The quarterly portfolios were delivered on a separate page using the same table format depicted in Fig. 2, and only 6 participants submitted the same portfolio for both horizons.[26] The 3-months and annual arbitrages were run under drastically different market conditions (Fig. 1). The returns on MARKET, SIZE and BTM were mostly negative at the first 3 months of the arbitrage years, representing the stronger collapse of riskier stocks at the peak of the crisis. The annual premia, however, were positive or mixed, representing the steep recovery in subsequent months. Similarly to the annual case, the bottom-line results for the short-run arbitrage were positive but statistically marginal. The quarterly return averaged at 2.7% with median 1.8% ($p = 0.08$) and the profitability rate was 55%. The correlation between 3-months and annual performance was close to zero ($\rho = -0.04$), suggesting that short-run profitability could not predict yearly success. Since the subset of eligible portfolios and market conditions are different, we separately analyze the quarterly arbitrages in Sonsino and Shavit (forthcoming).[27] The next paragraphs summarize the experiment, outlining the major similarities and discrepancies in annual and quarterly results.

Self-confidence emerges as the strongest predictor of performance in both shorter and longer arbitrages. Confidence levels were separately elicited for each task at the last screen of the program. The two scores naturally exhibit high correlation ($\rho = 0.8$), but it is interesting to observe that confidence sorts the best performers in both cases although profits show close to zero correlation. The participants at the top confidence quartile (CONF $> 75\%$) earned 10% on the 3-months arbitrages, while earning 26% on the annual portfolios. The overlap between the 2 groups is smaller than 50% (only 11 subjects showing CONF $> 75\%$ for both horizons), which illuminates again the strength of the result. CONF showed significant effect on risk-adjusted ALPHA in the shorter arbitrage, but the superior performance of the confident participants in the annual task also followed from successful loading on momentum. The profitability rates of the most confident participants (61%, 72%)

[25] To identify the variables that affected loading styles and alpha cross-section, we have run regressions with model selection on the estimated coefficients. Since the analysis revealed only few consistent effects, we discuss the conclusions directly skipping technical details.

[26] The average overlap between the 3-months and annual portfolios was 32.7% long vs. 26.2% short ($N = 117$; $p < 0.05$).

[27] The sample for the monthly task consists of 130 portfolios compared to 117 annual portfolios. The differences arise because of the removal of stocks that stopped trading along the yearly arbitrage. The joint sample consists of 113 observations. The correlation remains insignificant when returns are normalized with respect to Δ(TA220). The correlation in ALPHA's is negative -0.17 ($p = 0.08$; $N = 113$).

were significantly lower than their average confidence levels (84%, 86%) in both cases, in line with the wide literature on investors' overconfidence. While preceding research proposes that overconfidence drives investors to excessive disadvantaged trading, the current results demonstrate that when the trading-task is controlled, prior-confidence may strongly correlate with eventual performance even when traders strongly over-estimate their abilities. The subjective confidence scores that the arbitrageurs delivered did not correlate with skills related measures, login times or expectations. The competence related measures were constantly removed in model selections when CONF was accounted. Anecdotally, the results propose that consulting the most confident experts may prove profitable, at least with young, motivated consultants of the type employed in the current design.

Another common result for the shorter and longer arbitrages is negative partial correlation between the time spent on selected screens along the program and eventual payoffs. In the 3-months task, the effect is captured by T(INTRO), the time spent on the introductory pages where the concept of convergence arbitrage was illustrated (pages 2–3 in the script). In the longer arbitrage, the relevant measure is T(ARB), the time spent on the annual arbitrage table. The coefficient of correlation between T(ARB) and T(INTRO) was close to zero ($\rho = -0.01$), but both variables negatively correlated with skills related measures such as years of education, finance knowledge and industry experience (e.g., correlations with EXPERIENCE, $\rho = -0.16$ for T(INTRO) and $\rho = -0.09$ for T(ARB)). Again, the direct competence measures were removed in model selections, suggesting that login times capture the low skills effect on performance most effectively. The time-series analysis of the quarterly portfolios, in addition, revealed that the arbitrageurs with low T(INTRO) benefited from selling relatively riskier stocks (especially, high BTM stocks) that decreased more rapidly at the peak of the crisis. The parallel analysis of annual arbitrages, on the contrary, showed that the participants that spent more time on the arbitrage screen were hurt from holding short positions in small SIZE stocks while markets recuperated. In both cases, the quicker participants outperform others in terms of tuning their portfolio to subsequent market trends. It is worth noting that login times capture such differences in arbitrage styles, where direct skills-related controls fail.

The third variable showing significant interaction with arbitrage performance, in both cases, is our ad-hoc 0–3 measure of LOSS AVERSION. Interestingly, the sign of correlations reverse between tasks. Profitability rates decreased with loss aversion in the quarterly arbitrage ($R > 0$ rate 49% for the relatively loss-averse compared to 63% for the less loss-averse; $p = 0.04$), while eventual returns increase with loss aversion in the yearly condition (Section 3). The discrepancy is resolved in closer examination of arbitrage styles. The loss-averse participants delivered balanced portfolios in the 3-months task (mean volume-weighted BETA 1.09 long vs. 1.05 short; $p = 0.3$), switching to relatively aggressive stock selection for the annual arbitrage (mean weighted BETA 1.15 long vs. 0.97 short; $p < 0.05$). The less loss-averse, on the contrary, sold more aggressive stocks in the 3-monthly arbitrage (weighted BETA 1.09 long vs. 1.38 short; $p < 0.05$), but selected relatively balanced portfolios in the annual task (weighted BETA 1.08 long vs. 1.09 short; $p = 0.3$). The loss-accommodating types therefore benefited from holding short positions in aggressive stocks at the peak of the crisis, while the loss-averse gained from selective investment in relatively aggressive stocks in times of market recovery. The observation that an experimental (lottery-based) LOSS AVERSION measure captures subtle differences in arbitrage styles seems intriguing, especially in light of recent debates regarding the domain specificity of risk-preferences (e.g. Dohmen et al., 2011; Coppola, 2014).

Beyond such individual differences, the experimental arbitrageurs sell riskier stocks in terms of SIZE, BTM and MLP-related ratios in both 3-months and annual arbitrages. The selling of riskier stocks could be rationalized for the quarterly arbitrages, where 41% of the participants expected additional 3-months decline of TA100 and 26% expected modest (possibly selective) recovery of up to 5% in the index. The time-series estimations for the 3-months portfolios indeed confirmed that the arbitrageurs benefited from short-selling riskier stocks at the peak of the crisis. The mean β_{BTM} coefficient for the quarterly portfolios, for instance, was negative -0.11, while the quarterly premia on BTM-risk averaged at -8%. Portfolio-level calculations more accurately revealed that the selling of higher BTM stocks contributed, on average, about 1.1% to the quarterly returns. The selling of riskier stocks, however, decreased profitability in the annual assignment (Section 5). The crisis-prone annual positions, moreover, evidently contradict participants' expectations for market recovery along the arbitrage year. We attribute the persistent inclination to sell riskier stocks to misperception of financial risk (Shefrin and Statman, 1999; see also Shefrin, 1999, 2001; Ganzach, 2000). Affected by the ongoing crisis (Slovic et al., 2004), the participants choose to select safer stocks for investment while short-selling relatively riskier companies, although prices approach 5-years low records and recovery is expected for the coming year.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jbef.2014.01.004.

References

Amihud, Y., Goyenko, R., 2013. Mutual fund's R2 as predictor of performance. Rev. Financ. Stud. 26 (3), 667–694.
Beunza, D., Stark, D., 2004. Tools of the trade: the socio-technology of arbitrage in a Wall Street trading room. Ind. Corp. Change 13 (2), 369–400.
Carhart, M.M., 1997. On persistence in mutual-funds performance. J. Finance 52 (1), 57–82.
Clark, J., Friesen, L., 2009. Overconfidence in forecasts of own performance: an experimental study. Econ. J. 119 (534), 229–251.
Clark, E., Kassimatis, K., 2012. An empirical analysis of marginal conditional stochastic dominance. J. Bank. Finance 36 (4), 1144–1151.
Coppola, M., 2014. Eliciting risk-preferences in socio-economic surveys: how do different measures perform? J. Socio-Econ. 48, 1–10.
Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., Wagner, G.G., 2011. Individual risk attitudes: measurement, determinants, and behavioral consequences. J. Eur. Econom. Assoc. 9 (3), 522–550.

Doukas, J.A., Chansog, K., Christos, P., 2010. Arbitrage risk and stock mispricing. J. Financ. Quant. Anal. 45 (4), 907–934.
Fama, E.F., 1970. Efficient capital markets: a review of theory and empirical work. J. Finance 25 (2), 383–417.
Fama, E.F., French, K.R., 1992. The cross-section of expected stock returns. J. Finance 47 (2), 427–465.
Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financ. Econ. 33 (1), 3–56.
Fama, E.F., French, K.R., 2010. Luck vs. skill in the cross section of mutual fund returns. J. Finance 65 (5), 1915–1947.
Ganzach, Y., 2000. Judging risk and return of financial assets. Organ. Behav. Hum. Decis. Process. 83, 353–370.
Gatev, E., Goetzmman, W.M., Rouwenhorst, K.G., 2006. Pairs trading: performance of relative value arbitrage rule. Rev. Financ. Stud. 19 (3), 797–827.
Klayman, J., Soll, J.B., González-Vallejo, C., Barlas, S., 1999. Overconfidence: it depends on how, what, and whom you ask. Organ. Behav. Hum. Decis. Process. 79, 216–247.
Kuhnen, C.M., Knutson, B., 2011. The influence of affect on beliefs, preferences and financial decisions. J. Financ. Quant. Anal. 46 (3), 605–626.
Lichtenstein, S., Fischhoff, B., Phillips, L.D., 1982. Calibration of probabilities: the state of the art to 1980. In: Kahneman, D., Slovic, P., Tversky, A. (Eds.), Judgment Under Uncertainty: Heuristics and Biases. Cambridge Univ. Press, Cambridge, UK.
Newey, W.K., West, K.D., 1987. A simple positive definite heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.
Pole, A., 2007. Statistical Arbitrage. Algorithmic Trading Insights and Techniques. In: Wiley Finance Series.
Schleifer, Andrei., 2003. Inefficient Markets: An Introduction to Behavioral Finance. Oxford University Press.
Shefrin, H., 1999. Beyond Greed and Fear: Understanding Behavioral Finance and the Psychology of Investing. Harvard Business School Press, Boston.
Shefrin, H., 2001. Do investors expect higher returns from safer stocks than from riskier stocks? J. Psychol. Financ. Mark. 2 (4), 176–181.
Shefrin, H., Statman, M., 1999. Comparing expectations about stock returns to realized returns. Working paper, Santa Clara University.
Skala, D., 2008. Overconfidence in psychology and finance-an interdisciplinary literature review. Bank i Kredyt (4), 33–50.
Slovic, P., Finucane, M.L., Peters, E., MacGregor, D.G., 2004. Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. Risk Anal. 24 (2), 311–322.
Sonsino, Doron, Regev, Eran, 2013. Informational overconfidence in return prediction-more properties. J. Econ. Psychol. 39, 72–84.
Sonsino, D., Shavit, T., 2014. Short-run arbitrage in crisis markets—a field experiment. Ann. Financ. Econ. (forthcoming).
Subrahmanyam, A., 2008. Behavioral finance: a review and synthesis. Eur. Financ. Manage. 14 (1), 12–29.
Törngren, G., Montgomery, H., 2004. Worse than chance? Performance and confidence among professionals and laypeople in the stock market. J. Behav. Finance 5 (3), 148–153.
Von Lilienfeld-Toal, Ruenzi Stefan, 2014. CEO ownership, stock market performance, and managerial discretion. J. Finance (forthcoming).
Whistler, M., 2004. Trading Pairs: Capturing Profits and Hedging Risk with Statistical Arbitrage Strategies. John Wiley and Sons.