ThE Papers 06/03

Departamento de Teoría e Historia Económica

Universidad de Granada

Subject Pool Bias in Economics Experiments

Pablo Guillén, Harvard Business School

Robert F.Veszteg, Universidad de Navarra

Corresponding author: Pablo Guillen, Baker Library B90, Harvard Business School, Soldiers Field, Boston MA 02163, pguillen@hbs.edu, Phone: +1 617 495 6753, Fax: +1 617 495 5287
.

Subject Pool Bias in Economic Experiments*

Pablo Guillén and Róbert F. Veszteg†

27th February 2006

Abstract

In this paper we consider data from a large number of economic experiments, and look for demographic effects that may be a source of subject pool bias if not carefully accounted for in the subsequent statistical analysis. Our dataset contains information on 2,408 subjects and 597 experimental sessions from 74 experiments recorded over more than 2 years at an experimental laboratory. Using different estimation methods and model specifications, we identify the significant demographic determinants of personal earnings, and find that they account for less than 4% of the observed variation. Thus we deliver empirical evidence supporting the experimental method as monetary incentives, and therefore some kind of strategic behavior, seem to be more important than demographics in the laboratory. Exploiting the time-series nature of the data we also study some dynamic issues of the subject pool: we analyze the factors that influence subjects' decisions on returning to the laboratory.

Keywords: Experiments, Subject pool bias, Demographic characteristics

JEL Classification Numbers: C9

*We thank Antonia Atanassova, Al Roth, Alexander Gelber, Carmit Segal and Sarah Wittman for comments and precious help.

†Guillén: Harvard Business School, pguillen@hbs.edu; Veszteg: Universidad de Navarra, rveszteg@unav.es.

1 Introduction

A frequent criticism of experimental economics is that the behavior investigated in experiments is specific to the particular group of subjects, such as college students, who frequently participate in the laboratory. Is this a serious problem for the experimental method, and can it be rectified by including a wider range of subjects in experiments and/or controlling for a possible demographic bias in the statistical analysis? In this paper we consider data from a large number of economic experiments and look for demographic effects that may be a source of subject pool bias if not carefully accounted for in the subsequent data analysis.

The experimental method, though present from the early '30s in economics, has enjoyed a steady increase and wider acceptance in science over the last decades. Nevertheless it often receives harsh criticism that attacks its fundamentals. The essence of experiments is control as the experimenter is able to keep track of the environmnet, and also of the exogenous and endogenous variables except for a few unobservable ones. Falk and Fehr (2003) discuss briefly the most important lines of criticism including those that focus on low stakes, small number of participants, and unrealistic environments. They list references, and describe the most important results from experimental research that have been gathered to respond to critics. In this paper we wish to contribute to this discussion by studying the possible existence of subject pool bias. Our approach is novel in the sense that it is based on historical data rather than data recorded from a new and specifically design experiment.

The historical feature of the data implies that we lack precise information on the environment implemented in the laboratory. We focus on the monetary amount that subjects earn in the experiment, and try to establish connection between it and subjects' personal characteristics. This type of investigation usually does not appear in research outputs, except for those papers that report results from experiments designed precisely for measuring the gender, university, etc. effects. Roth et al. (1991) for example report results from bargaining and market behavior experiments run in four culturally very different countries: Israel, Japan, the United States of America, and Yugoslavia. They find important deviations from predicted behavior, but no payoff-relevant differences among countries.

The use of the money earned in the experiment as a proxy for the subject's performance relies on the fact that experiments in economics are based on monetary incentives. Participants usually earn a fixed amount (show-up fee) plus some additional, and in general more voluminous, amount that is determined by their actions and behavior in the experiment. After the critiques presented by Allen Wallis and Milton Friedman in 1942 pointing out that Thurstones's early experimental session in 1931 involved ill specified and hypothetical choices, experimenters started using real incentives, in the vast majority of cases money.[1]

We investigate whether certain groups of subjects systematically earn higher payoffs than others, using administrative data on 2,408 different subjects gathered at an experimental laboratory at a university in the Northeastern United States over the course of approximately two and a half years. We focus on payoffs since this is the easiest way of comparing behavior across large numbers of experiments, and therefore of deciding whether experiments (and the subsequent data analysis) in general should be structured to include a wider range of subjects (and control variables).

We find that while subjects' age and gender seem to have a significant effect on the final payoffs, experience and education do not. The divergence of payments between two groups is a necessary but not sufficient condition for the divergence in behavior. That is, if the payoffs of the two groups are signficantly different, we can conclude that their behavior must also have been different, but not the converse. Small differences in average payoffs, in fact, could correspond to large differences in behavior. Our finding that payoffs differ across different groups therefore provides evidence that behavior is significantly different across these groups. Since we lack precise information on the environment implemented in the laboratory, our results aggregate over large numbers of different types of experiments. This provides a possible caveat against generalizing the results of an experiment to too wide range of other possible subject pools.

Nevertheless, subjects' demographic characteristics account for less than 4% of the

[1] For more on the history of experiments in economics check Roth (1995).

observed variation in payoffs. We are aware of experimental results about important differences between the behavior of men and women.[2] The point we would like to make here is that once we consider a large variety of problems and situations, even if gender seems to mark a significant difference in subjects' performance, its absolute impact is negligible when compared to other variables. Experiment and experimental session dummies explain a large fraction (approximately 40%) of the observed variation. This gives us some confidence that subject pool effects are not very important in explaining payoffs. Nonetheless, it is worth noting that these small differences in average payoffs could correspond to larger differences in average behavior in any given experimental setup. We are also aware of the fact that a large fraction of the variance in payoffs may be explained by subject charactertics that we did not include explicitly in our regressions.[3]

In a very similar way to ours, Carbone (2005) attempts to establish a link between strategic behavior and demographic characteristics. Her analysis uses a unique experimental subject pool that participated in a life-cycle consumption experiment. She finds that demographics have no effect on observed behavior, which she interprets as a constructive message to economic theory, which often ignores the effect of demographics on behavior. Our analysis is very different, as we look at historical data on a very large number of experimental sessions, none of which was designed to investigate the link between demographics and behavior, rather than relying on data from one experimental session specifically designed for probing this link.

Some recent work has been done on studying how representative experimental findings are. These papers, however similar to ours in their philosophy, concentrate on different issues, namely external validity.[4] Harrison and List (2004) presents an extensive survey on the criticism that points at the impact of students being the most important group for building the subject pool for laboratory experiments. The paper proposes a taxonomy

[2]Gneezy et al. (2003), for example, found that women may be less effective than men in competitive environments.

[3]Nevertheless, in section 3 we do present results from regressions that control for personal fixed effects without explicitly pointing out which are the most important, i.e. influential, personal characteristics.

[4]For clarification purpuses, we include here the definition for external and internal validity from Brewer (2000). An experiment is said to possess *external validity* if the experiment's results hold across different experimental settings, procedures and participants. While an experiment is said to possess *internal validity* if it properly demonstrates a causal relation between two variables.

for field experiments and suggests that experiments be run in both ways. In this line, Benz and Meier (2006) are the first to directly compare how people (actually the same subjects) behave in the lab and in the field. They use a simple donation example and find that even if there is a significant positive correlation between the lab and field behavior, it is very small and therefore it seems to be difficult predict real-life behavior based on experimental data. List and Levitt (2005) study the question of what experimenters can learn about the real world using laboratory experiments. They develop a simple model and argue that being watched in the lab may distort subject's behavior in various ways. As for self-selection into the subject pool, List and Levitt (2005) believe that people "who have social preferences or readily cooperate with the experimenter and seek social approval" might more frequently volunteer for experimental studies. While recognizing the importance of this issue, we do not consider it in our analysis, because our data comes from the experimental lab and does not allow for comparisons with external, *out-of-the-lab* subject groups. On the other hand, we do wish to contribute to the point raised by List and Levitt (2005) on experiments that measure group differences. Since the observations in our analysis proceed from a large number of potentially very different experimental sessions, the bias introduced by subjects who tend to *please* the experimenter can be neglected. We look at statistical differences in the behavior of subjects from different groups according to their gender, race, and/or education; nevertheless the vast majority of the experimental sessions that produced the observations in our data was not especially designed to study such differences. It must be pointed out that our results do not directly support the external validity of laboratory experiments in economics. However, they do support their internal validity. In particular we can claim that experimental results seem to be only very slightly affected by spurious demographic variables.

This paper is organized as follows: Section 2 describes our dataset. Section 3 presents the details of the statistical analysis that was performed on it. Section 4 concludes. Tables are in the appendix.

2 Data

Our data set consists of 8,755 observations corresponding to 2,408 subjects who participated in 597 sessions from 74 different experiments. All data come from the same laboratory located at a university in the Northeastern United States. We use entries recorded after April, 2003, because that is the month when the laboratory started gathering participants' personal data on a regular basis. The latest observations that we include in this analysis were gathered in January, 2006. The data available includes various self-reported personal characteristics of the subjects, including their gender, age, and the university (if any) they are affiliated with, along with their payoff in the experiment.

In some experiments, the subject's payoff does not depend, or depends very little, on the subject's behavior. Since our objective is to check how important personal characteristics are in determining the final payment in the experiment, we have deleted data from those sessions with fixed payoffs or in which payments do not vary much. In particular, we decided to exclude all the sessions in which 80% or more of the participants receive the same amount of money.[5] We have not discovered important qualitative changes in the results when performing the same analysis using cutoffs of 50%, 90% and 100%, instead of 80%. We have also deleted records with zero registered payoff, and repeated entries, keeping the one with the highest payoff. These two categories are a result of faulty data entries: no subject actually received zero payoff, and no subject was paid more than once for the same participation. We found a total of 176 zero entries and 277 repeated entries. This leaves 8,755 observations in our dataset.[6]

[5]This selection is in line with the philosophy of regression analysis we perform on the data, since ignoring fixed and quasi fixed payments guarantees larger variance in our dependent variable. As a result, we effectively exclude most tournament experiments, and also those sessions with fixed payments in which a few of the subjects earned more money, due to the "early show-up fee" that rewards people who arrive at experiments early with an extra payment. Moreover, we present our statistical analysis considering both raw and standardized payoffs. The standardization process subtracts the (per session) average payment from individual payoffs, hence eliminates the common part, such as the fixed show-up fee.

[6]The laboratory has recently started collecting information on subjects' ethnic group. As there are only 4,559 (52.97%) entries that contain a value for this variable, in order to not reduce the number of observations in the analysis we have decided not to include this variable in the final data set. If we compare the mean payoffs across the nine ethnic groups using analysis of variance (ANOVA) tables we cannot reject the null hypothesis of them being equal. The p-value in this case is of 0.1398, and of 0.4940 if we compare standardized payoffs across sessions.

Apart from the recorded personal data such as age, gender, racial group, educational level (with intended major and the name of the college), and the basic characteristics of the experimental session (final payoff, type of experiment), we also created several variables for our empirical analysis. Of the 2,408 subjects in the data, 70% came more than once to the laboratory over the course of the period we investigate. In order to control for experience, i.e. training in experiments, we constructed three variables. EXPERIENCE_TOTAL counts the total number of occasions the subject appears in the database prior to the experimental session in question. EXPERIENCE_INCENTIVES counts only experiments that make the *80% cutoff* described earlier. The third experience variable, EXPERIENCE_ANY, is dichotomous and takes the value 1 if the subject appears in the database prior to the given record, and 0 otherwise. It turns out that subjects who participate in experiments have a long experience record: At the time of our analysis they have participated in more than 6 sessions on average (in almost 5 if we solely consider incentives based experiments). This number is in line with the usual concern on the validity of experimental results as explained in List and Levitt (2005) who discuss the problem of subjects' self-selection into experiments. Section 3.3 offers a detailed statistical analysis of return decisions.

Since we shall be comparing earning from different experiments conducted by different researchers and therefore different incentives schemes, we generated the standardized series of payments, ST_EARNING_AMT. We corrected by the mean and the variance estimated for the same session.[7]

The variables RETURN30/60/90/365 take value 1 if the subject returns to appear in an experiment in the subsequent 30, 60, 90 and 365 days, respectively.[8]

Tables 1, 2 and 3 display descriptive statistics on all the variables included in our analysis. The picture that these tables reflect are typical for laboratory experiments in economics. The vast majority (74.04%) of the subjects are college students, and almost 82% have experience in experiments by the time of participation. The largest share comes

[7]The formula that we applied for this correction is the following one: $ST\_EARNING\_AMT = \frac{EARNING\_AMT_{is} - mean(EARNING\_AMT)_s}{st.dev.(EARNING\_AMT)_s}$, where subindex i makes reference to subjects, while s to the experimental session.

[8]In order to eliminate the survivorship bias from our analysis we have cut the dataset eliminating observations from the last 30, 60, 90 and 365 days.

from the area of social sciences, though other specialities as humanities or natural sciences are also well represented. Participants earn roughly \$24 on average. The distribution of payoffs is positively skewed as income distributions tend to be. It shows a large variance, since its standard deviation is approximately \$9. The subject pool of the experimental laboratory that we study seems to be well-trained, as subjects have experience from more than 6 experiments on average (in almost 5 if we consider only incentive-based session).

3 Subject pool bias

This section present the main results of the statistical analysis that we performed on the database. The first subsection gives a preliminary insight into the structure of the data. It deals with parametric and non-parametric pairwise comparisons across different subject groups and complements the descriptives statistics reported in the previous section. The second subsection contains the main body of our numerical results. With the help of regression analysis it offers empirical evidence suggesting that personal characteristics, that are usually excluded from the analysis of experimental data, although may have significant influence on subjets' performance, they altogether account for less than 4% of the observed variation in monetary payoffs. Before proceeding to the results a comment is in order: in spite of its historical nature we treat our observations as panel data and use statistical tools accordingly. Nevertheless, the variables that proxy subjects' experience are able to capture some of the dynamics features in the determination of monetary payoffs.

3.1 Pairwise comparisons

We grouped our data according the categorical variables and performed both parametric and non-parametric comparisons between groups.[9] Tables 4 through 8 contain the detailed numerical results. Our tables report results on comparisons between groups taking into account payoffs, EARNING_AMT, and also standardized payoffs, ST_EARNING_AMT.

[9]Group means are compared using t-tests that handle possible differences in group variances. Distributions are compared by the non-parametric Kruskal-Wallis test.

Numbers above the diagonal refer to real money payoffs, while numbers below the diagonal refer to standardized ones. All differences are computed by subtracting the group mean in column from the group mean in row. The differences reported in the text are significant at least at 5% significance level if not stated otherwise.

It turns out that men tend to earn significantly more money than women, as male participants gained 44 cents more on average than females. This difference and also its significance resist standardization, i.e. correction for session mean and variance does not change its sign. The non-parametric test also suggests that the two groups according to gender are different at any usual significance level.[10]

As for subjects' education we treat those seven colleges that most frequently appear in the database separately, and group the others together under the name of "other". Little more than 70% of the observations belong to these seven. Subjects self-report their college status. Therefore, we can presume but not affirm that subjects who do not report any college are not students.

When comparing the average payment between the groups of university students and non students we observe that students earned 86 cents more in average terms, however the difference in standardized payoffs is merely of 0.0284 and loses significance with respect to the previously reported one. The corresponding p-value is 22,96%. We cannot discover any clear pattern when comparing the above mentioned seven colleges separately. Without controlling for other variables two effects seem to be persistent in the data both in money payments and standardized ones: subjects who have not reported college, i.e. non-students, earn significantly less, and interestingly MIT students seem to outperform everybody else.[11]

One may think that differences in the type of education are more important in determining payoffs than the name of the college itself. We compared the payoffs accordingly taking into account subjects' intended major. Table 5 reports the results. There is some evidence showing that students from social sciences outperform their fellows from humanities in monetary payoffs. The former group earned 63 cents more on average (6 standardized cents) than the latter. However we could not find any other meaningful and

[10] We refer to any significance level that is higher or equal than 1%.

[11] MIT is the abbreviation for Massachusetts Institute of Technology.

statistically significant difference.

We also compared payoffs across groups with different educational degrees. Our conclusion is that there are no significant differences especially if we consider standardized payoffs. However, those with a doctoral degree tend to earn less in absolute monetary terms.[12] This result is especially surprising if we note that this significant difference is fairly large, it amounts to \$1.4-2 on average. On the other hand we should be careful with the interpretation, since we have only 122 subjects with a doctoral degree in our data base. Therefore, it should be underlined that our sample is not representative in this sense, and this result can not be generalized.

Experiments on decision making usually involve hypothetical situations in which subjects are meant to solve some underlying numeric problems. Intuition suggests that the rules specific to incentive-based experiments in economics may give advantage to those participants that have participated before in any similar game. Therefore experience may exert important influence on payoffs. Once again, pairwise comparisons offer empirical support to this opinion, as those who have participated in any experiment before tend to earn significantly more than those who take part in their first experiment. The difference is of almost 75 cents in average terms, and of 0.0502 when considering standardized payoffs. Also based on a non-parametric test we can conclude that the two groups are significantly different with regard to their earnings.

Finally we wish to comment on the payoff comparison according to racial groups. As outlined in the previous section, due to technical reasons and the lack of sufficient data, we do not include categorical variables encoding race in our analysis. Pairwise comparisons shed some light on significant differences. It turns out that black participants earned roughly \$1.2 less than white or hispanic. Standardization eliminates these differences from the data, therefore we believe that ignoring ethnic groups does not introduce bias in our results.

[12] This difference, just like all the significant differences in absolute terms, disappears if we consider standardized payoffs.

3.2 Regression analysis: payoffs

Our main objective is to examine whether personal characteristics have any explanatory power in the determination of experimental earnings. And also whether they introduce any bias in the subject pool that is usually not controlled for in experimental studies. We also would like to separate the effects that these variables might have, therefore we proceed to regression analysis. This subsection deals with the determination of payoffs, while the following one concentrates on return decisions.

The categorical variables enter in form of dichotomous variables in the analysis. GENDER is the dummy for gender (1 represents male), while the variables U1 through U8 encode colleges.[13]

Tables 9 through 12 show the estimation results for a linear regression model that explains the observed variation in real money earnings. We use the standard ordinary least squares method (OLS) and also median regression (MEDIAN) as a robustness test.[14] The columns prefixed by SW report output from stepwise estimation that select the most influential regressors, or in other words the variables with the largest explanatory power. This procedure enters all the potential and available explanatory variables in the estimation and proceeds to the stepwise elimination of those that do not prove to be significant.[15] In order to account for unobservable subject, session or experiment related effects we studied several specifications of our regression model including so-called fixed effects. The tables in the appendix report these results in columns with the title ROBUST, since we used the OLS estimation method with robust variance estimates. We check for an age effect using both age (AGE) and it squared value (AGE2) as regressors as this is the usual model specification in the literature.

According to the information in tables 9 through 12 age has a negative effect on

[13]The control group is formed by non-student participants. The value 1 belongs to the following institutions. U1: other college; U2: University of Massachusetts; U3: Tufts University; U4: Northeastern University; U5: Boston College; U6: Boston University; U7: Harvard University; U8: MIT.

[14]The quantile regression methods, and in particular the median regression, take into account some flaws of the data itself and are more robust to outliers. The latter offers an estimation method of the conditional median function. Similarly to the OLS technique, the estimates of the median regression result from an optimization process: they minimize the sum of absolute deviations. As an advantage to the OLS method this method does not require distributional assumptions on the error term.

[15]We used 15% significance level in the elimination process.

payoffs. The coefficient estimates for these two variables give a negative net effect in the range form 18 to 73, except for the first four ROBUST estimation results in tables 11 and 12.[16] In those cases the net effect in question turns positive at age 42, 50, 47 and 52, respectively. Nevertheless, these estimated joint effects of AGE and AGE2 are not significantly different from zero, and we also should note that 95.99% of the subjects that appear in our database are aged 42 or less.

Experience does not seem to have a significant effect on earnings, nor individually when taking into account the three variables designed for capturing this effect, nor jointly. However the dummy variable EXPERIENCE_ANY that measures whether the subject has any experience in experiments or not is *close to being significant*. We shall return to this point when estimating the model with standardized payoffs.

In tables 9 and 10, from the list of colleges Boston College and MIT excel by their dummies having significant positive effect on earnings. Student coming from the former tend to earn as much as approximately $2.5-4 more on average then the others, while students from the latter make approximately $1-2 more. However, the variables associated to education fail to prove to have significant effect on earning if we use a joint test.

According to our regression analysis gender is a significant characteristic, since men earn on average roughly 50 cents more (per session) than women. Overall, in spite of the detected explanatory power of some of the personal characteristics, these account for less than 1% of the observed variation in real money payoffs.

If we reestimate our model allowing for subject fixed effects, the model performs slightly better as the adjusted R^2 increases to 4% once adjusted for the increased number of variables.[17] Age continually shows a significant negative sign, while the three variables attached to experience together are not significant. The changes in the absolute values of the coefficient estimates are difficult to address due to the presence of idiosyncratic terms in the model.

Results are more interesting if we enter the *experimental session* as control to the model. By allowing for experimental session related fixed effects the explanatory power

[16] These are the minimum and maximum ages for subjects that appear in the database.

[17] Dummies related to gender and education are dropped, because their effects are now captured by the fixed terms in the regression.

of the model increases to more than 40%. We wish to present this result as an important numeric rationale, apart from the intuitive verbal one presented before, for the standardization of monetary payoffs across sessions. By subtracting the mean payment and correcting by the variance in each session we clean the data from the session, i.e. experimental design, related effects. In spite of the guidelines of the experimental laboratory, researchers may differ according to the rules they apply regarding show-up fees, early show-up fees, and conversion rules that translate experimental monetary units into real money in their sessions. The regression models that control for session fixed effects suggest that apart from being an MIT student no other personal (demographic) characteristic has significant influence on earnings. However, if we look at the other specifications there seem to exist gender and age effects. Moreover the parameter estimates for these regressors do not suffer considerable changes in sign or absolute value from one model and/or estimation method to the other. Note that we also report results from regressions in which the experiment, i.e. a particular study, enters as a control variable. It is important to take into account that an experiment usually comprises several sessions which may correspond to different treatments or variations. In particular, AGE and GENDER are significant if we allow for experiment fixed effects in the analysis. Their importance is also confirmed by the analysis of standardized payoffs.

Tables 11 and 12 present results for the models that include the standardized payoff as the dependent variable. We observe that with this change fixed effects, related to subjects or experimental sessions, lose importance. The adjusted R^2 statistics takes the value of approximately 1%. This is an upper limit also for the other model specifications estimated with ST_EARNING_AMT. This confirms that personal characteristics have little importance on determination of payoffs in experiments. The primary forces in that are others, possibly the studied strategic behavior of subjects that is usually reported in research papers.

Age and gender keep their significance and do not alter their sign across different specifications and estimation methods, and also their absolute value appears to be robust to these changes. Younger participants tend to earn more and so do males compared to females. The joint effect of the experience variables improves on its significance, the corresponding p-value is equal to 0.164 that is comparable to the 0.871 from the model with EARNING_AMT as dependent variable. The eight dummies associated with education do not altogether play a significant role in determining standardized payoffs. Nevertheless, if the stepwise estimation method is implemented apart from age and gender there are two other explanatory variables that survive the elimination. MIT students excel from the subject pool by earning significantly more than other student or non-student participants. The regressor EXPERIENCE_ANY has a positive coefficient that confirms the intuitive fact that experience has a positive effect on payoffs. In other words, having participated in an experiment before is important and positively influences the money earned. However experience is not cumulative, as participation in more than one experiment in the past does not have additional effects.

Since experimenters are asked to fill in a questionnaire about the study they wish to perform in the laboratory, we have information on the type of the game that was played by participants. Experiments however often fall in more than one category. For this reason, instead of introducing new dummy variables into the analysis, we re-estimated the models with data from different types of experiments, and looked for variations in the results. The only remarkable novelty from this approach is the impact of experience on the monetary payoff. Especially because it turns out to be significant in the following types of games: public goods, coordination, decision-making under risk, organizational behavior and studies on altruism, fairness and reciprocity.[18]

Before closing the section two comments are in order. On one hand, as previously mentioned in the introduction, the laboratory has recently started to collect data on subjects' ethnic groups. We decided not to include this variable in our analysis, because it would have reduced the number of observations dramatically. *Only* 4,559 out of 8,755 entries contain a value for the ethnic group. Howeveer, we replicated the regression analysis described in this section with dummies controlling for the ethnic groups, and found that the new specification did not help to raise the value of the R^2 statistics over 1%. Moreover, we can not give clear interpretation to the alternative coefficient estimates, since most variables lose their explanatory power both individually and jointly with the new

[18]In order to save space we did not include the detailes regression results in the paper.

specification. On the other hand, our tests on heteroskedasticity can not deliver enough evidence to reject the null of homoskedasticity, especially if we consider standardized payoffs.[19] This is an important fact as it suggests that subjects in the same demographic group tend to behave in a homogeneous manner. More precisely, the variance of the payoffs does not seem to change from one group to the other according to any of the demographic variables that we consider.

3.3 Regression analysis: return decisions

The self-selection among subjects who show up in the experimental lab could introduce an important bias in the experimental analysis. In this subsection we show that it is not the case as we obtain a positive result that is similar to the one in the previous subsection. It turns out that although some personal characteristics have a significant effect on the return decision, they are not able to explain more than 2% of the decisions.[20] We have run logit regression, with and without subject and session specific fixed effects, in order to explain the decision on returning to the experimental lab in the next 30, 60, 90 and 365 days. As described before, dummy variables called RETURNXXX have been constructed to transform these decisions into numerical data. It is important to note that during this process we have ignored the last 30, 60, 90 and 365 days. With this we avoid the so called survivorship bias. The reason for it is that we can not know whether those observation belong to subject who wish to (and actually do) return to the lab in the future or not. Tables 13 and 14 offer the detailed estimation and test results. We observe that the payoff plays an important role, as it has a positive and significant effect on the logit. It is the payoff in US dollars and not its standardized version, since subject are usually paid individually and privately. That is, they do not have concise information to make intersubject comparisons. Age turns out to be significant in some specifications, however it net effect on the logit is ambiguous. Once again if we restrict our attention solely on the significant estimates and subjects aged 50 or less, we conclude that younger people tend to return in the long run. So do males. As for universities, in the long run many of

[19]Tables 9 through 12 report p-values for the White, and Breusch-Pagan tests for heteroskedasticity.

[20]To be precise, there is one regression on long term return decisions that reaches 5.8%. We shall come back to it later.

the dummy variables have significant positive effects. Some of them can be explained by simple geographic reasons, as the experimental lab whose data we are looking at is located closer to some universities than to others. On the other hand, laboratory experiments on decision making are known to build their subject pool using college students in general. The main reasons for this are that student subjects are available, easy to recruit, cheap, and can cope well with hypothetical decision making situations presented in the lab. One would expect a high correlation between the fact of being a student and the decision of returning to the lab on a voluntary basis. Pearson's χ^2 confirms such a relationship at any usual significance level, however Cramer's V statistics indicates that the association is very low.[21] List and Levitt (2005) argue that subjects self-select into experiments, and therefore people who are more interested in the announced research topic are more likely to participate. In line with this, as reported in table 3, the typical participant is a college student with a major in social sciences. Nevertheless logit regressions do not confirm these effects. The obtained educational level and the intended major do not have significant influence on any of the studied return decisions.[22]

It is interesting, although plausible, that the short-term return decisions seem to be more random than long-term ones. The R^2 statistics are much larger if we consider a possible returning in one year as compared to 30, 60 and 90 days. The logit regression of RETURN365 with subject specific fixed effect is the most capable of explaining variation in the dependent variable. In this case demographics and the money earned in the previous session account for almost 6% of the observed variation. However, the largest part of it is explained by other factors and variables.

[21]Our data confirms the sign of the suspected association, i.e. students tend to return, but Cramer's V is computed to be around 10%. It is equal to 10.63%, 10.52%, 10.02%, and 12.94% for the variables RETURN30/60/90 and 365 respectively.

[22]We have estimated the regressions with the categorical variables for education and/or major, but could not identify significant relations. Given these results, and due to considerations of length, the estimation outputs are omitted.

4 Conclusion

The validity of research results in experimental economics is often questioned based on an alleged subject pool bias. We use data coming from a single laboratory comprising 8,755 observations coming from a variety of incentive based economic experiments to find out that demographic differences can explain only 4% of the variations on payments in the best case. When controlling for experimental sessions we are able to explain roughly 40% of the variations on payments. We think that this indicates that some kind of strategic behavior seems to be more important than demographics in the laboratory.

Nevertheless, some of our pairwise comparisons of groups, e.g. between males and females, are statistically significant. Therefore, it seems to be a good practice to routinely study demographic, and in particular gender, effects in the data analysis. If one does not find them important after a careful statistical and economic analysis, they can be disregarded.

Our analysis also addresses the question of self-selection within the subject pool we analyze. Some demographic variables turn out to be significant when explaining the decision to participate again in a lab experiment. However, they are not able to explain more than 2% of the short term or roughly 6% of the long term decisions.

It has to be pointed out that our study does not directly support the external validity of economic experiments, as we do not have data on how the studied subjects behave in the field. However, our findings do support internal validity. In particular we claim that experimental results seem to be only slightly affected by spurious demographic variables.

References

[1] Benz, M. and Meier, S. (2006), Do People Behave in Experiments as in Real Life? - Evidence from Donations, mimeo.

[2] Brewer, M. (2000), Research Design and Issues of Validity, In Reis, H. and Judd, C. (eds), Handbook of Research Methods in Social and Personality Psychology. Cambridge: Cambridge University Press.

[3] Carbone, E. (2005), Demographics and Behaviour, *Experimental Economics*, 8: 217-232

[4] Falk, A. and Fehr, E. (2003), Why Labour Market Experiments?, *Labour Economics*, 10: 399-406

[5] Gneezy, U., Niederle, M. and Rustichini, A. (2003), Performance in Competitive Environments: Gender Differences, *The Quarterly Journal of Economics*, 3: 1049-1074

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

[7] List, J.A. and Levitt, S.T. (2005), What Do Laboratory Experiments Tell Us about the Real World?, mimeo

[8] Roth, A.E., Prasnikar, V., Okuno-Fujiware, M. and Zamir, S. (1991), Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh and Tokyo: an Experimental Study, *American Economic Review*, 81: 1068-1095

[9] Roth, A.E. (1995), Introduction to Experimental Economics, The Handbook of Experimental Economics, John H. Kagel and Alvin E. Roth, editors, Princeton University Press, 1995.

A Tables

Table 1: **Sample descriptives:** Continuous numerical variables.
All variables are measured in their natural units. AGE: years. EARNING_AMT: US dollars. ST_EARNING_AMT: standardized US dollars. EXPERIENCE: number of experimental sessions.

	AGE	EARN._AMT	ST_EARN._AMT	EXPERIENCE _TOTAL	EXPERIENCE _INCENTIVES
mean	23.56	23.80	0.00	6.47	4.80
std. dev.	7.14	8.85	0.97	7.78	5.87
skewness	3.20	1.95	0.25	2.07	2.16
kurtosis	14.87	13.42	3.42	8.27	9.06
min	18	0.00	-4.82	0	0
max	73	127.00	4.82	53	43
obs.	8706	8755	8755	8755	8755

Table 2: **Sample descriptives:** Categorical variables.
Number of observations (#), and proportion (%) of the population that belongs to each category listed on the left.

UNIVERSITY	#	%	RETURN	#	%
Boston College	54	0.62	30: yes	4810	54.94
Boston University	1073	12.26	30: no	3637	41.54
Harvard University	4171	47.64	30: N/A	308	3.52
MIT	355	4.05	60: yes	5034	57.50
Northeastern Univ.	194	2.22	60: no	2833	32.36
Tufts University	214	2.44	60: N/A	888	10.14
Univ. of Massachusetts	83	0.95	90: yes	4891	55.87
other	338	3.86	90: no	2383	27.22
not reported	2273	25.96	90: N/A	1481	16.92
EXPERIENCE_ANY			365: yes	3981	45.47
yes	7162	81.80	365: no	1018	11.63
no	1593	18.20	365: N/A	3756	42.49

Table 3: **Sample descriptives:** Categorical variables. (continued)
Number of observations (#), and proportion (%) of the population that belongs to each category listed on the left.

RACE	#	%	GENDER	#	%	EDUCATION	#	%
Asian	961	10.98	female	3923	44.81	Some High School	24	0.27
Black	396	4.52	male	4081	46.61	High School Diploma	688	7.86
Hispanic	280	3.20	not reported	751	8.58	Some College	4576	52.27
MR Black	25	0.29	MAJOR			Associate Level Degree	45	0.51
MR other	234	2.67	Engineering	451	5.15	Bachelor Level Degree	1621	18.52
Native American	3	0.03	Humanities	1049	11.98	Other Masters	588	6.72
White	2594	29.63	Natural Sciences	1095	12.51	MBA	97	1.11
other	66	0.75	Social Sciences	2823	32.24	Doctoral Level Degree	112	1.28
not reported	4196	47.93	other	368	4.20	not reported	1004	11.47
			not reported	2969	33.91			

Table 4: **Pairwise comparisons:** Gender, experience and students.
Difference between the average payoff across groups (row category - column category). Results for EARNING_AMT above the diagonal, for ST_EARNING_AMT below the diagonal. Superindex: significance for the t-test comparing sample means. Subindex: significance for the Kruskal-Wallis test. Difference: *Significant at 10%. **Significant at 5%. ***Significant at 1%.

EARNING_AMT ST_EARNING_AMT	Females	Males	not reported
Females	–	-0.4442^{**}_{***}	1.3747^{***}_{***}
Males	0.0518^{**}_{**}	–	1.8188^{***}_{***}
not reported	0.0339	-0.0179	–
	No experience	Experience	–
No experience	–	-0.7483^{***}_{***}	–
Experience	0.0502^{*}_{**}	–	–
	No college	College	–
No college	–	-0.8596^{***}_{***}	–
College	0.0284_{*}	–	–

Table 5: **Pairwise comparisons:** Intended major.
Difference between the average payoff across groups (row category - column category). Results for EARNING_AMT above the diagonal, for ST_EARNING_AMT below the diagonal. Superindex: significance for the t-test comparing sample means. Subindex: significance for the Kruskal-Wallis test. Difference: *Significant at 10%. **Significant at 5%. ***Significant at 1%.

EARNING_AMT ST_EARNING_AMT	Engineering	Humanities	Natural Sc.	Social Sc.	Other
Engineering	–	0.1062	0.1062	-0.3561	-0.0106
Humanities	-0.0422	–	0.1724	-0.6347^{*}_{***}	-0.2892
Natural Sc.	0.0424	0.0846^{**}_{*}	–	-0.4623_{*}	-0.1168
Social Sc.	0.0215	0.0637^{*}_{*}	-0.0209	–	0.3454
Other	0.0024	0.0445	-0.0400	-0.0192	–

Table 6: **Pairwise comparisons:** Racial groups.

Difference between the average payoff across groups (row category - column category). The abbreviation of MR stands for *multiracial.* Results for EARNING_AMT above the diagonal, for ST_EARNING_AMT below the diagonal. Superindex: significance for the t-test comparing sample means. Subindex: significance for the Kruskal-Wallis test. Difference: *Significant at 10%. **Significant at 5%. ***Significant at 1%.

EARNING_AMT ST_EARNING_AMT	Asian	Black	Hispanic	MR Black	MR Other	Native Am.	White	Other
Asian	–	1.1736^{**}_{***}	−0.0424	−0.7574	0.0445	1.2936	−0.1355	1.4789^{*}
Black	-0.1095^{*}_{*}	–	-1.2159^{*}_{***}	−1.9309	-1.1291^{*}_{***}	0.1200	-1.3091^{***}_{***}	0.3053
Hispanic	−0.0821	0.0274	–	−0.7150	0.0868	1.3359	−0.0932	1.5212^{*}
MR Black	0.0733	0.1828	0.1554	–	0.8018	2.0509	0.6218	2.2362
MR Other	−0.0240	0.0855	0.0581	−0.0973	–	1.2491	−0.1800	−0.1384
Native American	0.0471	0.1566	0.1292	−0.0262	0.0711	–	−1.4291	0.1853
White	−0.0295	0.0800	0.0526	−0.1028	−0.0055	−0.0766	–	1.6144^{**}
Other	0.1144	0.2239^{*}_{*}	0.1965	0.0411	0.1384	0.0673	0.1439	–

Table 7: **Pairwise comparisons:** Education.

Difference between the average payoff across groups (row category - column category). Results for EARNING_AMT above the diagonal, for ST_EARNING_AMT below the diagonal. Key to the abbreviations: SHS - Some High School; HSD - High School Diploma; SC - Some College; ALD - Associate Level Degree; BLD - Bachelor Level Degree; OM - Other Masters; DLD: Doctoral Level Degree. Superindex: significance for the t-test comparing sample means. Subindex: significance for the Kruskal-Wallis test. Difference: *Significant at 10%. **Significant at 5%. ***Significant at 1%.

EARNING_AMT ST_EARNING_AMT	SHS	HSD	SC	ALD	BLD	OM	MBA	DLD
SHS	–	−0.8739	−1.0912	−1.1363	−0.6620	−0.4516	−1.2171	0.7465
HSD	0.0618	–	−0.2173	−0.2624	0.2119	0.4223_{*}	−0.3431	1.6205^{***}
SC	0.0973	0.0355	–	−0.0451	0.4292_{***}	-0.6396_{***}	−0.1259	1.8377^{***}_{*}
ALD	−0.0378	−0.0997	−0.1351	–	0.4743	0.6847	−0.0808	1.8829
BLD	0.0690	0.0072	0.0283	−0.1068	–	0.2104	−0.5551	1.4085^{**}
OM	0.0555	−0.0063	−0.0418	0.0933	−0.0135	–	−0.7654	1.1982^{***}
MBA	0.9777	0.0359	0.0004	0.1355	0.0287	0.0422	–	1.9636^{**}
DLD	−0.2162	−0.0835	−0.1189	0.0162	−0.0906	−0.0771	−0.1193	–

Table 8: **Pairwise comparisons:** Colleges.
Difference between the average payoff across groups (row category - column category). Results for EARNING_AMT above the diagonal, for ST_EARNING_AMT below the diagonal. Superindex: significance for the t-test comparing sample means. Subindex: significance for the Kruskal-Wallis test. Difference: *Significant at 10%. **Significant at 5%. ***Significant at 1%.

EARNING_AMT ST_EARNING_AMT	Boston Coll.	Boston Univ.	Harvard Univ.	MIT	Northe. Univ.	Tufts Univ.	Univ. of MA	no college
Boston Col.	–	2.1110_{**}	2.1013_{**}	1.2027	2.5587^{*}_{**}	1.4105	2.7119^{*}_{**}	2.9326^{**}_{***}
Boston Univ.	−0.1361	–	−0.0097	-0.9084_{*}	0.4477	−0.7006	0.6008	0.8216^{**}_{***}
Harvard Univ.	−0.0930	0.0431	–	-0.8986^{*}_{*}	0.4574	−0.6909	0.6105	0.8313^{***}_{***}
MIT	−0.0321	0.1040^{**}_{**}	0.0609_{*}	–	1.3560^{*}	0.2078	1.5092	1.7299^{***}_{***}
Northe. Univ.	−0.1860	−0.0499	-0.0930_{*}	-0.1539^{**}_{**}	–	−1.1483	0.1531	0.3739
Tufts Univ.	−0.0985	0.0377	−0.0054	−0.0663	0.0876	–	1.3014	1.5222^{**}_{**}
Univ. of MA	−0.0936	0.0425	−0.0005	−0.0615	0.0924	0.0049	–	0.2208
no college	−0.1307	0.0054	-0.0376_{**}	-0.0985^{**}_{***}	0.0554	−0.0322	−0.0371	–

Table 9: **Regression analysis:** Payoffs.

Dependent variable: EARNING_AMT. Regressors: as defined in the main text. Estimation method in second row. OLS: ordinary least squares. MEDIAN: quantile or median regression. SW: regression with stepwise deletion of insignificant regressors using 15% significance level. ROBUST: regression with SUBJECT, SESSION or EXPERIMENT specific fixed effects and robust variance estimation. The last four rows contain p-values that correspond to the White, and Breusch-Pagan heteroskedasticity test, and to joint significance tests of coefficients. Coefficient: *Significant at 10%. **Significant at 5%. ***Significant at 1%.

	OLS	OLS	OLS	OLS	MEDIAN	SW OLS	SW MED.	ROBUST	ROBUST	ROBUST	ROBUST
AGE	−0.284***	−0.285***	−0.284***	−0.274***	−0.205**	−0.254***	−0.262***	−3.719***	−4.126***	−4.097***	−3.322***
AGE2	0.004***	0.004***	0.004***	0.003***	0.003**	0.003***	0.003***	0.014	0.020	0.023	0.023
EXPERIENCE											
_TOTAL	0.033	0.011	–	–	0.016	–	0.022	0.538***	0.147***	–	–
_INCENTIVES	−0.030	–	0.013	–	0.003	–	–	−0.529**	–	0.165***	–
_ANY	−0.036	–	–	0.043	0.982***	–	0.962***	−0.327	–	–	−0.194
GENDER	0.433**	0.432**	0.433**	0.440**	0.537***	0.442**	0.600***	–	–	–	–
U1	−0.471	−0.468	−0.469	−0.489	−0.016	–	–	–	–	–	–
U2	−0.123	−0.145	−0.147	−0.054	0.765	–	–	–	–	–	–
U3	0.493	0.482	0.479	0.496	0.716	–	–	–	–	–	–
U4	−0.240	−0.237	−0.233	-0.222	-0.067	–	–	–	–	–	–
U5	2.258*	2.260*	2.258*	2.230*	3.878***	2.340*	3.579***	–	–	–	–
U6	0.005	0.009	0.013	0.015	0.290	–	–	–	–	–	–
U7	−0.207	−0.203	−0.199	−0.196	0.469	–	––	–	–	–	–
U8	1.012*	1.011*	1.011*	1.004*	1.594***	1.115**	1.158**	–	–	–	–
CONSTANT	28.275***	28.258***	28.246***	28.099***	23.798***	27.676***	25.067***	102.055***	107.555***	105.620***	87.932***
CONTROL	–	–	–	–	–	–	–	SUBJECT	SUBJECT	SUBJECT	SUBJECT
R^2	0.004	0.004	0.004	0.004	–	0.004	–	0.284	0.283	0.282	0.281
adj./pseudo R^2	0.002	0.002	0.002	0.002	0.004	0.003	0.004	0.038	0.037	0.036	0.034
White χ^2	0.029	0.002	0.002	0.002	–	0.160	–	–	–	–	–
Breusch-Pagan χ^2	0.126	0.138	0.114	0.024	–	0.027	–	–	–	–	–
F-test											
_GLOBAL	0.006	0.002	0.002	0.003	–	0.000	–	0.000	0.000	0.000	0.000
_AGE AGE2	0.002	0.002	0.002	0.003	–	0.001	–	0.000	0.000	0.000	0.000
_EXPERIENCE	0.871	–	–	–	–	–	–	0.000	–	–	–
_U1 - U8	0.173	0.176	0.179	0.184	–	–	–	–	–	–	–

Table 10: **Regression analysis:** Payoffs. (continued)
Dependent variable: EARNING_AMT. Regressors: as defined in the main text. Estimation method in second row. ROBUST: regression with SUBJECT, SESSION or EXPERIMENT specific fixed effects and robust variance estimation. The last four rows contain p-values that correspond to the White, and Breusch-Pagan heteroskedasticity test, and to joint significance tests of coefficients. Coefficient: *Significant at 10%. **Significant at 5%. ***Significant at 1%.

	ROBUST	ROBUST	ROBUST	ROBUST	ROBUST	ROBUST	ROBUST	ROBUST
AGE	−0.099	−0.097	−0.099	−0.094	−0.156*	−0.152*	−0.155*	−0.141
AGE2	0.001	0.001	0.001	0.001	0.002	0.002	0.002	0.002
EXPERIENCE								
_TOTAL	−0.035	0.007	–	–	−0.064	0.015	–	–
_INCENTIVES	0.052	–	0.011	–	0.103	–	0.024	–
_ANY	0.236	–	–	0.258	0.155	–	–	0.232
GENDER	0.219	0.224	0.223	0.222	0.365**	0.369**	0.367**	0.375**
U1	0.339	0.332	0.336	0.323	0.049	0.037	0.047	0.005
U2	0.060	0.082	0.060	0.140	0.325	0.402	0.359	0.529
U3	0.416	0.444	0.438	0.437	0.235	0.276	0.264	0.284
U4	0.207	0.206	0.207	0.200	0.131	0.120	0.121	0.125
U5	0.750	0.750	0.755	0.733	1.178	1.172	1.181	1.134
U6	0.071	0.070	0.072	0.062	0.160	0.146	0.150	0.140
U7	0.385	0.381	0.383	0.374	0.276	0.259	0.264	0.253
U8	1.032**	1.035**	1.035**	1.031**	1.356***	1.358***	1.358***	1.352***
CONSTANT	24.973***	25.118***	25.135***	24.909***	25.839***	25.908***	25.942***	25.625***
CONTROL	SESSION	SESSION	SESSION	SESSION	EXPERIMENT	EXPERIMENT	EXPERIMENT	EXPERIMENT
R^2	0.453	0.453	0.453	0.453	0.338	0.338	0.338	0.338
adj./pseudo R^2	0.408	0.408	0.408	0.408	0.331	0.331	0.331	0.331
White χ^2	–	–	–	–	–	–	–	–
Breusch-Pagan χ^2	–	–	–	–	–	–	–	–
F-test								
_GLOBAL	0.076	0.066	0.063	0.045	0.002	0.003	0.002	0.002
_AGE AGE2	0.214	0.209	0.200	0.207	0.059	0.057	0.051	0.073
_EXPERIENCE	0.608	–	–	–	0.179	–	–	–
_U1 - U8	0.482	0.481	0.480	0.486	0.248	0.236	0.240	0.228

Table 11: **Regression analysis:** Standardized payoffs.

Dependent variable: ST_EARNING_AMT. Regressors: as defined in the main text. Estimation method in second row. OLS: ordinary least squares. MEDIAN: quantile or median regression. SW: regression with stepwise deletion of insignificant regressors using 15% significance level. ROBUST: regression with SUBJECT, SESSION or EXPERIMENT specific fixed effects and robust variance estimation. The last four rows contain p-values that correspond to the White, and Breusch-Pagan heteroskedasticity test, and to joint significance tests of coefficients. Coefficient: *Significant at 10%. **Significant at 5%. ***Significant at 1%.

	OLS	OLS	OLS	OLS	MEDIAN	SW OLS	SW MED.	ROBUST	ROBUST	ROBUST	ROBUST
AGE	−0.006	−0.006	−0.006	−0.004	−0.013	−0.004**	−0.007***	−0.124	−0.148	−0.138	−0.153
AGE2	0.000	0.000	0.000	0.000	0.000	–	–	0.003	0.003*	0.003*	0.003*
EXPERIENCE											
_TOTAL	0.003	0.003*	–	–	0.001	–	–	0.034*	0.000	–	–
_INCENTIVES	−0.001	–	0.003*	–	0.002	–	–	−0.047*	–	−0.002	–
_ANY	0.043	–	–	0.056*	0.123***	0.055*	0.132***	0.031	–	–	0.024
GENDER	0.042*	0.042*	0.042*	0.043*	0.027	0.046**	–	–	–	–	–
U1	−0.007	−0.006	−0.006	−0.011	0.045	–	–	–	–	–	–
U2	0.039	0.032	0.030	0.053	0.192	–	–	–	–	–	–
U3	0.004	0.005	0.004	0.006	0.069	–	–	–	–	–	–
U4	−0.040	−0.038	−0.037	−0.038	-0.005	–	–	–	–	–	–
U5	0.133	0.136	0.136	0.128	0.305	–	–	–	–	–	–
U6	−0.013	−0.010	−0.009	−0.012	0.065	–	–	–	–	–	–
U7	0.021	0.023	0.024	0.022	0.090*	–	–	–	–	–	–
U8	0.097*	0.097*	0.097*	0.096	0.247***	0.084**	0.171**	–	–	–	–
CONSTANT	0.038	0.066	0.065	0.010	−0.075	0.016	−0.068	1.300	1.641	1.416	1.710
CONTROL	–	–	–	–	–	–	–	SUBJECT	SUBJECT	SUBJECT	SUBJECT
R^2	0.003	0.003	0.003	0.003	–	0.002	–	0.267	0.265	0.265	0.265
adj./pseudo R^2	0.001	0.001	0.001	0.001	0.004	0.002	0.003	0.014	0.013	0.013	0.013
White χ^2	0.879	0.701	0.704	0.902	–	0.293	–	–	–	–	–
Breusch-Pagan χ^2	0.072	0.022	0.018	0.163	–	0.068	–	–	–	–	–
F-test											
_GLOBAL	0.079	0.067	0.071	0.055	–	0.001	–	0.255	0.417	0.395	0.353
_AGE AGE2	0.146	0.147	0.157	0.203	–	–	–	0.355	0.243	0.230	0.230
_EXPERIENCE	0.164	–	–	–	–	–	–	0.294	–	–	–
_U1 - U8	0.717	0.728	0.732	0.724	–	–	–	–	–	–	–

Table 12: **Regression analysis:** Standardized payoffs. (continued)
Dependent variable: ST_EARNING_AMT. Regressors: as defined in the main text. Estimation method in second row. ROBUST: regression with SUBJECT, SESSION or EXPERIMENT specific fixed effects and robust variance estimation. The last four rows contain p-values that correspond to the White, and Breusch-Pagan heteroskedasticity test, and to joint significance tests of coefficients. Coefficient: *Significant at 10%. **Significant at 5%. ***Significant at 1%.

	ROBUST	ROBUST	ROBUST	ROBUST	ROBUST	ROBUST	ROBUST	ROBUST
AGE	−0.010	−0.010	−0.010	−0.009	−0.007	−0.007	−0.007	−0.006
AGE2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
EXPERIENCE								
_TOTAL	0.003	0.003	–	–	0.003	0.003	–	–
_INCENTIVES	−0.001	–	0.003	–	−0.001	–	0.003	–
_ANY	0.053	–	–	0.063**	0.045	–	–	0.056
GENDER	0.049**	0.050**	0.050**	0.050**	0.044**	0.045**	0.045**	0.045**
U1	−0.012	−0.012	−0.011	−0.016	−0.008	−0.007	−0.007	−0.013
U2	0.053	0.046	0.044	0.069	0.042	0.036	0.034	0.058
U3	0.003	0.005	0.004	0.004	0.004	0.005	0.005	0.006
U4	−0.037	−0.035	−0.034	−0.035	−0.032	−0.030	−0.029	−0.030
U5	0.157	0.159	0.159	0.152	0.144	0.147	0.147	0.140
U6	−0.006	−0.004	−0.003	−0.006	−0.009	−0.006	−0.005	−0.008
U7	0.027	0.029	0.030	0.028	0.024	0.026	0.027	0.025
U8	0.103	0.104	0.104	0.103	0.100	0.100	0.100	0.099
CONSTANT	0.084	0.119	0.119	0.057	0.057	0.086	0.085	0.030
CONTROL	SESSION	SESSION	SESSION	SESSION	EXPERIMENT	EXPERIMENT	EXPERIMENT	EXPERIMENT
R^2	0.010	0.009	0.009	0.009	0.004	0.004	0.003	0.004
adj./pseudo R^2	−0.072	−0.072	−0.072	−0.072	−0.007	−0.007	−0.007	−0.007
White χ^2	–	–	–	–	–	–	–	–
Breusch-Pagan χ^2	–	–	–	–	–	–	–	–
F-test								
_GLOBAL	0.065	0.060	0.064	0.041	0.061	0.053	0.057	0.043
_AGE AGE2	0.103	0.103	0.110	0.146	0.109	0.110	0.118	0.158
_EXPERIENCE	0.198	–	–	–	0.195	–	–	–
_U1 - U8	0.716	0.726	0.728	0.709	0.710	0.718	0.721	0.704

Table 13: **Regression analysis:** Return decisions.
Dependent variable: RETURNXXX as specified in the first row. Regressors: as defined in the main text. Estimation method in second row. LOGIT: logit regression without and with SUBJECT, SESSION or EXPERIMENT specific fixed effects. The last three rows contain p-values that correspond to a χ^2 global significance test, and to F-test for the joint significance of the coefficients. Coefficient: *Significant at 10%. **Significant at 5%. ***Significant at 1%.

	RETURN30				RETURN60			
EARNING_AMT	0.010***	0.012***	0.002	0.005*	0.010***	0.011***	0.004	0.006*
AGE	-0.037*	0.508***	−0.068***	−0.051**	−0.030	0.502**	−0.065***	−0.052**
AGE2	0.000	−0.005	0.001**	0.001**	0.000*	-0.004	0.001***	0.001**
G1	0.155***	–	0.231***	0.197***	0.146***	–	0.183***	0.171***
U1	0.094	–	0.080	0.073	0.130	–	0.199	0.135
U2	−0.309	–	−0.210	−0.256	0.176	–	0.450	0.323
U3	0.049	–	−0.020	0.055	0.099	–	0.066	0.102
U4	0.131	–	0.083	0.134	0.459**	–	0.546**	0.490**
U5	0.147	–	0.132	0.113	0.134	–	0.068	0.064
U6	0.270***	–	0.233**	0.306***	0.288***	–	0.249**	0.319***
U7	0.185***	–	0.024	0.135*	0.214***	–	0.043	0.126
U8	−0.054	–	−0.030	0.020	−0.014	–	0.016	0.059
CONSTANT	0.509	–	–	–	0.603**	–	–	–
CONTROL	–	SUBJECT	SESSION	EXPERIMENT	–	SUBJECT	SESSION	EXPERIMENT
pseudo R^2	0.007	0.006	0.008	0.007	0.005	0.007	0.006	0.005
χ^2-test	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
F-test								
_AGE AGE2	0.074	0.000	0.000	0.005	0.128	0.000	0.028	0.056
_U1 - U8	0.044	–	0.376	0.102	0.039	–	0.060	0.045

Table 14: **Regression analysis:** Return decisions. (continued)
Dependent variable: RETURNXXX as specified in the first row. Regressors: as defined in the main text. Estimation method in second row. LOGIT: logit regression without and with SUBJECT, SESSION or EXPERIMENT specific fixed effects. The last three rows contain p-values that correspond to a χ^2 global significance test, and to F-test for the joint significance of the coefficients. Coefficient: *Significant at 10%. **Significant at 5%. ***Significant at 1%.

	RETURN90				RETURN365			
EARNING_AMT	0.010***	0.012***	0.006	0.009**	0.017***	0.017***	0.007	0.011**
AGE	−0.031	−0.097	−0.058**	0.050**	−0.049	−0.377	−0.088**	−0.059
AGE2	0.001*	0.004	0.001**	0.001**	0.001*	−0.033	0.001**	0.001**
G1	0.193***	–	0.203***	0.201***	0.197**	–	0.230**	0.235***
U1	−0.051	–	0.084	−0.022	0.169	–	0.258	0.180
U2	0.064	–	0.202	0.188	0.603	–	0.660	0.649
U3	−0.046	–	−0.029	−0.047	0.552**	–	0.257	0.488*
U4	0.449**	–	0.447**	0.436**	0.988***	–	0.969***	0.982***
U5	0.191	–	0.086	0.070	0.030	–	−0.221	−0.033
U6	0.294***	–	0.272**	0.311***	0.436***	–	0.363**	0.427***
U7	0.176**	–	−0.025	0.044	0.568***	–	0.351***	0.437***
U8	−0.012	–	0.022	0.033	0.124	–	0.123	0.112
CONSTANT	0.723*	–	–	–	1.170**	–	–	–
CONTROL	–	SUBJECT	SESSION	EXPERIMENT	–	SUBJECT	SESSION	EXPERIMENT
pseudo R^2	0.006	0.003	0.006	0.006	0.019	0.058	0.014	0.014
χ^2-test	0.000	0.001	0.001	0.000	0.000	0.000	0.000	0.000
F-test								
_AGE AGE2	0.007	0.504	0.018	0.015	0.005	0.000	0.012	0.014
_U1 - U8	0.039	–	0.080	0.050	0.000	–	0.033	0.003