Response to Reviewers

Reviews (Round 1)

Dear Dr. Rahal,

we thank you for the detailed feedback and comments on our manuscript submission. Unfortunately, revision of the manuscript took longer than expected due to teaching obligations of all authors and significant revisions regarding predictions and experimental design being necessary. Here, we provide a point-by-point response to your main points and each of the reviewer's comments. In particular, we have revised the introduction by including a more detailed literature review, the prediction section by updating the equation, using terms from previous research, and providing further information in the form of visualizations of predictions, and finally the experimental design by aligning treatments more to improve comparability. We hope that you will consider this revised manuscript for Stage 1 acceptance.

Dear Dr. Zickfeld,

thank you for your submission "The Effect of Individual and Group Punishment on Individual and Group-Based Dishonesty" to PCI RR, for which I have now received three independent reviews. Based on these reviews and my own reading of your manuscript, I would like to invite you to revise the proposal. There is much to like about the manuscript already, but I will highlight the most salient opportunities for further improvement below:

- Clarify the proposed hypotheses, particularly their derivation from equation 1 (review criterion 1B)

We have provided further details to clarify the proposed hypotheses and now spend more time on explaining how the equation is derived and what this predicts for each of the six treatments. For this we have updated Table 1 to provide a more detailed explanation and include an additional figure (Figure 1) showing the predicted expected outcomes for the different treatments (see below). Based on feedback from R2 we have aligned equations and terms more in line with previous literature on tax evasion games (Alm & Malézieux, 2021). Overall, we think that this has improved the clarity of our predictions and how hypotheses were derived. Note, that based on the updated equations and predictions one of the initial hypotheses was changed.

Table 1. Overview of predictions of expected company income $E(I)$ and expected gain from dishonesty for levels of fully dishonest ($R_i = 0$), moderately dishonest ($R_i = 500$), and honest ($R_i = 1000$) reporting across the manipulations of payoff structure, punishment risk, and punishment type.

		Punishment Risk (Audit Rate, p)						Declared Income (R_i)
		0%		30%				
				Punishment Type				
				Individual (R_I)		*Group* (R_{Gr})		
		Expected Company Income $E(I)$	Expected Gain from Dishonesty	I_E	Exp. Gain	I_E	Exp. Gain	
Payoff Structure	*Individual* (R_I)	(A) *Ind-No* $E(I) = I - tR_I$		(C) *Ind-Ind* $E(I) = p\ (I - tR_I) + (1-p)\ \{I - tR_I - f[t(I - R_I)]\}$		(E) *Ind-Gr* $E(I) = p\ (I - tR_I) + (1-p)\ \{I - tR_I - f[t(I - R_{Gr})]\}$		
		1000	1.33	850	1.13	900	1.29	0 (FD)
		875	1.17	800	1.07	800	1.14	500 (MD)
		750		750		700		1000 (H)
	Group (R_{Gr})	(B) *Gr-No* $E(I) = I - tR_{Gr}$		(D) *Gr-Ind* $E(I) = p\ (I - tR_{Gr}) + (1-p)\ \{I - tR_{Gr} - f[t(I - R_I)]\}$		(F) *Gr-Gr* $E(I) = p\ (I - tR_{Gr}) + (1-p)\ \{I - tR_{Gr} - f[t(I - R_{Gr})]\}$		
		916.66	1.50 (1.10 * l)	766.66	0.92	816.66	1.04	0 (FD)
		875	1.43 (1.05 * l)	800	0.96	800	1.02	500 (MD)
		833.33		833.33		783.33		1000 (H)

Note. FD = fully dishonest ($R_i = 0$), MD = moderately dishonest ($R_i = 500$), H = honest ($R_i = 1000$); l (loyalty parameter) = 1.36.

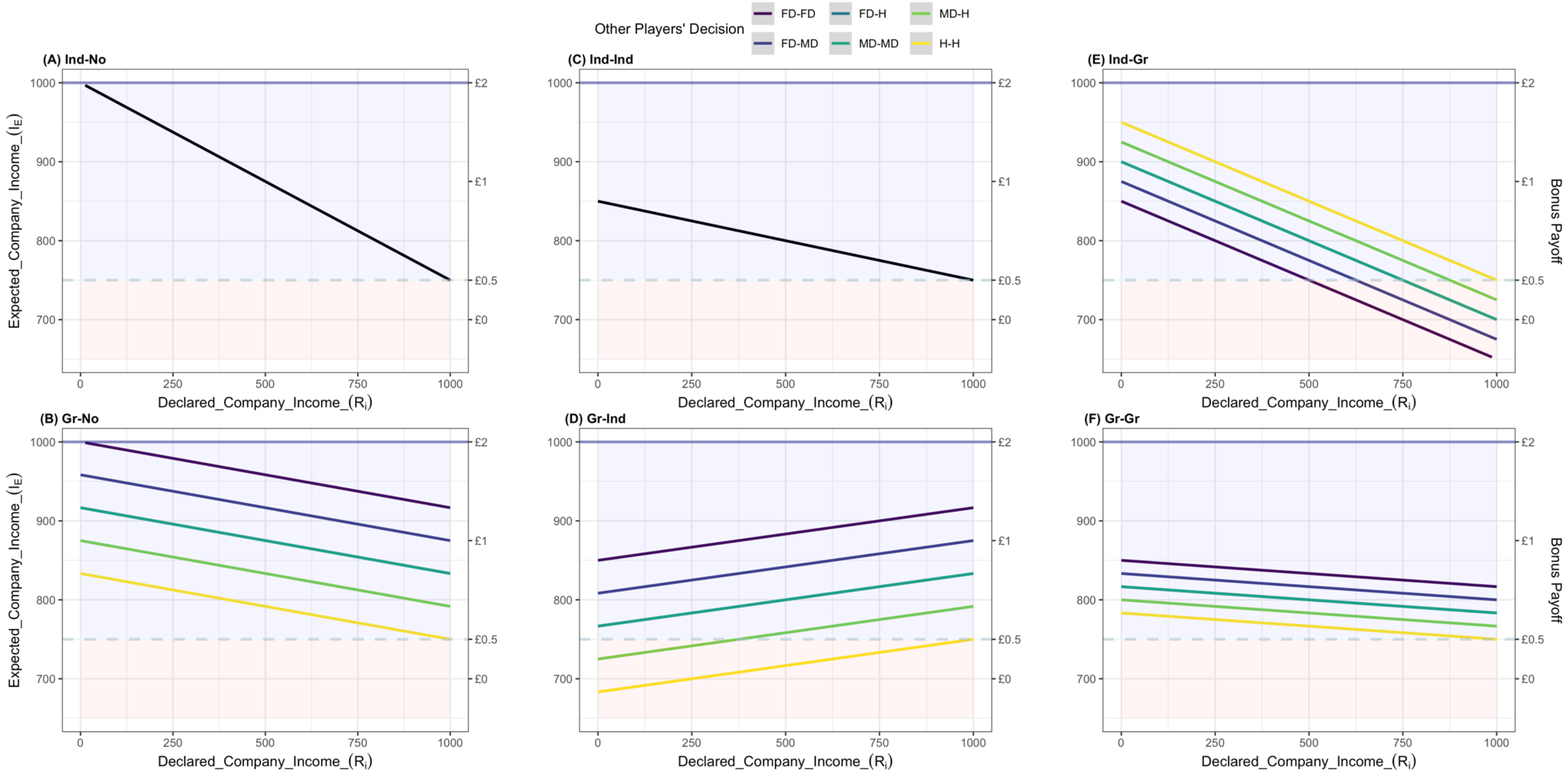

Figure 1. Simulated expected company income (I_E) for levels of fully dishonest ($R_i = 0$), moderately dishonest ($R_i = 500$), and honest (($R_i = 1000$) reporting and combinations of group members reporting (FD = fully dishonest, MD = moderately dishonest, H = honest) across the manipulations of payoff structure, punishment risk, and punishment type. Blue box and lines refer to bonus payoff > £0, while red box refers to bonus payoff < £0. Thick blue line refers to bonus payoff of £2, blue box refers to bonus payoff of £1, dashed blue line refer to bonus payoff of £0.5 and red box refers to bonus payoff of £0.

- Clarify the rationale of the intended experimental design, particularly the payoff scheme (review criterion 1C).

Based on comments by R2 we have revised the experimental design, now providing a full-factorial design of the factors and keeping the audit rate constant across treatments by performing individual audits in the individual punishment and group audits in the group punishment treatments.

- Consider whether the position of the additional manipulation check within the procedure should be adjusted (review criterion 1E)

Based on comments by R3 we now suggest another measure of commitment before the game task that is intended as a manipulation check. In addition, we provide more detailed information in the design table (Table 2) when this manipulation check is considered successful.

These issues fall within the normal scope of a Stage 1 evaluation and can be addressed in a comprehensive round of revisions.

Warmest,

Rima-Maria Rahal

Reviewer 1:

The abstract is well-written and synthesizes the main ideas of the registered report.

In the introduction, I understand what the authors mean by saying that some share of the tax revenue is lost due to tax fraud. However, it is also worth noting that a much higher share of taxes is not collected because very rich companies usually make a deal (usually corrupt deals) with international governments to pay very low percentages of taxes. Although it is important to pay attention to the lost revenue due to tax fraud among individual citizens or small companies, it is also important to notice that much more is lost when the billionaire companies pay an insignificant share of their incomes in taxes.

We thank you for your suggestion and agree that in many situations another issue might be companies trying to reduce company taxes by establishing their headquarters in locations with lower tax rates (i.e., tax havens). Those deals you are mentioning are typically on an institutional level (i.e., including local governments). However, the current focus and method is on evading taxes, which characterizes an individual level problem (company level), which institutions typically want

to reduce. We will acknowledge this in the discussion, and we have already added a footnote to acknowledge this:

"The current manuscript focuses on organizational tax fraud. There are different potentially dishonest practices that might negatively impact society such as paying a low percentage of taxes through so-called *tax havens*, which is not the focus here." (footnote 1)

Minor detail: in the last paragraph of page 3, authors used a comma instead of a point before "on the other hand…"

Thank you for catching this, we have changed it accordingly.

The methodological propositions of the authors are sound but I have a theoretical concern that should be addressed by the authors at some point. Establishing a 30% chance of being caught makes sense according to previous literature and it is well-operationalized in the study design. However, how does it apply to real-life? When people decide whether to act dishonestly or not, they usually do not make calculations of the percentage of being caught, rather they usually make decisions based on a lot of distinct factors (e.g., how many people have been caught in their surroundings etc). What real-life settings would correspond to a "30% chance of being caught"?

We agree that it is important to consider the likelihood of punishment compared to real-life events. We mainly selected 30% based on the previous study by Fochmann et al. (2021). Likelihoods of tax auditing in organizations vary across regions. In some regions, likelihoods might be small, while some have even introduced auditing probabilities of 100% within the first 5 years of establishment of a company. Likelihoods also differ per size of company, with smaller companies being less likely to be audited. Looking at actual numbers, the Internal Revenue Service plans to audit large companies with a likelihood of 22.6% in 2026 (IRS, 2024: https://www.irs.gov/newsroom/irs-releases-strategic-operating-plan-update-outlining-future-priorities-transformation-momentum-accelerating-following-long-list-of-successes-for-taxpayers). Similarly, the likelihood of large companies being audited in Germany was 22% in 2022 (Bundesministerium der Finanzen, 2023: https://www.bundesfinanzministerium.de/Monatsberichte/2023/10/Inhalte/Kapitel-3-Analysen/3-2-steuerliche-betriebspruefung-2022.html). These numbers are not too far away from 30% as used in the current study and to be in line with the study by Fochmann et al. (2021) we decided to keep this percentage.

Furthermore, regarding your point that "when people decide whether to act dishonestly or not, they usually do not make calculations of the percentage of being caught," we agree that this reflects real-life decision-making, where the probability of being caught is often ambiguous. However, the alternative would be to provide no information about this probability at all. In reality, people typically form some intuitive sense of the likelihood of being caught, even if it is imprecise. In line with Becker's (1968) theory, we expect people to at least implicitly calculate the expected utility of their actions, taking into account the likelihood and severity of punishment as tradeoff to the expected benefit.

We have added a short reflection on real-life settings:

"Audit rate is based on previous study by Fochmann et al. (2021). We acknowledge that in real-life settings audit rates can differ considerably based on company size and region. Some regions apply

or plan to apply audit rates that are similar to the current one (Bundesministerium der Finanzen, 2023; IRS, 2024).” (footnote 3)

As a sidenote, I think that it is important for us, researchers of dishonest acts, to think about other possibilities of curbing corruption that do not rely exclusively on punishment. This is because those who are punished in a corporate or governmental environment are usually scapegoats, and the corrupt/dishonest structure keeps working despite the punishment of the scapegoat. Group pressure may be a promising venue for this.

We agree that punishment might have negative side effects. The main focus of the current research is on punishment, but we will discuss alternative strategies as highlighted in previous research in the Discussion section. We have also added a measure on self-reported stress to evaluate the potential negative impact of audits and punishments.

Regarding the experimental design, I considered it hard to understand the bonus payments. It may also be hard to explain it for the participants in the experimental design, so the sample will be likely biased toward highly-educated individuals. Furthermore, UK is a place where people are highly educated, so findings may not be generalizable to many countries. Authors should point this out in the discussion section of the future manuscript.

This is an important point, and we will discuss the constraints of generalizability of our sample in the Discussion section when submitting the full manuscript. We have attempted to simplify instructions further by cutting information and providing a slider to illustrate the game outcomes (see Supplementary Note 2) and conducted two pilot tests to gauge comprehension of the instructions (Supplementary Note 4) and simplified them further based on these results.

Reviewer 2:

The submitted registered report addresses a relevant and novel research question. Personally, I would be very interested to see the results of the suggested experiment. Furthermore, I believe that tax compliance is a very good practical example to assess the impact of individual versus group punishment. However, I struggle with the theoretical placement of the research question and the theoretical decision model. Therefore, I have a couple of comments and questions regarding the theoretical argumentation and their fit to the experimental design. Probably most of my questions are the result of my personal shortcomings in understanding the provided material and I apologize for that. Nevertheless, the points raised need clarification before I can issue my final assessment. Overall, I hope the authors will find my feedback helpful to further develop their research.

1) To regard tax compliance as a mere question of dishonesty does not fully capture the problem. Fochmann et al. (2021) explain to some length that tax compliance combines

dishonesty with risk-taking. I believe the submitted report would benefit from taking this more specific description of the tax compliance decision into consideration. There is a large literature on risk-taking by groups that currently is barely considered by the authors. See for example the work by Gary Charness and co-authors.

We have provided more background and a short discussion on the literature of risk-taking summarizing that studies have shown diverging effects whether group decisions become more or less risky.

"Punishment has been mostly tested for individual decisions focusing on retribution of the individual, though recent theories have highlighted the importance of social context and influence (Pratt et al., 2017; Weisel & Shalvi, 2022). Indeed, dishonesty, here defined as *distorting the true state to acquire profits* (Zickfeld et al., 2024), seems to increase when individuals collaborate with others (Leib et al., 2021) or feel commitment towards "partners in crime" (Zickfeld et al., 2024). Such *collaborative dishonesty* (Kocher et al., 2018; Weisel & Shalvi, 2015) has been theorized based on differences in risk-taking between individuals and groups (Fochmann et al., 2021; Jiang & Villeval, 2024). Under increasing risks of punishment, dishonest behavior becomes riskier in that decision makers might face potential penalties and reduced profits. However, there is mixed evidence in whether groups perform riskier decisions than individuals with some observing a *risky shift* (Isenberg, 1986) and others observing a *cautious shift* in groups (Shupp & Williams, 2008). There is accumulating evidence that groups perform more rational decisions (Charness et al., 2007) and based on utility theory, dishonesty can be considered the more rational choice in situations in which potential benefits outweigh their costs (i.e., punishment). Responsibility for and conformity with group members might be additional factors influencing risk taking in groups (Charness & Jackson, 2009).

Indeed, empirical evidence suggests that group decision making increases dishonesty because loyalty and prosocial concerns and norms (i.e., dishonesty benefiting the group or other group members) can overwrite the motivation to tell the truth (Leib et al., 2021; Zickfeld et al., 2024). Based on these findings, in settings in which dishonesty cannot be detected, such loyalty norms should always appeal to dishonesty to benefit the group (as it presents the more rational economic and social choice). Once the risk of punishment increases, group decisions might become more nuanced as dishonesty could incur negative outcomes on the group when detected. Individual punishment becomes more costly in group settings since transgressors are more difficult to detect (Miceli & Segerson, 2007). In such settings, group punishment —i.e., the collective retribution of a community based on few transgressors—can become a viable tool because it punishes the dishonest person with definite probability (Miceli & Segerson, 2007).

On the downside, group punishment also incurs costs on innocent individuals (Pereira & Prooijen, 2018). Indirect evidence suggests that loyalty concerns in group settings can undermine potential effects of punishment or whistleblowing behaviors (Batolas et al., 2023; Jiang & Villeval, 2024; Rullo et al., 2024; Solaz et al., 2019). Loyalty concerns with the risk of potential group punishment can appeal to maximizing the group's benefit either through dishonesty or highlighting the responsibility and the negative costs that dishonesty can cause on the group outcome. Social responsibility indeed seems to reduce risk-taking behavior (Charness & Jackson, 2009). Importantly, there is limited

direct evidence on whether interventions tested in individual settings, such as introducing the risk of punishment, are also applicable and effective in group settings and whether group punishment can be more effective than individual punishment." (pp. 5-7)

In addition, we have included a direct measure of general risk taking/aversion tendencies in the proposed study and an analysis investigating the moderating effect of this. Based on discussions and recommendations by Alm and Malézieux (2021) of assessing risk taking in tax evasion games we have included a one-item self-report scale based on a previous study (Casal et al., 2016).

"Risk Aversion. We will include a one-item measure on risk aversion ("Generally speaking, would you characterize yourself as someone who is willing to take risks, or as someone who is avoiding risks?") on a scale from 1 (absolute risk aversion) to 9 (absolute risk seeking) as employed in a previous study (Casal et al., 2016). While there exist multiple measures to assess risk taking or risk aversion, research on tax evasion has recommended employing simple single item scales (Alm & Malézieux, 2021)." (pp. 28-29)

2) Many experimental studies have investigated the effect of punishment, i.e., different audit and penalty rates, on individual tax compliance. None of these studies is currently mentioned in the literature overview on the effects of punishment. This literature seems to be very relevant to the suggested study. It provides ample causal evidence on the effects of punishment and it is perhaps more closely related to the experimental design than the literature on punishment in general that is currently considered by the authors. For an overview of relevant experimental studies see Alm and Malézieux (2021).

We failed to review the relevant studies as summarized by Alm and Malézieux (2021). We have now included a review of this literature as it is really relevant to the suggested design. Similar to the already reviewed literature the findings are more complex. In the meta-analysis Alm and Malézieux (2021) find that increasing audit rate (risk of punishment) increases tax compliance and this interacts positively with penalty rate (severity of punishment) on the decision whether or not to evade taxes (extensive margin). However, for tax evaders (intensive margin) there is a negative association with audit rate and compliance, suggesting a more nuanced effect of punishment.

"In economics, a large body of literature has investigated the impact of punishment, referred to as audits, in so-called tax evasion games, tasks that model real world tax reporting (Alm & Malézieux, 2021). Across two meta-analyses (Alm & Malézieux, 2021; Blackwell, 2010), risk of punishment (i.e., audit probability) was associated with increased tax compliance (i.e., honesty) and positively interacted with severity of punishment (i.e., fine/penalty rate). However, when considering dishonest reporters only, increased risk of

punishment (and its interaction with fine size) predicted honesty negatively. Importantly, many studies employing tax evasion games do not compare risk of punishment to a control condition (i.e., no risk of punishment) and therefore evidence is mostly correlational. " (p. 4)

3) In my opinion, the first contribution is overstated. This contribution has already been made by Bonfim and Silva (2019) as well as Fochmann et al (2021). Regarding Fochmann et al. (2021): It is true that punishment is not manipulated in the underlying experiment. However, the authors do manipulate the group setting. Therefore, the effect of the group context on decision making can be assessed exactly. The effect of the risk of punishment can be assessed in comparison to prior studies on dishonesty that did not include risk. This is the contribution the authors make. The current paper verifies this finding by integrating a "no risk" and a "risk" treatment in one study. Nevertheless, this has been done by Bonfim and Silva (2019) (On a sidenote: The reference list includes a wrong title in reference to Bonfim & Silva, 2019)). The way I see it, the main contribution of this paper really lies in the manipulation of punishments and therefore payoffs at the group versus the individual level. This contribution should be highlighted.

We have revised the contribution statement section and now make clear that the first contribution is to replicate the effect of collaborative dishonesty and the second contribution (originally listed as the first contribution) is to replicate previous studies (Bonfim & Silva, 2019). Replication is a cornerstone of science and important for knowledge generation and therefore, we think that it is an important contribution of the current study to replicate previous findings by Bonfim and Silva (2019). We have also highlighted that the third contribution, crossing these factors, is the main contribution and has most likely not been tested yet.

"Thereby, we provide three central contributions. First, we aim to replicate the finding that group-based (vs. individual) decision-making increases dishonesty (Leib et al., 2021; for studies using tax evasion see Lohse & Simon, 2021; Matthaei & Kiesewetter, 2020). Second, we aim to replicate the effectiveness of punishment in group (vs. individual) settings (Bonfim & Silva, 2019). Third, we investigate the effectiveness of individual or group punishment for individual- or group-based dishonesty by crossing these factors. To the best of our knowledge, the full interaction between individual- and group-based decision making and individual or group punishment has not yet been tested systematically." (pp. 8-9)

We have also corrected the title in the reference list for Bonfim & Silva (2019).

Regarding Fochmann et al. (2021), as the reviewer notes this study manipulates individual vs. group decision making, but not punishment. Therefore, it does not provide any causal

evidence on the effect of punishment. Also comparing the findings with previous work do not allow for any causal claims in this study, which is why we do not list it for the second contribution.

4) The effect of payoff differentials, as they result in the suggested individual punishment scenarios, on joint decision making has been previously investigated in the context of tax declarations. Relevant papers are Lohse and Simon (2021), as well as Matthaei and Kiesewetter (2022). These studies should be taken into account when assessing the contribution of the current study.

We have added the studies by Lohse and Simon (2021) and Matthaei and Kiesewetter (2020 - we couldn't find an article dated 2022) highlighting previous studies investigating the effect of individual vs. group-based decision making for tax evasion settings.

"Thereby, we provide three central contributions. First, we aim to replicate the finding that group-based (vs. individual) decision-making increases dishonesty (Leib et al., 2021; for studies using tax evasion see Lohse & Simon, 2021; Matthaei & Kiesewetter, 2020). Second, we aim to replicate the effectiveness of punishment in group (vs. individual) settings (Bonfim & Silva, 2019). Third, we investigate the effectiveness of individual or group punishment for individual- or group-based dishonesty by crossing these factors. To the best of our knowledge, the full interaction between individual- and group-based decision making and individual or group punishment has not yet been tested systematically." (pp. 8-9)

5) Unfortunately, for me personally, the predictions are very hard to follow. I for one do not see how the expected payoff equals YL, or 0.5, in case of honest reporting according to Equation 1. Does Equation 1 truly represent the expected payoff for the individual or does Equation 1 represent the additional expected payoff in case of dishonest reporting? These two options are clearly not identical. I kindly ask the authors to clarify this. In addition, I suggest to state the expected payoff function in terms of true income, declared income, tax rate, audit and penalty rate as it is commonly done in tax compliance studies.

We have revised the prediction section carefully and now provide a more detailed and step-by-step discussion of our predictions. To simplify presentation, we now present three different equations, following classic calculations as used in the literature on tax evasion (Allingham & Sandmo, 1972; Alm & Malézieux, 2021). We have also changed the terms, now referring to concepts typically used in studies on tax compliance.

"Predictions

Based on expected utility theory and the deterrence model of crime (Becker, 1968) we can derive the expected income function for each experimental treatment based on common equations employed in the tax evasion literature (Allingham & Sandmo, 1972;

Alm & Malézieux, 2021). To align with the literature on tax evasion, we refer to the fixed amount of company income as the *true income* (*I*), the reported amount of company income by the individual or group as the *declared company income* (*R*), the share of the declared company income paid as taxes as the *tax rate* (*t*), the tax rate times the declared income as the applicable tax (*T*), the risk of punishment as the *audit rate* (*p*), and the severity of punishment as the *penalty rate* (*f*). Across the study, the true company income (*I*) is fixed at 1000, the tax rate (*t*) at 25%, and the penalty rate (*f*) at 2 (referring to the lost amount of taxes plus a fine of the same size) based on Fochmann et al. (2021; see Method or Figure 1 for more details). The applicable tax (*T*) equals the declared company income (*R*) times the tax rate (*t*).

Equation 1 depicts the *company success* (i.e., company income after taxes) when not being caught underreporting the company income (or when reporting the true income). The company success when not being caught (I_N) is determined by the difference between the true company income (I = 1000) and the declared company income (*R*) times the tax rate of 25% (*t*). For example, for a declared company income of 0 the company success is 1000 (1000 – 0 * 0.25).

Equation 1

$$I_N = I - t * R$$

Equation 2 depicts the *company success* when being caught (I_C). In this case, the penalty equals twice the unpaid taxes. If the declared company income equals the true company income (i.e., honest reporting, $R = I$) the penalty equals 0 and thus Equation 1 applies. However, if the declared company income differs from the true company income (i.e., dishonest reporting, $R < I$) the penalty equals two times the difference in true and declared company income times the tax rate (i.e., the unpaid tax). For instance, declaring a company income of 0 and being audited a penalty of two times 25% of 1000 (i.e., 500) is applied and the company success when being caught (I_C) is 500.

Equation 2

$$I_C = I - t * R - f * [t * (I - R)]$$

To calculate the *expected company income E*(*I*) we use Equation 3, which sums the probability of an audit (*p*) times the company success when being caught (I_C) and the probability of not being audited (1-*p*) times the company success when not being caught (I_N). In case of audit rate being 0% or honest reporting, Equation 3 is identical to Equation 1. However, in case of underreporting the true company income, expected company income needs to consider the probability of an audit and the company success when being caught. In this case, reporting a declared company income of 0 results in an expected company income *E*(*I*) of 850 (0.3 * 500 + 0.7 * 1000).

Equation 3

$$\text{Expected company income } E(I) = p * I_C + (1\text{-}\ p) * I_N$$

In the current study, we implement a factorial design manipulating punishment risk (audit rate 0% vs. 30%), punishment type (individual vs. group), and payoff structure (individual vs. group; see Method and Figure 2 for more details). Below and in Table 1, we provide an example for each combination of treatments and for reporting the company income as honest (*R* = 1000), moderately dishonest (*R* = 500), and fully dishonest (*R* = 0). [1] As the control treatment, we implement a situation in which the audit rate is 0% and there is no risk of punishment (see Table 1, A). In this situation, the expected company income is based on Equation 1. We further manipulate the payoff structure across treatments, with incomes either based on individual reports (R_I) or group reports (R_{Gr}) - with the group report being the *average* of the three individual reports in a group (Figure 1). Importantly, in the group payoff structure reports are first made independently and individually without the possibility for communication and only later averaged for the group report. We implement five rounds of the task to model temporal effects over time. However, to control for possible learning or expectations effects participants receive no feedback after the specific round, but only after completing all five rounds. Therefore, they will not be able to know the other participant's reports, the group average report, or whether an audit occurred before the end of the game. This is to ensure control of potential learning effects and models real-world situations in which there is often a temporal delay of possibly implemented audits." (pp. 9-12)

We have also revised Table 1 to provide a more detailed overview of the specific predictions with regard to each treatment.

6) Changing the wording and probability display from risk of punishment to probability of not-being punished decreases clarity and makes it harder to follow the explanations. I suggest to follow a common wording.

We have changed this aspect when revising the prediction section and now focus on the audit rate or the probability of being punished.

7) Similarly, I do not understand how payoffs displayed in Table 1 are derived. From what I see in cell "Individual / 0% punishment risk", Table 1 shows the expected additional payoff from being dishonest. Again, this is not the same as the overall expected payoff. Statements such as "In the individual payoff/no punishment treatment, the risk of being punished is zero and the gain from full dishonesty is £1.5, thereby resulting in an expected payoff of 1.5." sadly do not help the confusion between "gain from dishonesty" and total "expected payoff".

We have further clarified how Table 1 was derived and now focus on the ratio between expected payoff for dishonesty and expected payoff of being honest. If this is 1, expected payoff is identical for being honest or dishonest, if this is below 1 expected payoff is higher for being honest, and if it is above 1 expected payoff is higher for being dishonest. Further, we have revised Table 1 and provide more detail on how we derive the predictions for each

treatment. This is also visualized with the help of Figure 1 that provides further details in the expected payoff for different levels of reporting.

8) As far as I understand the experimental setup the audit is always applied at the individual level. From the description of the experiment I see two different audit and punishment scenarios:

a. Individual punishment: Each individual declaration is subject to an audit risk of 30% and only the individual that reported too little income is punished.

b. Group punishment: Each individual declaration is subject to an audit risk of 30% but every member of the group is punished for every individual that reported too little income.

There is no treatment where only the group decision, i.e., the median of the individual decisions, is subject to an audit risk and penalties are based on the comparison of the group declaration to the true income. My understanding matches the statement on p. 17 that "Audits are individually applied across all treatments for each round." However, when reading the experimental instructions (Supplementary Note 2 p. 8) I see instructions which describe that only the group income will be compared to the true income in case of an audit. The only way I can see these instructions matching the description in the main text is if each individual declaration is replaced with the group declaration, i.e., the median individual declaration and then an individual audit is applied, in the group payoff/ group punishment treatment. If this is indeed the correct interpretation this has to be clarified in the main text but especially also in the experimental instructions. Otherwise, participants will not understand what will happen in the experiment.

We have further clarified the instructions and revised them to fit the purpose of the experiment. Thank you for raising these concerns and allowing us to further reflect on the proposed experimental setup. We have now changed the design so that the individual punishment is based on an audit of the individual report and imposes a penalty on the individual, while the group punishment is based on an audit of the group and imposes a penalty on the group. This keeps the audit rate identical at 30% for both the individual and group punishment treatments and corrects a previous flaw in the design.

"The risk of punishment is implemented by varying a 0% or 30% audit rate. Audits are constant across rounds. Thus, for each round participants face the exact same risk of an audit. Further, we vary the type of punishment by implementing an individual punishment or group punishment. In the individual punishment treatments, each individual report is audited, and the penalty is applied at the individual level. In the Ind-Ind treatment this means that each group member independently faces an audit rate of 30% for their individual report and penalties are applied to the individual payoff. In the Gr-Ind treatment this means that each group member independently faces an audit rate of 30% for their individual report and penalties are also applied to the individual payoff. However, individual punishment does not affect the group payoff which is still calculated on the

individual reports (even if an individual audit occurs). In the group punishment treatments, each group is audited with an audit rate of 30% and the penalty is applied at the group level. In the group payoff – group punishment treatment this means that the group average is audited at a 30% rate and a penalty is applied to each individual payoff. In the Ind-Gr treatment this means that the group is audited at a 30% rate and the penalty is applied to each individual payoff." (p. 27)

9) Related to my previous comment, participants cannot understand that there is an increased audit risk in case of group punishment given the current experimental instructions. This has to be made very explicit.

We have changed the setup based on your feedback and now the audit rate is constant across the individual and group punishment (30%)

10) Not including a treatment in which only the group decision is subject to an audit is a choice made by the experimenter. Nevertheless, this choice may be problematic for several reasons:

1) It is not fully clear which treatment serves as a baseline comparison for the group payoff/ group punishment treatment. This treatment seems to differ in more than one aspect from the other treatments as

a. The individual's declaration has externalities as the median choice determines the individual's payoffs.

b. The individual being audited has externalities as all group members will have to pay a fine.

c. The audit risk increases for the individual.

In my understanding, it is not possible to differentiate which of the three changes, A, B, or C, causes a change in the individual reporting decision. Nevertheless, this may be related to my trouble to fully understand the experimental setup.

This was indeed a flaw of the design, and we have corrected this by applying the same audit rate for the group punishment. Now, the designs are more comparable, and it is possible to experimentally determine what drives changes in the outcome. For example, for the group payoff/group punishment as mentioned the group payoff effect (A) can be compared to the individual payoff/group punishment treatment and the group punishment effect (B) can be compared to the group payoff/individual punishment treatment. C does not apply anymore as the audit risk is identical across the four experimental treatments.

2) The design is odd given that in reality the group decision is what would be reported to financial authorities and subject to an audit risk not the underlying individual decisions.

Especially, the scenario of group payoff but individual punishment feels odd given this background. As far as I understand the experimental setup in this treatment, the individually reported income will be replaced by the median of the individual declarations and then the individual faces an audit risk for this declaration. This refers to a situation were someone else makes a decision but the subject faces the risk of this decision alone, e.g., in case of joint tax reporting partner A fills in the tax return but partner B has to pay the fine in case of an audit. This is possible for sure but seems rather constructed.

We agree that the group payoff/individual punishment is a more uncommon situation. However, our aim is to compare and isolate the effect of each treatment, which affords a full factorial manipulation of the factors. We now provide a detailed reflection on the design and how it relates to situations in real life and provide an example for each combination when introducing the combination. This shows that each combination of factors occurs in real life and has negative impacts. For the group payoff/individual punishment we now provide the following example:

"This setting is similar to when a parent company has subsidiaries and prepares consolidated financial statements. The group benefits from the retained earnings that are higher if a subsidiary evades taxes and can redistribute these within the group. The tax evasion of the subsidiary is punished individually, as it retains its status as a separate legal entity. Such issues pertain especially in international tax settings and are often achieved through transfer pricing, where profits can be shifted to lower tax countries (Diller et al., 2024; see also IRC Section 482, 1504). " (pp. 18-19)

3) The design choice further limits comparability to previous studies. This choice should be motivated and explained in more detail. The explanation currently included on page 17-18 is not convincing given the tax framing of the study.

As mentioned in the previous comments, we have now revised the experimental setup, which makes it more comparable to previous findings.

11) The experimental setup is complicated and participants are required to do quite a bit of math to understand the outcome of their decision. Therefore, I suggest to include additional examples in the instructions. Could it be easier for the participants to abandon the 20% share of group income and to simply split income by 3? I was wondering if the 20%-share may be confused by participants with the 1 out 5 decisions that is chosen for payoff in the end.

We have streamlined the instructions and now provide a slider that provides participants with an example of bonus payoffs at different income reports. We have also abandoned the 20% share of the company income, which was originally based on Fochmann et al. (2021) and now calculate individual payoff directly on the company success. This removes one step in the calculation of payoff and simplifies the design. At the same time, the individual

payoff is a fraction of the company success, highlighting individual payoff as a function of company success. We have also conducted two pilot tests (Supplementary Note 4) to gauge difficulty and comprehension of instructions and have further revised and simplified instructions based on the feedback.

12) The demographics are not mentioned in the analysis plan. Are these going to be added as control variables in the regressions? I specifically wonder about asking for nationality – do you plan to investigate differences regarding heterogeneity between groups? I do not see a control for individual risk preferences. I suggest to add this.

We have added an additional exploratory analysis controlling for the demographics. Based on experiences with sampling on Prolific, we don't expect many participants to be of a different nationality than the UK. This is often less than 10% of the sample based on our experiences. Nationalities are also not sampled systematically to allow any cultural impact of risk aversion or cheating propensity.

"Demographics. We repeat the model in H3/H5 with gender (male: 0, female: 1), age (mean centered), and nationality (other: 0, UK: 1) as additional predictors and their interaction with payoff structure and punishment. We only include nationality if at least 10% of responses report a nationality different than the UK. " (p. 31)

Based on your comment, we have included a direct measure of general risk taking/aversion tendencies in the proposed study and an analysis investigating the moderating effect of this. Based on discussions and recommendations by Alm and Malézieux (2021) of assessing risk taking in tax evasion games we have included a one-item self-report scale based on a previous study (Calas et al., 2016).

"Risk Aversion. We will include a one-item measure on risk aversion ("Generally speaking, would you characterize yourself as someone who is willing to take risks, or as someone who is avoiding risks?") on a scale from 1 (absolute risk aversion) to 9 (absolute risk seeking) as employed in a previous study (Casal et al., 2016). While there exist multiple measures to assess risk taking or risk aversion, research on tax evasion has recommended employing simple single item scales (Alm & Malézieux, 2021)." (pp. 28-29)

References

Alm, J., & Malézieux, A. (2021). 40 years of tax evasion games: A meta-analysis. Experimental Economics, 24(3), 699–750.

Bonfim, M. P., & Silva, C. A. T. (2019). Inhibitory Mechanisms on Dishonesty of Groups and Individuals. Journal of Accounting, Management and Governance, 22(2), 205-226.

Fochmann, M., Fochmann, N., Kocher, M. G., & Müller, N. (2021). Dishonesty and risk-taking: Compliance decisions of individuals and groups. Journal of Economic Behavior & Organization, 185, 250-286.

Lohse, T., & Simon, S. A. (2021). Compliance in teams – Implications of joint decisions and shared consequences. Journal of Behavioral and Experimental Economics, 94, 101745.

Matthaei, E., & Kiesewetter, D. (2022). Group Decisions and Asymmetric Payoffs: Risky Tax Avoidance in the Laboratory. Available at SSRN: https://ssrn.com/abstract=3626982

Reviewer 3:

Short summary
The Effect of Individual and Group Punishment on Individual and Group-Based Dishonesty is a stage 1 RR proposing to investigate the effect of punishment (structure) and payoff structure on dishonest behavior.
The authors provided a clear introduction of the existing literature and then detailed the hypothesis generation process, followed by sample size planning and a clear illustration of experimental procedures, with the main task being a tax evasion game. The main IVs are payoff structure (individual vs. group) and punishment (no vs. individual vs. group). The authors also planned to measure a few variables including feeling of commitment, guilt and anger, as well as humility-honesty.

The main hypotheses are:
1) Group payoff increases dishonesty (i.e., non-compliance in the tax evasion game) compared to the individual payoff (baseline) for the no punishment treatment;
2) A main effect of risk of punishment. Risk of punishment (vs. no punishment) reduces dishonesty (i.e., non-compliance);
3) A main effect of type of punishment. Group punishment will show stronger effects in reducing dishonesty compared to individual punishment; and
4) H40: No interaction effect of type of punishment and payoff structure. H41: An interaction effect of risk and type of punishment and payoff structure. Group payoff increases dishonesty compared to individual payoff, but only for no punishment or individual punishment. For group punishment, individual payoff increases dishonesty compared to group payoff.

The analysis plan is consistent with the hypotheses overall. I also thank the authors providing the details of power analyses and experimental materials in the Supplementary.

However, I do have a few comments that I feel the authors should consider before the IPA. Below please find my specific comments. I hope that they are helpful and apologize in advance if I have misunderstood parts of the study.

Comments

Abstract
1. ‘High-powered’ is not a precise term. Whether or not a sample offers high power also depends on the specific tests and the expected effect sizes. Perhaps this term can be removed (Minor).

This statement was based on our predictions and assumptions and the different power simulations. However, we agree that if the effect sizes are really small, we most likely do not have high power. Therefore, we removed the statement from the abstract.

Introduction
2. In the abstract and the introduction, the authors mentioned the severity of a test quite often. From my perspective, the proposed study examines the presence of punishment (yes. Vs. no) but not the severity, unless here we consider that the group punishment condition is more severe than the individual punishment condition. However, these two conditions differ not only in terms of severity.

We have removed the focus on severity from the abstract. However, we feel that for a full review we need to at least mention the effects of severity, especially since many studies show that risk and severity interact (e.g., Alm & Malézieux, 2021). To make it clear that we focus on risk (and not severity), we have added an explicit statement:

“Due to the complex nature of the effects of punishment, our study primarily focuses on varying punishment risks while keeping punishment severity constant. ” (p. 5)

3. In Zickfeld et al. 2023, the effect is reported in terms of Hedge’s g, not Cohen’s d. The conversion between the two could be made clearer in the manuscript

We have changed this to “ A recent meta-analysis observed an overall effect of *Hedges’ g* = 0.17 of commitment on dishonesty (Zickfeld et al., 2024; identical to *Cohen's d* for large sample sizes as in this case). ” (p. 16) to be more precise. For larger sample sizes both g and d are mostly identical.

Method
4. The authors stated that ‘In the unlikely case that we must exclude more than

20% of participants based on the registered exclusion criteria, we will collect another round of participants to fill up the original sample size.’. Does that mean that the authors will only do another round of collection when the exclusion is 20% or higher? Then it could mean that for some tests, you don’t have sufficient power. Why not recruit participants until your valid sample size reaches the planned sample size?

This decision was made as we will first recruit *N* = 630 participants and perform exclusions once collection is finished. As it is difficult to foresee the number of exclusions we included the criteria to sample more participants in case of more exclusions. We have now changed this specification so that we stop data collection once we retain *N* = 630 participants AFTER exclusions.

“We will recruit UK-based participants via Prolific, equally balancing men and women and screening for English language fluency. Participants will be paid £1.5 as a base payment and might earn an additional bonus payment between 0 and £2 based on their or their group’s reporting. We will stop data collection once we have recruited 630 participants after exclusions. Participants will be automatically screened out during recruitment when a) failing at least one comprehension check twice, b) failing two attention checks, or c) not being matched with a partner. Based on previous studies and two pilots on comprehension (Supplementary Note 4) we expect around 30% of participants to be screened out or excluded based on these criteria. Recruitment is stopped once 630 participants are retained after exclusions.” (p. 24)

5. For the simulated power analyses, if the total sample is 630. Then the sample for H1 would be 210 (2 no punishment conditions). Therefore, the power analysis should report the results with a sample of 210 instead of 630 (as reported in the design Table). The same principle applies to other hypotheses. Could the authors confirm that the specific sample sizes for different hypotheses are all sufficiently powered?

The different hypotheses are sufficiently powered given the current design and we have provided now more detail and a specific analysis for each hypothesis separately in the Supplementary Note 1. The original power analysis for H1 was already performed for a sample of 210 (and a total sample of 630 including all treatments) but this was indeed not fully clear from the information. We have now provided more detail on the specific sample sizes and effect sizes estimated for the different tests (Supplementary Note 1).

6. In the procedures, the authors propose to measure feeling of commitment at the end of the tax task as manipulation check. The manipulation check might be influenced by the task outcomes. what about measuring it before the formal task but after the practice?

This measure was originally not intended as a manipulation check but we agree that it could make sense to use it. We now suggest assessing commitment before the tax game and also after the tax game and provide analyses for both separately and combined.

“Social Commitment. Right before and after the tax report game, participants will complete a commitment measure adapted from Zickfeld et al. (2023). Participants will be asked how much they feel committed to their two group members on a scale from 1 to 7 (not committed at all to very much committed).” (p. 28)

“Manipulation check. We will compute a multilevel model with social commitment as the outcome and payoff structure as the predictor. We will repeat the model by adding punishment (contrast coded: punishment risk -2/3 (no punishment) vs. 1/3 (individual punishment) vs. 1/3 (group punishment); punishment type: 0 (no punishment) vs. –½ (individual punishment) vs. ½ (group punishment)) in a second step. We will compute two models, one with social commitment before the game as the outcome and one with social commitment after the game as the outcome. We follow up with post-hoc comparisons comparing each treatment to the baseline (Ind-No). Using Tukey adjustment to adjust for multiple comparisons. In addition, we will run one model with difference in social commitment (pre vs. post) as the outcome and payoff structure, punishment, and their interactions as predictors.” (p. 30)

7. On the analytical side, what would count as successful manipulation? It would also be nice to include the manipulation check test in the design table.

We have added more information to the design table (Table 2) summarizing the analysis we will perform for the manipulation check and also when the manipulation is considered successful. If the group-based treatments show statistically higher social commitment compared to the baseline (Ind-No) we consider the manipulation to be successful (more information detailed in Table 2).

8. Stage 1 report should not include exploratory analysis. Therefore, I suggest to only keep the planned analyses for manipulation check and the testing of H1 to H4. The authors can of course still run the additional analysis in Stage 2 but report them as exploratory.

We were not aware of this convention to not include exploratory analyses in the Stage 1 report. We have previously published Registered Reports including exploratory analyses included at Stage 1. We leave this to the editor to decide how we should deal with this and

would be open to leave the exploratory analysis or remove them and only add them at Stage 2.

9. Exploratory analyses on actual punishment, shouldn't it be that actual punishment from the previous round influence behavior in this round?

Yes this is correct. We have now changed the design in a way that participants don't receive feedback on potential punishment before the end of all five rounds to control for potential learning effects. Therefore, this analysis was removed, and the comment does not apply anymore.

10. Exploratory analysis on moral emotions: perhaps it would also be interesting to add a measure of shame, given the theoretical relevance of shame vs. guilt? (e.g., https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6143989/)

We agree that including shame would be interesting to include in the current study. However, since we mainly focus on a previous study investigating emotions in unethical behavior that focused on guilt and anger (Motro et al., 2018) and to avoid increasing the length of the study further we have decided not to include additional items on shame.

11. For the mediation analysis, if the moral emotions are not measured after each round of the game but at the end of the tax task. I wonder if it is appropriate to run such a mediation analysis. It's likely the case that the task outcomes influenced the emotions rather than the other way around.

We agree that this is a disadvantage and acknowledge that the measures could be influenced by post-treatment bias. We have now removed the mediation analysis and will only focus on regression analyses between the variables and emphasize that we cannot conclude anything about the direction of the relationship between moral emotions and dishonesty.

"Moral Emotions. We conduct two multilevel regression models with moral anger or guilt as the dependent variable. As predictors we will add payoff structure (individual: 0; group: 1) and punishment (contrast coded: punishment risk -2/3 vs. 1/3 vs. 1/3; punishment type: 0 vs. –½ vs. ½) and their interactions. Further, we will conduct an ordered beta regression model with compliance as the DV and both moral anger and guilt (mean centered) as predictors. " (p. 31)

12. Design table: it seems that for rejecting H0, you are using NHST, but for not rejecting H0, you are using minimal effect testing against SESOI. However, it could be case that the test is statistically significant but smaller than the SESOI. See here for recommendation on testing under a unified framework:

https://psycnet.apa.org/fulltext/2024-76241-001.pdf
https://royalsocietypublishing.org/doi/full/10.1098/rsos.221328

This was indeed mixed up for the design table. We have now updated the table and refer to null regions for interpretation of hypotheses based on Smiley et al. (2023). For example:

“H1$_0$ is rejected if the main effect of payoff structure is statistically significant, the 95% CI is not entirely included in the equivalence bounds, and compliance is higher for individual payoff compared to group payoff.

H1$_0$ is not rejected if the 90% TOST CI of the main effect of payoff structure is entirely included within the equivalence bounds.” (Table 2)

13. In the supplementary materials, ‘For the experimental treatments we expected a small effect in the individual payoff/individual punishment treatment. Given previous findings in the dishonesty literature (Gerlach et al., 2019; Leib et al., 2021; Zickfeld et al., 2023), we set our smallest effect size of interest (SESOI) for these main effects at d = +/- .15. Employing this effect for the group payoff/individual punishment treatment would suggest an increase in compliance of 5% (d ~ .16). We expected a somewhat stronger effect for the individual payoff/individual punishment treatment and set this at an increase of 10% (d = .33) or a compliance of 75%.’ The effect of individual payoff/individual punishment treatment appeared twice with different effect size expectations. Could the authors check?

Thank you for alerting us. We have checked this and corrected this part when revising the Supplementary Note 1. It now reads:

“For the experimental treatments, we focused on individual payoff/individual punishment (Ind-Ind), group payoff/individual punishment (Gr-Ind), individual payoff/group punishment (Ind-Gr), and group payoff/group punishment (Gr-Gr, Supplementary Table 1). Overall, we expect more compliance in the punishment treatments compared to the no punishment treatments. The highest compliance is expected for Gr-Ind, which shows the lowest expected gain from dishonesty (0.92). The lowest compliance is expected for Ind-Gr, which shows the highest expected gain from dishonesty for the experimental treatments (1.29). Since the expected gain is only slightly lower compared to the no punishment treatments, we expected a small effect in the Ind-Gr treatment. Given previous findings in the dishonesty literature (Gerlach et al., 2019; Leib et al., 2021; Zickfeld et al., 2023), we set our smallest effect size of interest (SESOI) for these main effects at *d* = +/- .15. Employing this effect for the Ind-Gr treatment would suggest an increase in compliance of 5% (*d* ~ .16), resulting in a compliance of 70% (the lowest compliance among the punishment

treatments. We expected a slightly lower expected gain for Ind-Ind (1.13) and therefore a somewhat stronger effect for the Ind-Ind treatment on compliance and set this at an increase of 10% (*d* = .33) or a compliance of 75%. We expect again a slightly lower expected gain for Gr-Gr (1.04) and therefore a slightly higher compliance, setting this to an increase in 15% (*d* = .50) from baseline or at 80%. Finally, we expect the lowest expected gain for Gr-Ind (0.92), setting this to an increase of 20% (*d* = .66) and a compliance of 85%. Therefore, we expect no interaction effect between type of punishment and payoff structure, as individual punishment should be more effective regardless of payoff structure. An overview of the expected tax compliance for each treatment is provided in Supplementary Table 1 and Supplementary Figure 1." (pp. 2-3)