Online Appendix to "Selection, Payment, and Information Assessment in Social Audits: A Behavioral Experiment"

O.1. Payment Manipulations

We introduced several elements in our experimental design to make the source of the Audit Fee clear in the Paying (P) and Choosing and Paying (CP) treatments: (1) We have a series of paying screens in these two treatments. After being matched with a Firm, Auditors saw a waiting screen that read "Waiting for payment. Please wait while you are being paid by the 1st [2nd] Firm". In the meantime, Firms had to actively click a button to pay the Auditors. In contrast, in treatment N, Auditors skipped these screens. (2) The matching screens were different across treatments. After being matched with a Firm, as well as paid by it, Auditors in treatments P and CP were explicitly informed that "The 1st [2nd] Firm has paid you 60 tokens for conducting the audit." In treatment N, on the other hand, they were informed that "You have received a payment of 60 tokens for conducting the audit." In the latter case, it was also made clear to them in the instructions that this payment did not come from the Firm. (3) Finally, we included a mandatory comprehension question after the instructions, which participants had to correctly answer before starting the experiment. This question stated: "The Firm pays the audit fee (60 tokens) to the Auditor out of the Firm's initial [N: 70, P and CP: 130] tokens." The right answer was False for N, and True for P and CP.

O.2. Additional Design Choices

First, we do not consider a step where the Auditor must choose and exert some costly effort to obtain the audit information. We make this choice for two main reasons. (1) Though some effort is needed to gather information in practice, in most cases there is limited room for auditors to deviate from a contracted level of detail, and hence effort, in their processes. As auditors with decades of experience described it in a recent study, "auditing firms do not have a whole lot of influence on the methodology. We can influence it in a very limited way;" and "audit schemes define the timeframes or have guidance. Put guidance in inverted commas because there's a kind of push to minimize time" (Kashyap 2022). If an auditor deviates from the agreed level of detail, audited companies can raise complaints (Dreier 2023). (2) From a design perspective, allowing Auditor participants to choose the level of effort, and thus the quality of the information obtained, would introduce endogeneity in the latter. This, in turn, would make our main research questions—to understand how social auditors assess noisy information, and how this key aspect of the audit is affected by the hiring scheme—much more difficult to address. For example, if being chosen by the Firm led to a higher effort and much better quality of information, then we would be unable to compare participants' information assessment under equal conditions in the P, N, and CP treatments.

Second, we randomly form new 4-person (2 Auditors and 2 Firms) groups each round from within the 12-24 participants in the session , as opposed to keeping groups constant over time, for several reasons: i) Fixed groups would introduce information asymmetry, as Firms would know the plate composition that a matched Auditor observed, but they would not know what information the other, non-matched, Auditor saw in the same round. ii) Keeping group composition constant would similarly risk introducing status quo bias (Samuelson and Zeckhauser 1988): all else equal, Firms are likely to prefer a "known" Auditor over

an unknown one. And iii) the strength of reciprocal behavior between Firms and Auditors is likely to be influenced by the number of rounds in which they have interacted. Though long-term reciprocal behavior may exist in practice, in our setting it would lead to a confound factor: players would likely interact more often when the Firm can select the Auditor (treatment CP) than when matching is random (treatments N and P). Our approach addresses these concerns, while keeping the presence of reputational concerns—especially in CP. The use of random rematching between rounds, accompanied by the display of players' past decisions, is common in the study of reputation in experimental economics (e.g., Bolton et al. 2005, Bartling et al. 2012).

O.3. Multiple Price List (Task 3)

Table O.1 presents the Multiple Price List table used to measure risk preferences in Task 3.

Table O.1 Task 3: Multiple Price List

Decision	Option A	Option B
1	20 tokens with a probability of 1/10,	39 tokens with a probability of 1/10,
2	20 tokens with a probability of 2/10,	39 tokens with a probability of 2/10,
3	20 tokens with a probability of 3/10,	39 tokens with a probability of 3/10,
4	20 tokens with a probability of 4/10,	39 tokens with a probability of 4/10,
5	20 tokens with a probability of 5/10,	39 tokens with a probability of 5/10,
6	20 tokens with a probability of 6/10,	39 tokens with a probability of 6/10,
7	20 tokens with a probability of 7/10,	39 tokens with a probability of 7/10,
8	20 tokens with a probability of 8/10,	39 tokens with a probability of 8/10,
9	20 tokens with a probability of 9/10,	39 tokens with a probability of 9/10,
10	20 tokens with a probability of 10/10,	39 tokens with a probability of 10/10,
In all cases:	16 tokens otherwise	1 token otherwise

O.4. Additional Summary Statistics

Tables O.2 and O.3 present summary statistics from the control tasks and demographic information for participants in the role of Auditor in studies 1 and 2. *T2ForOther* and *T2ForNPO* correspond to the number of tokens given out to the other player and the NPO, respectively, out of the endowed 30 tokens. Similarly, *T3RowSwitch* indicates the first row at which participants changed their lottery decision in Task 3 from A to B, with a greater value indicating a higher risk aversion. For our binary variable *RiskAverse*, participants are considered as risk averse if they switch later than row 5. Across these control variables, we find no statistical differences between treatments in Study 1 ($p > 0.1$), and differences in demographics are similarly not significant (Chi-squared test $p > 0.15$). In Study 2, there are significant differences between treatments for T2ForOther ($p < 0.05$) and Female ($p < 0.01$), with significantly fewer female participants and less generous towards other participants in CP than in P and N. However, we control for gender, tokens for others, and the other listed control variables in our regression analyses.

Table O.2 Auditor's Summary Statistics (Study 1)

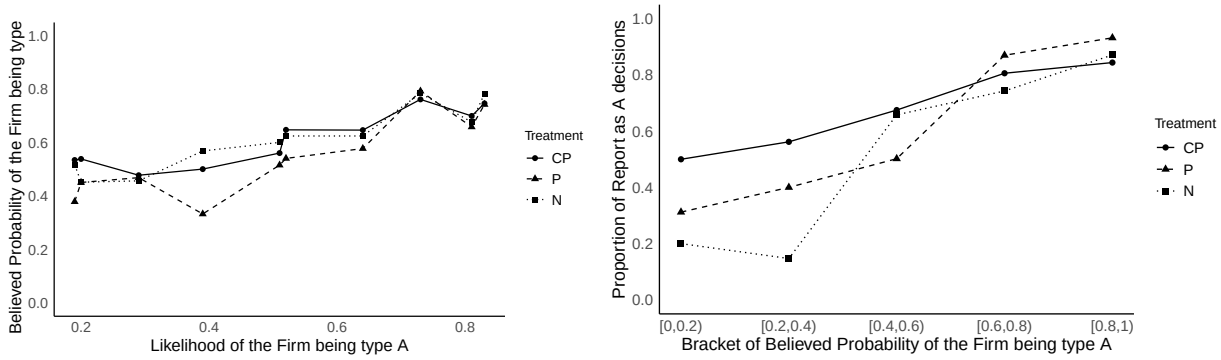
Statistic	CP	P	N
ReportA	0.74 (0.44)	0.60 (0.49)	0.64 (0.48)
BelievedProbA	0.63 (0.21)	0.56 (0.24)	0.63 (0.23)
LikelihoodA	0.54 (0.23)	0.54 (0.23)	0.54 (0.23)
T2ForOther	8.25 (4.49)	6.44 (4.18)	7.82 (5.40)
T2ForNPO	7.97 (4.83)	8.38 (6.13)	7.50 (6.50)
T3RowSwitch	6.56 (1.81)	6.41 (1.84)	6.71 (2.01)
Female	0.69 (0.47)	0.47 (0.51)	0.65 (0.49)
Age (in years)	20.19 (1.15)	19.91 (1.46)	20.44 (3.64)

Table O.3 Auditor's Summary Statistics (Study 2)

Statistic	CP	P	N
ReportA	0.69 (0.46)	0.64 (0.48)	0.62 (0.49)
BelievedProbA	0.64 (0.23)	0.64 (0.21)	0.57 (0.26)
LikelihoodA	0.54 (0.23)	0.54 (0.23)	0.54 (0.23)
T2ForOther	6.40 (4.64)	7.40 (4.47)	7.57 (4.32)
T2ForNPO	7.00 (5.23)	8.00 (4.47)	7.67 (5.70)
T3RowSwitch	0.53 (0.50)	0.53 (0.50)	0.60 (0.49)
Female	0.53 (0.50)	0.77 (0.42)	0.63 (0.48)
Age (in years)	20.20 (2.28)	19.80 (4.63)	20.27 (1.81)

Note: Mean (standard deviation) for each variable, 6 round-level observations per participant.

Figure O.1 The role of believed probability that the Firm is of Type A



(a) Believed probability by *LikelihoodA* (b) Report decision by range of believed probability

O.5. Additional results for Study 1

Figure O.1 visually reflects the relationship between *BelievedProbA* and both *LikelihoodA* and *ReportA* in Study 1. First, O.1a summarizes the average value of *BelievedProbA* as a function of *LikelihoodA* and treatment. We observe that beliefs are, as expected, affected by the plates' information. However, with the exception of the case where *LikelihoodA* is approximately equal to 0.4, we do not find clear differences in belief by treatment. Then, Figure O.1b shows the average report decisions (i.e., the average of *ReportA*) for different brackets of *BelievedProbA*, using 0.20 intervals between 0 and 1. We observe different behavior for the first two brackets, where the chances of reporting the Firm as Type A directionally increase from N to P to CP, versus the three higher brackets, where those differences do not exist. That is, report decisions of Type A (i.e., decisions that are beneficial to the Firm) are more frequent in CP than in the other two treatments only when the Auditor believes that the opposite type (B) is more likely to be true. Conversely, this difference disappears for higher believed probabilities; in particular, when the Auditor believes the Firm to most likely be of Type A. Overall, the results in Figure O.1 suggest that what drives the difference between CP and the other treatments is not an inflated belief about the Firm being of a "good" type, but rather, what Auditors choose to report when they believe the Firm to most likely be "bad".

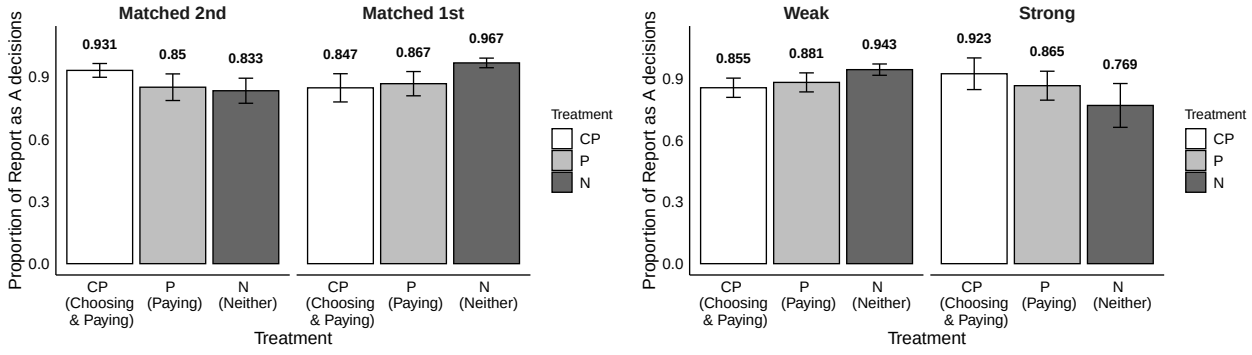
Table O.4 corresponds to the complete version—including variables for control tasks, age, gender and round—of Table 4 in the main body of the paper. See Appendix O.4 for definitions of *T2ForOther*, *T2ForNPO*, and *RiskAverse*. Figure O.2 summarizes Auditors' report decisions, separately for the two reciprocity variables studied in §5.2.1, when the Bayesian posterior of a Firm being of Type A is 60% or greater. Regardless of matched order or feelings of obligation towards the Firm, the Auditor is *not* significantly more likely to report the Firm as A in CP than in the other two treatments ($p > 0.1$) in these high-likelihood cases.

Table O.4 Regression results for Study 1 (with control variables)

	ReportA			BelievedProbA
	(1)	(2)	(3)	(4)
Intercept	−2.658**	−2.125*	−0.291	0.153
	(1.075)	(1.131)	(1.157)	(0.120)
TreatmentP	−0.860***	−1.755***	−1.511**	−0.154***
	(0.279)	(0.651)	(0.689)	(0.053)
TreatmentN	−0.673**	−1.649**	−1.683**	−0.038
	(0.265)	(0.643)	(0.731)	(0.052)
LikelihoodA	6.193***	4.731***		0.387***
	(0.537)	(0.888)		(0.058)
BelievedProbA			2.782***	
			(0.841)	
LikelihoodA×TreatmentP		2.056		0.173**
		(1.283)		(0.080)
LikelihoodA×TreatmentN		2.251*		0.063
		(1.311)		(0.080)
BelievedProbA×TreatmentP			2.037*	
			(1.157)	
BelievedProbA×TreatmentN			2.041*	
			(1.180)	
Round	0.081	0.079	0.022	0.011**
	(0.063)	(0.064)	(0.056)	(0.004)
T2ForOther	−0.019	−0.018	−0.009	−0.003
	(0.022)	(0.023)	(0.021)	(0.003)
T2ForNPO	−0.026	−0.028	−0.010	−0.002
	(0.019)	(0.019)	(0.018)	(0.002)
RiskAverse	0.448*	0.462**	0.347	0.016
	(0.229)	(0.231)	(0.211)	(0.026)
Female	0.060	0.062	0.060	0.012
	(0.229)	(0.231)	(0.212)	(0.026)
Age	0.026	0.030	−0.028	0.012**
	(0.045)	(0.047)	(0.047)	(0.005)

Note: Obs.: 600. Standard error in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Figure O.2 Report decisions by treatment, matched Firm, and self-reported reciprocity (LikelihoodA $\geq$ 60%)



(a) Report decision by matched firm

(b) Report decision by self-reported measure

Note: Error bars represent the standard error of the mean.

O.6. Firm Behavior

Even though the main focus of the present study is to better understand Auditors' report decisions and beliefs, in this section we present additional results regarding the behavior of Firms in the Social Audit

Game. Table O.5 presents summary statistics for these participants. It is worth noting that the *ReportA* variable has a different meaning for Firms (compared to Auditors), as they are asked to state whether they *believe* that their matched Auditor would report them Type A. Interestingly, we do not find any statistically significant differences between treatments for report decisions or beliefs ($p > 0.1$). In other words, Firms do not anticipate that Auditors will be more lenient towards them in CP than in the other two treatments.

Table O.5 Firm's Summary Statistics (Social Audit Game)

Statistic	CP	P	N
ReportA	0.65 (0.48)	0.64 (0.48)	0.64 (0.48)
BelievedProbA	0.59 (0.23)	0.59 (0.24)	0.60 (0.23)
LikelihoodA	0.54 (0.23)	0.54 (0.23)	0.54 (0.23)

Note: Mean (standard deviation) for each variable, 6 round-level observations per participant.

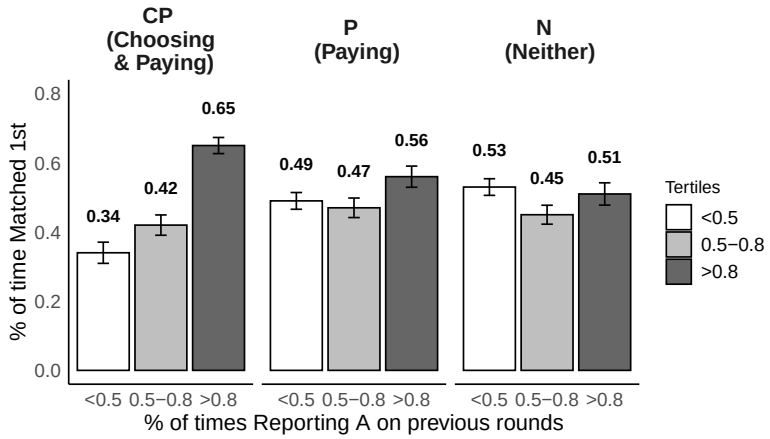
Regarding Firms' preferences in the only treatment where they could choose an Auditor (CP), most Firms (85%) agreed or strongly agreed in the post-experiment survey that they preferred Auditors with a reputation of reporting Type A more often. This is confirmed by their behavior in the Social Audit Game, which we analyze via the following analysis. First, for every round after round 1, we computed the proportion of times that each Auditor reported A in all previous rounds. As a reminder, this corresponds to the historical information that Firms observed before choosing an Auditor. Then, we divided this variable into three groups, each corresponding to one third of the distribution (i.e., tertiles): reporting A less than 50%, between 50 and 80%, or more than 80% of the time. Finally, for each of these three groups, we computed the proportion of times that Auditors were matched with the 1st Firm in the focal round. As Figure O.3 shows, Auditors in N and P were matched with the 1st Firm between 45% and 56% of the time regardless of previous decisions, which corresponds to the implemented random matching. In each of these two treatments, the differences between the three tertiles are not statistically significant. In contrast, Auditors in CP who report A infrequently (first and second tertiles, i.e., less than 80 % of the time) are penalized by being matched with the 1st Firm less often, only $34 - 42\%$ of the time. Meanwhile, Auditors who reported A most often (more than 80 % of the time) were chosen by the 1st Firm 65% of the time. This difference between the top tertile and the other two for Auditors in CP is statistically significant ($p < 0.05$). This corroborates that when they are able to, Firms indeed favored Auditors with a reputation of reporting A in past rounds.

O.7. Regression analyses for Reciprocity

Table O.6 shows the results from regression analyses for the effect of reciprocity in both studies. We focus on the cases where we find differences between treatments ($LikelihoodA < 0.60$). The first (last) two columns show the results for Study 1 (Study 2). In the first and third columns, we study differences based on being matched with the 2nd Firm vs. being matched with the 1st Firm (the omitted baseline level for the dummy variable). In the second and fourth columns, we analyze differences based on having self-reported weak feelings of obligation towards the Firm (defined as not agreeing with the relevant Likert-scale item) vs. having strong feelings of obligation (the omitted baseline level).

In Study 1, we find that conditional on being matched with the 1st Firm or having a strong feeling of obligation towards the Firm, Auditors report the Firm as A significantly less often in P and N than in

Figure O.3 Auditor's likelihood of being matched with the 1st Firm as a function of past reports



Note: Error bars represent the standard error of the mean.

Table O.6 Regression results: Report decisions by reciprocity measures (LikelihoodA below 60%)

	Study 1		Study 2	
	(1)	(2)	(3)	(4)
Intercept	−0.026	0.484	1.768**	2.387**
	(1.215)	(1.235)	(0.887)	(0.985)
TreatmentP	−1.681***	−1.376**	−0.712*	−1.033*
	(0.428)	(0.582)	(0.432)	(0.619)
TreatmentN	−1.285***	−1.160**	−0.403	−0.829
	(0.407)	(0.524)	(0.418)	(0.557)
Matched2nd	−0.692*		−0.314	
	(0.407)		(0.416)	
WeakFeltObligation		−1.316***		−0.971**
		(0.443)		(0.486)
TreatmentP×Matched2nd	1.405**		0.530	
	(0.568)		(0.595)	
TreatmentN×Matched2nd	1.035*		0.084	
	(0.559)		(0.598)	
TreatmentP×WeakFeltObligation		0.844		0.847
		(0.662)		(0.736)
TreatmentN×WeakFeltObligation		0.802		0.640
		(0.639)		(0.709)
Round	−0.058	−0.054	−0.301***	−0.315***
	(0.071)	(0.072)	(0.079)	(0.080)
T2ForOther	−0.020	−0.012	−0.014	−0.012
	(0.024)	(0.024)	(0.030)	(0.031)
T2ForNPO	−0.019	−0.024	−0.029	−0.026
	(0.021)	(0.022)	(0.026)	(0.026)
RiskAverse	0.319	0.249	−0.117	−0.183
	(0.253)	(0.260)	(0.251)	(0.274)
Female	0.011	−0.119	0.099	0.101
	(0.252)	(0.261)	(0.267)	(0.280)
Age	0.056	0.056	−0.019	−0.022
	(0.052)	(0.053)	(0.035)	(0.038)

Note: Obs.: 336 in Study 1, 300 in Study 2. Standard error in parentheses. *p<0.1; **p<0.05; ***p<0.01.

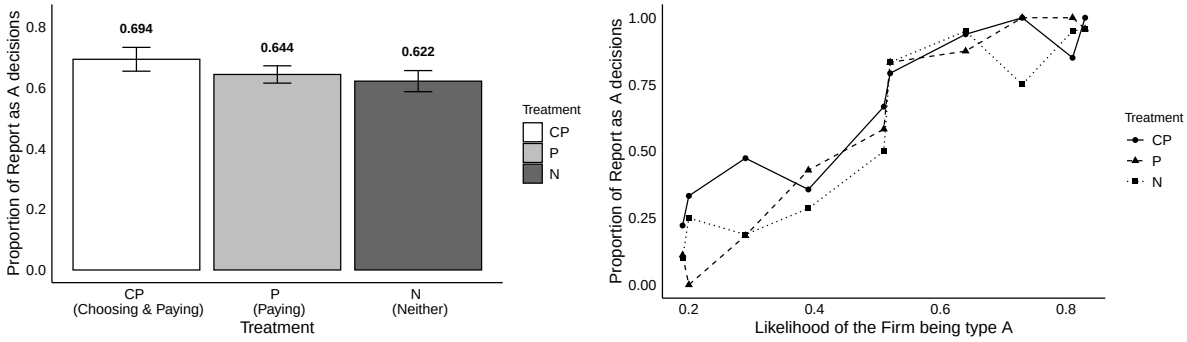
CP, as evidenced by the significant treatment dummy variables. Conversely, these differences are no longer significant when we condition on being matched with the 2nd Firm or on participants having a weak feeling of obligation towards the Firm (in these cases, we compute the treatment effect as the sum of the main

effect and the interaction between the reciprocity variable and the relevant treatment dummy). In Study 2, treatment differences are directionally similar but weaker in magnitude and significance. Auditors report the Firm as A significantly less often in P than in CP based on both reciprocity variables. Similarly, Auditors in N report the Firm as A less often than in CP, but these differences are only close to significant in the case of reported feelings of obligation ($p = 0.34$ in column (3) and $p = 0.13$ in column (4)). As in Study 1, when we condition on Auditors being matched with the 2nd Firm or having weak feelings of obligation towards the Firm, differences are far from significant ($p > 0.45$).

O.8. Study 2 Results

Figures O.4a and O.4b summarize Auditors' report decisions in Study 2. Treatment differences are similar to those in Study 1, but smaller in magnitude. The proportion of Report A decisions continues to be higher in CP (69.4%) than in P (64.4%) or N (62.2%), but these differences are not statistically significant ($p = 0.33$ and $p = 0.18$, respectively); and the difference between P and N continues to be highly non-significant, $p = 0.61$). However, as we discuss below, we recover significant differences when we control for other variations across treatments, such as gender and control-task decisions. Moreover, these aggregate analyses mask some differences that exist based on the value of *LikelihoodA*. As shown in Figure O.4b, we find an overall greater tendency to report the Firm as A in CP for low values of the posterior probability. However, and consistent with treatment effects being less pronounced in Study 2 than in Study 1, these differences are now statistically significant only for the three lowest values of *LikelihoodA* that we consider, i.e., when the plates' information suggests that there is less than a 1/3 chance of the Firm being Type A. Averaging over these cases, we find that Auditors are significantly more likely to report A in CP than in P or N ($p < 0.05$). Conversely, and consistent with Hypothesis 3, these differences are not significant for posteriors greater than 1/3 ($p > 0.40$). In both cases, differences between P and N are not statistically significant ($p > 0.40$).

Figure O.4 Study 2 - Summary of Report Decisions by Treatment and LikelihoodA



(a) Report decision by treatment

(b) Report Decision by *LikelihoodA* and treatment

To formally test our hypotheses, we run the same regression analyses as in Study 1, which we summarize in Table O.7. Column (1) shows that the coefficient on *LikelihoodA* is positive and highly significant ($8.539, p < 0.01$), indicating that the more the information suggests the Firm is of Type A, the more likely the Auditor

is to report the audited party as A. That is, we find support for Hypothesis 1. Relatedly, we also find clear deviations from the rational benchmark: as can be seen in Figure O.4b, when *LikelihoodA* is less than 0.5, Auditors report the Firm as Type A 42% of the time in CP, 20% in P, and 21% in N (significantly different from 100 in CP and from 0 in P and N; χ^2 test, $p < 0.01$). When *LikelihoodA* is greater than 0.5, Auditors report the Firm as Type A in between 84 and 89% of the cases depending on the treatment (significantly different from 100; χ^2 test, $p < 0.01$).

After controlling for the posterior, *LikelihoodA*, we continue to find that Auditors report the Firm as A significantly more often in CP (the baseline/omitted level) than in treatments P and N. However, the differences are more nuanced than in Study 1. In Table O.7, there is a main negative effect of N on *ReportA* (column 1). However, when its interaction with *LikelihoodA* is incorporated (column 2), neither the negative main effect of N nor the positive interaction term are statistically significant ($p = 0.21$ and $p = 0.52$, respectively), though both remain directionally the same as in Study 1. Conversely, the effect of the dummy variable associated with treatment P is statistically significant only when its interaction with *LikelihoodA* is also considered, in which case both values are significant (column 2). In other words, we find support for the frequency of *ReportA* being lower in P and N than in CP, and for this difference decreasing as *LikelihoodA* increases—though the latter effect is statistically significant only in P. Another interesting difference is that in Study 2, we observe a negative and significant effect of round on *ReportA* (first two columns in the regression table). This stands in contrast to Study 1, where round has a positive and significant effect on beliefs about the Firm being type A—which in turn positively affect *ReportA*. In other words, when there is a very strong incentive to be lenient (Study 1), participants become directionally more lenient over time; when these incentives decrease (Study 2), participants appear to learn to be less lenient as time progresses.

Next, we explore the potential roles of motivated reasoning and reciprocity in explaining treatment differences. Similar to Study 1, we do not find significant differences in beliefs between treatments, as shown in column 4 of Table O.7. Thus, motivated reasoning is unlikely to explain our results. For the study of reciprocity, we repeat the two analyses conducted for Study 1; for consistency, we focus on the cases where *LikelihoodA* is less than 0.6. First, we compare average *ReportA* by treatment, separately when the Auditor is matched with the 2nd or the 1st Firm, as shown in Figure O.5a. We continue to find that when Auditors are matched with the 1st Firm, they report A significantly more more often in CP than in P ($p = 0.05$). Though qualitatively similar, results are non-significant when we compare CP against N ($p = 0.26$). Finally, differences continue to be not significant between treatments when Auditors are matched with the 2nd Firm. Second, in Figure O.5b we explore differences between treatments by self-reported reciprocal concerns. Auditors who felt a strong obligation towards the Firm report the Firm as A significantly more in CP than in P ($p = 0.10$) and directionally but not significantly when compared against N ($p = 0.16$). In contrast, when Auditors indicated a weak sense of obligation towards the Firm, there are no significant differences. We also conducted a formal regression analysis using both of these variables for Study 2. The results are summarized in columns (3) and (4) of Table O.6 in Section O.7 and show similar results.

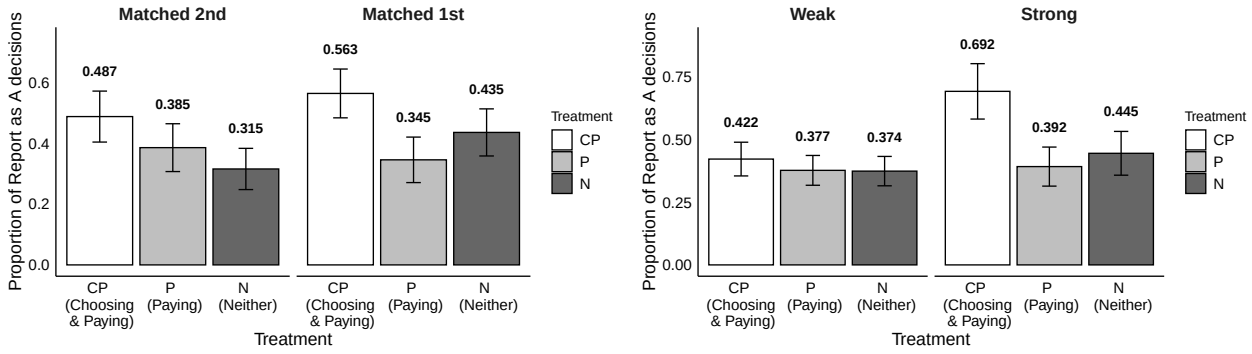
Finally, to better understand any potential differences between our two studies, we conduct a regression analysis with observations from both of them. We add a dummy variable that is equal to 1 for Study

Table O.7 Study 2 - Regression results: effects of Treatment and LikelihoodA on reports and beliefs

	ReportA			BelievedProbA
	(1)	(2)	(3)	(4)
Intercept	−1.657*	−0.728	−1.073	0.340***
	(0.891)	(1.041)	(0.952)	(0.093)
TreatmentP	−0.401	−2.402***	−1.073	0.026
	(0.323)	(0.923)	(0.869)	(0.058)
TreatmentN	−0.562*	−0.957	0.367	−0.078
	(0.319)	(0.769)	(0.717)	(0.057)
LikelihoodA	8.539***	7.087***		0.475***
	(0.773)	(1.097)		(0.064)
BelievedProbA			4.847***	
			(0.935)	
LikelihoodA×TreatmentP		4.474**		−0.054
		(1.914)		(0.091)
LikelihoodA×TreatmentN		0.991		0.017
		(1.535)		(0.091)
BelievedProbA×TreatmentP			1.378	
			(1.369)	
BelievedProbA×TreatmentN			−0.716	
			(1.171)	
Round	−0.155**	−0.162**	−0.083	−0.00001
	(0.074)	(0.075)	(0.061)	(0.005)
T2ForOther	0.003	−0.003	−0.017	0.002
	(0.031)	(0.032)	(0.028)	(0.003)
T2ForNPO	−0.024	−0.024	−0.011	0.0001
	(0.027)	(0.027)	(0.024)	(0.003)
RiskAverse	0.010	0.007	0.167	−0.027
	(0.264)	(0.273)	(0.237)	(0.025)
Female	−0.111	−0.103	−0.278	0.016
	(0.280)	(0.284)	(0.248)	(0.026)
Age	−0.037	−0.050	−0.023	0.002
	(0.037)	(0.041)	(0.034)	(0.004)

Note: Obs.: 540. Standard error in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Figure O.5 Study 2 - Report decisions by matched Firm and by self-reported reciprocity (LikelihoodA < 60%)



(a) Report decision by matched firm

(b) Report decision by self-reported measure

2 and we include its interactions with treatment variables and the posterior probability, *LikelihoodA* (3-way interactions are omitted for simplicity, but we confirm that they are not statistically significant). The results are shown in Table O.8. First, we do not find any significant coefficients for the interactions between treatment and Study 2 dummies. This suggests that, though the magnitude of the differences between CP and P, and between CP and N, are smaller in Study 2, this difference is not statistically significant. Instead,

the main difference between the studies is that there is a lower tendency to report the Firm as A in Study 2 (significant negative effect of the dummy *Study2*). In addition, we also observe a positive and significant interaction between *Study2* and *LikelihoodA*, so differences between studies disappear, as expected, as the posterior probability indicates a high chance of type A. In other words, the main impact of Study 2 is that it significantly increases how often Auditors report the Firm as B when the evidence suggests so, i.e., for low values of *LikelihoodA*.

Table O.8 Regression results: Report decisions by treatment and study

Intercept	−0.735	TreatmentN×Study2	0.110
	(0.754)		(0.390)
LikelihoodA	4.682***	TreatmentP×Study2	0.621
	(0.758)		(0.408)
TreatmentN	−1.375***	Round	−0.023
	(0.513)		(0.048)
TreatmentP	−2.155***	T2ForOther	−0.015
	(0.555)		(0.017)
LikelihoodA×TreatmentN	1.657*	T2ForNPO	−0.020
	(0.985)		(0.015)
LikelihoodA×TreatmentP	2.699***	RiskAverse	0.234
	(1.037)		(0.167)
Study2	−1.266***	Female	−0.059
	(0.459)		(0.170)
LikelihoodA×Study2	2.250***	Age	−0.012
	(0.861)		(0.028)

Note: Dep. variable: ReportA; Obs.: 1,140. Standard errors in parentheses. *p< 0.1; **p< 0.05; ***p< 0.01.

O.9. Theoretical Models

In this section, we conduct a formal analysis of Auditors' optimal decision-making. We first assume hyper-rational agents, and then we complement it with a framework where we introduce behavioral regularities.

For simplicity, we consider the problem faced by the Auditor in round $r < 6$, where they are affected by the potential impact of their round-r decision on (i) the expected utility in the focal round r and (ii) the chance of being matched with the 1st Firm in round $r+1$—which in turn affects their profit in that round via the cost of conducting the audit, c. (For simplicity, we hereafter refer to the two matching options as being matched 1st or matched 2nd.) We believe that this approach constitutes a reasonable simplification because in our experiment only one round is randomly selected for payoff. Therefore, even though an Auditor's decision may have impacts on all subsequent rounds (particularly when it comes to the chances of being matched 1st or 2nd, as discussed below), only one of these rounds will have an impact on their earnings.

O.9.1. Rational Benchmark

First, we define the factors that affect the Auditor's profit function (π_A). The main decision variable in each round, δ, is whether to report the Firm as Type A ($\delta = 1$) or Type B ($\delta = 0$). The Auditor starts with an initial endowment (i_A), receives a fixed audit fee (f), and incurs a cost for conducting the audit. This cost is smaller if matched 1st than if matched 2nd ($c_1 < c_2$), with only one of the two outcomes being possible each round. We define c_r as the realized cost in the focal round r, which is sunk; and we define the probability of being matched 1st *in round* $r+1$ as α, with $\alpha \in [0,1]$ possibly depending on their report decision. Finally,

the Auditor has to consider their expected cost of being penalized, which is given by the combination of: (i) the penalty cost p; (ii) the chance of being penalized if misreporting, q; and (iii) the probability that they misreport (i.e., the report not matching the true Firm type). To capture the latter, we consider the term $(\delta(1-\mu)+(1-\delta)\mu)$, where $\mu \in [0,1]$ represents the Bayesian posterior probability of the Firm being Type A—which we refer to in the paper as *LikelihoodA*. As a result, the Auditor's profit is:

$$\pi_A = i_A + f - c_r - qp(\delta(1-\mu)+(1-\delta)\mu) + \alpha c_1 - (1-\alpha)c_2 \tag{O.1}$$

We are particularly interested in the *difference in profit* between reporting A versus B. We begin with the scenario where Auditors are randomly matched with a Firm, i.e., the Firm cannot choose its auditor. This corresponds to the situation in **Treatments P and N**. In this scenario, whether the Auditor incurs the cost c_1 or c_2 in round $r+1$ is random, and each situation has a 50% probability. Therefore, the Auditor's profit with respect to their report decision is:

$$\pi_A(\delta=1) - \pi_A(\delta=0) = qp(2\mu-1)$$
$$\pi_A(\delta=1) - \pi_A(\delta=0) > 0 \quad \Leftrightarrow \quad \mu > 1/2$$

Condition 1 *When Auditors are randomly matched to the Firm (Treatments N and P), their predicted hyper-rational behavior is to report the Firm as A (B) when LikelihoodA is above (below) 0.5.*

That is, from a hyper-rational perspective, Auditors find it optimal to report the type (A or B) that is most likely to be the Firm's true type based on the information observed in the audit. This approach minimizes the risk of receiving a penalty for an incorrect report.

We now turn to the Chosen and Paid scenario (Treatment CP), where the Firm can choose which Auditor to work with. First, we define the Firm's profit function (π_F) in a given round, which depends on the Auditor's report decision (δ). The Firm starts with an initial endowment (i_F), and its profit depends on whether it is reported as Type A ($\delta=1$) or not ($\delta=0$):

$$\pi_F = i_F + \delta \times a + (1-\delta) \times b \tag{O.2}$$

If reported as A, the Firm receives a reward ($a>0$); if not, they incur a penalty ($b<0$). Thus, under a hyper-rational assumption on the part of the Firm, too, they strictly prefer to be reported as A.

Based on the above, the Auditor knows that given the opportunity, the Firm will select the Auditor that they believe is most likely to report them as A. Even though they cannot know this future decision for sure, Firms can form beliefs about this likelihood from Auditors' previous report decisions. Specifically, if the two Auditors in the 1st Firm's group have different historical report decisions, then we assume that the Firm will strictly prefer the one who has reported A more often. Otherwise, if the two Auditors have reported A equally often, then we assume each of them has a 50% chance of being chosen by the 1st Firm (i.e., matched 1st). Given this, in equilibrium the Auditor's optimal decision is to *always report A*. To show this, if an Auditor follows (deviates from) this strategy, then their chance of being matched 1st in the next round is equal to 0.5 (0). As a result, their profits when reporting A compared to when reporting B are equal to:

$$\pi_A(\delta=1) = i_A + f - c_r - qp(1-\mu) - 0.5c_1 - 0.5c_2$$

$$\pi_A(\delta=0) = i_A + f - c_r - qp\mu - c_2$$
$$\pi_A(\delta=1) - \pi_A(\delta=0) = 0.5(c_2 - c_1) + qp(2\mu - 1)$$
$$\pi_A(\delta=1) - \pi_A(\delta=0) > 0 \quad \Leftrightarrow \quad \mu > \frac{1}{2} - \frac{c_2 - c_1}{4qp}$$

In other words, an Auditor could be better off deviating from the proposed equilibrium if $\mu \leq \frac{1}{2} - \frac{c_2-c_1}{4qp}$. However, given our experimental parameters, $c_2 - c_1 = 50$ in Study 1 (25 in Study 2), $q = 0.05$, and $p = 25$, the right-hand side of the condition is negative in both studies, and as a result, a hyper-rational Auditor always prefers to report A in treatment CP.

Condition 2 *When Auditors are chosen and paid by the Firm (Treatment CP), their predicted hyper-rational behavior is to always report A.*

O.9.2. Incorporating Behavioral Regularities

Next, we consider a situation where Auditors exhibit some behavioral regularities that have been extensively documented in the literature. Specifically, we consider an Auditor who, in addition to their own profits: (i) considers the potential impact of their decision on the NPO's payoff (e.g., due to altruism); (ii) incurs a psychological penalty (benefit) for misreporting (accurately reporting); and (iii) exhibits reciprocity towards the Firm for their role (possibly passive) behind the Auditor's payoff. Note that the latter behavior may also be accompanied by pure altruism towards the Firm, but for simplicity we do not differentiate the two (in practice, they would both push the Auditor's report decision in the same direction, i.e., towards reporting A more often as shown below). At the end of the section, we also briefly discuss the impact of bounded rationality on our predictions, particularly as it relates to the use of the Bayesian posterior in Auditor's optimal decisions. Throughout this section, we limit our study to the Auditor's decisions under the assumption that all other players continue to behave as hyper-rational agents; the study of equilibria where agents form beliefs about the degree of behavioral preferences on the part of other players is outside the scope of the present work. It is also worth noting that the behavioral economics literature has discussed several ways in which our behaviors of interest could be modeled. The identification of *which* specific modeling approach best captures our participants' behavior is similarly beyond our scope.

To begin, we define the NPO's payoff function in a given round in Equation O.3. This is important because altruism suggests that Auditors may take this quantity into consideration when making their decisions. The NPO starts with an initial endowment (i_N) and its profit depends on whether the Auditor reports the Firm as Type A ($\delta = 1$) or Type B ($\delta = 0$):

$$\pi_N = i_N + \delta\left(t\mu + u(1-\mu)\right) + (1-\delta)\left(t(1-\mu) + u\mu\right) \tag{O.3}$$

If the Firm is reported accurately, then the NPO receives a reward ($t > 0$); if not, they incur a penalty ($u < 0$). Thus, it strictly benefits from the report being accurate; that is, for the Firm to be reported as A ($\delta = 1$) if its true type is A (μ) and as B ($\delta = 0$) if its true type is B ($1 - \mu$). Formally:

$$\pi_N(\delta=1) - \pi_N(\delta=0) = (t-u)(2\mu - 1)$$
$$\pi_N(\delta=1) - \pi_A(\delta=0) > 0 \quad \Leftrightarrow \quad \mu > 1/2$$

An altruistic Auditor will incorporate the NPO payoffs into their decision-making. In particular, we assume that the Auditor derives utility not only from their own profit, but also from the NPO's outcome, scaled by a parameter λ, e.g., $U_A = \pi_A + \lambda\pi_N$. For example, (i) if $\lambda = 0$, the Auditor is purely self-interested and does not consider the NPO's welfare at all. (ii) If $\lambda = 0.5$, two dollars for the NPO are valued the same as one dollar for themselves. (iii) If $\lambda = 1$, the Auditor is indifferent between one dollar for themselves and for the NPO. It is worth noting that the value of λ, and each behavioral parameter in this section, is specific to each individual Auditor i, i.e., it should be noted as λ_i. However, we omit the subscripts for each of notation.

In addition, we incorporate a psychological penalty (benefit) for misreporting (accurately reporting). If misreporting, the Auditor incurs a disutility ($d < 0$); otherwise, they derive an additional utility ($e > 0$). Note that with this formulation, and given our experimental setup, these *accuracy concerns* and altruism towards the NPO will push decisions in the same direction, as the NPO benefits (is hurt by) accurate (inaccurate) report decisions. Similar to the parameter λ, we define γ as the weight that determines how much the Auditor's utility is affected by their concern for accuracy.

Finally, we incorporate into the Auditor's behavioral model their reciprocity towards the Firm for their role in generating their payoffs. Following a simplified version of the model in Falk and Fischbacher (2006), we consider reciprocity to be the product of the following elements: (i) the outcome term Δ, which measures the impact the Firm has on the Auditor's payoff; (ii) the degree of intentionality the Auditor attributes to the Firm's action, $\theta \in [0, 1]$; and (iii) the reciprocation term σ, which captures how much the Auditor's action affects the Firm's payoff. For simplicity, we assume this last term to be equal to δ, i.e., the decision to report the Firm as Type A. That is, we assume that the Auditor can receive a (positive) reciprocity-based measure of utility when reporting the Firm as Type A, since this benefits the Firm's payoff; and that they do not incur any reciprocal utility when they report Type B.[15] Similarly, rather than explicitly computing the term Δ, we simply observe that this is higher when matched with the 1st Firm than when matched with the 2nd Firm, as this matches the payoff ranking for the Auditor; and that its value is nonzero in the latter case, as the Auditor still earns a positive payoff even when matched with the 2nd Firm. Note that our results would remain qualitatively similar if we incorporated reciprocity through a more complex term, as long as such a term was directly proportional to the three quantities of interest.

In other words, reciprocity in the Auditor setting can be understood as a feeling of obligation to favor the Firm in return for a favorable action, and it is expected to be stronger when the Auditor believes the Firm's help was intentional. The intentionality of this action is most clearly present in CP when matched with the 1st Firm, so this situation should lead to a stronger reciprocal feeling than in other cases. However, some intentionality may still be attributed to the Firm's action of paying the audit fee, which occurs in both CP and P. Similarly, a (lower) baseline level of intentionality could even be assumed in treatment N, as the Auditor may view their payoffs as being the consequence of being matched with the player in the role of the Firm. It is worth noting that, though less intuitive, the non-zero value of the intentionality factor when

[15] In our experiment, there is no reason to negatively reciprocate the Firm: even when matched with the 2nd Firm (the least advantageous situation for the Auditor), that Firm plays no active role in the matter. Our theoretical predictions would be qualitatively similar, however, if we included such a negative-reciprocity term.

no alternative action is available to the other player (in our case, the Firm) is explicitly incorporated in the model by Falk and Fischbacher (2006), and it has been documented in the literature (Charness 2004). Finally, a non-zero baseline level of the intentionality parameter can also allows us to incorporate, as discussed above, a general level of concern on the part of the Auditor for the Firm's payoff without having to separately include an altruism component towards the Firm.

To summarize, the Auditor's utility with behavioral regularities can be written as:

$$\begin{aligned} U_A = {} & i_A + f - c_r - qp(\delta(1-\mu) + (1-\delta)\mu) + \alpha c_1 - (1-\alpha)c_2 \\ & + \lambda(i_N + \delta\,(t\mu + u(1-\mu)) + (1-\delta)\,(t(1-\mu) + u\mu)) \\ & + \gamma\,\delta\,(e\mu + d(1-\mu)) + (1-\delta)\,(e(1-\mu) + d\mu) + \rho\,\Delta\,\theta\,\delta, \end{aligned}$$

where ρ is the Auditor's individual weight of reciprocity-based utility, similar to λ in the case of altruism towards the NPO. In words, the rest of the reciprocity term is the product of how much the Firm benefits the Auditor (Δ), how much the Auditor feels that this benefit is intentional (θ), and whether the action taken by the Auditor can indeed be seen as reciprocating/benefiting the Firm (δ).

Building on the above, we proceed to analyze the *difference in utility* between reporting A and reporting B. We start with the situation in **treatments P and N**, with the same considerations as in the hyper-rational benchmark. In this scenario, the Auditor's utility differences are:

$$\begin{aligned} & U_A(\delta=1) - U_A(\delta=0) = (qp + \lambda(t-u) + \gamma(e-d))\,(2\mu-1) + \rho\,\Delta\,\theta \\ & U_A(\delta=1) - U_A(\delta=0) > 0 \quad \Leftrightarrow \quad \mu > \frac{1}{2} - \frac{\rho\,\Delta\,\theta}{2\big(qp + \lambda(t-u) + \gamma(e-d)\big)} \end{aligned}$$

A similar condition can be found in the case of **treatment CP**, where similar to the rational benchmark, an additional term must be added for the differential in future audit costs that are expected if the Auditor reports A instead of B:

$$\begin{aligned} & U_A(\delta=1) - U_A(\delta=0) = 0.5(c_2 - c_1) + (qp + \lambda(t-u) + \gamma(e-d))\,(2\mu-1) + \rho\,\Delta\,\theta \\ & U_A(\delta=1) - U_A(\delta=0) > 0 \quad \Leftrightarrow \quad \mu > \frac{1}{2} - \frac{(c_2 - c_1) + 2\rho\,\Delta\,\theta}{4\big(qp + \lambda(t-u) + \gamma(e-d)\big)} \end{aligned}$$

In all treatments, the introduction of the behavioral regularities has the following effects: (1) Reciprocal concerns (as well as a baseline level of concern for the Firm's payoff) generate a *decrease* in the probability threshold for μ, pushing it closer to 0. That is, reciprocity increases the range of posterior values (*LikelihoodA*) for which the Auditor prefers to report A. (2) Altruism towards the NPO and concerns for accurately reporting the Firm's type attenuate the degree to which reciprocity (in all cases) and consideration for future payoffs (in CP) can push the threshold away from 0.5. In other words, altruism and accuracy concerns help to *increase* the posterior threshold, pushing it closer to 0.5 and hence expanding the values of μ for which the Auditor prefers to report B (or having no effect in P and N if $\rho\,\delta\,\theta = 0$). In all cases, because the terms on the RHS of the threshold condition are nonzero, the effects of behavioral regularities are expected to be stronger (i.e., to be more likely to make a difference in a subject's preference for reporting A vs. reporting B) for low values of μ, and their effects should disappear for high enough values of this posterior probability—particularly once μ is greater than 0.5.

It is also worth noting that, while altruism and accuracy concerns are expected to have similar strength across treatments, this is not the case for reciprocal concerns. Following the earlier discussion, Δ is larger when matched with the 1st Firm than when matched with the 2nd Firm and in Study 1 than in Study 2 (since the cost differential is larger in Study 1); and θ is largest when matched with the 1st Firm in CP than in all other cases, and may be larger in P than in N if intentionality is affected by paying the audit cost.

That is, reciprocity should contribute to an Auditor having a lower threshold (above which they prefer to report the Firm as Type A) in CP than in P or N, and possibly in P than in N.

Finally, though we do not include them explicitly in our models, we note the following. First, bounded rationality can prevent Auditors from accurately estimating posterior probabilities, leading to deviations from the rational and behavioral models; e.g., we can expect some people to report the firm as A even when *LikelihoodA* is above 0.5. This is part of why, in our empirical analyses—particularly when we study some of the motives behind treatment differences—we focus on cases where the posterior probability is less than 0.6, as this includes a few cases where *LikelihoodA* is equal to 0.51 and 0.52 (see Table 1). Second, motivated reasoning has the potential to exacerbate differences between the three treatments, as beliefs about the likelihood of Type A could be inflated in CP compared to P and in P compared to N. To account for these belief-related effects, in our empirical analyses we study: (i) report decisions based on our measure of Auditors' beliefs, rather than posterior, as well as their decisions based on the proportion of damaged plates as a robustness check; and (ii) the potential effect of treatment on beliefs.

O.10. Regression Results with Behavioral Regularities

We run logistic regressions for Study 1 incorporating variables that capture the main behavioral regularities discussed in the theoretical model. The results are presented in Table O.9. We use *ReportA* as the dependent variable with random effects at the individual level. We control for the amount of tokens the participant gives to the NPO in Task 2, as a proxy for altruistic concerns, as well as the self-reported question about obligation towards the Firm, which we binarize as we do in our reciprocity analyses (with a high sense of reciprocity as the baseline level for the dummy variable; see §O.7). In addition, we include the answer to another post-experiment survey question in which Auditors indicated, on a Likert scale from 1 to 5, how much they agreed with the statement: "I felt an obligation to report the Firm's type as accurately as possible." Based on this question, we classify Auditors as having strong (or weak) accuracy concerns depending on whether they agreed (or did not agree) with the statement. The baseline is a weak concern for accuracy, so the dummy variable *AccuracyConcern* is equal to 1 when participants agree with the statement. We also include controls for round, control tasks, and demographic information. Based on our theoretical models, we include interactions between *LikelihoodA* and the three behavioral-regularity variables, since their effects may be more pronounced at low posterior values; and we also include interactions between treatment dummies and the reciprocity variable, as our theory posits that the effect of reciprocity (feeling of obligation towards the Firm) should be largest in CP—the baseline level of the treatment dummies.

For low values of *LikelihoodA*, Auditors with strong accuracy concerns are significantly less likely to report the Firm as A; i.e., they adhere more closely to the information suggested by the plates observed in the

Table O.9 Regression Results: Report decisions

Intercept	0.199	Round	0.088
	(1.320)		(0.066)
TreatmentP	−1.780**	Female	−0.047
	(0.822)		(0.245)
TreatmentN	−1.997**	Age	0.028
	(0.800)		(0.049)
LikelihoodA	1.741	RiskAverse	0.380
	(1.521)		(0.241)
LikelihoodA×TreatmentP	0.880	T2ForOther	−0.012
	(1.369)		(0.023)
LikelihoodA×TreatmentN	1.064	WeakFeltObligation	−2.709***
	(1.429)		(0.656)
AccuracyConcern	−1.015*	LikelihoodA×WeakFeltObligation	2.789**
	(0.599)		(1.178)
LikelihoodA×AccuracyConcern	2.237*	TreatmentP×WeakFeltObligation	1.061*
	(1.182)		(0.610)
T2ForNPO	−0.098*	TreatmentN×WeakFeltObligation	1.474**
	(0.052)		(0.606)
LikelihoodA×T2ForNPO	0.144		
	(0.099)		

Note: Dep. variable: ReportA; Obs.: 600. Standard errors in parentheses. *p< 0.1; **p< 0.05; ***p< 0.01.

audit. Similarly, as the amount of tokens transferred to the NPO increases, Auditors report the Firm as A significantly less often: at low posterior values, this is consistent with altruism towards the NPO, since a report of B when the reality is Type A would hurt the NPO's payoff. In addition, Auditors with a weak (as opposed to strong) sense of obligation towards the Firm are significantly less likely to report the Firm as A. However, this effect is much stronger in the baseline (treatment CP) than in treatments P or N, as evidenced by the positive interactions between these treatment dummies and *WeakFeltObligation*—a finding that is consistent with reciprocity. The differences between P and N when it comes to this variable are not statistically significant. Finally, we observe that the magnitude of the three behavioral effects of interest diminish as *LikelihoodA* increases, with the effect being significant for reciprocity and accuracy concerns. In other words, we find support for Auditors' report decisions being influenced by the three behavioral variables discussed in our theoretical model. This result strongly suggests that the deviations from the rational benchmark are not due simply to noise or participants' bounded rationality.