Catch Me If You Can: Testing the reduction of compound lotteries axiom in a tax compliance experiment

Michele Bernasconi
Dipartimento di Economia, Università "Ca' Foscari", Cannaregio 873, Venezia, Italy
E-mail: bernasconi@unive.it

Juliana Bernhofer
Faculty of Economics and Management, Free University of Bozen-Bolzano, Italy
E-mail: juliana.bernhofer@unibz.it

Abstract

We test the Reduction of Compound Lotteries Axiom (RCLA) in an experiment on tax compliance. We disentangle the compound probability of audit and detection and test it against a more realistic situation with a probability of audit and a subsequent probability of detection. Various experiments have shown that abstract framing often leads to violations of the RCLA; our framed set-up also reveals statistically significant departures from the RCLA. We find that subjects comply more in the two-stage lottery set-up than they do in the one-stage equivalent and that violations are compatible with subjects applying different weighting functions to one-stage lotteries versus two-stage lotteries: inverse S-shaped - likelihood insensitive - in the former set-up; and significantly less insensitive in the latter. Violations substantially decline with subjects' experience but not at the same rate for all groups of subjects.

Keywords: reduction of compound lotteries, rank dependent utility, bomb crater effect, laboratory experiments

JEL Classification: H26, D03, C91, J16

[*]We acknowledge the financial support of the Center for Experimental Economics at Fondazione Università Ca' Foscari and GAM investment. We thank two anonymous referees and the editor for very helpful suggestions.

1 Introduction

How does the reduction of compound lotteries axiom (RCLA), one of the fundamental underpinnings of expected utility theory, relate to the field of tax compliance? The answer to this question lies in considering as two independent entities the probability of a random audit and the probability of the evaded tax's being detected. Is the taxpayer indifferent between facing a single audit-detection probability and a situation in which audit and detection are presented separately? The underlying assumption of a theoretical model that considers only a single probability p must be either that detection of the undeclared income is guaranteed in the case of an audit, or that taxpayers respect the RCLA. The first relies on the strong conjecture that there is no information asymmetry in the audit procedure, whereas the latter refers to a specific attitude of the taxpayer, which we test in the present study.

The first in the literature of tax compliance to provide a theoretical framework were Allingham and Sandmo (1972) and Yitzhaki (1974) who developed adapted versions of Becker's (1968) formulation of crime and punishment. In the original Beckerian model, the decision concerning whether to commit a crime is driven by economic considerations, with potential outlaws weighing expected benefits in terms of monetary and psychic income against the expected costs of a pecuniary sanction. Those considering such options are assumed to be *homines oeconomici*, acting in a self-interested and perfectly rational manner. Over the years, the approach has been extended in many directions, both nesting the model into a broader social welfare analysis that seeks to provide an optimal public policy response to evasion behaviour, and adapting the approach to new developments, including many obtained by the behavioural economic literature (reviews and references in various surveys, e.g. Andreoni et al., 1998; Kirchler, 2007; Hashimzade et al., 2013; Alm, 2019). However, to the best of our knowledge, none of these analyses have directly tested the reduction axiom in that context.

The RCLA classifies the decision maker as rational and able to reduce multi-stage compound lotteries mentally into single-stage lotteries by multiplication of probabilities. However, numerous studies, such as Bar-Hillel (1973), Kahneman and Tversky (1979), Bernasconi (1992), and Harrison et al. (2015), have outlined that the axiom does not always hold in experiments. Therefore, a number of alternative approaches have been developed (including Kahneman and Tversky, 1979; Segal, 1990; Dillenberger, 2010). Deviations from expected utility often occur because people's perceptual and cognitive limitations prevent them from understanding probabilities. Rank-dependent utility, originally introduced by Quiggin (1982) and studied by many others (e.g. references in Diecidue and Wakker, 2001), is among the class of non-expected utility models that can account for violations of the RCLA, given that the probabilities attached to outcomes are warped in a nonlinear mental process.

Generally, laboratory experiments that test the reduction axiom are embedded in a neutral setting in which subjects face a lottery choice problem that is presented to them as a gamble. While a context-free approach might be alluring when testing axiomatic decision theories, as it allows fundamental principles to be controlled, the lack of a realistic framing might fail to provide insights for applications of those models to specific contexts (e.g. discussions in Alm et al., 1992, and Abbink and Hennig-Schmidt, 2006, versus Laury and Taylor, 2008, and Voors et al., 2012). Our investigation takes a step towards building an experimental assessment of the RCLA to reduce the gap between axiomatic analyses and economic applications.

In our study, we frame the lottery decision as a choice regarding tax compliance and divide the overall probability of being audited and detected into the probability of being audited and the separate probability of being detected. Our null hypothesis is that participants behave according to models of rational choice. If this hypothesis were to be accepted, we would not find any treatment effects in the single-stage set-up that differ from those of the two-stage equivalent. Our results show that, in the first 14 periods of the experiment, subjects do not reduce lotteries according to the reduction axiom, while they comply more in the two-stage treatment, revealing an aversion to waiting for detection. In terms of rank-dependent utility, our results are compatible with an inverse S-shaped weighting function with likelihood insensitivity in the one-stage treatment and a less insensitive weighting function that represents a more risk averse and pessimistic subject in the two-stage treatment.

We also find that subjects exhibit a different response to audit and detection in post-audit and post-fine responses. Running a second phase of the experiment for another 14 periods reveals that violations of the reduction axiom substantially decline, even if, depending on the field of study, some experimental subjects still violate the axiom. We document the underlying learning process, but conclude that it does not allow us to dismiss lightly the relevance of the reduction violations in the field of tax compliance.

In section 2 we review the classic model of tax evasion decision. Then we consider the literature on the reduction axiom, focussing on rank-dependent utility as a notable non-expected utility alternative that can explain violations of the reduction axiom in the tax compliance case. In section 3 the experiment is described in detail and results are given in section 4. Section 5 presents the concluding discussion.

2 Theoretical background

2.1 The economics-of-crime approach to tax evasion

In Allingham and Sandmo (1972)'s classic economics-of-crime approach to tax evasion, a typical taxpayer with gross income $y > 0$ is required to pay taxes on reported income according to a proportional tax rate τ. The tax authority does not observe the taxpayer's gross income directly but, with probability $p \in (0, 1)$, the tax authority conducts an audit that detects any concealment of income with certainty. If audited and found to have concealed part of her income, the taxpayer must pay the taxes evaded plus a sanction. In Allingham and Sandmo (1972), sanctions were computed on evaded income, whereas Yitzhaki (1974) noted that sanctions are commonly determined as a percentage of the evaded taxes. In the combined Allingham-Sandmo-Yitzhaki model, the taxpayer maximizes the following expected utility (EU) function with respect to the income x to declare:

$$EU = (1 - p)U(y - \tau x) + pU(y - \tau x - \tau\phi(y - x)) \tag{1}$$

where $\phi > 1$ is the fine rate, proportional to the evaded tax.

Solving the decision problem provides several insights on the determinants of the tax evasion decision, particularly the sensible result that compliance depends on enforcement. A wide body of literature on the impact of sanctions and the probability that they will be imposed has emerged.[1] Several empirical studies have been conducted using various approaches, including natural field data, laboratory data, and data from field experiments. In a recent survey Alm (2019, p. 367) reported an estimated elasticity between declared income and audit rate of between 0.2-0.4, although this small

[1] At the same time, some features of the basic model in equation (1) are considered not fully convincing. Accordingly, the literature has extended the approach in several directions to generalise it, to accommodate elements like endogenous income, progressive tax rules, other institutional settings, and to enrich the model by introducing as determinants of the compliance decision various ethical and/or psychological factors (review and references to the literature in e.g. Andreoni et al., 1998; Kirchler, 2007; Hashimzade et al., 2013).

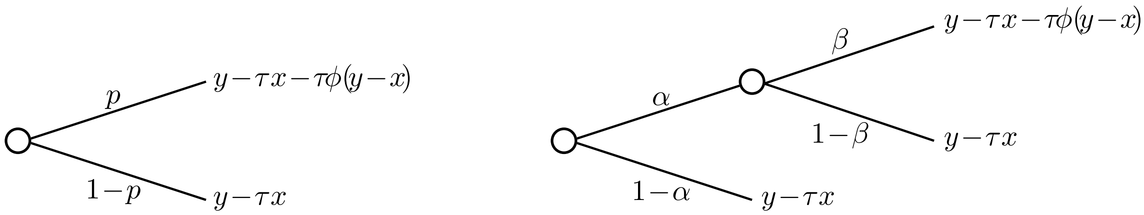
level of elasticity varies across studies. Alm (2019) also remarked that "both laboratory experiments and field data have generally found that the impact of increased audits is non-linear, so that the deterrent effect seems to diminish (and may even be reversed) with higher audit rates." Therefore, not only objective audit rates but also their perception affects compliance behaviour.

Cognitive considerations also seem to matter for post-audit responses. Mittone (2006), Maciejovsky et al. (2007), Guala and Mittone (2005), Mittone et al. (2017) studied the lagged effect of previous audits on future compliance behaviour and all found evidence of negative post-audit responses. Mittone (2006) called this effect the 'bomb-crater effect' since it recalls a fallacy that a bomb crater is a good place to hide in wartime because bombs do not hit twice in the same place. An urge for 'loss repair', another explanation for negative post-sanction responses, has been mentioned in Andreoni et al. (1998) and studied in depth in Maciejovsky et al. (2007).

Part of the literature has focussed on the institutional arrangement of the audit process, including analyses of strategic, rather than random, audit rules, the effect of delayed audit feedback, and responses to information on forthcoming audit probabilities (see reviews in Andreoni et al., 1998; Alm, 2019). As it is not always the case that an investigation automatically leads to detection of the full amount of undeclared income (e.g., Feinstein, 1991; Snow and Warren, 2005; Kleven et al., 2011), Rablen (2014) moved away from fine rates, arguing that the trade-off between audit probability and audit effectiveness is a more policy-relevant aspect of tax compliance, given the limited ability of tax authorities to influence the fine rate itself. Given this 'wiggle room', a well-funded authority should focus on performing a higher number of less rigorous audits, whereas authorities with a lower per-capita budget should increase the effectiveness of each audit.

The relevance of the monitoring process to tax compliance also underpins our idea. We focus on the impact of overall detection probability's involving several stages on peoples' attitudes toward reporting income. Accordingly, we represent the tax monitoring action as a two-stage process that distinguishes a first stage characterised by a pure audit probability α and a second stage that is characterised by a pure detection probability β. Figure 1 illustrates the difference between the one-stage and two-stage processes. Clearly, when $p = \alpha\beta$, the two processes are probabilistically equivalent. However, for the taxpayer to behave equivalently in the two situations, we must ap-

Figure 1: One-stage versus two-stage monitoring process



a) One-stage process (O_{1S}) b) Two-stage process (O_{2S})

peal to the RCLA of the expected utility model. While acknowledging the higher level of realism of fractional detection models, as in Rablen (2014), we take an all-or-nothing approach to compare our results directly with the classic model of tax evasion and to get a more clear-cut idea of the underlying behavioral mechanisms with respect to the reduction axiom.

2.2 The reduction of compound lotteries axiom and rank-dependent utility

The RCLA goes back to the early derivations of expected utility (EU) by von Neumann and Morgenstern (1944). It was then taken up in all other EU derivations, either explicitly or implicitly.

There is evidence that people violate the axiom in several contexts. Kahneman and Tversky (1979) provided an early example of violations of RCLA occurring when people simplify a multistage decision problem by evaluating each stage in isolation, rather than by considering the whole problem. Violations of RCLA can be due to cognitive limitations (Nebout and Dubois, 2014; Harrison et al., 2015; Prokosheva, 2016; and references therein) and related to a sense of anxiety or psychological displacement that arises from the sequential resolution of uncertainty (Palacios-Huerta, 1999). In such cases, decision makers may prefer any two-stage lottery to be resolved in a single stage (Dillenberger, 2010). Others found mixed directions of the violations. For example, Zimmermann (2014) found that about half of the subjects that violated RCLA showed preference for one-shot resolution of uncertainty, and about half showed preference for more gradual resolution.

More generally, violations of RCLA imply violations of the EU and need to be handled by non-expected utility models. A popular non-expected utility model that accounts for several phenomena outside the domain of expected utility is rank-dependent utility (RDU). The class of RDU models was introduced in the literature by Quiggin (1982) and developed further by several authors (see Diecidue and Wakker, 2001, for an intuitive derivation of the theory). RDU differs from the EU because it weighs the outcome probabilities of a lottery nonlinearly, with the weight attached to each outcome depending only on the probability of the outcome and its ranking position.

Formally, in RDU the outcomes of a one-stage lottery $X = (x_1, p_1, ..., p_n, x_n)$ are ordered from highest to lowest: $x_1 > x_2 > ... > x_n$. The lottery is then evaluated as $RDU(X) = \sum_{i=1}^{n} \pi_i U(x_i)$, where π_i is the decision weight for outcome x_i:

$$\pi_i = w(p_1 + ... + p_i) - w(p_1 + ... + p_{i-1}) \tag{2}$$

for the nonlinear probability weighting function $w : [0, 1] \to [0, 1]$, increasing, onto, and with $w(0) = 0$ and $w(1) = 1$.[2] Risk aversion, that is, aversion to a mean-preserving spread, holds everywhere in the theory if and only if u is concave and w is convex on all domain $[0, 1]$.

The shape of the weighting function is pivotal for predictions of the theory in relation to people's perceptions of probabilities. For this reason, RDU has already been applied to the study of tax evasion decisions (Hashimzade et al., 2013, for references);[3] as far as we know, however, not to distinguish between the one-stage and two-stage monitoring processes shown in Figure 1.

We discuss the shape of the weighting function after applying RDU to the tax compliance problem. To this end, we denote the taxpayer's net income in case of no-audit/no-sanction as $NC = y - \tau x$ and the taxpayer's net income in case of sanction as $C = y - \tau x - \tau\phi(y - x)$, and we write the two-stage lottery process in Figure 1 as $O_{2S} = ((C, \beta; NC, 1 - \beta), \alpha; NC, 1 - \alpha)$, and the one-stage process as $O_{1S} = (C, p; NC, 1 - p)$.

[2] In the early derivation of RDU, the weighting function was sometimes written as $g(p) = 1 - w(1 - p)$, in which case $\pi_i = g(p_i + ... + p_n) - g(p_{i+1} + ... + p_n)$. Clearly, the formulations $w(p)$ and $g(p)$ are equivalent, implying, among other things, that the weight for the worst outcome x_n is $\pi_n = g(p_n) = 1 - w(1 - p_n)$, and the weight for the best outcome x_1 is $\pi_1 = w(p_1) = 1 - g(1 - p_1)$. For these reasons, g is sometimes also called the 'badnews weighting function' and w the 'goodnews weighting function' (Diecidue and Wakker, 2001). Even though we use the form $g(q)$ for some expressions (see Proposition 1), the use of w is by far the most common in the literature, so we use it in the general discussion of RDU.

[3] In fact, more generally, the effect of non-linearity of audit probability on lie deterrence found in tax compliance literature can also be found in other monitoring processes, including for example for monitoring probability in information transmission literature (e.g. Behnk et al., 2018.)

In the one-stage lottery process, RDU can be applied directly to O_{1S}. Moreover, given $\alpha\beta = p$, the value of O_{2S} under RCLA is also equal to the value of O_{1S}, so we will refer to it interchangeably. The value is given by:

$$RDU(O_{1S}) = w(1-\alpha\beta)u(NC) + [1 - w(1-\alpha\beta)]u(C) = RDU_{RCLA}(O_{2S}) \tag{3}$$

On the other hand, if taxpayers violate RCLA, they can evaluate the two-stage lottery by separately weighing the probabilities of the various stages. Segal (1990) showed that this procedure is equivalent to a form of backward induction in which an individual values a multi-stage lottery by computing the certainty equivalents of each lottery stage one at a time (the certainty equivalent method, CEM). Applying the method to the two-stage lottery O_{2S} gives:

$$RDU_{CEM}(O_{2S}) = w(1-\alpha)u(NC) + [1-w(1-\alpha)][w(1-\beta)u(NC) + (1-w(1-\beta))u(C)] \tag{4}$$

Clearly, when $w(\cdot)$ is the identity function, RDU reduces to expected utility, and the two expressions $RDU_{RCLA}(O_{2S})$ and $RDU_{CEM}(O_{2S})$ are also equal. However, when $w(\cdot)$ is not the identity, $RDU_{RCLA}(O_{2S})$ and $RDU_{CEM}(O_{2S})$ will usually be different. The following proposition applies:

Proposition 1. Let incomes reported under RCLA and under CEM be given by x_{RCLA} and x_{CEM}, respectively. Then, denoting $g(q) = 1 - w(1-q)$:

$$x_{RCLA} \gtreqless x_{CEM} \iff g(\alpha\beta) \gtreqless g(\alpha)g(\beta) \tag{5}$$

Proof. In Appendix.

The proposition is an instance of a more general result obtained by Segal (1987, Lemma 4.1), who showed that, for a general weighting function g, the condition $g(x)g(y) < g(xy)$ is necessary and sufficient for the elasticity of g to be increasing (see also Segal, 1990, p. 367).

The elasticity of a function is clearly related to its concavity/convexity, even if neither property is strictly sufficient. In particular, when the function is convex, elasticity is more likely to be increasing, whereas when the function is concave, elasticity is more likely to be decreasing. Thus, in correspondence of g convex, which means concave w, it is more likely that $x_{CEM} < x_{RCLA}$; the opposite holds in correspondence of g concave, which means convex w.

As indicated above, the early literature on RDU especially emphasized the convexity of w in association with risk aversion. While it has also been occasionally documented by the empirical literature, more often the experimental evidence has suggested an inverse S-shape: namely, a weighting function concave for small probabilities and convex for high probabilities (see Diecidue and Wakker, 2001, and Wakker, 2010, for references to the large literature on the probability weighting function).

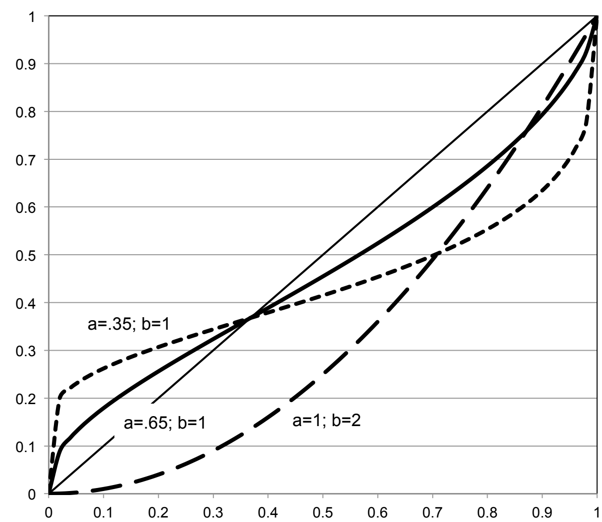
Concavity/convexity of the weighting function is also linked to the decision maker's optimism/pessimism. In particular, since concavity suggests that a decision maker increases the probability weight put on an outcome when the outcome improves its ranking position, concavity is associated with optimism. The symmetric argument is that convexity is linked to pessimism. An inverse S-shaped weighting function suggests that, at low probabilities, people are optimistic about high outcomes and pessimistic about low outcomes.

Another characteristic of an inverse-S weighting function is that it implies that a decision maker is oversensitive to probability changes when they occur close to the end points 0 and 1~~,~~ and is relatively insensitive toward changes in intermediate probabilities. This property is also termed likelihood insensitivity (e.g. Wakker, 2010). For example, it can explain the evidence, alluded to above, that the impact of audits is high at low probabilities, but it quickly diminishes with intermediate and high audit rates. Figure 2 shows examples of weighting functions with different degrees of likelihood insensitivity based on a parametric form proposed in Prelec (1998).

Studies have also shown variability in the probability weighting function. In fact, the weighting function is not a subjective probability but a distortion of a given probability (Gonzalez and Wu, 1999), so it can be affected by the contexts and conditions of perception, including the decision maker's expertise and experience with similar lotteries, and the realizations of these lotteries (Hogarth and Einhorn, 1990; Tversky and Wakker, 1995; Gayer, 2010). Clearly, then, if probabilistic sensitivity is also affected by the one-stage or two-stage resolution of lotteries, it can be another source of violations of RCLA.

To sum up, we test the reduction axiom in the context of tax compliance. Our null hypothesis is that decision makers behave in line with the RCLA by being indifferent between the single-stage and the two-stage lotteries (Figure 1). Our alternative hypothesis is that the RCLA is violated, in which case we would observe compliance rates that

Figure 2: Examples of weighting functions $w(q)$



The diagram shows three weighting functions based on the parametric form $w(p) = \exp(-b(-\ln(p))^a)$ from Prelec (1998). In the expression, the parameter a determines the curvature of the probability weighting function, with probability insensitivity decreasing when $a < 1$.

differ in the two environments. Although our focus is not on estimating the probability weighting function, the arguments presented above are useful in obtaining predictions in this case. Under a convex weighting function, which is consistent with a decision maker's pessimistic attitude, it is likely that compliance will be higher in the two-stage process than in the one-stage process; under an inverse S-shaped weighting function, which suggests concavity at small probabilities, the opposite is more likely. Another question with RDU concerns whether a decision maker's probabilistic sensitivity is the same under the one-stage monitoring process as it is under the two-stage monitoring process.

3 The experiment

We ran a tax compliance laboratory experiment using the software z-Tree (Fischbacher, 2007). We distinguished two main types of treatments: participants in the one-stage lottery control treatments faced a standard tax compliance decision given by a sequence of periods in which they decided how much gross income to report in a situation in which the audit rate p, the tax rate τ, the fine rate ϕ, and the gross income y were exogenously given; the two-stage lottery treatments were similar, except the audit probability α and the detection probability β were given separately and featured the same percentage p in compound terms as in the one-stage lottery treatments.

Gross income per period was drawn randomly from a discrete uniform distribution ranging from 80 to 160 *Experimental Currency Units* (ECU) in steps of 10, and total profits were converted at a rate of 180 ECU= 1€. We opted for a pay-off mechanism that rewards each single period over a random-lottery incentive procedure. Scholars like Starmer and Sugden (1991), Harrison and Swarthout (2014), and Harrison et al. (2015) have analysed problems that relate to biases that could arise with the use of the random-lottery incentive mechanism (RLIM), or the "1-in-K" payment method (where $K > 1$). The RLIM might create distortions when decision theory's basic axioms are tested experimentally, because a procedure that randomly draws one period to be paid out to the subjects at the end of the experiment adds an additional stage to the lottery choice task and might influence the subject's decision.[4] On the other hand, the "1-in-1" pay-off mechanism could introduce income effects. To minimize this risk, subjects are shown their total payoff only at the end of the experiment. In addition, our analysis controls for the accumulated experimental wealth up to the beginning of the current experimental period and for the subjects' income in the period itself. Having acknowledged the weaknesses of both methods, we find that the period-wise mechanism has no direct interference with the hypotheses we wish to test.

3.1 Experimental parameters

We conducted six sessions, in each of which participants played both the one-stage and the two-stage lottery treatments in two phases of fourteen periods each: phase 1 (periods 1-14) and phase 2 (periods 15-28). The compound probabilities were different in the two phases of each session. Table 1 provides an overview of the six sessions and the corresponding parameters. In all treatments, the tax rate τ and the fine rate ϕ were held constant at 30 percent and twice the evaded taxes, respectively, whereas the overall probability p of being detected varied between sessions and phases and was

[4] For example, in work about preference reversals and the independence axiom of EU, Holt (1986) describes how the use of RLIM could lead to a preference reversal if the independence axiom is not satisfied. This reversal is due to a dilution of choices as a consequence of the additional probabilistic stage. The author concludes that the RLIM does not elicit true preferences if the reduction principle holds. Harrison and Swarthout (2014) and Harrison et al. (2015) provide similar critiques, but other scholars find more support for RLIM (e.g. Starmer and Sugden, 1991). However, as we were conducting an experiment on the test of the RCLA, we could not assume *ex ante* that EU was valid.

Table 1: Sessions and parameters

Session	τ	fine rate ϕ	Income y	subjects	Phase 1 (periods 1-14)			Phase 2 (periods 15-28)		
					α	β	p	α	β	p
1	0.3	2	u[80,160]	18	0.2	1	0.2	0.9	$1/9$	0.1
2	0.3	2	u[80,160]	15	0.9	$1/3$	0.3	0.1	1	0.1
3	0.3	2	u[80,160]	13	0.5	0.4	0.2	0.3	1	0.3
4	0.3	2	u[80,160]	16	0.9	$1/9$	0.1	0.2	1	0.2
5	0.3	2	u[80,160]	18	0.1	1	0.1	0.9	$1/3$	0.3
6	0.3	2	u[80,160]	15	0.3	1	0.3	0.5	0.4	0.2

One-stage lottery treatments are with $\beta = 1$, and the two-stage lottery treatments with $\beta < 1$.

either 10%, 20% or 30% in compound terms.

Subjects were not told about the two phases of the experiment nor about how many rounds they were going to play.[5] They were told only that, during the experiment, the probability of an audit could be modified, in which case they would be informed of the change. After period 14, participants were provided with a new set of instructions for the other 14 periods. Those who had just concluded the one-stage (two-stage) lottery treatment then had to decide in a two-stage (one-stage) lottery environment.[6]

3.2 Procedures and experimental flow

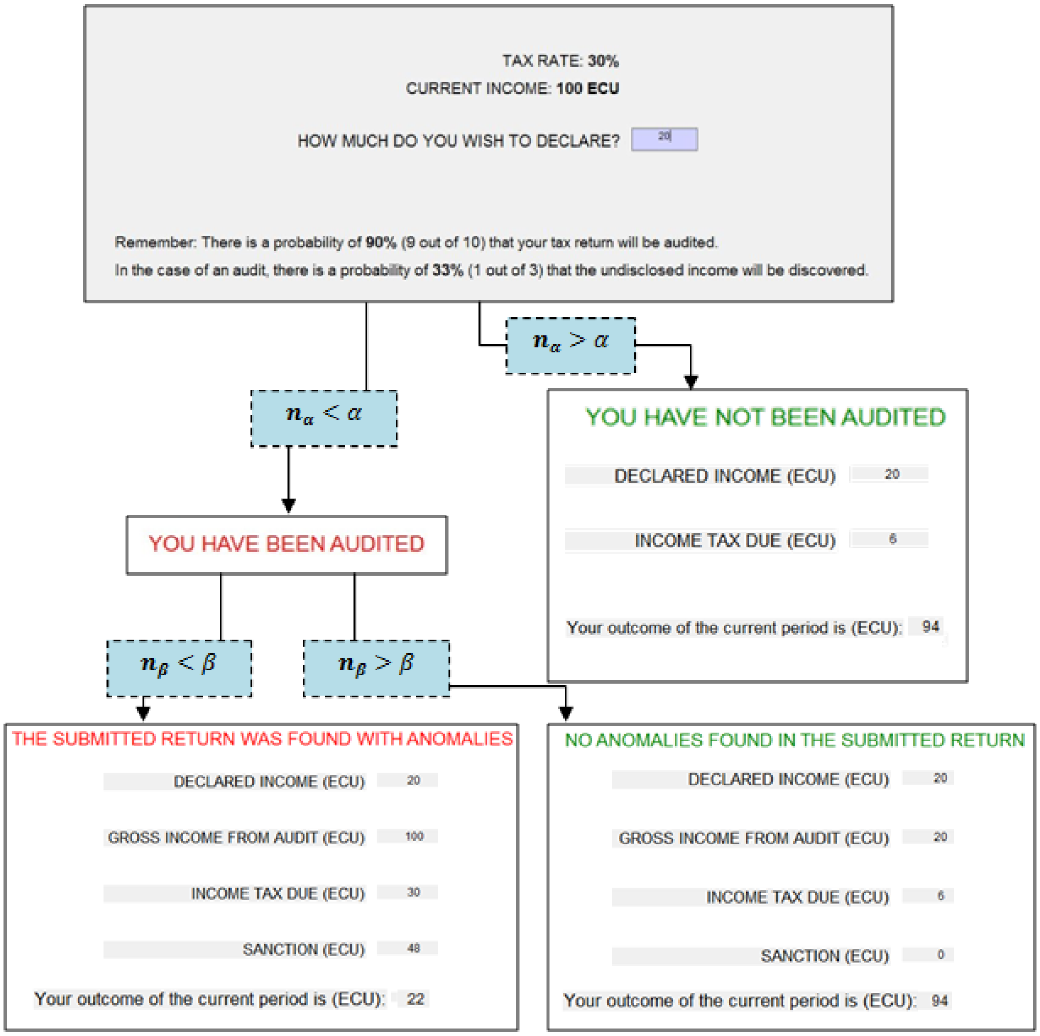
The experiment was conducted at the Ca' Foscari Laboratory for Experimental Economics in Venice, Italy. The subject pool consisted of 95 undergraduate and master students at the university, who were divided into six session groups. They were recruited through the Orsee platform (Greiner, 2015), on which participants previously registered to participate in economic experiments. To enhance the experimental design's consistency and minimize potential latent cultural confounders, we selected only subjects who were fluent in Italian, and we used the Italian language throughout the process of recruitment, instruction and experimental programming.

At the beginning of each session, subjects were seated in front of one of the laboratory computers, separated by cardboard dividers to prevent them from observing or being observed by other participants. Once seated in their cubicles, they read the printed instructions, as the experimenter read them aloud. Figure 3 shows the English

[5]This was done to reduce the possibility of endgame effects or extreme choices at the end of experiments, as observed by Selten and Stoecker (1986) and analysed in greater depth by Reuben and Suetens (2012).

[6]Instructions of the experiment are available from the authors.

Figure 3: Experimental flow - the two-stage treatment



translation of the experimental flow of the two-stage treatment.

On the first screen of the experiment flow, subjects were informed about their gross income for the period, were reminded of the tax and detection parameters, and were asked to take a decision about the level of income to declare. They were also given a calculator that they could use if they wished. Once they submitted their decisions, a random draw n_α (n_p in the one-stage treatment) from 0 to 1 determined whether an audit was to be performed. When $n_\alpha < \alpha$ ($n_p < p$), an audit was performed and a blinking text lasting eight seconds appeared on the screen informing the participant of an audit in progress. In the one-stage treatment, if no evasion was detected, the final outcome of the period - gross income minus taxes - was displayed. If, on the other hand, the declared amount was found to be lower than the actual gross income, the participant was shown a summary containing the sanction that resulted from the audit. In the two-stage treatment, a second background process drew a new random number n_β between 0 and 1. When n_β was lower than the detection rate β and the declared amount x was lower than the gross income y, the participant was shown the screen containing the sanction. If $n_\beta > \beta$ or $x = y$ (full compliance), the participant was informed that no anomalies had been found. At the end of the experiment, the participants completed a short questionnaire that asked about demographics, risk aversion, and fiscal understanding.

4 Results

We collected 2660 income reports from 95 subjects over 28 periods. A bit less than half of the participants were male (47%) and 72% were students in economics. The rest were from other fields of study, mainly the humanities and linguistics. Summary statistics for the participants' data are reported in Appendix.

In the following we first present descriptive evidence of the results based on treatment averages. This general overview is only suggestive. Robust statistical evidence is obtained from regression analyses presented afterward.

The compliance rate was calculated as the ratio between declared income and total gross income. Under the assumption of risk neutrality, a utility-maximizing agent should declare zero income. Overall, nearly a third of all income reports (29%, 767 reports) were honest, fully declaring the assigned gross income; around the same percentage (27%, 729 reports) declared zero income.

The histogram of all compliance rates is shown in Figure 4. The two peaks at 0 and 1 are the corner solutions that indicate censored-type behaviour. The remaining 1164 reports of partial compliance are distributed between 0 and 1 (excluded). The overlaid kernel density function highlights an average tendency toward declaring more than 50% of gross income.

Table 2 summarizes the average compliance rates in the two phases for the six treatment variations. In phase 1, consisting of periods 1-14, the average compliance rate is 54.8% in the one-stage treatment and 67.3% in the two-stage treatment. We perform a between-subject Mann-Whitney U test on individual mean compliance rates, relating

Figure 4: Histograms of overall compliance and partial compliance with kernel density function

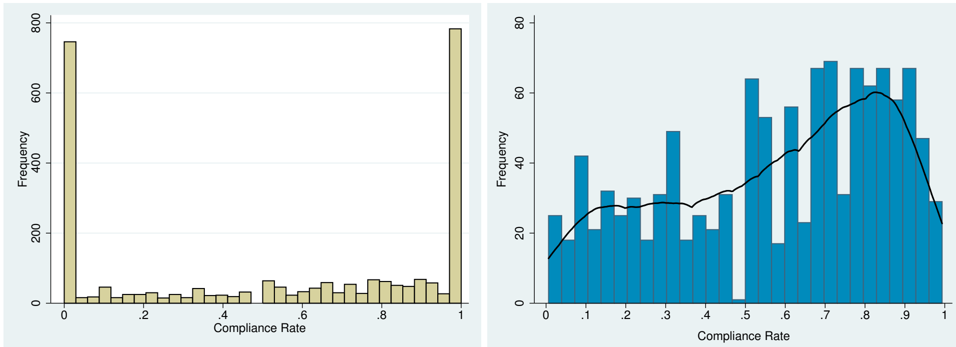


Table 2: Average compliance rates by treatment

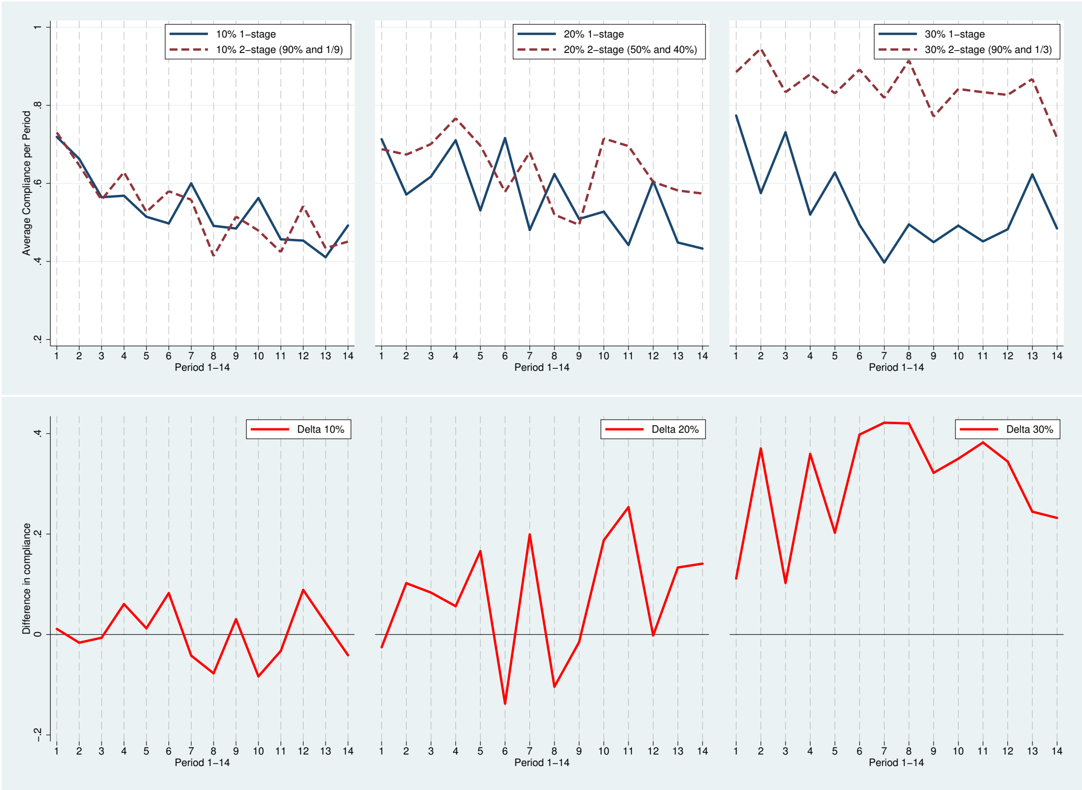
		Phase 1 (periods 1-14)		Phase 2 (periods 15-28)	
		One-stage L.	Two-stage L.	One-stage L.	Two-stage L.
Overall		.5480	.6725	.4859	.4708
Compound detection rate p (CDR)	10%	.5342	.5348	.3870	.3315
	20%	.5664	.6406	.4337	.4663
	30%	.5426	.8470	.6645	.6140

44 independent observations from the one-stage treatment to 51 average individual compliance rates from the two-stage treatment.[7] The test rejects the null hypothesis of equality in means at the $p < 0.05$ level (z=-2.374; p= 0.0176), implying rejection of the RCLA overall. Considering the impacts by levels of compound detection rate (CDR), data for the one-stage treatment indicate that the average compliance rates are not responsive to variations in the one-stage probability. On the other hand, in the two-stage treatment, average compliance rates increase monotonically in the overall compounded probabilities, indicating behavioural differences between the one-stage and two-stage lottery treatments depending on the overall probability of being detected.

In phase 2 of the experiment (periods 15-28), the compliance rates in the two-stage lottery treatments decline significantly with respect to phase 1, both with respect to the overall average (47.08% versus 67.25%, z=3.081; p=0.0021) and in the correspondence of the three levels of CDR. In the one-stage lottery treatments, the overall average compliance does not decline in phase 2 with respect to phase 1 (48.59% versus 54.80%, z=0.911, p=0.3624), but contrary to phase 1, the compliance rates increase monotoni-

[7] A t-test is not applicable, as compliance rates are bounded between 0 and 1, so they cannot be assumed to be normally distributed. The Shapiro-Wilk test of normality clearly rejects the normal distribution.

Figure 5: Average compliance rates per period and differences - Periods 1-14

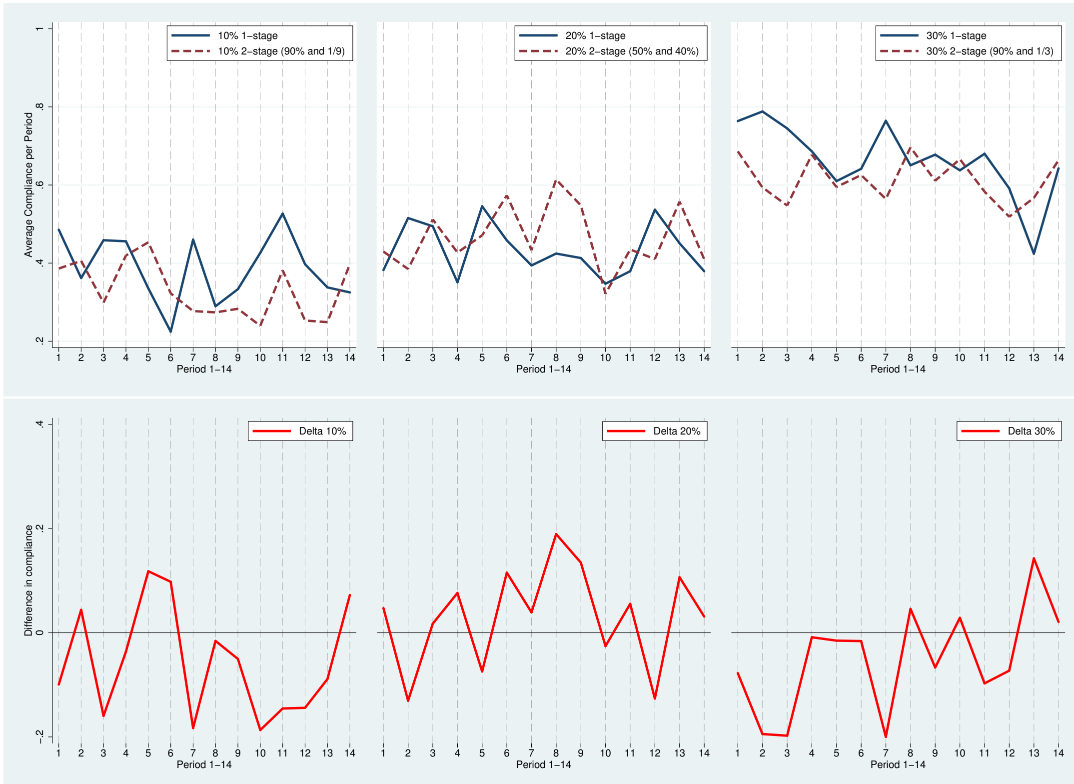


cally in the CDR.

Further evidence on the dynamic patterns of the average compliance rates can be obtained by looking at the data from a period-wise perspective. Figure 5 compares the average compliance rates for each decision period of the one-stage and the two-stage treatments for all three compound parameter values, 10%, 20% and 30%, considering only phase 1 (1330 reports). Differences between the period averages of the one-stage and the two-stage treatments are increasing as they move from the 10% to the 20% levels, though the latter exhibits noisier behaviour.

The graphic representations of the average compliance rates per period in phase 2 (Figure 6) show that they are lower than they are in phase 1 and that they increase with higher audit rates. Moreover, compliance does not differ between treatments and control groups, indicating that no RCLA violations are detectable in the second phase, and confirming that the probabilistic sensitivity changes between phases in both the one-stage and two-stage lottery treatments.

Figure 6: Average compliance rates per period and differences - Periods 15-28



4.1 Regression analysis

The observations from the overall averages are suggestive, but the results should be fully examined by appropriate econometric techniques. This section describes a more robust analysis, presenting the results of random-effects Tobit models, where we take account of individual-specific effects. The dependent variable in all the regressions is *Declared income*, censored from the left at zero and from the right at the gross income endowment of the period. We divide the analysis into three parts: first, we consider the data from periods 1-14, corresponding to phase 1 of the experiment; then we look at periods 15-28 for phase 2 of the experiment; finally we analyse the data, pooling all the observations to investigate the differences between the two phases.

4.1.1 Regression analysis of phase 1 (periods 1-14)

The results of phase 1 are shown in Table 3. We start with a simple model (1). The first control is gross *Income*, the experimental endowment of the period, drawn from a range of 80 to 160 ECU. Its effect is positive. The second control in the model is *CDR*, namely the compound detection rate, which refers to the overall probability of finale detection, taking in the experimental treatments the values of 10%, 20% and 30%. The core of our research is to determine whether subjects are indifferent between being presented with a one-stage tax compliance problem and a two-stage lottery. The *Two-Stage L.* dummy, our treatment variable, takes the value of 1 when the subject was asked to decide in the two-stage lottery environment and 0 for the one-stage set-up. The evidence from model (1) is that the overall probability of getting caught evading taxes (*CDR*) has a positive effect on compliance. The variable *Two-Stage L.* also has a statistically significant effect on compliance. Thus, the regression shows that the RCLA is indeed violated in phase 1.

In model (2) we investigate the source of the violations. We include the interaction *CDR * Two-Stage L.* to determine whether the treatment effect is due to the general propensity to increase compliance, affecting the intercept, or to a change in the reaction to the compound detection probability, as the descriptive evidence suggests. The regression result strongly supports the latter interpretation: the interaction term is large and positive, while the effect of the *Two-Stage L.* dummy alone becomes negative and marginally significant (at the 10% level). The effect of *CDR* alone becomes non-significant, which means that the *CDR* has no effect in the one-stage lottery treatment.

As previously indicated, the purpose of our experimental investigation is not to obtain a direct estimate of the subjective weighting function. Nevertheless, the investigation allows us to interpret the evidence in terms of people's perceptions and distortions of probabilities. The results are consistent with an inverse S-shaped weighting function, suggesting likelihood insensitivity (resulting from a form of optimism) in the one-stage treatment, and are consistent with a weighting function suggesting less insensitivity (possibly prone to pessimism and convexity) in the two-stage lottery treatment.

The other models in Table 3 add various controls that improve substantially the goodness of fit, as indicated by the log-likelihood of the regressions. Starting with model (3), the variable *Caught evading in (t-1)* is a dummy that equals 1 if the participants were detected and had to pay a fine in the previous period (t-1), and zero otherwise. The effect of the control is negative and significant. As indicated, this evidence is consistent with various previous studies and can be explained either in terms of the 'bomb crater effect' (e.g. Mittone, 2006) or as being due to a form of the 'loss

Table 3: Random-effects Tobit model in phase 1 (periods 1-14)

	(1)	(2)	(3)	(4)	(5)
Dependent variable: Declared income					
Income	0.344***	0.346***	0.213**	0.216**	0.217**
	(0.091)	(0.091)	(0.094)	(0.095)	(0.095)
CDR (compound detection rate)	18.595**	-4.341	-4.681	-2.219	-0.632
	(8.761)	(11.581)	(9.958)	(9.622)	(9.572)
Two-Stage L.	32.011**	-62.602*	-47.299	-31.214	-32.609
	(14.375)	(35.661)	(31.414)	(31.313)	(31.008)
CDR * Two-Stage L.		48.268***	43.454***	36.919**	35.567**
		(16.873)	(14.705)	(14.532)	(15.203)
Caught evading in (t-1)			-28.808***	-29.144***	-29.245***
			(6.609)	(6.666)	(6.668)
Audit and not caught in (t-1)			-23.028***	-24.591***	-24.671***
			(7.528)	(7.768)	(7.767)
Seconds for decision			0.381***	0.380***	0.379***
			(0.123)	(0.124)	(0.124)
Risk attitude			-24.830***	-17.411***	-15.253***
			(5.162)	(5.573)	(5.698)
Tax morality			-4.206	-4.443	-4.280
			(3.277)	(3.219)	(3.205)
Econ. students female				9.241	0.391
				(13.895)	(17.151)
Non-econ. students				37.080**	45.135**
				(15.997)	(20.093)
Econ. st. female * CDR * Two-Stage L.					13.505
					(13.130)
Non-econ. st. * CDR * Two-Stage L.					-5.920
					(11.874)
Wealth			-0.067*	-0.072*	-0.073*
			(0.039)	(0.039)	(0.039)
Trend (period 1-14)			4.177	4.632	4.724
			(3.581)	(3.585)	(3.578)
Constant	-7.136	37.214	147.540***	107.457***	98.297***
	(22.326)	(26.650)	(33.746)	(36.312)	(36.670)
ρ	0.520	0.497	0.421	0.391	0.384
Observations	1330	1330	1235	1209	1209
Log-likelihood	-4207.251	-4203.267	-3809.524	-3745.812	-3744.870
Wald χ^2	24.359	33.295	115.937	120.807	124.348

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

repair effect' (Maciejovsky et al., 2007). Our experimental design allows us to separate to some extent the former effect from the ~~the~~ latter because, in the two-stage lottery, when partial compliers are audited but not caught evading, the tax audit is not successful and subjects do not have to pay a fine. Therefore, there is no motivation for loss repair, and the negative reaction in this case is a pure 'bomb crater effect'. To capture this effect, the regression includes a dummy for *Audit and not caught evading taxes in (t-1)*. The effect of the variable is negative and statistically significant, indicating that, in phase 1, a 'bomb crater effect' is operating in addition to any effect that is due to an audit with a fine.

We also control for the time spent to take the decision, which is given by *Seconds for decision*. The effect of the variable is positive, which means that individuals who take more time to make a decision are more likely to comply.

The next control is a measure of self-reported *Risk attitude*, elicited from the final questionnaire. We asked a general risk question, as in Dohmen et al. (2011), that is, "ow do you deal with potentially risky situations?" Our variable ranges from 1 ("I always avoid risky situations and choose the safer option") to 7 ("I am willing to take risks and always choose the riskier option"). We opted for a questionnaire approach to obtain an index for risk propensity rather than an incented risk approach (e.g. Holt and Laury, 2002) to ensure consistency and comparability with other personal controls that were also obtained from the questionnaire, including *Tax morality*. We find that individuals who self-report a more positive attitude toward risk declare less income, which is in line with previous studies (Bazart and Bonein, 2014; Bernasconi et al., 2014). On the other hand, the regression does not indicate a statically significant impact for the index of self-reported tax morality.[8]

We include a time *Trend* and a variable *Wealth* that measures the profit accrued to each subject up to the current experimental period. The latter variable controls for changes in their willingness to make risky income declarations in the experiment because of accumulation of profit. The time trend variable is not significant, while *Wealth* exhibits a slightly significant negative effect, which is in line with a hypothesis of decreasing absolute risk aversion because of an income effect. An additional latent effect that could be responsible for the negative impact of *Wealth* is the 'house-money effect' described in, for example, Thaler and Johnson (1990). Accordingly, a gain from a risky operation (evasion) could induce the subjects to decrease their risk aversion with respect to the extra amount offered 'by the house' that they can reinvest.

The last two models, shown in Table 3, investigate the effect on compliance of gender and the field of study. Previous studies have shown that women tend to have

[8]Similar to the question on risk attitude, the question on tax morality was asked as follows: "On a scale from 1 to 7, to what degree do how you agree with the statement that a citizen will not pay taxes if s/he has a limited sense of morality?" Even though, in the existing literature on tax compliance, the terms 'tax morale' and 'tax morality' are often used interchangeably, we chose the term 'morality' based on the Merriam-Webster Dictionary's definition as "conformity to ideals of right human conduct", as opposed to the definition of "morale" therein as "moral principles, teachings, or conduct".

greater aversion to tax evasion and corruption (Torgler and Valev, 2010, and references therein). Regarding the field of study, starting from Ames and Marwell (1981), various studies have shown that economics students often behave differently from other students in experiments, exhibiting a significantly higher overall enthusiasm about profit maximization, regardless of their choices' moral implications. On the other hand, in choice experiments that require mathematical skills, the evidence is more ambiguous, as the decisions of economics students do not always differ from those of non-economics students (Rubinstein, 2006).

Most of our participants (72%) were economics students. Most of the other participants, from the humanities and linguistics, were female (76% of non-economics students). To study the impact of gender and field of study, we construct two dummies: one for female economics students (*Econ. students female*) and one for non-economics students (*Non-econ. students*), without distinguishing in the latter between male and female since there are few of the latter. Thus, the coefficients on the two dummies measure the differences in the two sociological controls with respect to the compliance behaviour of male economics students.

Model (4) shows that female economics students behave no differently from male economics students and that both male and female economics students comply significantly less than non-economics students do.

In model (5) we investigate the effects of gender and the field of study in connection to the observed violations of RCLA. To this purpose, we interacted the same two sociological controls *Econ. students female* and *Non-econ. students*, with the variable *Two-Stage L. * CDR*. Here, we see no difference in the behaviours of male economists, female economists, and non-economists, so all participants in phase 1 of the experiment violate the RCLA in a similar way.[9]

4.1.2 Regression analysis of phase 2 (periods 15-28)

The results of repeating the same analysis for phase 2 (periods 15-28) are shown in Table 4. Considering the first four models, we found that the main difference with respect to phase 1 is that the impact of the two-stage lottery treatment vanishes. In particular,

[9] A regression that also includes the interaction of the two sociological controls with the two-stage lottery dummy alone does not provide statistically significant evidence since it dilutes the effects of the two-stage treatment over too many controls.

not only is the *Two-Stage L.* dummy not significant, but its interaction with the CDR (namely *CDR*Two-Stage L.*) is no longer statistically significant. On the other hand, the variable *CDR* alone is statistically significant and has the expected sign. Therefore, the regressions indicate that participants as a whole no longer violate the RCLA, and they respond equivalently to the CDR, whether presented as one-stage lottery or as two-stage lottery. However, in model (5), in which the variable *CDR * Two-Stage L.* is interacted with the two sociological controls *Econ. students female* and *Non-econ. students*, some differences emerge among the groups. In particular, while the regressions confirm that there are no differences in behaviour between male and female economics students, non-economics students still violate the axiom and exhibit more sensitivity to *CDR* in the two-stage lottery treatment than they do in the one-stage lottery treatment.

Explanations for the evidence here are mainly related to learning effects. One explanation that we find appealing and consistent with the RDU development rests on the possible latent effects of participants' probability weighting functions, which adjust over time depending on experience and expertise. For example, as Wakker (2010, p. 204) emphasized, people do not deviate from expected utility only because of extra pessimism or the like; they also do not adequately understand probability because of their perceptual and cognitive limitations. Under this hypothesis, subjects in our experiment could have adjusted their weighting functions dynamically, that with repetitions and experience become more linear, making subjects more similar to expected utility maximizers who also behave equivalently with the one-stage and two-stage lotteries. Nevertheless, not all individuals learn and correct their perceptual limitations at the same speed. The result here is that economics students learn how to deal with two-stage lotteries more quickly than non-economics students do.

Other differences between the models estimated in Table 4 for phase 2 and those estimated above for phase 1 are considered below in an analysis that pools all data. Here we remark on the most relevant differences. First, we see that the effect of gross *Income* becomes insignificant. The variable *Caught evading in (t-1)* is still significant, whereas *Audit and not caught evading taxes in (t-1)* is not. This result seems to reject the occurrence of a 'bomb crater effect' in relation to an audit without a fine in phase 2. (However, see the pooled model below). The impact of *Seconds for decision* continues to be positive, while the self-reported *Risk attitude* continues to be negative. The control for *Tax morality* is confirmed here as not statistically significant.

Table 4: Random-effects Tobit model in phase 2 (periods 15-28)

	(1)	(2)	(3)	(4)	(5)
Dependent variable: Declared income					
Income	0.166	0.166	0.112	0.104	0.093
	(0.117)	(0.117)	(0.122)	(0.125)	(0.125)
CDR (compound detection rate)	49.427***	44.711**	41.627**	41.036**	31.406*
	(13.181)	(19.912)	(17.944)	(18.168)	(17.811)
Two-Stage L.	-1.690	-18.202	-10.030	-12.257	-24.145
	(21.455)	(56.635)	(51.382)	(52.069)	(50.429)
CDR * Two-Stage L.		8.311	8.309	7.748	2.241
		(26.393)	(23.744)	(24.003)	(26.062)
Caught evading in (t-1)			-27.698***	-28.060***	-28.024***
			(9.009)	(9.195)	(9.190)
Audit and not caught in (t-1)			-10.837	-11.180	-10.234
			(9.109)	(9.230)	(9.214)
Seconds for decision			1.224***	1.239***	1.249***
			(0.279)	(0.284)	(0.283)
Risk attitude			-26.498***	-23.059**	-18.105**
			(8.202)	(9.293)	(8.993)
Tax morality			0.304	-1.277	-3.457
			(5.310)	(5.385)	(5.233)
Econ. students female				9.406	30.755
				(23.624)	(31.514)
Non-econ. students				9.229	-34.855
				(27.056)	(32.103)
Econ. st. female * CDR * Two-Stage L.					-10.552
					(19.341)
Non-econ. st. * CDR * Two-Stage L.					45.904**
					(19.547)
Trend (period 1-14)			4.102	4.550	5.514
			(4.732)	(4.818)	(4.830)
Wealth			-0.051	-0.056	-0.066
			(0.049)	(0.050)	(0.050)
Constant	-66.270**	-56.967	24.042	19.509	40.504
	(33.607)	(44.675)	(52.376)	(58.209)	(56.630)
ρ	0.605	0.605	0.553	0.540	0.514
Observations	1330	1330	1235	1209	1209
Log-likelihood	-3588.242	-3588.193	-3290.183	-3230.389	-3226.302
Wald χ^2	16.125	16.236	63.087	61.660	71.400

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Models (4) and (5) show that, other than for the violations of RCLA discussed above, non-economics students do not behave differently from economics students in phase 2. All the other variables in Table 4, including the *Trend* and the variable *Wealth*, are not statistically significant.

4.1.3 Regression analysis of pooling data of the two phases (periods 1-28)

Table 5 provides a more accurate analysis of the impacts of the covariates in the two phases of the experiments. We estimate random-effects Tobit models with individual-specific effects over all 28 periods, interacting the various regressors with dummies to highlight the effects of the variables in the two phases. We estimate three model specifications, including all the controls investigated previously. Model (1) includes all variables except those that control for the sociological groups. Models (2) and (3) take economics students as the reference category and estimate the impacts of non-economics students: model (2) controls only for the effect of non-economics students on the intercept, while model (3) introduces the interaction of the CDR with the two-stage lottery treatment. We do not include controls for female economics students since all previous regressions showed that their behaviour is indistinguishable from that of their male colleagues.

By increasing the number of observations per subject, the longitudinal analysis sharpens the impact of several controls and provides some additional insights. First, all three models confirm that the CDR is statistically significant and with the expected sign in phase 2 of the experiment (*CDR*Phase 2*) and in the two-stage lottery treatment of phase 1 (*CDR*Two-Stage L.*Phase 1*). The CDR is not among the relevant predictors in the one-stage lottery treatment of phase 1 since the control *CDR* alone is not statistically significant.

The models also show that the impact of *CDR*Phase 2* is significantly greater than both *CDR* alone and *CDR * Two-Stage L. * Phase 1* ($p < 0.01$ in all the three models), which demonstrates that the sensitivity to the total CDR increases significantly in phase 2 of the experiment. This evidence is consistent with the notion that, with repetition and experience, the phenomenon of probability insensitivity weakens substantially and may even vanish for some subjects, but not for all.

Model (3) shows some further effects of the CDR's interaction with the two-stage lottery dummy in phase 2. In particular, we see that the variable *CDR * Two-Stage L. * Phase 2*, which is not statistically significant in models (1) and (2), becomes marginally significant with a negative sign in model (3). At the same time, the interaction with non-economics students (namely *Non-econ. st. * CDR * Two-Stage L. * Phase 2*) is highly significant and positive. The same interaction for Phase 1, *Non-econ. st. * CDR * Two-Stage L. * Phase 1*, remains not significant. This result confirms different behaviour in phase 2 of the experiment between economics and non-economics students, with the latter group still committing violations of the RCLA.

Pooling the data from the two phases substantiates the evidence previously obtained. The gross *Income* assigned in each period is not statistically significant in either phase of the experiment. However, the *Wealth* accumulated during the experiment makes participants more willing to take risky reporting decisions. The effect is smaller in phase 2 and partially contrasted by a positive linear trend. Nevertheless, overall, the pooled analysis shows that these income effects, which may include a 'house money effect', could have contributed to changing subjects' willingness to incur the risk of being detected between the two phases, leading them to increase their evasion in phase 2 (see dummy *Phase 2*).

Two effects that are stable between the two phases are the negative impact of *Caught evading taxes in (t-1)*, which can be due to either a 'loss repair effect' or a 'bomb crater effect' (or both), and the effect of an audit without a fine, namely *Audit and not caught evading taxes in (t-1)*, which can be explained as a 'bomb crater effect'. While the evidence on the first control was expected since the effect was documented by the separate regressions for both phases of the experiment (Tables 3 and 4), the pooled model confirms the second effect also in phase 2, which the subsample regression did not detect.

In all three models, phase 2 shows a statistically significant increase in the time taken for the compliance decision. The controls for *Risk attitude* and *Tax morality* offer some additional insights. *Risk attitude* is confirmed to be negative, with, according to model (1), a lower if only marginally significant effect in phase 2; and *Tax morality* is a marginally significant predictor in model (1) but has no significant effect in model (2) once we control for the academic discipline. Model (2) also confirms that non-economics students tend to comply more than economics students do, although the effect is smaller in phase 2, even if at only a marginally significant level. One explanation for these effects could lie in that, with learning and experience, the differences between economics and non-economics students becomes more dependent on the subjects' personal characteristics (perhaps such as those linked to tax morality and risk attitude), rather than on general traits and noise.

Table 5: Random-effects Tobit model - Pooled sample of Phases 1 and 2 (periods 1-28)

	(1)	(2)	(3)
Dependent variable: Declared income			
Income	0.159	0.163	0.156
	(0.109)	(0.111)	(0.111)
Income * Phase 2	-0.042	-0.049	-0.063
	(0.153)	(0.156)	(0.156)
CDR (compound detection rate)	5.422	7.331	4.334
	(6.439)	(6.492)	(6.512)
CDR * Phase 2	41.294***	38.655***	38.524***
	(9.563)	(9.700)	(9.727)
CDR * Two-Stage L. * Phase 1	23.371**	19.373**	20.032**
	(9.477)	(9.752)	(10.054)
CDR * Two-Stage L. * Phase 2	-14.876	-14.652	-17.953*
	(9.428)	(9.516)	(9.699)
Two-Stage L.	-1.446	9.739	0.731
	(23.057)	(23.647)	(23.829)
Two-Stage L. * Phase 2	44.159	30.873	32.740
	(31.983)	(32.053)	(32.412)
Caught evading in (t-1)	-42.598***	-42.742***	-41.219***
	(7.663)	(7.756)	(7.726)
Caught evading in (t-1) * Phase 2	5.398	5.160	5.375
	(10.995)	(11.163)	(11.132)
Audit and not Caught in (t-1)	-35.932***	-38.006***	-35.869***
	(8.457)	(8.729)	(8.725)
Audit and not Caught in (t-1) * Phase 2	13.788	15.038	16.871
	(11.602)	(11.843)	(11.849)
Seconds for decision	0.337**	0.352***	0.366***
	(0.134)	(0.136)	(0.135)
Seconds for decision * Phase 2	1.002***	0.978***	0.988***
	(0.253)	(0.256)	(0.255)
Risk attitude	-26.203***	-18.755***	-20.649***
	(5.376)	(5.847)	(5.998)
Risk attitude * Phase 2	5.640*	2.763	5.719
	(3.110)	(3.492)	(3.561)
Tax morality	-3.802	-3.962	-4.270
	(3.435)	(3.382)	(3.427)
Tax morality * Phase 2	3.642*	3.159	2.478
	(2.029)	(2.066)	(2.090)
Non-econ. students		34.628**	21.329
		(15.333)	(17.914)
Non-econ. students * Phase 2		-17.977*	-29.475
		(9.335)	(17.954)
Non-econ. st. * CDR * Two-Stage L. * Phase 1			14.494
			(9.376)
Non-econ. st. * CDR * Two-Stage L. * Phase 2			22.565***
			(8.566)
Wealth	-0.145***	-0.148***	-0.146***
	(0.027)	(0.028)	(0.028)
Wealth * Phase2	0.023**	0.022**	0.022**
	(0.010)	(0.010)	(0.010)

Continued

Table 5 continued

Trend (period 1-28)	11.169***	11.427***	11.299***
	(2.492)	(2.518)	(2.521)
Phase 2	-298.486***	-278.113***	-277.017***
	(47.171)	(48.647)	(48.923)
Constant	134.102***	98.508***	114.848***
	(31.371)	(34.183)	(35.010)
ρ	0.354	0.331	0.340
Observations	2470	2418	2418
Log-likelihood	-7171.773	-7048.453	-7037.645
Wald χ^2	390.905	390.238	392.105

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

5 Concluding discussion

Inspired by the evidence against the reduction of compound lotteries axiom in various contexts, we investigated the reduction axiom in a tax compliance decision problem in which we separated the audit rate and the detection rate in a two-stage lottery process. This approach is a natural extension of the classic Allingham and Sandmo (1972) model, where agents take compliance decisions based on an overall one-stage audit/detection lottery.

A theoretical section focussed on rank-dependent utility as the most notable generalisation of expected utility to illustrate several ways in which the reduction axiom can be violated. There is now plenty of evidence that people‘s sensitivity to probabilities can be affected by their perceptual and cognitive limitations, captured by the probability weighting function of rank-dependent utility. We derived predictions for taxpayers' compliance behaviour that depend both on the shape of the probability weighting function and on whether they respect or violate the reduction axiom.

We conducted a tax evasion experiment to compare compliance in the one-stage control treatment to compliance in the two-stage lottery treatment. In the first phase of the experiment, with inexperienced participants, subjects violated the reduction axiom, behaving differently in the one-stage treatment than in the two-stage lottery set-up. In particular, participants were sensitive to the overall detection probability used in the experiment (ranging between 10% and 30%) only in the two-stage lottery treatment, not in the one-stage treatment. The evidence is consistent with the hypothesis that participants use different weighting functions in the one-stage lottery set-up than they do in the two-stage set-up, such that the function is inverse S-shaped, more optimistic, and likelihood insensitive in the former set-up, and less insensitive and inverse

S-shaped in the latter. We also found that both economics and non-economics students committed similar violations of the reduction of compound lotteries axiom, although the latter group, on average, reported a greater share of their income.

In the second phase of the experiment, with more experienced subjects, the sensitivity to the overall detection rate increased substantially for all participants in both treatments. In fact, economics students behaved similarly in the two set-ups and no longer violated the reduction axiom in the second phase of the experiment. For non-economics students, convergence occurred at a lower rate, so they still committed some violations of the axiom even in the second phase.

Notwithstanding the learning process and even given this evidence, we believe that there are valid reasons to handle the reduction of compound lotteries axiom in tax compliance decisions with care and that it is advisable to separate the audit rate from the detection rate when considering the weights and cognitive distortions with which people perceive probability.

Learning may have occurred in our experiments because of several factors, not all of which are equally relevant to a parallel with the field. Participants played several tax evasion lotteries in one hour or so, which could have offered them the opportunity to infer a frequentist perception of the probability of being caught, even in the two-stage situations.

In the field, the experience of tax reporting decisions with the chance of being audited is spread over a much longer period. In fact, in the real world, taxpayers often don't know when an auditing process starts and may not know when it finishes. In addition, participants' perception of the detection rate of undeclared income is likely to be influenced by factors like the credibility of institutional efficiency, the level of self-evaluation, and exogenous institutional determinants. In such contexts, taxpayers may have much greater difficulty making inferences about the probability of being caught and treat complex multi-stage monitoring processes as equivalent to simple, one-shot lottery games.

The experiments showed that other dimensions may interact with violations of the reduction axiom. We found that self-reported risk attitudes affect subjects' decisions, and we obtained marginal evidence for an index of self-reported tax morality. These and other personal traits may also interact with the learning process, suggesting that, in the real world, where experience requires longer time to build, the effect of learning could be weak or even disappear for some people.[10]

Finally, we found that separate reactions may be induced by post-audit and post-detection when two classic effects of the behavioural literature on tax evasion, the 'bomb crater effect' and the need for 'loss repair', are studied. We found that reactions to audits and fines are significant but also that the mere experience of having been audited impacts negatively on compliance in the subsequent period. An important policy message emerges from this result: "if you audit, do it well". In particular, tax authority should not disregard the negative effect after an unsuccessful audit when allocating the tax agency budget between the costs of the number of audits and that for careful examination of the tax files.

[10]Future research may also analyse how well predictions and experimental findings like those reported in this paper align with the initial wave of empirical studies, based on administrative data, that have looked at taxpayers' dynamic response to audits (as in papers by Advani et al., 2017; DeBacker et al., 2015; Mazzolini et al., 2017).

Appendix

A Proof of Proposition 1.

Proposition 1. Let incomes reported under RCLA and under CIA be given by x_{RCLA} and x_{CEM}, respectively. Then, denoting $g(q)=1-w(1-q)$:

$$x_{RCLA} \gtreqless x_{CEM} \iff g(\alpha\beta) \gtreqless g(\alpha)g(\beta) \tag{6}$$

To see why, notice that the first-order conditions in the two cases of $V_{CIA}(O_{2S})$ and $V_{RCLA}(O_{2S})$ are:

$$\text{FOC under } V_{RCLA}(O_{2S}): \quad \frac{g(\alpha\beta)}{1-g(\alpha\beta)} = \frac{u'(NC)}{\phi u'(C)}$$

$$\text{FOC under } V_{CIA}(O_{2S}): \quad \frac{g(\alpha)g(\beta)}{1-g(\alpha)g(\beta)} = \frac{u'(NC)}{\phi u'(C)}$$

Denote the right-hand term of both expressions as a function of reported income x, namely $\Gamma(x) = \frac{u'(NC)}{\phi u'(C)}$. Clearly, the two first-order conditions imply that, when $g(\alpha)g(\beta) > (<) g(\alpha\beta)$, then $\Gamma(x)$ under CIA must be greater (lower) than $\Gamma(x)$ under RCLA. The proposition follows, as:

$$\frac{\partial\Gamma(x)}{\partial x} = \frac{-t[u''(NC)u'(C) + \phi u''(C)u'(NC)]}{\partial[u'(C)]^2} > 0$$

where the sign follows from $u'(\cdot) > 0$ and $u''(\cdot) < 0$.

B Summary Statistics

Table 6: Summary Statistics (Panel Data)

Variable		Mean	Std. Dev.	Min	Max	Observations
Declared Income	overall	64.82	52.35	0.00	160.00	N = 2660
	between		29.83	5.00	131.43	n = 95
	within		43.12	-44.97	191.42	T = 28
Income	overall	120.65	23.34	80.00	160.00	N = 2660
	between		5.16	108.21	132.50	n = 95
	within		22.77	68.86	172.43	T = 28
Seconds (Decision)	overall	12.79	19.13	0.00	164.00	N = 2660
	between		9.92	2.79	59.07	n = 95
	within		16.39	-29.82	159.25	T = 28
Female	overall	0.53	0.50	0.00	1.00	N = 2660
	between		0.50	0.00	1.00	n = 95
Econ Discipline	overall	0.72	0.45	0.00	1.00	N = 2604
	between		0.45	0.00	1.00	n = 93
Risk Propensity	overall	2.89	1.21	1.00	6.00	N = 2660
	between		1.22	1.00	6.00	n = 95
Tax Morale	overall	3.96	1.87	0.00	6.00	N = 2660
	between		1.88	0.00	6.00	n = 95
Trust	overall	2.32	1.22	0.00	5.00	N = 2660
	between		1.22	0.00	5.00	n = 95

The panel data summary statistics were obtained by using the STATA® command *xtsum*. For the within-subjects component, the between-subjects mean $\bar{x}_i$ is subtracted from the value x_{it} and the global mean $\bar{x}$ is added back in order to get comparable results. Within minimum and maximum values are calculated as deviations from individual means (note: to obtain deviation values, the global averages have to be subtracted from the reported minimum and maximum values).

References

Abbink, K. and Hennig-Schmidt, H. (2006). Neutral versus loaded instructions in a bribery experiment. *Experimental Economics*, 9(2):103–121.

Advani, A., Elming, W., Shaw, J., et al. (2017). The dynamic effects of tax audits. Technical report, Institute for Fiscal Studies.

Allingham, M. G. and Sandmo, A. (1972). Income Tax Evasion: A Theoretical analysis. *Journal of Public Economics*, 1(3-4):323–338.

Alm, J. (2019). What motivates tax compliance? *Journal of Economic Surveys*, 33(2):353–388.

Alm, J., McClelland, G. H., and Schulze, W. D. (1992). Why do people pay taxes? *Journal of Public Economics*, 48:21–38.

Ames, E. and Marwell, G. (1981). Economist Free Ride, Does Anyone Else? Expriments on the Provision of Public Goods. *Journal of Public Economics*, 15(3):295–310.

Andreoni, J., Erard, B., and Feinstein, J. (1998). Tax compliance. *Journal of Economic Literature*, 36(2):818–860.

Bar-Hillel, M. (1973). On the subjective probability of compound events. *Organizational Behavior and Human Performance*, 9(3):396–406.

Bazart, C. and Bonein, A. (2014). Reciprocal relationships in tax compliance decisions. *Journal of Economic Psychology*, 40:83–102.

Becker, G. S. (1968). Crime and Punishment: An Economic Approach. *Journal of Political Economy*, 76(2):169–217.

Behnk, S., Barreda-Tarrazona, I., and García-Gallego, A. (2018). Punishing liars?how monitoring affects honesty and trust. *PloS one*, 13(10):e0205420.

Bernasconi, M. (1992). Different frames for the independence axiom: An experimental investigation in individual decision making under risk. *Journal of Risk and Uncertainty*, 5(2):159–174.

Bernasconi, M., Corazzini, L., and Seri, R. (2014). Reference dependent preferences, hedonic adaptation and tax evasion: Does the tax burden matter? *Journal of Economic Psychology*, 40(0):103–118.

DeBacker, J., Heim, B. T., Tran, A., and Yuskavage, A. (2015). Once bitten, twice shy? the lasting impact of irs audits on individual tax reporting. *Journal of Financial Economics*, 117(1):122–138.

Diecidue, E. and Wakker, P. P. (2001). On the intuition of rank-dependent utility. *Journal* *of Risk and Uncertainty*, 23(3):281–298.

Dillenberger, D. (2010). Preferences for one-shot resolution of uncertainty and allais-type behavior. *Econometrica*, 78(6):1973–2004.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences. *Journal of the European Economic Association*, 9(3):522–550.

Feinstein, J. S. (1991). An Econometric Analysis of Income Tax Evasion and Its Detection. *The RAND Journal of Economics*, 22(1):14–35.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Gayer, G. (2010). Perception of probabilities in situations of risk: A case based approach. *Games and Economic Behavior*, 68(1):130–143.

Gonzalez, R. and Wu, G. (1999). On the shape of the probability weighting function. *Cognitive psychology*, 38(1):129–166.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association*, 1(1):114–125.

Guala, F. and Mittone, L. (2005). Experiments in economics: External validity and the robustness of phenomena. *Journal of Economic Methodology*, 12(March):495–515.

Harrison, G. W., Martínez-Correa, J., and Swarthout, J. T. (2015). Reduction of compound lotteries with objective probabilities: Theory and evidence. *Journal of Economic Behavior and Organization*, 119:32–55.

Harrison, G. W. and Swarthout, J. T. (2014). Experimental payment protocols and the Bipolar Behaviorist. *Theory and Decision*, 77(3):423–438.

Hashimzade, N., Myles, G. D., and Tran-Nam, B. (2013). Applications of behavioural economics to tax evasion. *Journal of Economic Surveys*, 27(5):941–977.

Hogarth, R. M. and Einhorn, H. J. (1990). Venture theory: A model of decision weights. *Management science*, 36(7):780–803.

Holt, C. A. (1986). Preference Reversals and the Independence Axiom. *The American Economic Review*, 76(3):508–151.

Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. *American economic review*, 92(5):1644–1655.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):pp. 263–292.

Kirchler, E. (2007). *The economic psychology of tax behaviour*. Cambridge University Press.

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S. S., and Saez, E. (2011). Unwilling or Unable to Cheat? Evidence From a Tax Audit Experiment in Denmark. *Econometrica*, 79(3):651–692.

Laury, S. K. and Taylor, L. O. (2008). Altruism spillovers: Are behaviors in context-free experiments predictive of altruism toward a naturally occurring public good? *Journal of Economic Behavior and Organization*, 65(1):9–29.

Maciejovsky, B., Kirchler, E., and Schwarzenberger, H. (2007). Misperception of chance and loss repair: On the dynamics of tax compliance. *Journal of Economic Psychology*, 28(6):678–691.

Mazzolini, G., Pagani, L., and Santoro, A. (2017). The deterrence effect of real-world operational tax audits. Technical report, University of Milano-Bicocca, Department of Economics.

Mittone, L. (2006). Dynamic behaviour in tax evasion: An experimental approach. *The Journal of Socio-Economics*, 35(5):813–835.

Mittone, L., Panebianco, F., and Santoro, A. (2017). The bomb-crater effect of tax audits: Beyond the misperception of chance. *Journal of Economic Psychology*, 61:225–243.

Nebout, A. and Dubois, D. (2014). When Allais meets Ulysses: Dynamic axioms and the common ratio effect. *Journal of Risk and Uncertainty*, pages 1–31.

Palacios-Huerta, I. (1999). The aversion to the sequential resolution of uncertainty. *Journal of Risk and uncertainty*, 18(3):249–269.

Prelec, D. (1998). The Probability Weighting Function. *Econometrica*, 66(3):497–527.

Prokosheva, S. (2016). Comparing decisions under compound risk and ambiguity: The importance of cognitive skills. *Journal of Behavioral and Experimental Economics*, 64:94–105.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, 3(4):323–343.

Rablen, M. D. (2014). Audit probability versus effectiveness: The beckerian approach revisited. *Journal of Public Economic Theory*, 16(2):322–342.

Reuben, E. and Suetens, S. (2012). Revisiting strategic versus non-strategic cooperation. *Experimental Economics*, 15(1):24–43.

Rubinstein, A. (2006). A sceptic's comment on the study of economics. *Economic Journal*, 116(510):1–9.

Segal, U. (1987). The Ellsberg Paradox and Risk Aversion: An Anticipated Utility Approach. *International Economic Review*, 28(1):175–202.

Segal, U. (1990). Two-stage lotteries without the reduction axiom. *Econometrica: Journal of the Econometric Society*, 58(2):349–377.

Selten, R. and Stoecker, R. (1986). End behavior in sequences of finite Prisoner's Dilemma supergames A learning theory approach. *Journal of Economic Behavior and Organization*, 7(1):47–70.

Snow, A. and Warren, Jr, R. S. (2005). Tax evasion under random audits with uncertain detection. *Economics Letters*, 88(1):97–100.

Starmer, C. and Sugden, R. (1991). Does the Random-Lottery Incentive System Elicit True Preferences? An Experimental Investigation. *American Economic Review*, 81(4):971–978.

Thaler, R. H. and Johnson, E. J. (1990). Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. *Management science*, 36(6):643–660.

Torgler, B. and Valev, N. T. (2010). Gender and Public Attitudes Toward Corruption and Tax Evasion. *Contemporary Economic Policy*, 28(4):554–568.

Tversky, A. and Wakker, P. (1995). Risk attitudes and decision weights. *Econometrica: Journal of the Econometric Society*, pages 1255–1280.

von Neumann, J. and Morgenstern, O. (1944). *Theory of games and economic behavior*, volume 60. Princeton university press Princeton.

Voors, M., Turley, T., Kontoleon, A., Bulte, E., and List, J. A. (2012). Exploring whether behavior in context-free experiments is predictive of behavior in the field: Evidence from lab and field experiments in rural Sierra Leone. *Economics Letters*, 114(3):308–311.

Wakker, P. P. (2010). *Prospect theory: For risk and ambiguity*. Cambridge university press.

Yitzhaki, S. (1974). A note on Income Tax Evasion: A theoretical analysis. *Journal of Public Economics*, 3:201–202.

Zimmermann, F. (2014). Clumped or piecewise? evidence on preferences for information. *Management Science*, 61(4):740–753.