

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Do experienced subjects bias experimental results? Evidence from 16 laboratories in six countries

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Guerra, A., Harrington, B., Steinmo, S., D'Attoma, J. (2022). Do experienced subjects bias experimental results? Evidence from 16 laboratories in six countries. *ECONOMIC AND POLITICAL STUDIES*, Published online, 1-15 [10.1080/20954816.2022.2110244].

Availability:

This version is available at: <https://hdl.handle.net/11585/893056> since: 2022-12-08

Published:

DOI: <http://doi.org/10.1080/20954816.2022.2110244>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

Guerra, A., Harrington, B., Steinmo, S., & D'Attoma, J. (2022). Do experienced subjects bias experimental results? Evidence from 16 laboratories in six countries. *Economic and Political Studies*, 1-15.

The final published version is available online at:

<https://doi.org/10.1080/20954816.2022.2110244>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)

When citing, please refer to the published version.

Do Experienced Subjects Bias Experimental Results? Evidence from 16 Laboratories in Six Countries

Alice Guerra · Brooke Harrington · Sven Steinmo · John D'Attoma

Abstract

This paper addresses an area of growing concern for laboratory researchers: are subjects' behaviours affected by prior experiences in laboratory experiments? We address the question with a large and highly diverse international dataset, and an operationalization strategy that allows our findings to cohere with previous work while shedding new light for future research. The findings presented here are drawn from original data gathered as part of one of the largest tax compliance experiments ever conducted, involving more than 3,000 participants in six countries, across 16 different laboratories. Our results reveal that subjects' behaviour correlates with their past experimental experiences, in a way that could bias results and compromise a study's external validity; however, this change in behaviour due to experience occurs only after subjects have participated in at least two previous laboratory experiments. Our findings have implications not just for tax compliance research, but for allocation experiments more generally, and for participant recruitment.

Keywords: Methodology; Experimental experience; Laboratory experiment; Tax compliance

JEL classification: B41; C9; D9; H26

PsycINFO classification: 2260; 2300

[1] Alice Guerra (Corresponding Author)

Department of Economics, University of Bologna, via Angherà 22, Rimini, Italy.

ORCID: 0000-0003-1956-0270

e-mail: alice.guerra3@unibo.it

[2] Brooke Harrington

Department of Sociology, Dartmouth College, Hanover, New Hampshire, USA

e-mail: brooke.harrington@dartmouth.edu

[3] Sven Steinmo

Department of Political Science, University of Colorado, USA

e-mail: steinmo@colorado.edu

[4] John D'Attoma

Department of Finance and Accounting and Tax Administration Research Centre, University of Exeter, UK

e-mail: johndattoma@gmail.com

1. Introduction

Experimental methods have gained increased currency across the social sciences, but some remain sceptical about their external validity (e.g. Levitt and List 2007b; Al-Ubaydli et al. 2013; Czibor et al. 2019; List 2020). A central concern, particularly in the case of laboratory experiments, is whether subjects’ responses in the lab reflect their real-world behaviour (e.g. Lynch 1982; Winer 1999; Levitt and List 2007a; Alm et al. 2015; Al-Ubaydli et al. 2017).

A critical and largely understudied aspect of the external validity issue is whether subjects modify their behaviour in the laboratory after participating in previous experiments (Dengler-Roscher et al. 2018). This question is of special interest for public goods researcher where cheating pays off, including in tax compliance, dictator and ultimatum games (e.g. Marwell and Ames 1979; Rand 2018; Dengler-Roscher et al. 2018; Conte et al. 2019).

This paper addresses whether past experience in the laboratory affects subjects’ behaviour *across* experiments. Previous studies addressing this issue have been limited by their small sample sizes, as well as by disparities in operationalization and localization (e.g. Marwell and Ames 1980; Isaac et al. 1984; Benndorf et al. 2017). Our study addresses these shortcomings.

First, we draw upon a much larger and more geographically diverse dataset than has been available to previous laboratory researchers working on this problem. Our unique original dataset was collected as part of one of the largest controlled tax compliance experiments ever conducted (Pampel et al. 2019), encompassing 3,000 individuals in six countries, across 16 different laboratories; the scale and diversity of our data permits for more reliable generalization than has been possible with earlier work. Second, we operationalize subjects’ laboratory experience in ordinal terms, offering a more fine-grained and generalizable analysis than the crude and inconsistent binaries used in previous studies.

This analysis offers two key insights. First, we show that individuals with prior experience as laboratory research subjects *do* behave in distinctive ways that can bias experimental results. Second, a detailed analysis of a subsample of our data reveals a threshold level of experience to be meaningful for behavioural change: individuals change their behaviour after their second laboratory experience, but the changes cease from the sixth experimental study onward.

2. Related Literature

To date, research investigating whether subjects learn from experience and change their behaviour across laboratory experiments has fallen into three categories. The first group finds no significant difference in the behaviors of more and less experienced subjects (Marwell and Ames 1980; Isaac et al. 1984; Bolton 1991; Xue et al. 2017; Medda et al. 2021). A second group of studies suggests that experience *does* matter. For instance, more-experienced subjects are less likely to reciprocate and contribute to public goods than less-

experienced subjects (Matthey and Regner 2013; Conte et al. 2019; Jouxte1 2019).¹ In dictator games, experienced subjects perform better and make more selfish allocations in a dictator game, compared to their inexperienced counterparts (Dengler-Roscher et al. 2018). In coordination games, experienced subjects are significantly less trusting and more likely to behave selfishly (Schmidt et al. 2020).²

A third set of studies answers the question whether experience matters in experimental laboratory subjects with “it depends.” One subset of this research suggests that the effects of prior experience are contingent on experimental task conditions. For example, Capraro and Cococcioni (2015) show that experienced participants are more cooperative than first-timers, but only under time pressure, while Benndorf et al. (2017) show that the behaviour differs between inexperienced and experienced subjects in a trust game and a risk elicitation task, but not in the other decision settings (a beauty contest, an ultimatum game, a travellers’ dilemma, and a single-player lying task). The second subset in this category focuses on increments of experience, to determine whether each additional laboratory experience had a measurable impact on subjects’ behaviour. Here again, the results are mixed: for example, Matthey and Regner (2013) find a negative, marginal effect of each past laboratory experience on subjects’ cooperative behaviour, while Capraro and Cococcioni (2015) find the opposite, and Benndorf et al. (2017) find no marginal effects at all.

The mixed findings may stem from two sources of fragmentation in the literature: tightly limited sampling and inconsistent operational definitions. In the first case, previous studies testing the effects of subjects’ prior laboratory experiences have relied on relatively small numbers of individuals at single lab sites. For example, Matthey and Regner (2013) reach their conclusions about experience effects by combining information from four studies—involving 64, 192, 127 and 288 observations—conducted at one lab in Jena, Germany (see Table S1 in the Supplemental).

A second source of fragmentation in the literature is the lack of consistent operational definitions of “experienced” vs “inexperienced” participants in laboratory experiments. In some experiments, only participants with *zero* prior experience in the lab (e.g. first-timers) are classified as inexperienced, while other work includes in that category subjects who have participated in up to *five* previous laboratory experiments. The “experienced” category is even less clearly defined, ranging from one to 20 prior participations in laboratory research. We address this problem with an operationalization strategy that treats subjects’ prior laboratory experience as an ordinal variable, producing a fine-grained analysis that allows our findings to be compared to previous research while going beyond the limitations of those studies (see Table S2 in the Supplemental).

¹ This laboratory finding tracks that of research conducted online: Meyers et al. (2020) show that inexperienced participants made more contributions in public good game than experienced participants.

² See also Benson and Faminow (1988), noticing that prior laboratory experience affects subjects’ tacit cooperation (i.e. collusion) choices.

3. Data and Methodology

3.1 Data

Our data derive from three tax compliance experiments, conducted between 2014 and 2018 in 16 laboratories across six countries. We merged the data from these studies, which share the same first-round baseline treatment described in Section 3.2.³ The final dataset contains 3,266 observations: 74 in Denmark; 684 in Italy; 443 in Romania; 501 in Sweden; 590 in the UK; and 974 in the US.⁴ Table S3 in the Supplemental provides descriptive statistics, including the list of the 16 laboratories. Approximately 48.3% of subjects were male; subjects' average age was 23 years old, and 89.0% were students.⁵

Our analysis indicates that previous participation in laboratory experiments affects subjects' behaviour. However, since this study was not originally designed to explore the behavioural effects of prior laboratory experiences, we lack information about the types of experiments our subjects participated in previously, and on the possible mechanisms behind our results. See Section 5 for a discussion.

To measure (1) *whether* past experimental experience is correlated with subjects' compliance choices, and (2) *how much* experience it takes to make a difference in behaviour, we use the variable *Experience*, which we measure in the post-experiment questionnaire and operationalize in two ways. First, we operationalize *Experience* as a dummy variable coded as 1 if a subject reported *any* previous participation in laboratory experiments ("Experienced"); and 0 if a subject reported no previous experiences ("Inexperienced" or "first-timers"). Of our 3,266 subjects, a total of 2,110 (64.6%) were "experienced," while the remaining 1,156 (35.4%) were "inexperienced" (see Table S4 in the Supplemental). The highest percentages of experienced subjects were in Denmark (83.8%) and the UK (81.9%); the lowest percentages of experienced subjects were in Romania (23.7%) and the US (57.9%).

In our pooled sample, the first-timers had a mean age of 22.106 years (Std. Dev. 5.953, min 18, max 85, N=1,155), while experienced participants had a mean age of 23.824 years (Std. Dev. 7.464, min 18, max 76, N=2,108). Males comprised 49.4% of experienced participants, and 46.4% of first-timers. Balance checks—with *p* values computed following Chiapello (2018)—indicate that the participant groups are balanced in terms of gender ($p=.103$), but not in terms of age ($p<.001$). In our parametric analysis, we account for unbalanced characteristics between experienced vs inexperienced in two ways: (1) by controlling for *Age* (along with gender and performance in the clerical task) and including its interaction with the variable *Experience*; (2) by conducting robustness checks with the Propensity Score Matching technique. This technique allows us to check whether

³ Details on the differences among the three experiments—including instructions after the first-round baseline treatment—are available upon request.

⁴ The original dataset contained 3,320 observations, from which we had to drop 54 observations due to missing information about previous participation in experiments and other personal characteristics.

⁵ As the vast majority of the sample consisted of students, a natural extension of this research is to investigate whether our results are affected by differences between disciplines (e.g., economics vs. psychology).

inexperienced subjects are similar to their experienced counterparts in observable terms—age, gender, and performance in the clerical task—or differ from them in terms of compliance behaviour.

Tables S7 and S8 in the Supplemental shows the proportion of experienced and inexperienced participants by lab country and region respectively, providing their distribution between and within each location. Chi-squared tests show that the proportion of experienced subjects is significantly different among the six countries ($\chi^2(5)=501.333$, $p<.001$), as well as among regions ($\chi^2(15)=677.962$, $p<.001$). For this reason, our parametric analysis controls both for lab country fixed effects and lab region fixed effects. As robustness checks, we also conducted regressions for each country separately.

To analyse *how much* laboratory experience influences subjects' behaviour, we operationalize *Experience* as an ordinal variable that measures the number of times an experienced subject previously participated in experiments. This variable takes integer values from 1 through 4, where 1 represents *one* prior laboratory experience, 2 represents *two* prior laboratory experiences, 3 represents cases of *three to five* past laboratory experiences, and 4 represents cases of *six or more* past laboratory experiences. These values correspond to the framing of answers participants could choose in the post-experimental questionnaire. After being asked whether they had participated in experiments previously, those who answered yes were offered the following response options: "Once," "Twice," "3-5 times," "More than 5 times," "More than 10 times." In our analysis, the ordinal variable *Experience* is truncated at 4 because the percentage of subjects with more than ten previous experiences was very small (65 out of 1,608 observations, or 2.4%).

For this analysis, we consider a subset of our data—the experienced group, 1,608 observations out of 3,266—to see whether the number of experiments in which they participated is linked to their behaviour in our tax experiment. We did this for two reasons. First, in one of our three experiments on tax compliance (Guerra and Harrington 2018, which gathered data from 180 participants, 74 in Denmark and 106 in Italy), we did not ask participants about the *number* of experiments they already participated in, but only *whether* they had participated in lab studies previously (yes or no). Second, there was some missing data, partly due to non-response by participants and partly because the early rounds of the experiments in Italy and the UK did not include the question about the number of past participations.

Table S5 reports summary statistics for the ordinal variable *Experience*, whose average is 2.378 with a median of 2 (Std. Dev. 1.180, min 1, max 4). Among the six countries represented in our data, the mean was highest in Italy (2.628) and the UK (2.513), and lowest in Romania (1.500) and the US (2.139). The level of experience varied widely within countries: for example, the mean was highest in Oxford (3.258) and RHUL (2.909), and lowest in Bucharest (1.500) and Essex (1.688). ANOVA tests confirm that experience levels differ significantly across the countries and regions where the experimental labs were located, at $p<.01$. For this reason, our parametric analyses control for both national

and regional fixed effects. As robustness checks, we also conducted regressions for each country separately.

3.2 Design and Procedures

All of our experiments were identical in the first round of the tax compliance game, and in the procedure through which they were conducted. Our experimental design (Andrighetto et al. 2016; Zhang et al. 2016; Guerra and Harrington 2018; Ottone et al. 2018) followed the basic elements of most tax compliance experiments (for extensive reviews, see Torgler 2002; Alm 2019; Alm and Malézieux 2020). The game comprises different stages, each divided into three rounds. At the beginning of each stage, subjects performed a data entry task: copying a string of characters from a piece of paper to the screen.⁶ For each row subjects copied correctly, they received 10 points. Next, they observed their earnings on the screens of their lab workstations and were asked to report them for tax purposes; they were informed in advance of the tax rate, audit probability, and redistribution policy. Here, we focus on the first round of the tax compliance game—which is the only one identical in all our experiments.

In that round, participants were informed that they were free to report any amount; it would be taxed at a rate of 30%, with a 5% audit probability and a penalty of double the taxes owed if true earnings were found to be understated. There was no redistribution of tax revenues. We only revealed the result of an audit at the conclusion of the experiments and did not provide participants with information about others' choices, nor whether others were audited. After the tax game, we asked participants to complete a questionnaire about their demographics and previous participation in experimental research. Following the conventions of experimental economics, participants were not deceived, and they were compensated in cash at the end of each session.

Most experiments were programmed using *zTree* (Fischbacher 2007).⁷ Participants were recruited through the online platform ORSEE (Greiner 2015) and performed all tasks via computer. Once subjects arrived at the lab, we randomly assigned them to a desk with a computer terminal and privacy screens shielding their computers from the view of other participants. We read aloud the experimental instructions (Supplemental), which were also displayed on each terminal screen. Participants were informed that their decisions during the experiment – as well as their final payments – would be kept confidential and anonymous. We informed them that during the experiment they would earn *points* on the basis of their own choices, choices of others in the experiment, and by chance. These points were converted at the end of the experiment into the local currency. As it is usual in economics experiments, in each lab region the exchange rate was set so that the average

⁶ Participants had the chance to practice the task, and were informed that the performance in the practice task had no effects on the subsequent tasks, nor in the final payment.

⁷ A few experiments were programmed using *Behaviory*. There are no substantial differences in user interfaces between *zTree* and *Behaviory*. The experiments were identical in all ways, except for that *Behaviory* was conducted through an online server whereas *zTree* had a local server. There is no reason that our results would vary by *zTree* and *Behaviory*.

hourly payment to subjects would be approximately 1.5 times the local minimum hourly wage for student employment. Each subject received a show-up fee for participation.⁸

3.3 Econometric Specification

Our analyses test whether and to what extent prior experience participating in laboratory experiments correlates with individuals' willingness to declare their true earnings for tax purposes. We estimate these associations using linear regression models specified as follows:

$$Compliance\ Rate_i = \alpha + \beta Experience_i + \mathbf{X}'_i \gamma_1 + \mathbf{Z}'_L \delta + \varepsilon_i \quad (1)$$

where *Compliance Rate_i* is our dependent variable, defined as the ratio of individual *i*'s reported points to their actual earned points in the clerical task. Potential values range from 0 (full tax evasion) to 1 (full tax compliance). This variable has a bimodal distribution, with large spikes at 0 and 1 (Fig. S1 in the Supplemental). Because of this distribution, following standard econometric analyses of tax compliance (Alm and Mal  zieux 2020), we added an analysis of experience effects on the extensive and intensive margins of compliance rates.

Specifically, we consider two different yet related dependent variables. To estimate the experience effect on the *extensive* margin, we consider the *probability* of being tax compliant, i.e. $\Pr(Compliance\ Rate_i > 0)$. In this case, the dependent variable is operationalized as a dummy coded 1 if subject *i* declared some positive earnings, and 0 if zero earnings were declared. Given the binary nature of this dependent variable, we estimate Linear Probability Models (LPM). To measure the experience effect on the *intensive* margin of compliance rate, we consider the *amount* of tax compliance conditional upon being tax compliant, i.e. $Compliance\ Rate_i | (Compliance\ Rate_i > 0)$. In this case, we estimate Ordinary Least Squares (OLS) regressions.⁹

Our key explanatory variable is *Experience_i*, which we measure in the post-experiment questionnaire, and operationalize in two ways. First, we operationalize it as a *dummy variable* coded as 1 if subject *i* participated in experiments previously, 0 otherwise. Second, to estimate the incremental effects of previous laboratory experience on subjects' compliance choices, we operationalize *Experience_i* as an *ordinal variable*, representing subjects' previous experience participating in experiments.

The vector \mathbf{X}'_i includes individual *i*'s gender (male = 1), age, and the number of rows correctly copied in the clerical task.¹⁰ \mathbf{Z}'_L is a vector of dummy variables for lab location fixed effects. The error term ε_i denotes unobserved characteristics determining compliance decisions. All standard errors are robust against heteroskedasticity and clustered at the individual level. We conduct regressions over the pooled sample, as well as for each country individually.

⁸ For more detailed information about average final payments, we refer to our publications (Andrighetto et al., 2016; Zhang et al., 2016; Guerra and Harrington, 2018).

⁹ As robustness checks, we estimated Probit regressions for the extensive margin, and Tobit regressions for the intensive margin. All results are consistent across models. We have reported here the estimates from linear regressions to ease interpretation of the coefficients; the untabulated estimates are available upon request.

¹⁰ We control for performance in the clerical task following Dengler-Roscher et al. (2018).

4.1 Does Past Experience Matter?

Compliance varies widely between experienced and inexperienced subjects. In our pooled sample, subjects with previous lab research experience declared an average of .471 of their earnings (Std. Dev. .442, median .428, $N=2,104$), while first-timers declared an average of .663 (Std. Dev. .416, median 1, $N=1,152$).¹¹ *T*-tests show this difference is significant at the $p<.01$ level. This result holds even at the level of the six individual countries where the experiments were conducted (Table S6). The column “Difference” reports the disparity in compliance rates between first-timers and experienced subjects. In each lab country, compliance rates average higher among first-timers versus experienced subjects. Denmark is an exception, possibly due to small sample size and rarity of first-timers.

These results are confirmed by regression analyses distinguishing between extensive and intensive margins, and controlling for observable characteristics (*Age*; *Male*; *Rows*), lab country and region fixed effects. Table 1 reports estimates of linear models specified in Equation 1. In each panel, the key explanatory variable is the dummy *Experience*. The dependent variable in Panel **A** is *Compliance Rate*: the ratio of subjects’ declared income to their total earnings. In Panels **B** and **C**, we investigate experience effects on the extensive and intensive margins of compliance rates, respectively. For each panel, we report estimates from five regression models: Column (1) reports estimates of the basic model with observations pooled by lab countries; Column (2) adds control variables (*Age*; *Male*; *Rows*); Column (3) includes the interactions *Age#Experience* and *Male#Experience*; Column (4) adds lab country fixed effects (“FE”); Column (5) replaces lab region FE with lab country FE. Standard errors are robust to heteroskedasticity and clustered at the individual level.

Table 1 HERE

In all regressions, the coefficient of *Experience* is negative and statistically significant. As shown in Column 1 of Panel **A**, any previous participation in a laboratory experiment is associated with an average 19.2% decrease in subjects’ tax compliance. This result is robust to controls over observable characteristics (Column 2), interactions (Column 3), lab country FE (Column 4), and lab region FE (Column 5). Panels **B** and **C** further show that previous laboratory experience is negatively associated both with the probability of being tax compliant (17.7% reduction, shown in Panel **B**) and the extent of subjects’ tax compliance (7.3% reduction, shown in Panel **C**). These results hold even when considering each country separately (Table S9 in the Supplemental), although Denmark is again exceptional.

As a robustness check, we analyse whether inexperienced subjects who are similar to experienced ones in terms of age, gender, and number of rows correctly copied in the clerical task differ in average compliance behaviour. We estimate Average Treatment Effect (ATE) and Average Treatment Effect on Treated (ATT) on *Experience* by

¹¹ Note that the total number of observations here is 3,256 (instead of 3,266) because ten subjects earned zero points in the first round of the clerical task. This has produced ten missing values.

propensity-score matching. Results are robust: the coefficient of *Experience* is negative and statistically significant at $p < .001$, for ATE and the ATT.

4.2 How Much Does Experience Matter?

Table 2 shows the effect of varying experience levels on compliance. The key explanatory variable is now ordinal: *Experience* here represents the number of times experienced subjects previously participated in experiments. The dependent variable in Panel **A** is Compliance Rate. In Panel **B**, it is the probability of being tax compliant (i.e. a dummy variable coded as 1 if Compliance Rate > 0). In Panel **C**, it is the extent of subjects' tax compliance (conditional upon Compliance Rate > 0). Each panel reports estimates from five regression models: Column 1 is the basic model with observations pooled across lab countries; Column 2 adds control variables (Age; Male; Rows); Column 3 adds the interactions *Age#Experience* and *Male#Experience*; Column 4 adds lab country fixed effects (FE); Column 5 replaces lab region FE with lab country FE. In all regressions, standard errors are robust and clustered at the individual level.

Table 2 HERE

The coefficient of *Experience* is negative and statistically significant at $p < .05$ (or better) in Columns 1 and 2 of Panel **A**, and Columns 1, 2, 3, and 4 of Panel **B**.¹²

In Table 3, we conduct OLS regressions to estimate the *incremental* effects of each laboratory experience on subjects' compliance. The estimates suggest a threshold model, in which behaviour only changes significantly after subjects have participated in more than two laboratory experiments. Among subjects who have participated in three to five prior lab studies, the influence of past experience is consistently negative and significant across all five models. For subjects who have participated in six or more studies, the impact of prior experience is negative in all models, but significant only in two (Columns 1 and 2).

Table 3 HERE

5. Conclusion

Our results show past experience makes a significant difference in laboratory subjects' behaviour. In our tax compliance experiments, first-timers were significantly more compliant than others. Previous experimental experience had a decisively negative impact on compliance at both the extensive and intensive margins. The findings are robust to controls for personal characteristics, as well as fixed effects by lab country and region. Results are consistent across alternative specifications (linear regressions; propensity score matching). Moreover, they hold even when considering each of the six countries independently.

These findings suggest biased results may occur in research that rewards cheating but does not control for subjects' prior experience with laboratory experiments. Repeat subjects

¹² Instead, *Experience* is insignificant in Panel **C**, where the DV is compliance. Future research could investigate whether this is due to low variance in experience levels within regions.

adjust their behaviour to reap rewards from cheating. *Their choices are consistently and significantly different from those of subjects entering the lab for the first or second time*: less-experienced subjects do not usually know in advance that they can earn money by behaving selfishly or dishonestly in public goods games.

The distinctive behaviour patterns of these two groups—experienced vs inexperienced experimental subjects—may affect many studies; it is therefore important that tax compliance experiments control for this variable going forward. Failure to do so may produce misleading results, as our own data show. Indeed, omitting prior experimental experience from data analyses or recruitment processes may limit not just external but also *internal* validity of results, when experience is not balanced across experimental treatments.

We acknowledge three limitations to our analysis. First, our data only show a correlation between subjects' behaviour in the lab and their record of past participation in experiments; our experiments were not designed to assess underlying reasons for this. The behavioural difference we observe between experienced and inexperienced subjects may be due to self-selection: subjects who enrol in multiple studies, and thus become more experienced, may view experiments primarily as an opportunity to earn money, leading to more frequent cheating on their part (Casari et al. 2007; Guillén and Veszteg 2012). This interpretation is consistent with the recent findings of Schmidt et al. (2020). Alternatively, subjects' preferences may change because of repeated participations (Brosig-Koch et al. 2017). Future research is needed to disentangle those factors and cleanly isolate those driving the behavioural difference between experienced vs inexperienced subjects.

A second limitation of this study is that our instrument for measuring experience is insufficiently specific. Future studies should examine more closely the thresholds for changes in participants' behaviour. It would be particularly valuable to identify the mechanisms driving change and stability: why do individuals adjust their behaviour after their third experimental research participation, but not before? Possible explanations include self-selection: some individuals may be more inclined to return to the laboratory once they realize that they can adjust their behavioural strategies to increase personal gain.

Finally, our findings are limited by a lack of data about the types of experiments our subjects participated in previously. Recent evidence suggests this may be significant: for example, prior experience in social dilemma games may have a stronger negative effect on tax compliance than does experience with games centered on market power (Dengler-Roscher et al. 2018; Conte et al. 2019). Future research should examine whether and how different types of experimental games impart different lessons to subjects.

We conclude with a couple of final remarks. First, drawing on a large, geographically diverse sample has both benefits and potential costs. Even when experiments share a design, subject pools and laboratories will differ, introducing unobservable characteristics and sources of variation. This deserves further research attention. Second, we acknowledge that the levels of subject responses (in the context here, the level of compliance) may differ by subject type (in the context here, experienced vs inexperienced subjects), even if the

responses of the different subject types to policy parameters is largely the same. For the issue of external validity, these comparative-static-type responses, which we are not able to examine here with our data, are of utmost importance to be investigated in future studies.

Despite these limitations, we believe this study offers significant implications for future tax compliance research and other public goods experiments in which cheating pays off. First, researchers using this design should consider recruiting first-timers as laboratory research subjects (e.g., Huck et al. 2004; Lohse 2016).¹³ If this is impractical, researchers should at least *control* in their empirical analyses for subjects' past experiences in experiments. Indeed, it is increasingly common for data on this to be gathered in post-experimental questionnaires (e.g. Houser et al. 2012; Rand and Kraft-Todd 2014; Rand et al. 2014; Chaudhuri et al. 2016; Kesternich et al. 2018; Guerra and Harrington 2018; Weimann and Brosig-Koch 2019; Jouxte 2019).

Increasingly, researchers require that subjects have no prior experience with laboratory experiments, or at least no prior experience with the specific experiment being conducted (e.g., Sarin and Weber, 1993; Chowdhury et al. 2017). The practice suggests growing recognition that experience matters and that researchers are taking steps to reduce any potential impact. This aligns with the main take-away message from our paper, on the necessity to control for subjects' prior experience whenever possible. Finally, laboratory administrators should periodically and frequently renew their labs' subject pool. This practice has already been adopted by a few internationally-recognized research institutions, such as the Laboratory for Research in Behavioural Experimental Economics (LINEEX) at the University of Valencia.

Funding. The authors acknowledge the Danish Research Council for its support of the “Mind the Gaps” research project (Grant 4003-00026B), the European Research Council for its support of the “Willing to Pay?” research project (Advanced Grant, Agreement N°295675–WillingToPay), and the support of the Tax Administration Research Centre (TARC).

Acknowledgements. The authors are grateful to James Alm, Yongzheng Liu, Andre Hartmann, Benno Torgler, Christoph Kogler, and the participants of the 2022 International Workshop on “Economic and Behavioral Aspects of Tax Compliance” for useful comments.

¹³ Meyers et al. (2020) make a similar suggestion for experiments using online platforms.

References

- Alm, J. (2019). What motivates tax compliance? *Journal of Economic Surveys*, 33(2), 353-388.
- Alm, J., Bloomquist, K. M., McKee, M. (2015). On the external validity of laboratory tax compliance experiments. *Economic Inquiry*, 53(2), 1170-1186.
- Alm, J., Malézieux, A. (2020). 40 years of tax evasion games: A meta-analysis. *Experimental Economics*, 1-52.
- Al-Ubaydli, O., List, J. A. (2013). On the Generalizability of Experimental Results in Economics. In Frechette, G., Schotter, A., *Methods of Modern Experimental Economics*, Oxford University Press
- Al-Ubaydli, O., List, J. A., Suskind, D. L. (2017). What Can We Learn from Experiments? Understanding the Threats to the Scalability of Experimental Results. *American Economic Review P&P*, 107(5), 282–286.
- Andrighetto, G., Zhang, N., Ottone, S., Ponzano, F., D’Attoma, J., Steinmo, S. (2016). Are some countries more honest than others? Evidence from a tax compliance experiment in Sweden and Italy. *Frontiers in Psychology*, 7.
- Benndorf, V., Moellers, C., and Normann, H. T. (2017). Experienced vs. inexperienced participants in the lab: do they behave differently? *Journal of the Economic Science Association*, 3(1), 12-25.
- Benson, B. L., Faminow, M. D. (1988). The impact of experience on prices and profits in experimental duopoly markets. *Journal of Economic Behavior & Organization*, 9(4), 345-365.
- Bolton, G. E. (1991). A comparative model of bargaining: Theory and evidence. *The American Economic Review*, 1096-1136.
- Brosig-Koch, J., Riechmann, T., Weimann, J. (2017). The dynamics of behaviour in modified dictator games. *PloS ONE*, 12(4): e0176199.
- Casari, M., Ham, J. C., and Kagel, J. H. (2007). Selection bias, demographic effects, and ability effects in common value auction experiments. *American Economic Review*, 97(4), 1278-1304.
- Chaudhuri, A., Li, Y., Paichayontvijit, T. (2016). What’s in a frame? Goal framing, trust and reciprocity. *Journal of Economic Psychology*, 57, 117-135.
- Chiapello, M. (2018). BALANCETABLE. *Technical Report*, Boston College, Department of Economics.
- Chowdhury, S. M., Jeon, J. Y., and Saha, B. (2017). Gender differences in the giving and taking variants of the dictator game. *Southern Economic Journal*, 84(2), 474-483.
- Conte, A., Levati, M. V., and Montinari, N. (2019). Experience in Public Goods Experiments. *Theory and Decision*, 86(1), 65-83.
- Czibor, E., Jimenez-Gomez, D., List, J. A. (2019). The Dozen Things Experimental Economists Should Do (More Of), *Southern Economic Journal*, 86 (2): 371-432
- Dengler-Roscher, K., Montinari, N., Panganiban, M., Ploner, M., Werner, B. (2018). On the malleability of fairness ideals: Spillover effects in partial and impartial allocation tasks. *Journal of Economic Psychology*, 65, 60-74.
- Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2), 171-178.
- Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1), 114-125.

- Guerra, A., and Harrington, B. (2018). Attitude–behaviour consistency in tax compliance: A cross-national comparison. *Journal of Economic Behavior & Organization*, 156, 184-205.
- Guillén, P., and Veszteg, R. F. (2012). On “lab rats”. *The Journal of Socio-Economics*, 41(5), 714-720.
- Houser, D., Vetter, S., and Winter, J. (2012). Fairness and cheating. *European Economic Review*, 56(8), 1645-1655.
- Huck, S., Normann, H. T., and Oechssler, J. (2004). Two are few and four are many: number effects in experimental oligopolies. *Journal of Economic Behavior & Organization*, 53(4), 435-446.
- Isaac, M., Walker J. M., and Thomas S. H. (1984). Divergent Evidence on Free Riding: An Experimental Examination of Possible Explanations. *Public Choice*, 43, 113-149.
- Jouxte, J. (2019). Voluntary contributions of time: Time-based incentives in a linear public goods game. *Journal of Economic Psychology*, 75, 102139.
- Kesternich, M., Lange, A., Sturm, B. (2018). On the performance of rule-based contribution schemes under endowment heterogeneity. *Experimental Economics*, 21(1), 180-204.
- Levitt, S. D., List, J. A. (2007a). What do laboratory experiments measuring social preferences reveal about the real world?. *Journal of Economic Perspectives*, 21(2), 153-174.
- Levitt, S. D., List, J. A. (2007b). On the generalizability of lab behaviour to the field. *Canadian Journal of Economics/Revue canadienne d'économie*, 40(2), 347-370.
- List, J. A. (2020). Non est disputandum de generalizability? A glimpse into the external validity trial. *NBER Working Paper*, (w27535).
- List, J. A., Lucking-Reiley, D. (2000). Demand reduction in multiunit auctions: Evidence from a sports card field experiment. *American Economic Review*, 90(4), 961-972.
- Lohse, J. (2016). Smart or selfish—When smart guys finish nice. *Journal of Behavioral and Experimental Economics*, 64, 28-40.
- Lynch, J. G. Jr., (1982). On the external validity of experiments in consumer research. *Journal of Consumer Research*, 9, 225–240.
- Marwell, G., and Ames, R. E. (1979). Experiments on the Provision of Public Goods. I. Resources, Interest, Group Size, and the Free-Rider Problem. *American Journal of Sociology*, 84(6), 1335-1360.
- Marwell, G., and Ames, R. E., (1980). Experiments on the provision of public goods. II. Provision points, stakes, experience and the free rider problem. *American Journal of Sociology*, 85(4), 926-937.
- Matthey, A., Regner, T. (2013). On the independence of history: experience spill-overs between experiments. *Theory and Decision*, 75, 403-419.
- Medda, T., Pelligra, V., and Reggiani, T. (2021). Lab-Sophistication: Does Repeated Participation in Laboratory Experiments Affect Pro-Social Behaviour?. *Games*, 12(1), 18.
- Meyers, E. A., Walker, A. C., Fugelsang, J. A., Koehler, D. J. (2020). Reducing the number of non-naïve participants in Mechanical Turk samples. *Methods in Psychology*, 3, 100032.

- Ottone, S., Ponzano, F., Andrighetto, G. (2018). Tax compliance under different institutional settings in Italy and Sweden: an experimental analysis. *Economia Politica* 1–36.
- Pampel, F., Andrighetto, G., Steinmo, S. (2019). How institutions and attitudes shape tax compliance: a cross-national experiment and survey. *Social Forces*, 97(3), 1337-1364.
- Rand, D. G. (2018). Nonnaïvety may reduce the effect of intuition manipulations. *Nature human behaviour*, 2(9), 602-602.
- Rand, D. G., Kraft-Todd, G. T. (2014). Reflection does not undermine self-interested prosociality. *Frontiers in Behavioral Neuroscience*, 8, 300.
- Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., Green, J. D. (2014). Social Heuristics Shape Intuitive Cooperation. *Nature Communications*, 5, Article number: 3677.
- Sarin, R. K., and Weber, M. (1993). Effects of ambiguity in market experiments. *Management Science*, 39(5), 602-615.
- Schmidt, R., Schwierén, C., Sproten, A., (2020) Norms in the lab: Inexperience vs experience participants., *Journal of Economic Behavior & Organization*, 173, 239-255.
- Torgler, B. (2002). Speaking to theorists and searching for facts: Tax morale and tax compliance in experiments. *Journal of Economic Surveys*, 16(5), 657-683.
- Weimann, J., and Brosig-Koch, J. (2019). The Experiment from a Statistical Perspective. In: *Methods in Experimental Economics*, pp 169–28, Eds: Weimann, J., and Brosig-Koch, J. Springer Texts in Business and Economics. Springer, Cham. https://doi.org/10.1007/978-3-319-93363-4_4.
- Winer, R. S. (1999). Experimentation in the 21st century: The importance of external validity. *Journal of the Academy of Marketing Science*, 27, 349–358.
- Xue, L., Sitzia, S., and Turocy, T. L. (2017). Mathematics self-confidence and the “prepayment effect” in riskless choices. *Journal of Economic Behavior & Organization*, 135, 239-250.
- Zhang, N., Andrighetto, G., Ottone, S., Ponzano, F., Steinmo, S. (2016). ‘Willing to pay?’ Tax compliance in Britain and Italy: An experimental analysis. *PLoS One* 11 (2).

Tables

Table 1: Effect of Experience on Compliance

| (A) DV: Compliance Rate | (1) | (2) | (3) | (4) | (5) |
|-----------------------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| Experience (0/1) | -.192*** (.016) | -.186*** (.015) | -.229*** (.052) | -.146*** (.016) | -.151*** (.017) |
| Intercept | .664*** (.012) | .756*** (.036) | .789*** (.048) | .487*** (.055) | .723*** (.050) |
| N | 3,256 | 3,252 | 3,252 | 3,252 | 3,252 |
| (B) DV: Pr(Compliance Rate>0) | | | | | |
| Experience (0/1) | -.177*** (.016) | -.167*** (.015) | -.080 (.049) | -.124*** (.016) | -.127*** (.017) |
| Intercept | .809*** (.012) | .962*** (.036) | .626*** (.068) | .651*** (.062) | .896*** (.049) |
| N | 3,266 | 3,262 | 3,262 | 3,262 | 3,262 |
| (C) DV: Compliance (Compliance>0) | | | | | |
| Experience (0/1) | -.073*** (.013) | -.081*** (.013) | -.140** (.044) | -.071*** (.014) | -.075*** (.015) |
| Intercept | .821*** (.010) | .797*** (.032) | .739*** (.072) | .693*** (.067) | .825*** (.043) |
| N | 2,258 | 2,255 | 2,255 | 2,255 | 2,255 |
| Controls | NO | YES | YES | YES | YES |
| Interactions | NO | NO | YES | NO | NO |
| Lab Country FE | NO | NO | NO | YES | NO |
| Lab Region FE | NO | NO | NO | NO | YES |

Notes: The table reports LPM estimates, with data pooled by lab country and robust standard errors clustered at the individual level. The dependent variable (DV) in Panel (A) is *Compliance Rate*, operationalized as the ratio of subjects' declared income to their total earned income; the variable takes values between 0 and 1, inclusive. The DV in Panel (B) is *Compliance Rate* operationalized as a dummy variable coded as 1 if *Compliance Rate*>0, and 0 otherwise. The DV in Panel (C), is *Compliance Rate* conditional upon *Compliance*>. The key explanatory variable is the dummy variable *Experience*, coded as 1 if a subject has previously participated in other experiments prior to the current one, and 0 otherwise. Column (1) reports estimates of the basic model with observations pooled by lab countries; Column (2) adds control variables (Age; Male; Rows); Column (3) includes the interactions Age#Experience and Male#Experience; Column (4) adds lab country fixed effects ("FE"); Column (5) replace lab region FE to lab country FE. Abbreviations: DV for Dependent Variable; FE for Fixed Effects.

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 2: Effect of Experience Level on Compliance

| | (1) | (2) | (3) | (4) | (5) |
|---|--------------------|-------------------|-------------------|-------------------|-------------------|
| (A) DV: Compliance Rate | | | | | |
| Experience (ordinal) | -.026** (.009) | -.023* (.009) | -.017 (.009) | -.017 (.009) | -.010 (.010) |
| Intercept | .533*** (.025) | .601*** (.057) | .561*** (.061) | .561*** (.061) | .575*** (.068) |
| N | 1603 | 1603 | 1603 | 1603 | 1603 |
| (B) DV: Pr(Compliance Rate>0) | | | | | |
| Experience (ordinal) | -.034*** (.010) | -.030** (.010) | -.022* (.010) | -.022* (.010) | -.013 (.010) |
| Intercept | .718*** (.026) | .847*** (.059) | .766*** (.064) | .766*** (.064) | .771*** (.069) |
| N | 1608 | 1608 | 1608 | 1608 | 1608 |
| (C) DV: Compliance (Compliance>0) | | | | | |
| Experience (ordinal) | -.000 (.009) | -.002 (.009) | -.003 (.009) | -.002 (.009) | -.003 (.009) |
| Intercept | .744*** (.023) | .727*** (.055) | .756*** (.060) | .763*** (.062) | .756*** (.060) |
| N | 1018 | 1018 | 1018 | 1018 | 1018 |
| Controls | NO | YES | YES | YES | YES |
| Interactions | NO | NO | YES | NO | NO |
| Lab Country FE | NO | NO | NO | YES | NO |
| Lab Region FE | NO | NO | NO | NO | YES |

Notes: The table reports OLS estimates, with data pooled by lab country and robust standard errors clustered at the individual level. The dependent variable (DV) in Panel (A) is *Compliance Rate*, operationalized as the ratio of subjects' declared income to their total earned income; the variable takes values between 0 and 1, inclusive. The DV in Panel (B) is *Compliance Rate* operationalized as a dummy variable coded as 1 if *Compliance Rate*>0, and 0 otherwise. The DV in Panel (C), is *Compliance Rate* conditional upon *Compliance*>0. The key explanatory variable is the variable *Experience*, which measures the number of times an experienced subject previously participated in experiments; this variable takes integer values from 1 through 4, where 1 represents one prior laboratory experience, 2 represents two prior laboratory experiences, 3 represents cases of three to five past laboratory experiences, and 4 represents cases of six or more past laboratory experiences. Column (1) reports estimates of the basic model with observations pooled by lab countries; Column (2) adds control variables (Age; Male; Rows); Column (3) includes the interactions Age#Experience and Male#Experience; Column (4) adds lab country fixed effects ("FE"); Column (5) replace lab region FE to lab country FE. Abbreviations: DV for Dependent Variable; FE for Fixed Effects.

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 3: Effects of Each Experience Level on Compliance

| DV: Compliance Rate | (1) | (2) | (3) | (4) | (5) |
|---------------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| Two Experiences | -.012 (.031) | -.001 (.030) | -.004 (.030) | -.004 (.030) | .003 (.030) |
| Three to Five Experiences | -.071* (.030) | -.067* (.028) | -.066* (.028) | -.066* (.028) | -.060* (.028) |
| Six or More Experiences | -.067* (.030) | -.057* (.028) | -.036 (.029) | -.036 (.029) | -.012 (.031) |
| Intercept | .507*** (.019) | .579*** (.057) | .551*** (.060) | .551*** (.060) | .570*** (.068) |
| N | 1603 | 1603 | 1603 | 1603 | 1603 |
| Controls | NO | YES | YES | YES | YES |
| Interactions | NO | NO | YES | NO | NO |
| Lab Country FE | NO | NO | NO | YES | NO |
| Lab Region FE | NO | NO | NO | NO | YES |

Notes: The table reports OLS estimates, with data pooled by lab country and robust standard errors clustered at the individual level. The dependent variable (DV) is *Compliance Rate*, operationalized as the ratio of subjects' declared income to their total earned income; the variable takes values between 0 and 1, inclusive. The key explanatory variables are dummy variables representing the number of times an experienced subject previously participated in experiments. The baseline is *One Experience*. Column (1) reports estimates of the basic model with observations pooled by lab countries; Column (2) adds control variables (Age; Male; Rows); Column (3) includes the interactions Age#Experience and Male#Experience; Column (4) adds lab country fixed effects ("FE"); Column (5) replace lab region FE to lab country FE. Abbreviations: DV for Dependent Variable; FE for Fixed Effects.

* $p < .05$, ** $p < .01$, *** $p < .001$