

How resolving enforcement ambiguity reduces tax compliance

Eric Floyd, Michael A Hallsworth, John A. List, Robert D. Metcalfe, Kristian Rotaru, Ivo Vlaev¹

Abstract

Tax authorities often leave the consequences of non-compliance unspecified. Enforcement letters warn of further action without stating what it is, and penalty notices give maxima set case by case. We examine whether specifying those consequences raises or lowers compliance, and why, in four experiments spanning both behavioral margins of the tax gap, from the field to physiological measurement. In a field experiment with HM Revenue and Customs, letters randomized to more than 200,000 taxpayers with outstanding income-tax debts increased payment when they left the consequences unspecified, but not when they specified a financial or an administrative consequence. In an online payment experiment with 2,050 taxpayers across fourteen conditions, of whom 1,194 formed the factorial design carrying the principal test, resolving the maximum fine lowered the stated likelihood of paying, whether the revealed maximum was modest or severe, which statistical updating about the penalty does not accommodate and ambiguity aversion does. An incentivized experiment with 827 participants reproduced the effect on an income-reporting decision with money at stake, extending it to the reporting margin. A laboratory experiment with 239 participants, using eye tracking and pupillometry, found that enforcement messages differed in the autonomic arousal they produced and not in the visual attention they received, and that pupillary response accounts for part of the association with compliance. Ambiguity about enforcement consequences, over and above penalty magnitude, is therefore a determinant of compliance, operating through affective response rather than belief revision. The findings have direct implications for how tax authorities disclose penalty schedules.

Keywords: tax compliance; enforcement ambiguity; penalty disclosure; ambiguity aversion; field experiment; pupillometry

¹ Floyd: University of California, San Diego; Hallsworth: BIT; List: University of Chicago; Metcalfe: Columbia University; Rotaru: Monash University; Vlaev: University of Warwick.

1. Introduction

Every tax authority collects less than it is owed. The Internal Revenue Service decomposes its gross tax gap into non-filing, under-reporting, and underpayment. It defines underpayment as “tax that was reported on time, but not paid on time”, and voluntary compliance as taxes paid voluntarily and on time (Internal Revenue Service, 2024). Authorities respond partly by writing to taxpayers, and they are seldom precise about what non-compliance costs. Letters warn that failure to pay “may result in further action” without naming the action. Penalty schedules state maxima “up to” an amount determined case by case. The criteria for audit selection are withheld. The practice is deliberate and defended on the record. In *Roberts v. Internal Revenue Service* the agency argued that “one of the tools in the arsenal of the IRS which promotes voluntary compliance is the uncertainty in the minds of the taxpayers as to just how much overstepping of the boundaries of strict compliance will bring down the enforcement authority of the agency” (Reinganum & Wilde, 1988, p. 794). Whether that imprecision raises or lowers compliance, and through what psychological process, is the question this paper investigates.

We study the margin where a taxpayer owes an acknowledged amount and has not paid it, and the margin where a taxpayer decides how much income to declare. Both are compliance in the administrative sense that the authorities themselves apply. The OECD defines taxpayer obligations as registration, timely filing, complete and accurate reporting, and payment on time (OECD, 2014, p. 60). A taxpayer failing any of them may be considered non-compliant. In the United Kingdom, the tax gap for 2024 to 2025 is provisionally estimated at £59.2 billion, 6.4 percent of theoretical tax liabilities (HM Revenue and Customs, 2026). The authority’s illustrative decomposition by behavior attributes £5.7 billion (10 percent) to non-payment, tax that is due but not ultimately collected. Timely payment is likewise central to the authority’s

administration strategy (HM Revenue and Customs, 2021). In the United States, underpayment is projected at \$94 billion of a \$696 billion gross tax gap for tax year 2022 (Internal Revenue Service, 2024). Non-payment of an acknowledged debt and deliberate under-declaration are therefore two canonical components of the tax gap.

We focus on non-payment for three reasons. First, it is where the enforcement communication we study occurs. Enforcement letters go to taxpayers who already owe a known amount, so the field evidence on message wording concerns the payment margin. Second, the margin is understudied relative to its share of the tax gap. Hallsworth (2014) defines compliance over filing, accurate reporting, and paying tax owed on time. The definition takes in revenue declared but not paid on time, around 10 percent of the United States tax gap at the time of writing. Third, the field-experimental literature itself has raised the distinction between the margins as a substantive question. Hallsworth et al. (2017), reporting enforcement-letter experiments with the same tax authority as our Experiment 1, consider whether decisions about declaring tax differ substantially from decisions about paying tax. They conclude that the possibility “points towards the need for more theoretical and empirical work on tax payment rather than tax declaration”.² The payment margin has begun to receive direct attention in accounting research (Hurwitz & McGowan, 2024; Robitaille et al., 2021). We bring the enforcement-communication question to it.

² Hallsworth et al. (2017) is also the published source of the eleven-point response scale used in Experiments 2 and 4. The paper reports a laboratory rating study whose first item reads, word for word, “How likely are you to pay the money to the government? Please give your answer out of 10, where 0 means you are certain not to pay and 10 means you are certain to pay”, as one of five items the authors describe as structured around the determinants of the Allingham and Sandmo model. Experiment 2 modifies the object of payment to “the outstanding tax obligation”; Experiment 4 reproduces the 2017 wording. The relation between Experiment 1 and the studies reported in that paper is set out in Section 3.1 and in Online Appendix C.

The payment decision is nonetheless distinctive. A taxpayer who does not pay an acknowledged debt conceals nothing, so the psychology of deception does not arise³. Research on past-due consumer debt shows, however, that repayment reflects more than the capacity to pay. A one-sentence moral appeal reduced credit card delinquency by as much as a substantial cash rebate (Bursztyn et al., 2019). Payments anchor on the stated minimum payment beyond what liquidity constraints explain (Keys & Wang, 2019). Repayments are allocated across debts by a balance-matching heuristic rather than cost minimization (Gathergood et al., 2019). Such non-pecuniary determinants reach tax debt: among United States tax delinquents, raising the social visibility of delinquency and reminding debtors of financial penalties both raised payment (Perez-Truglia & Troiano, 2018). This literature has not asked whether specifying the consequence of continued non-payment matters: that choice belongs distinctively to a creditor which also sets and enforces the penalty, and it is the choice our experiments manipulate.

Even so, we do not assume that findings on under-declaration transfer to the payment margin. The design tests the question. Experiment 2 holds the financial context of the debt fixed, so ability to pay cannot vary with the information manipulation. Experiment 3 places the same four information conditions onto an incentivized income-declaration decision, the canonical evasion margin, with real money at stake. If both margins show the same pattern, the phenomenon attaches to the enforcement communication rather than to features of either decision. One limit is stated at the outset: our laboratory measure on the payment margin is a stated likelihood rather than an observed payment; the behavioral evidence on that margin comes from Experiment 1, and the incentivized evidence from Experiment 3.

³ The creditor is the state, which unlike a commercial lender sets the terms of the debt, adjudicates disputes about it, and enforces it; and the capacity to pay constrains the decision in a way that the decision to under-declare does not.

Theory does not settle whether specifying an enforcement consequence raises or lowers compliance. The models on which enforcement policy is built take the enforcement parameters as known to the taxpayer (Allingham & Sandmo, 1972; Becker, 1968). They address what an authority should set rather than what it should disclose. Where relevant and knowable information is missing, however, decision-makers face ambiguity rather than risk and are commonly averse to it (Camerer & Weber, 1992; Ellsberg, 1961). An unspecified consequence may therefore weigh more heavily than a specified one of the same expected cost. The disclosure question was opened analytically by Beck and Jung (1989a, 1989b) and pursued experimentally by Beck et al. (1991, 1992). Those studies examine reporting when taxpayers are uncertain about their taxable income or the audit rule. That line establishes that the effect of taxpayer uncertainty is parameter-dependent rather than signed and locates the mechanism in the agency's equilibrium audit response.

Direct evidence is limited and runs in both directions. Tan and Yim (2014) find that an auditing rule announcing only a bound on the number of audits produces higher compliance than an announced constant probability when the resulting strategic uncertainty is high. The same rule suppresses the “bomb crater” effect, the immediate decline in compliance after a tax audit. Reinganum and Wilde (1988) show analytically that a unique positive amount of enforcement uncertainty maximizes equilibrium compliance. Ambiguity has, however, been found to raise non-compliance where detection is close to certain. Midgette et al. (2021) report that an ambiguous compliance threshold raised violations by 27 to 28 percentage points. Loughran et al. (2011) find ambiguity encouraging offending at high perceived detection probabilities. Whether specifying enforcement consequences helps or harms collection therefore remains an open empirical question.

Four experiments answer this question, moving from the field to physiological measurement and covering both margins, payment and reporting. In collaboration with HM Revenue and Customs, we randomized enforcement letters to more than 200,000 taxpayers with outstanding Self-Assessment debts. The letters compare a warning of unspecified further action with otherwise similar messages specifying a financial or an administrative cost. Two online experiments then manipulate whether the maximum fine is stated. The first operates on the payment margin, with prior beliefs elicited before any disclosure. The second uses an incentivized income-declaration decision with real money, a real undisclosed audit, and real penalties. A laboratory experiment records pupil dilation and visual fixations while participants read enforcement messages. Section 2.6 sets out how the four studies fit together.

Three findings emerge. In the field, leaving the consequence unspecified raised payment. The warning of further action increased payment within the administrative window by 4.4 percentage points, from 42.1 percent in the control group to 46.5 percent ($p < 0.01$, Cohen's $d = 0.09$). A structurally parallel message of comparable length referring to psychological relief rather than enforcement did not raise payment. The effect is attributable to the unspecified consequence rather than to salience or message length. For a message intervention in a national administrative population this is substantial: the information-based class is the weakest in meta-analytic evidence on choice architecture (Mertens et al., 2022). Second, in the laboratory, specifying the consequence reduced compliance. Revealing the maximum fine lowered the stated likelihood of paying relative to leaving it unstated, at modest and severe maxima alike. Belief updating does not accommodate this. Among participants whose elicited prior understated the fine, disclosure was unfavorable news that should have raised compliance, yet compliance fell. The same pattern appeared, attenuated, when the decision was an incentivized

declaration rather than a stated likelihood (one-tailed $p = 0.001$, $d = 0.21$, $N = 827$), extending the result to the reporting margin. Third, autonomic arousal differed across the enforcement messages while visual attention did not, a dissociation consistent with affect rather than attention.

The paper makes three contributions. To the economics of tax enforcement, we show that what an authority discloses is itself an enforcement instrument. The closest antecedents are the taxpayer-uncertainty studies of Beck and Jung (1989a, 1989b) and Beck et al. (1991, 1992). We extend that line by manipulating communicated penalty information while holding audit information constant and undisclosed. Because the applied penalty is drawn independently of the declaration, the equilibrium audit response those papers identify cannot produce our result. We also test it in the field at scale. To the behavioral literature on ambiguity, we document ambiguity aversion in a consequential and institutionally embedded decision. The demonstration spans the loss domain, two compliance margins, and real monetary incentives, with physiological evidence on the process. To practice, we identify a cost in transparency initiatives that would state worst-case penalties as specific amounts. The specification removes a worst case that taxpayers otherwise supply for themselves. We draw no conclusion about an optimal disclosure policy: the design compares stating a maximum against not stating one and does not evaluate combinations of disclosure and severity.

The paper proceeds as follows. Section 2 develops the theory, states four hypotheses, and sets out how the experiments fit together. Sections 3 to 6 report the four experiments in turn, and Section 7 discusses the implications, the scope of the compliance construct, and the limitations.

2. Theory and hypothesis development

Tax authorities seldom tell taxpayers what will follow from non-compliance. Enforcement letters warn of “further action” without naming it. Penalty schedules state maxima “up to” an amount determined case by case. Audit selection criteria are withheld. The practice is deliberate. Bergolo et al. (2023) observe that audit-probability information is much harder to obtain than other decision-relevant information because tax authorities “seem to prefer to conceal this information”. Vossler and Gilpatric (2018) note that agencies employing endogenous audit rules do not advertise their procedures. Nor is the practice peculiar to tax administration. Firms differ in transparency about the parameters of their control systems, and some penalize employees for sharing target information (Arnold et al., 2024).

The taxpayer’s resulting information state is impoverished. Widiyanto et al. (2024) report that 73.9 percent of their survey respondents were unaware of imprisonment sanctions and about half were aware of administrative fines. Audit coverage for individuals in that system is 0.4 percent. Nagin (2013) reports that only two dated surveys of public knowledge of statutory penalties exist and that sanction-risk perceptions are unstable and poorly calibrated. Taxpayers therefore hold vague beliefs about what enforcement would cost them, and authorities decline to correct those beliefs.

2.1 Ambiguity about enforcement consequences

We use *ambiguity* in the sense adopted in Camerer and Weber’s (1992, p. 330) review: uncertainty about probability, created by missing information that is relevant and could be known. The definition fits enforcement communication directly. The maximum penalty is relevant information that the authority holds and chooses not to state. Ellsberg (1961) supplies the underlying concept: the operative quality is “the ambiguity of this information, a quality

depending on the amount, type, reliability and ‘unanimity’ of information”. His construct concerns the information supporting judgments of likelihood. In an enforcement setting the magnitude of the consequence is also left unstated. Ellsberg’s analysis does not speak to that feature, and Section 2.5 addresses it from the affective literature.

Two points of terminology follow. First, *ambiguity* denotes the taxpayer’s information state and *ambiguity aversion* the preference that gives that state behavioral consequences.⁴ Second, *uncertainty* is retained only where a cited literature uses it as its own term, where it names an experimental condition or a measured item, or where it names a treatment in the field experiment. It does no theoretical work in this paper.

Classical deterrence models take the enforcement parameters as known to the taxpayer and address what an authority should set rather than what it should disclose (Allingham & Sandmo, 1972; Becker, 1968). The analytical work that does treat the taxpayer’s information yields conditional, non-monotonic prescriptions rather than a case for transparency. Bebachuk and Kaplow (1992) show that the maximal-sanction result fails when individuals perceive the enforcement probability with error, because a high sanction magnifies that error in perceived expected sanctions. Reinganum and Wilde (1988) make the authority’s tolerances endogenous. In the case they describe as policy-relevant, a unique positive amount of uncertainty maximizes both equilibrium compliance and equilibrium revenue. They conclude that refusing to disclose explicit criteria may be appropriate while extensive secrecy is unwarranted (p. 798). Calford and DeAngelo (2023) make revelation itself the choice variable. In their model, an agency seeking deterrence should shroud its resource allocation when those it monitors are uncertainty

⁴ We adopt no single label for the authority’s decision to create or preserve the state, because we are aware of no established term whose definition covers an enforcement authority withholding penalty information; we describe the decision instead.

averse (their term) and reveal it when they are uncertainty seeking. Closest to our premise, Beck and Jung (1989b, pp. 482-483) conclude from an accounting-journal model that the agency “may not have strong incentives to support tax and penalty simplification”. They add that unless the anticipated audit probability exceeds 0.5, greater uncertainty about audit costs will likely produce less aggressive reporting. On that basis, “a rationale is also provided for the tax agency’s nondisclosure of information about their audit strategy.”

Across these results the sign of the prescription depends on how the regulated population responds to ambiguity. This paper estimates that response.

2.2 The explanations to be separated

Three explanations could produce a compliance response to enforcement messages that leave consequences unspecified. They differ in what each predicts once the consequence is stated. Under *ambiguity aversion*, a taxpayer who cannot bound the worst case treats an ambiguous prospect as worse than a specified one. Stating the maximum therefore lowers compliance whatever number is stated (Camerer & Weber, 1992; Ellsberg, 1961). Under *belief updating*, behavior follows the direction of the news relative to the taxpayer’s prior belief, so a stated penalty above (below) the prior should raise (lower) compliance (Snow & Warren, 2007).

Under an *attention* explanation, message effects track engagement with the message rather than the content of what is disclosed (Lynch & Andiola, 2019). The three are not mutually exclusive, and the designs reported below separate them. Every participant’s prior belief about the maximum is elicited before any disclosure, so the updating prediction is testable at the individual level. Visual attention is recorded directly by eye-tracking in Experiment 4. Evidence against the updating prediction is not evidence against Bayesian updating as a computation. In the two-level framework of Sommer et al. (2024), updating proper is approximately Bayesian

while the upstream processes of evidence search, acceptance, and integration are heuristic. What our designs can locate is whether an unstated maximum is represented and weighted differently from a stated one. A precursor study, reported in full in Online Appendix A and referred to below as Study A, supplies the premise that taxpayers fill unspecified enforcement consequences with fines rather than with imprisonment, which is why fine information is the object manipulated in Experiments 2 to 4.

2.3 Ambiguity and compliance in the field

That the wording of an enforcement letter changes payment behavior at scale is established. Hallsworth et al. (2017) demonstrate it in the same tax system and on the same margin as Experiment 1. Bergolo et al. (2023) and Holz et al. (2023) demonstrate it for firms, Bott et al. (2020) for foreign-income reporting, and Castro and Scartascini (2015) for a municipal tax. Gil et al. (2024) demonstrate it for amnesty take-up and Yogama et al. (2024) for tax-penalty payment. That enforcement messages outperform appeals to morality or reciprocity is close to settled (Hallsworth, 2014) and has been reproduced repeatedly (Bott et al., 2020; Gil et al., 2024; Widiyanto et al., 2024; Yang et al., 2024).

What has not been tested is whether an enforcement message is more effective when it leaves the consequence unspecified than when it specifies one. Three studies approach the question without answering it. The effective deterrence letter in Yogama et al. (2024) warns that failure to settle “may result in heightened actions” by the authority, an escalation it does not set out. The most durable treatment in Yang et al. (2024) names a class of consequence with no magnitude attached: the debt will be recorded and will affect the taxpayer’s personal interests. That message outperformed one that quantified the overdue fine. Abdelnabi et al. (2024) conclude that where financial penalties or legal action are rarely applied, “messages that

have vaguely defined consequences can be particularly effective”. However, no message combination in that study significantly improved filing on average, and its compliance margin is the filing of returns rather than payment. We therefore treat its conclusion as suggestive rather than as corroboration.

We expect a message that leaves the consequence unspecified to raise compliance for two reasons. First, a taxpayer facing an enforcement consequence is in the domain of losses, where ambiguity aversion has been demonstrated directly. Loughran et al. (2011) find that respondents prefer a precise .30 probability of a loss to an equally valued range of .20 to .40. Second, an unspecified consequence leaves the worst case to be imagined. Anticipatory emotions such as fear, anxiety, and dread respond to a risky situation in ways that diverge from its cognitive assessment, and the emotional reaction often drives behavior where the two diverge (Loewenstein et al., 2001). Both reasons attribute the effect to the absence of a stated consequence rather than to the presence of an additional sentence. Experiment 1 therefore includes a message of parallel structure and comparable length whose content is psychological relief rather than enforcement, as a comparison for message salience.

H1. Enforcement messages that leave consequences ambiguous produce higher compliance than otherwise similar messages that specify costs.

2.4 Revealing the maximum penalty, and the boundary of the effect

The nearest antecedents are the analytical and experimental studies of taxpayer uncertainty by Beck and Jung and by Beck, Davis, and Jung. That line does not sign the effect. Beck and Jung (1989a) find that greater tax-liability uncertainty raises reported income under some parameter conditions and lowers it under others. Their Proposition 6 (p. 12) states that under risk neutrality an increase in liability uncertainty reduces (increases) reported income when $p/(1 - p)$

$< (>) 2/q$, for audit probability p and penalty rate q . For parameter ranges commonly occurring in the United States, they conclude, such an increase is “much more likely to lead to a reduction in reported income” unless taxpayers are highly risk averse. Beck and Jung (1989b) obtain the opposite sign in their Proposition 4 (p. 479), where greater liability uncertainty raises the equilibrium audit benefit and the anticipated audit probability and produces less aggressive reporting. Their footnote 17 records that the two models differ because the audit probability is fixed in the earlier one and endogenous in the later one. The direction is conditional in a second respect. Proposition 4 holds where monetary penalties are proportional to the deficiency and are transferred to the agency. Under fixed, nontransferable penalties, Proposition 8 (p. 481) yields divergent effects. In that case “the magnitude of potential penalties borne by taxpayers is independent of the tax liability and the tax agency has diminished incentives to perform audits”. Their Table 1 (p. 482) tabulates the effects under both penalty structures and notes that divergent signs depend on the specific parametric relations. Yoon et al. (2011) reach a third indeterminacy once litigation is modeled.

The two experimental members of that line locate the mechanism. Beck et al. (1991) manipulated taxable-income uncertainty and found no effect on the reporting fractile ($p = 0.879$). The effects on reported income were interactions with the audit probability and with the penalty rate (both $p < 0.001$), whose sign follows the threshold stated in their H3a (p. 540). Beck et al. (1992) then allowed the audit rule to respond to the report and found that reducing taxable-income uncertainty made low reports more frequent, but only under strategic auditing. The mechanism throughout is the agency’s equilibrium audit response rather than a preference over ambiguity. The object of the uncertainty is the taxpayer’s own liability or the agency’s audit rule, not a penalty the authority communicates.

Three incentivized studies outside that line show that removing an enforcement-relevant uncertainty can lower compliance for a different reason. Vossler and Gilpatric (2018) found that reducing taxpayers' liability uncertainty raised evasion rather than reducing it. They attributed the result to improved information reducing "maximum possible penalties", which lowers the risk of very negative outcomes from an audit. Agranov and Buyalskaya (2022) held the monitoring probability and the expected fine constant and found that a scheme communicating only the minimum fine deterred lying better than either a scheme communicating the maximum or full information, with beliefs about one's own fine as the measured channel. Snow and Warren (2005) supply the analytical statement of the preference at issue: ambiguity about the audit probability raises compliance among the ambiguity averse and lowers it among ambiguity lovers.

We expect revealing the maximum penalty to lower compliance, because a stated maximum bounds a worst case that is otherwise left to be imagined. Experiments 2 and 3 hold audit information constant and undisclosed, draw the applied penalty independently of the declaration, and elicit each participant's prior belief about the maximum before any disclosure. The audit-response channel identified by Beck, Davis, and Jung is closed by design, and the belief-updating prediction is separately testable.

H2. Revealing the actual maximum penalty reduces compliance relative to leaving the maximum ambiguous.

Whether the magnitude revealed also matters is a separate question, and the evidence is mixed. Three literatures suggest that it should matter little. Once an outcome category carries strong negative affective meaning, variation in its parameters is weighted lightly and responses acquire an all-or-none character (Slovic et al., 2004). In the field, Bergolo et al. (2023) varied

the communicated penalty rate from 15 to 68 percent and the communicated audit probability from 2 to 25 percent. The estimated elasticities were -0.033 and -0.063 respectively, neither statistically distinguishable from zero and both far below the calibrated values. In criminology, the certainty of apprehension deters more effectively than the severity of the consequence (Nagin, 2013, 2017). Two further mechanisms limit the effect of a severe stated maximum. It may not be believed, and it may not be imposed, since willingness to punish falls as the fine rises when guilt is uncertain (Feess et al., 2018).

This prediction is not without tension. Friesen (2012) finds that increases in severity deter more than equivalent increases in the probability of punishment, in a design displaying the inspection probability and the fine in each round. Expected-utility reasoning under risk aversion implies that result. H3 is accordingly stated as a claim about revealing a previously unstated maximum rather than as a claim that magnitude never matters. It is tested on the hypothetical payment margin, where the two revealed maxima differ by a factor of more than five.

Experiment 3 examines its boundary under real monetary stakes as a secondary question.

H3. The reduction in compliance produced by revealing the maximum penalty does not depend on the magnitude revealed: a modest and a severe revealed maximum reduce compliance similarly.

2.5 Affect rather than attention

The reasons given for H1 and H2 attribute the effect to what a taxpayer feels rather than to what a taxpayer computes. That attribution can be measured rather than assumed. Uncertainty is experienced subjectively as anxiety and is associated with heightened noradrenergic activity. Clear goals and belief structures constrain the experience by narrowing the set of competing possibilities (Hirsh et al., 2012). A stated maximum is such a structure. The clinical literature

draws the same distinction. Fear is a circumscribed response to imminent danger, whereas anxiety is apprehension “about a situation in which the outcome may be uncertain” and its expression is broad and diffuse (Hyde et al., 2019). Leaving the maximum unstated is a manipulation of anxiety in this sense rather than of fear.

Both halves of the prediction have established measurement mappings. Bradley et al. (2008) recorded pupil diameter together with skin conductance and heart rate during picture viewing. Following an initial light reflex, dilation was larger for emotionally arousing pictures irrespective of whether they were pleasant or unpleasant. Dilation covaried with skin conductance rather than with heart-rate change, which supports sympathetic mediation. Pupil dilation therefore indexes arousal rather than valence. In accounting research, eye-tracking metrics predominantly proxy attention to stimuli, through the count and duration of fixations within an area of interest (Lynch & Andiola, 2019). Rotaru et al. (2025) record that pupillometry entered accounting research only after 2019 and remains uncommon. Hecht et al. (2025) demonstrate the measure in a management-accounting task, and Wang et al. (2010) established the pairing of the two measures in incentivized decisions.

Neither measure identifies emotion on its own. Pupil dilation also tracks cognitive effort (Alós-Ferrer et al., 2021), and Mathôt (2018) states that in terms of pupil size the effects of arousal and of mental effort appear similar. The prediction is therefore stated as a dissociation between two channels rather than as a main effect in one. Were enforcement messages to differ in the processing they demand, the fixation measures would be expected to move with the pupil. Were they to differ in the affect they provoke, arousal should differ while attention does not. Rose et al. (2022) report that second pattern for two data-visualization formats and present the flat fixation measures as a substantive result rather than as a failed manipulation, the

precedent followed here. Because post-experimental questionnaire measures used as mediators are particularly prone to validity threats (Asay et al., 2022), we treat the self-report process measures as convergent evidence rather than as formal mediators. The physiological measure of Experiment 4 is analyzed as a mediator, and Section 6 reports the indirect effect.

H4. Enforcement-message effects are accompanied by heightened autonomic arousal, not by differences in visual attention or belief updating.

2.6 Overview of the experiments

Four experiments test the hypotheses. They cover both behavioral margins of the tax gap and move from the field to physiological measurement (Figure 1). Experiment 1 randomizes enforcement letters to more than 200,000 taxpayers holding outstanding Self-Assessment debts and tests H1 on payment behavior recorded administratively. Experiment 2 brings the same contrast into an online panel on the same margin, adds the conditions that separate the fact of disclosure from the magnitude disclosed, and tests H2 and H3. Experiment 3 transfers the four information conditions to an incentivized income-reporting decision with real money, a real undisclosed audit, and real penalties. It tests H2 on the reporting margin and examines H3 as a secondary question. Experiment 4 records pupil dilation and fixations while participants read the messages, and tests H4. The sequence is cumulative. Each experiment answers the sharpest objection available to the one before it: field realism, then isolation of the operative feature, then generality of the compliance margin, then the process.

[Insert Figure 1 about here]

3. Experiment 1: Field evidence from HMRC

3.1 Setting, sample, and randomization

In collaboration with HM Revenue and Customs (HMRC), we randomized enforcement letters to 204,945 UK taxpayers in England, Wales and Northern Ireland. All held standalone Self-Assessment income-tax debts of between £250 and £50,000 as at 1 February 2013 and had not responded to the initial reminder statement.⁵ Letters were issued over seven working days between 5 and 13 March 2013. The primary outcome is whether the taxpayer paid within 23 days of receiving the letter, the administratively relevant window before follow-up action. The sample was randomized into twelve groups (mean $n = 17,079$) by taxpayer reference number. Background characteristics were balanced across the groups (Online Appendix C, Table OA-C1).

Hallsworth et al. (2017) reported experiments in the same recovery process. They inserted social norm and public services messages into the authority's standard reminder letters to Self-Assessment debtors, on debts outstanding in August 2011 and in August 2012, and measured payment within the first 23 days. Experiment 1 is a separate randomization conducted later in the same process. Its sample is defined on debts outstanding at 1 February 2013, and its treatments vary the content of enforcement communication rather than social information.

Online Appendix C sets out the correspondence between the two designs.

⁵ This total corrects the previous submission, which stated 204,936 in its text and reported two group sizes, for the Contact Now and Contact Now Information conditions, that were 34 and 10 lower than the totals its own date-of-issue table gave for the same conditions. We adopt the date-of-issue figures, because each of the twelve group totals in that table reconciles with its seven daily issue counts. The corrected group sizes are reported in Online Appendix C and sum to the total stated here.

3.2 Treatments

All letters contained the amount owed and instructions for payment. Each treatment letter added one short message in bold typeface. The messages varied three features of the communication. Messages on financial costs stated the monetary consequence of the timing of payment: the “Interest Today” message read, “We are charging you interest daily. You will save money if you pay now.” Messages on administrative effort framed prompt payment as reducing inconvenience: the “Less Effort” message read, “Paying now means less hassle because we won’t contact you about this again.” A third class left the enforcement consequence unspecified: the “Costs Further Action” message read, “If you pay this debt now, you can stop us taking any further action.” The “Costs Peace” message read, “If you pay this debt now, you won’t have to think about it anymore.” It is structurally parallel to “Costs Further Action” but refers to psychological relief rather than enforcement. It provides the comparison for message salience. Full message texts and group assignments are given in Online Appendix C.

3.3 Results

[Insert Table 1 about here]

Table 1 reports the treatment effects. The message that left the enforcement consequence unspecified was the most effective treatment: “Costs Further Action” increased payment within the administrative window by 4.4 percentage points, from 42.1 percent in the control group to 46.5 percent ($p < 0.01$, Cohen’s $h = 0.09$). The estimated effect of the “Costs Peace” message was -0.8 percentage points (standard error 0.5 percentage points; Table 1). The comparison locates the effect of “Costs Further Action” in the unspecified enforcement consequence rather than in salience or message length. The message that combined a financial cost with the unspecified consequence, “Interest & Action”, raised payment by 1.7 percentage points ($p =$

0.002). The messages stating financial costs or administrative effort alone produced estimated effects of between -0.3 and 0.8 percentage points (Table 1). The intervention accelerated approximately £5.16 million in tax payments within the 23-day window. Heterogeneity analyses by debt size and payment history, reported in Online Appendix C, are directionally consistent with the main effect.

3.4 What Experiment 1 establishes and what it cannot

Experiment 1 supports H1. The message warning of unspecified further action raised payment. Otherwise similar messages specifying a financial or an administrative cost did not, and neither did a message of parallel structure without enforcement content. The design cannot identify which feature of the ambiguous message produces the effect, or the psychological process involved. A field letter presents the unspecified consequence together with the legal formality of an official communication, and no belief or process measures can be collected from its recipients. Experiments 2 to 4 isolate these elements. The field data were analyzed on HMRC premises under the authority's data-access requirements. The experiment is reported in this revision with presentation changes only.

4. Experiment 2: Revealing the fine on the payment margin

4.1 Design, sample, and procedure

Experiment 2 was conducted in two periods of data collection with UK taxpayers recruited through Prolific (£1 compensation; participants in prior related studies excluded). Fourteen conditions were collected in total, and 2,050 taxpayers provided the dependent variable across them. The first period, from which the principal analyses are drawn, comprised the eight factorial conditions and the two further £2,500-debt conditions described below, and accounts for 1,490 of those taxpayers. The second period, described at the end of this subsection,

accounts for the remaining 560. Of the 1,195 taxpayers who completed the factorial conditions (50.2 percent female; mean age 41), 1,194 provided the dependent variable and form the analysis sample for the factorial design. Each participant read a hypothetical scenario in which they held an outstanding tax balance and then answered the item “How likely are you to pay the outstanding tax obligation to the government? (0 = Certain not to pay, 10 = Certain to pay)”. The eleven-point scale and its anchors are the published instrument reported in Hallsworth et al. (2017), whose item asks about “the money”. Experiment 2 modifies the object of payment to the outstanding tax obligation, and the dependent variable is the stated likelihood of paying it.

The design crossed the size of the debt (£500 or £2,500), the type of fine described (minimum or maximum), and the information provided. In the Revealed conditions the participant’s prior belief about the fine was elicited and the actual fine was then disclosed. In the No-Reveal conditions the prior belief was elicited and no fine was disclosed. Fine information is the object manipulated in this design because taxpayers fill unspecified enforcement consequences with fines rather than with imprisonment, the premise stated in Section 2.2 and supplied by Study A (Online Appendix A). The fine values were calibrated in a pilot study conducted for this experiment (N = 199), in which each taxpayer estimated the minimum and the maximum penalty for one of the two debt levels. For the £2,500 debt the mean estimated maximum was £1,960.55, from which £1,900 was adopted as the revealed maximum and £800 as the corresponding minimum (£1,600 and £700 for the £500 debt). Because every participant’s prior belief is measured before any disclosure, the design can separate belief updating from ambiguity aversion at the individual level.

Two further between-subjects conditions, collected in the first period with the factorial conditions, complete the maximum-fine design at the £2,500 debt: an Uncertain condition (n =

147) and an Elevated condition ($n = 149$). In the Uncertain condition, participants were told that the maximum fine is set case by case and may increase with the severity of the non-compliance. In the Elevated condition, a £10,000 maximum was revealed. The No-Reveal condition mentions the maximum and never resolves it, and the Specific condition reveals the £1,900 maximum. These four cells separate the fact of disclosure from the magnitude disclosed (Section 2.6), and they are tabulated together in Table 3.

In the second, later period of data collection, four further conditions were collected at the £500 debt level: an Uncertain condition ($n = 138$), an Elevated condition revealing the £10,000 maximum ($n = 139$), and a second pair of revealed values, a £1,200 maximum ($n = 142$) and a £600 minimum ($n = 141$). These conditions are reported as supplementary analyses in Section 4.3, and their tabulation is in Online Appendix D.

4.2 Results for the factorial design

All participants correctly recalled the size of the tax debt they had been shown.

[Insert Table 2 about here]

Descriptive statistics are reported in Table 2, Panel A, and the analysis of variance in Panel B. Consistent with H2, the only significant effect in the design is the main effect of information: participants who were told the actual fine reported a lower likelihood of paying than participants for whom the fine was left unresolved (8.27 against 8.70; $F(1, 1186) = 10.00$, $p = 0.002$). Neither the debt manipulation ($F(1, 1186) = 3.44$, $p = 0.064$) nor the fine-type manipulation ($F(1, 1186) = 0.67$, $p = 0.413$) was significant, and no interaction was: fine type $\times$ information, $F(1, 1186) = 1.57$, $p = 0.211$; debt $\times$ information, $F(1, 1186) = 0.49$, $p = 0.484$; the three-way interaction, $F(1, 1186) = 0.00$, $p = 0.967$.

Simple-effects tests locate the reduction. Within the maximum-fine conditions, disclosure lowered the stated likelihood of paying from 8.73 to 8.14 (Welch $t = 3.07$, $p = 0.002$). Within the minimum-fine conditions the corresponding change, from 8.67 to 8.41, was not significant (Welch $t = 1.40$, $p = 0.163$). Because the fine type $\times$ information interaction is not significant, we do not claim that disclosure matters more for maximum-fine than for minimum-fine information. What the design supports is that disclosure lowered compliance overall, and that the reduction within the maximum-fine conditions is significant on its own (Welch $t = 3.07$, $p = 0.002$).

A further feature of Table 2, Panel A concerns the dispersion of the responses rather than their means. The standard deviations in the disclosure cells (2.45-2.80) are uniformly larger than those in the no-disclosure cells (1.93-2.01). Revealing the fine lowered the average response and widened the distribution of responses. This pattern is consistent with individual differences in the response to ambiguity.

4.3 The four maximum-fine conditions

[Insert Table 3 about here]

The pattern across the four maximum-fine cells is the principal result of the paper. It is reported in Table 3 and shown in Figure 2, panel (a). The stated likelihood of paying is highest where the maximum remains unresolved and lower where it has been revealed. The planned ambiguity contrast $\{1, 1, -1, -1\}$ is significant ($F(1, 604) = 12.84$, one-tailed $p = 0.0002$, $d = 0.29$). Participants reported a higher likelihood of paying when the maximum was left ambiguous than when it was revealed (8.722 against 8.033). The residual between-cells variance is not significant ($F(2, 604) = 0.03$, $p = 0.970$), and the contrast leaves $q^2 = 0.5$ percent of the

systematic between-cells variance unexplained. The difference between the two ambiguity-preserved cells is 0.052 (No-Reveal 8.697, Uncertain 8.748; Welch $p = 0.822$).

Two features of this pattern identify ambiguity aversion rather than belief updating. The first is the comparison stated in H3, between the two revealed maxima. The Specific and Elevated conditions state maxima that differ by a factor of more than five, £1,900 against £10,000, and the responses they produced differ by 0.041 (Specific 8.013, Elevated 8.054; 95 percent confidence interval [-0.568, 0.650], Welch $p = 0.895$). Consistent with H3, the reduction in the stated likelihood of paying accompanied the revelation of the maximum rather than the magnitude revealed.

The second feature is the direction of the response relative to the news received. For each participant in a disclosure condition, the fine disclosed can be compared with the fine the participant had estimated moments earlier. Among participants who had under-estimated the maximum, the disclosure was unfavorable news, and belief updating predicts that compliance should rise. The stated likelihood of paying instead fell, from 8.73 to 8.04 (difference -0.697, Welch $t = 3.02$, $p = 0.003$). Among participants who had over-estimated the maximum, for whom the disclosure was favorable news, the change was in the same direction and not significant (8.71 to 8.41; difference -0.306, Welch $t = 0.92$, $p = 0.360$). The belief $\times$ information interaction is not significant ($F(1, 595) = 0.87$, $p = 0.352$), so we do not claim that the reduction is larger in one subgroup than in the other. The claim is narrower: in the subgroup for which belief updating makes a directional prediction, behavior moved in the opposite direction.

Individual-level belief analyses are reported in Online Appendix D.⁶

⁶ The corresponding analysis within the minimum-fine conditions is weaker and we do not rely on it. There the reduction appears among participants who had over-estimated the minimum (8.92 to 8.33; difference -0.597, Welch $t = 1.82$, $p = 0.070$) rather than among those who had under-estimated it (8.56 to 8.45; $p = 0.633$), and the

The conditions collected in the second period permit a counterpart of this four-cell comparison at the £500 debt. Together with the factorial design's two £500-debt maximum-fine cells (No-Reveal, and the revealed £1,600 maximum), the second period's Uncertain and Elevated conditions form the same four-cell structure. The ambiguity contrast replicates: the pooled ambiguity-preserved cells average 8.652 ($n = 276$) against 8.160 in the pooled resolved cells ($n = 288$), a difference of 0.492 ($F(1, 560) = 5.36$, one-tailed $p = 0.011$, $d = 0.19$). The omnibus F-test across the four cells is not significant ($F(3, 560) = 2.14$, $p = 0.094$). Among the revealed values, the four magnitude comparisons are likewise not significant (£1,200 against £1,600, $p = 0.644$; £600 against £700, $p = 0.271$; £1,600 against £10,000, $p = 0.498$; £1,200 against £10,000, $p = 0.795$), consistent with H3. These supplementary analyses are consistent with the principal analyses. Cell-level means and standard deviations are reported in Online Appendix D.

4.4 What Experiment 2 establishes

The results are consistent with H2 and H3 on the hypothetical payment margin, in the principal analyses and in the supplementary £500-debt comparisons. Revealing the maximum fine lowered the stated likelihood of paying the outstanding tax obligation, at a modest and at a severe revealed maximum alike. The reduction appeared in the subgroup for which the disclosure was unfavorable news and belief updating predicts the opposite movement. The reconciliation of these results with Study A, including where the two designs differ, is reported in Online Appendix D. Experiment 2 establishes the pattern on a stated likelihood of payment.

belief $\times$ information interaction is again not significant ($F(1, 591) = 1.51$, $p = 0.220$). We report this analysis for completeness and draw no conclusion from it about distinct pathways for minimum and maximum fines.

Whether the same pattern appears when the decision is an incentivized declaration with real consequences is the question Experiment 3 addresses.

5. Experiment 3: Revealing the fine on the reporting margin

5.1 Motivation

Payment and reporting are administratively distinct compliance obligations (OECD, 2014), and how much income to declare is the canonical decision of the tax-evasion literature (Allingham & Sandmo, 1972; Beck et al., 1991; Alm & Malézieux, 2021). Experiment 3 places Experiment 2's four maximum-fine conditions, wording unchanged, on an incentivized income-reporting decision and tests H2 on that margin. Incentivization matters because under-declaration trades tax retained now against an expected penalty later. Without stakes, a reporting task elicits a statement of intent rather than the decision the theory concerns. Incentivized reporting is also the standard instrument on this margin. The Beck, Davis, and Jung experiments and the seventy tax-evasion games meta-analyzed by Alm and Malézieux (2021) all attach money to the declaration, making Experiment 3 commensurable with that literature. The design also supplies an observed action where Experiment 2 supplies a stated likelihood.

5.2 Design, sample, and procedure

Participants were UK residents aged 21-65 with self-employment or business-income experience, recruited through Prolific. Each participant completed an incentivized declaration task as owner of a stylized business that had earned 10,000 tokens of taxable profit in addition to their other income. The 10,000 tokens were worth £1.00 of real bonus payment. The tax rate was 40 percent of the declared amount, corresponding to the UK higher-rate income-tax band.⁷

⁷ The task levies a single stylized income tax. Self-employed taxpayers additionally pay Class 4 National Insurance contributions (2 percent above the upper profits limit); a combined rate of approximately 42 percent would only strengthen the incentive to under-declare, and nothing in the design depends on this simplification.

Participants chose how much of the 10,000 tokens to declare. Each return faced a real audit draw with 25 percent probability, never stated to participants and constant across conditions. A checked, under-declared return paid the additional tax due plus the condition's maximum fine scaled by the share of income under-declared. Penalties were deducted only from the bonus, which could not fall below £0. The guaranteed participation payment was never at risk. All realized bonuses were paid.

The four between-subjects conditions varied only the maximum-fine notice, worded identically to Experiment 2. In No-Reveal the maximum is mentioned and never resolved. In Uncertain it is set case by case and may increase. Specific reveals a £1,900 maximum and Elevated a £10,000 maximum, both denominated in tokens. In the two ambiguity-preserved conditions, the penalty for audited under-declarers was drawn from a predetermined distribution spanning the resolved maxima with mean at their midpoint, independent of the declaration and never shown to participants. The resolved conditions applied their stated maxima. Prior beliefs about the maximum fine were elicited before any disclosure, with Experiment 2's wording up to two unit substitutions (tokens for pounds). Perceived audit likelihood, measured after the declaration and before the audit outcome, did not differ across conditions ($F(3, 823) = 1.45$, $p = 0.23$), indicating that the manipulation moved penalty ambiguity rather than audit-probability beliefs.

The disclosed £0 lower bound on the bonus bounds the material worst case, which the field setting leaves to be imagined without limit. The design is therefore a conservative test of H2.

5.3 Statistical approach

The primary test of H2 is the planned ambiguity contrast $\{1, 1, -1, -1\}$, under which the pooled ambiguity-preserved conditions are predicted to exceed the pooled resolved conditions. It is evaluated following Guggenmos et al. (2018) with its residual between-cells variance. The comparison between the Specific and Elevated conditions is a secondary question, reported as a difference with its confidence interval and exact p value. Under real monetary stakes a £10,000 maximum is materially worse than a £1,900 maximum. H3 is therefore tested on the hypothetical payment margin in Experiment 2, and Experiment 3 examines its boundary (Section 2.4). Welch-type heteroskedasticity-robust tests are the default (cell standard deviations range from 0.36 to 0.41).

5.4 Results

Of the 887 participants who completed the task, 827 (93.2 percent) passed both the attention check and the instructional manipulation check and constitute the primary sample (60 excluded: 32 failed the attention check only, 25 the instructional check only, and 3 both; approximately 207 participants per cell). Inferences are unchanged when all participants are retained or a consequence-comprehension check is additionally required (Online Appendix E, which also reports covariate-adjusted models with unchanged inferences).⁸ A separate recognition check asked whether a specific maximum-fine amount had been stated. Correct answers were common in the No-Reveal (96 percent), Uncertain (94 percent), and Specific (68 percent) conditions and substantially less common in the Elevated condition (42 percent). In that

⁸ With all 887 completed responses retained, the ambiguity contrast is $F(1, 883) = 11.23$, one-tailed $p = 0.0004$; requiring the consequence-comprehension check in addition ($N = 806$) gives $F(1, 802) = 7.85$, one-tailed $p = 0.0026$.

condition the stated 10,000-token maximum frequently was not recognized as a specific figure, which matters for interpreting the cell means below.

[Insert Table 4 about here]

The task produced substantial variation: 53.3 percent of participants declared their full income, and the remaining declarations spanned the measure's range. Figure 2, panel (b), plots the cell means of Table 4, Panel A, against the predicted pattern. Consistent with H2, the planned ambiguity contrast $\{1, 1, -1, -1\}$ is significant ($F(1, 823) = 9.52$, one-tailed $p = 0.001$; Welch-robust one-tailed $p = 0.001$; $d = 0.21$). Declarations were higher under ambiguous than under resolved maximum-fine information (0.721 against 0.636). The residual between-cells variance is also significant ($F(2, 823) = 3.12$, $p = 0.045$; $q^2 = 39.6$ percent), reflecting one departure from a two-level pattern. Declarations under the resolved Elevated maximum (0.684) lie with the ambiguity-preserved cells rather than with the Specific cell (Specific 0.590 against pooled ambiguity-preserved 0.721; Elevated against pooled ambiguity-preserved difference -0.038, $p = 0.26$). Ambiguity aversion predicts this pattern. Where the maximum is unspecified, participants act as if the penalty is large, so revealing a large maximum changes behavior little while revealing a moderate maximum lowers declarations. Consistent with this reading, restricting the sample to participants answering the recognition check correctly strengthens the contrast (0.116; $F(1, 617) = 10.44$, one-tailed $p = 0.0006$). It also lowers the Elevated mean to 0.665 and renders the residual between-cells variance non-significant ($p = 0.086$). Where the £10,000 disclosure was recognized, the condition behaved as a resolved maximum.

Consistent with Section 2.4, the comparison between the two resolved conditions is a secondary question. Experiment 2 addresses whether the magnitude revealed matters on the hypothetical payment margin. Under real incentives the two maxima are not interchangeable:

declarations were higher under the Elevated than under the Specific maximum (0.684 against 0.590; difference 0.093, 95 percent confidence interval [0.015, 0.172], Welch $p = 0.021$). We therefore neither claim that the reduction under real incentives is independent of the magnitude revealed nor interpret the two conditions as equivalent. The two one-sided tests (TOST) procedure appears, with no equivalence claim, in Online Appendix E.

The mechanism items are consistent with ambiguity aversion at the first stage. Felt uncertainty about the penalty was higher in the ambiguity-preserved than in the resolved conditions (6.26 against 5.49 on 0-10 scales; Welch $p = 0.0002$), while worry ($p = 0.21$) and perceived severity ($p = 0.24$) did not differ. Felt uncertainty does not, however, predict declarations. A formal mediation of the declaration effect through self-reported felt uncertainty is not supported in the incentivized task (a-path 0.774, $p < 0.001$; indirect effect -0.0007, 95 percent bootstrap confidence interval [-0.008, 0.007]). The items therefore provide convergent evidence that the manipulation moved the intended construct. The process question is assigned to Experiment 4.

A further item, a check on the stimuli rather than a mechanism, asked how realistic the notice's penalty information was. Notices that named a specific maximum were rated somewhat less realistic than those that did not (6.62 against 7.23; Welch $p = 0.0004$), but the declaration difference does not follow the credibility of the stimuli. Perceived realism is only weakly related to the amount declared ($r = 0.10$). The Elevated notice named the least plausible amount yet was rated no less realistic than the Specific notice, and slightly more so (6.80 against 6.45; Welch $p = 0.18$). This reverses the ordering a credibility explanation requires. The item is in any case measured after the declaration and records a consequence of the manipulation rather than a characteristic that could confound it (Online Appendix E).

[Insert Figure 2 about here]

5.5 What Experiment 3 establishes

Experiment 3 supports H2 on the reporting margin. The effect of leaving the maximum fine ambiguous is not an artifact of the payment margin or hypothetical stakes. It appears, as a modest effect ($d = 0.21$), on an incentivized reporting decision in the tradition of Beck, Davis, and Jung. It does so despite the disclosed £0 lower bound, which limits the material worst case, and the incomplete recognition of the stated maximum in the Elevated condition. Together with Experiment 2, the same information manipulation moves both of the tax gap's principal behavioral margins in the same direction. The process evidence is taken up in Experiment 4.

6. Experiment 4: Physiological evidence on the mechanism

6.1 Motivation

Experiments 1 to 3 establish the effect of leaving enforcement consequences unspecified on both compliance margins. Experiment 4 examines the process that produces it, separating the affective explanation stated in H4 from alternatives based on visual attention or perceived severity. It returns to the payment margin and measures the process while the participant reads the message rather than inferring it from the subsequent decision.

6.2 Design, sample, and procedure

Two hundred and fifty students with tax-filing experience at an Australian university took part for course credit. Eleven were excluded, six for failing the instructional manipulation check and five for not meeting the pupillometry data-quality threshold of at least 85 percent of pupil samples during presentation of the letter. Their exclusion does not qualitatively affect the behavioral or the physiological results. This leaves 239 participants for analysis, of whom 236 have fixation measures. Each participant read a hypothetical tax-debt scenario with one of eight

enforcement messages. These are the Study A set (Online Appendix A), together with a Fine Minimum condition (\$100)⁹ added in a second wave of data collection. The statistics below were derived from the merged two-wave dataset, whose eight message conditions comprise 41, 41, 19, 32, 38, 15, 39 and 14 participants. Treatment labels are standardized across tables.

The dependent variable is the stated likelihood of paying. The elicitation item reads: “How likely are you to pay the money to the government? Please give your answer out of 10, where 0 means you are certain not to pay and 10 means you are certain to pay.” The eleven-point scale, its anchors and this wording are the published instrument reported in Hallsworth et al. (2017). Experiment 2, by contrast, asks instead about the outstanding tax obligation (Section 4.1).

Eye movements and pupil size were recorded continuously throughout the experiment with a Tobii TX300 eye-tracking system at 300 Hz. Pupillary response was computed from the samples recorded while the participant’s gaze was fixated on the letter from the tax authority, the area of interest, identical in size across treatments. The response is expressed as a percentage change relative to baseline, following Hayes and Petrov (2016) and Hecht et al. (2025). The baseline is the mean pupil diameter over the full 60-second baseline period, not over a subinterval of it, and the measure analyzed below is the baseline-corrected pupil percentage change. That period was recorded at the start of the session, before any treatment message, while the participant viewed a fixation cross. Visual attention was measured by fixation duration and fixation count within the same area of interest, the two measures predominant in accounting research (Lynch & Andiola, 2019).

⁹ The dollar amount reflects the Australian setting of the laboratory experiment; the studies run with UK samples state amounts in pounds sterling.

The pupillary light reflex is modulated by stimulus luminance and precedes the arousal-related component of the pupillary response (Bradley et al., 2008). Pupillometry therefore requires low and controlled ambient light. The experiment was conducted in dimly lit booths, with viewing distance held at approximately 65 cm. Illuminance measured at the participant's viewing position was 17.3 lux for every one of the eight message conditions, so the light reaching the eye did not differ across treatments.

6.3 Results

[Insert Table 5 about here]

Pupillary responses differed across the eight enforcement messages ($F(7, 231) = 3.65, p < 0.001$), while visual attention did not (fixation count $F(7, 228) = 1.01, p = 0.42$; fixation duration $F(7, 228) = 1.37, p = 0.22$; $n = 236$). This dissociation is the principal result of Experiment 4 and is consistent with H4: the messages differed in autonomic arousal and not in the visual attention they received. Rose et al. (2022) report the same dissociation for two data-visualization formats, and present the absence of a fixation difference as a substantive result rather than as a failed manipulation.

On compliance (Table 5, Panel B), the two fine disclosures moved the stated likelihood of paying in opposite directions. Relative to the control letter, the minimum-fine message lowered it ($b = -0.849, p = 0.034$) and the maximum-fine message raised it ($b = 0.848, p = 0.026$; $R^2 = 0.087, N = 239$). In the pupillary regression against the control baseline (Panel C), only the minimum-fine message differed significantly, and it lowered arousal ($b = -4.367, p = 0.001$). The two largest positive coefficients, Jailtime non-specific (1.993, $p = 0.213$) and Fine maximum (1.715, $p = 0.184$), are not significant. The message resolving the penalty to a small

specified amount therefore produced lower autonomic arousal and a lower stated likelihood of paying.

We next examine whether pupillary response accounts for the association between message and compliance. The model is a single-mediator model (PROCESS Model 4). The independent variable is message framing, the treatment identifier entered as a single continuous predictor taking the values 1 to 8 across the eight conditions. Pupillary response is the mediator, and the stated likelihood of paying is the dependent variable. No covariates are included. The interval reported is a 95 percent percentile bootstrap interval on 5,000 resamples. The paths are shown in Figure 4. The path from message framing to pupillary response is negative ($a = -0.349$, $p = 0.031$). The path from pupillary response to the stated likelihood of paying is positive ($b = 0.246$, $p < 0.001$). Figure 4 also reports the indirect effect of message framing through pupillary response, -0.086 . Its 95 percent bootstrap confidence interval, $[-0.174, -0.002]$, does not include zero. The direct effect is 0.040 ($p = 0.128$). Online Appendix F reports the full indirect-effect analyses.

[Insert Figure 3 about here]

[Insert Figure 4 about here]

6.4 What Experiment 4 establishes

Experiment 4 supports H4. The messages differed in the autonomic arousal they produced, not in the visual attention they received. Pupillary response accounts for part of the association with the stated likelihood of paying.

Experiment 4's compliance pattern differs from Experiment 2's in one respect. The maximum-fine message raised the stated likelihood of paying relative to the control letter (0.848 , $p = 0.026$). Experiment 2 found instead that revealing the maximum lowered it relative

to conditions in which it was left ambiguous. The two results are consistent, because the designs answer different questions. Experiment 4 exposes each participant to a single message, with no elicitation of prior beliefs and no manipulation of whether the maximum is revealed. Its minimum-fine and maximum-fine differences therefore concern message content against a control letter, not the removal of ambiguity against a condition in which ambiguity is preserved. Experiment 4 contributes the process evidence of Section 6.3 rather than a further test of H2.

7. Discussion and conclusion

Across the four experiments the evidence is consistent. In the field, the message leaving further action unspecified was the most effective treatment tested. Messages specifying a financial or an administrative cost did not reliably raise payment. Neither did a structurally parallel message conveying psychological relief rather than enforcement. That pattern supports H1 and locates the effect in the unspecified consequence rather than message salience. In the laboratory, revealing the maximum fine lowered the stated likelihood of paying relative to leaving it unstated, supporting H2. The two revealed maxima lowered it by similar amounts although they differ by a factor of more than five, supporting H3 on the hypothetical payment margin. The same contrast appeared on an incentivized income declaration with real money, a real undisclosed audit, and real penalties, extending H2 to the reporting margin. Under real stakes the two maxima were not interchangeable, so H3 is established on the hypothetical margin alone. We draw no equivalence conclusion from any pair of cells. Physiologically, pupillary arousal differed across the eight enforcement messages while visual attention did not, supporting H4. Sections 3 to 6 report the magnitudes.

These results separate the three explanations of Section 2.2. Under belief updating, behavior follows the direction of the disclosed news relative to the prior (Snow & Warren,

2007). The individual-level test contradicts that prediction. For participants whose elicited prior understated the maximum fine, disclosure was unfavorable news and updating predicts increased compliance. The stated likelihood of paying instead fell (Section 4.3). For those who had overstated the fine, the change ran in the same direction and was not significant. The two subgroups do not differ reliably, so the claim is confined to the one for which updating makes a directional prediction. A field precedent exists: the audit-statistics letter of Bergolo et al. (2023) raised firms' tax compliance while lowering their perceived audit probability. Following Sommer et al. (2024), we read neither finding as evidence against Bayesian updating as a computation. What the designs locate is that an unstated maximum is represented and weighted differently from a stated one. The attention explanation is likewise not supported. The messages did not differ in the visual attention received, and arousal and mental effort produce similar effects on pupil size (Mathôt, 2018). Separating those two readings therefore rests on the unchanged fixation measures recorded alongside the reliable pupillary differences (Bradley et al., 2008; Lynch & Andiola, 2019). Experiment 3's self-report items converge. Felt uncertainty about the penalty was higher in the ambiguous than in the revealed cells, although it does not predict declarations and no formal mediation is supported. Following Asay et al. (2022), we treat those items as convergent process evidence and Experiment 4's physiological measure as the mediator. The explanation consistent with all four experiments is ambiguity aversion with affective content. An unspecified consequence is weighted more heavily than a specified one, and the weighting registers in autonomic arousal rather than in attention or the direction of news.

The one departure from the predicted pattern sharpens the construct. Under real stakes, declarations under the revealed £10,000 maximum were nearer the ambiguous cells than the

revealed £1,900 cell (Section 5.4). The recognition check indicates why. Only a minority of that cell registered the stated maximum as a specific figure, against majorities in every other condition. Among participants who did register it, the planned contrast strengthens and the residual between-cells variance is no longer significant. Salmon and Shniderman (2019), the closest counter-evidence, found at best weak evidence of ambiguity aversion in a tax-framed experiment. Their ambiguity concerned the detection probability, drawn from a stated set of named values, so the worst case was visible to their participants. Read together, the results suggest that the operative feature is whether the worst case is bounded rather than whether a number is present. Their ambiguous conditions bounded it and produced little aversion. Our revealed £10,000 condition bounded it where the figure was registered, and our ambiguous conditions leave it unbounded, where compliance is highest. That reading is consistent with the explanation that variation in the parameters of an affectively charged outcome is weighted lightly (Slovic et al., 2004).

Section 1 established, on the authorities' own definitions, that non-payment of an acknowledged obligation and deliberate under-declaration are both compliance obligations. What the experiments add is empirical. A hypothetical payment measure and an incentivized declaration responded to the same information manipulation in the same direction, suggesting that the phenomenon attaches to the enforcement communication rather than to either decision's features.

For tax enforcement research, the principal implication concerns the antecedent taxpayer-uncertainty literature. In Beck and Jung (1989b) and Beck et al. (1992), taxpayer uncertainty affects reporting through the agency's equilibrium audit response, and the effect's direction is parameter-dependent. Experiments 2 and 3 close that pathway by design. Audit

information is constant and undisclosed, the applied penalty is drawn independently of the declaration, and perceived audit likelihood did not differ across conditions (Section 5.2). The effect nonetheless appears. Communicated penalty information therefore operates on compliance through a preference-based component in addition to the strategic component those models identify. For the ambiguity literature, the results document ambiguity aversion in a consequential, institutionally embedded decision, in the loss domain, and under real monetary incentives. They do so in between-subjects designs where each participant evaluates a single information condition in isolation. Under the comparative ignorance hypothesis of Fox and Tversky (1995), ambiguity aversion is pronounced when a clear and a vague prospect are evaluated jointly and diminishes when either is evaluated alone, the evaluation our designs impose. If that hypothesis extends to enforcement information, our designs did not favor the effect, and joint evaluation of clear and ambiguous regimes could produce larger estimates.

For enforcement practice, the results identify a behavioral cost of penalty transparency but not an optimal disclosure policy. As stated in Section 1, the design compares stating a maximum against not stating one and does not evaluate combinations of disclosure and severity. The results indicate something narrower: transparency initiatives that would state worst-case penalties as specific amounts may reduce payment among populations resembling those studied, because the specification removes a worst case taxpayers otherwise supply for themselves. Whether an authority should act on this depends on considerations our design does not measure, and the analytical literature makes any prescription conditional. The value of withholding enforcement information depends on the objective pursued and the ambiguity attitudes of the regulated population (Calford & DeAngelo, 2023). Reinganum and Wilde (1988, p. 798) locate an interior optimum where refusing to disclose explicit criteria may be

appropriate while extensive secrecy is not. Fostered ambiguity also has welfare costs: it reduces the welfare of ambiguity-averse taxpayers, and the authority can neither categorize nor screen taxpayers by ambiguity preference, so the instrument cannot be targeted (Snow & Warren, 2005). Experiment 2's dispersion result is the empirical counterpart. Revealing the fine lowered the average response and widened its distribution, and dispersion in perceived sanctions produces over-deterrence of some and under-deterrence of others (Bebchuk & Kaplow, 1992). Field effects are further conditional on the population addressed. Deterrence messages did not improve compliance among firm owners or the highest earners in Yang et al. (2024), and no message combination improved filing on average in Abdelnabi et al. (2024). The legitimacy and procedural-justice consequences of withholding penalty information the authority possesses are not modeled here.

Four limitations bound these claims. First, the laboratory payment measure is a stated likelihood of paying rather than an observed payment, as stated at the outset. Neither Experiment 2 nor Experiment 4 provides a measure of taxpayers' actions on the payment margin. In both, the dependent variable is a likelihood judgment made within a hypothetical scenario rather than a payment decision with real consequences. The behavioral evidence on the payment margin comes from Experiment 1, where the outcome is whether the taxpayer paid within the administrative window. The incentivized behavioral evidence comes from Experiment 3, where the declaration determined real earnings. The two laboratory instruments word the payment item differently, and both wordings are reproduced in Online Appendix G. Second, the results speak to a specific parameter regime. Penalty information is communicated while audit information is constant and undisclosed, the audit probability is 25 percent where one is set, penalties are bounded, and every decision is single-shot. The antecedent literature

establishes that taxpayer-uncertainty effects are parameter-dependent (Beck & Jung, 1989a, 1989b). Ambiguity has been found to raise non-compliance where detection is close to certain (Loughran et al., 2011; Midgette et al., 2021). In repeated interaction taxpayers update audit beliefs from experience, which raises evasion over time (Snow & Warren, 2007), a dynamic our single-shot designs cannot address. Third, the finding that the revealed magnitude does not alter the reduction holds on the hypothetical payment margin only. Under real stakes the two maxima differed (Section 5.4), so H3 is not established for incentivized reporting. Fourth, the samples are online panel adults, students with tax-filing experience, and one national administrative population. None of the designs includes repeated play or reputation effects.

Tax authorities communicate enforcement while seldom stating what non-compliance will cost, and Sections 1 and 2 record that the practice is deliberate. The experiments indicate that the practice has behavioral consequences that the models underlying enforcement policy do not address. In the field, a letter warning of unspecified further action raised payment where otherwise similar letters stating specific costs did not. In the laboratory, revealing the maximum fine lowered both the stated likelihood of paying and incentivized income declarations, and the reduction did not follow the direction of the disclosed news. The process evidence locates the response in autonomic arousal rather than visual attention. Together, the results suggest that what an authority chooses to disclose about penalties is an enforcement instrument in its own right, and that within the parameter regime studied, stating the worst case reduced compliance rather than strengthening it. An authority weighing penalty transparency against its collection objectives faces a trade-off to which these experiments give empirical content on both margins of the tax gap.

References

- Abdelnabi, M., Afraz, N., Ahmed, T., Ayub, A., Makki, F., Said, F., Schiektekat, P., & Vlaev, I. (2024). Nudging tax filing through text message combinations: Evidence from Pakistan. *Behavioural Public Policy*, 10(3), 670-691.
- Agranov, M., & Buyalskaya, A. (2022). Deterrence effects of enforcement schemes: An experimental study. *Management Science*, 68(5), 3573-3589.
- Allingham, M. G., & Sandmo, A. (1972). Income tax evasion: A theoretical analysis. *Journal of Public Economics*, 1(3-4), 323-338.
- Alm, J., & Malézieux, A. (2021). 40 years of tax evasion games: A meta-analysis. *Experimental Economics*, 24(3), 699-750.
- Alós-Ferrer, C., Jaudas, A., & Ritschel, A. (2021). Effortful Bayesian updating: A pupil-dilation study. *Journal of Risk and Uncertainty*, 63, 81-102.
- Arnold, M. C., Artz, M., & Tafkov, I. D. (2024). The effect of target transparency on managers' target setting decisions. *Accounting, Organizations and Society*, 112, Article 101545.
- Asay, H. S., Guggenmos, R. D., Kadous, K., Koonce, L., & Libby, R. (2022). Theory testing and process evidence in accounting experiments. *The Accounting Review*, 97(6), 23-43.
- Bebchuk, L. A., & Kaplow, L. (1992). Optimal sanctions when individuals are imperfectly informed about the probability of apprehension. *The Journal of Legal Studies*, 21(2), 365-370.
- Beck, P. J., & Jung, W.-O. (1989a). Taxpayer compliance under uncertainty. *Journal of Accounting and Public Policy*, 8(1), 1-27.
- Beck, P. J., & Jung, W.-O. (1989b). Taxpayers' reporting decisions and auditing under information asymmetry. *The Accounting Review*, 64(3), 468-487.
- Beck, P. J., Davis, J. S., & Jung, W.-O. (1991). Experimental evidence on taxpayer reporting under uncertainty. *The Accounting Review*, 66(3), 535-558.
- Beck, P. J., Davis, J. S., & Jung, W.-O. (1992). Experimental evidence on an economic model of taxpayer reporting under strategic and non-strategic audits. *Contemporary Accounting Research*, 9(1), 86-112.
- Becker, G. S. (1968). Crime and punishment: An economic approach. In *The economic dimensions of crime* (pp. 13-68). Palgrave Macmillan.
- Bergolo, M., Ceni, R., Cruces, G., Giacobasso, M., & Perez-Truglia, R. (2023). Tax audits as scarecrows: Evidence from a large-scale field experiment. *American Economic Journal: Economic Policy*, 15(1), 110-153.
- Bott, K. M., Cappelen, A. W., Sørensen, E. Ø., & Tungodden, B. (2020). You've got mail: A randomized field experiment on tax evasion. *Management Science*, 66(7), 2801-2819.
- Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4), 602-607.

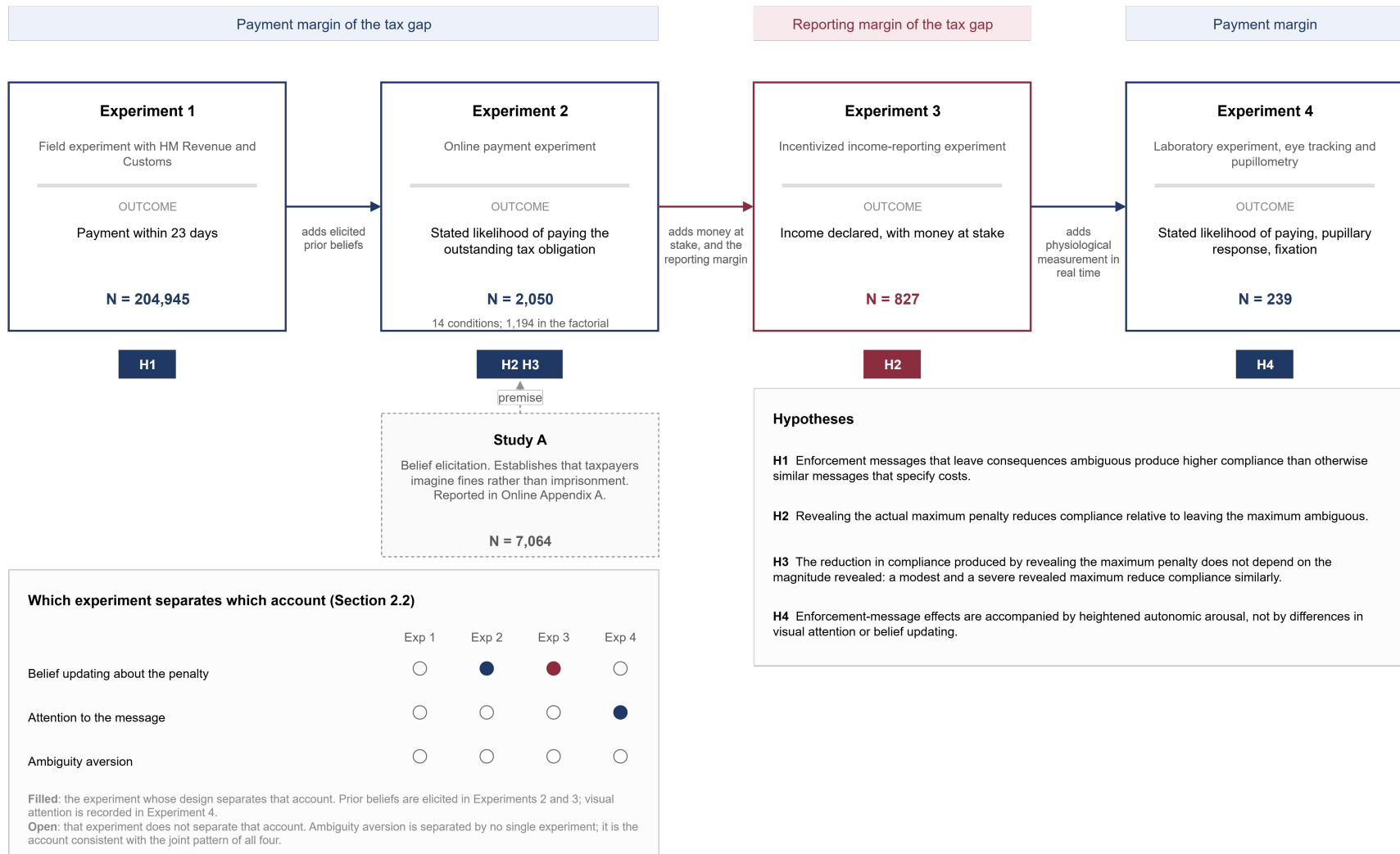
- Brandon, A., Ferraro, P. J., List, J. A., Metcalfe, R. D., Price, M. K., & Rundhammer, F. (2026). Do the effects of nudges persist? Theory and evidence from 38 natural field experiments. *Review of Economic Studies*. Advance online publication.
- Bursztyn, L., Fiorin, S., Gottlieb, D., & Kanz, M. (2019). Moral incentives in credit card debt repayment: Evidence from a field experiment. *Journal of Political Economy*, 127(4), 1641-1683.
- Calford, E. M., & DeAngelo, G. (2023). Ambiguity and enforcement. *Experimental Economics*, 26(2), 304-338.
- Camerer, C., & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, 5(4), 325-370.
- Castro, L., & Scartascini, C. (2015). Tax compliance and enforcement in the pampas evidence from a field experiment. *Journal of Economic Behavior & Organization*, 116, 65-82.
- Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 75(4), 643-669.
- Feess, E., Schildberg-Hörisch, H., Schramm, M., & Wohlschlegel, A. (2018). The impact of fine size and uncertainty on punishment and deterrence: Theory and evidence from the laboratory. *Journal of Economic Behavior and Organization*, 149, 58-73.
- Fox, C. R., & Tversky, A. (1995). Ambiguity aversion and comparative ignorance. *The Quarterly Journal of Economics*, 110(3), 585-603.
- Friesen, L. (2012). Certainty of punishment versus severity of punishment: An experimental investigation. *Southern Economic Journal*, 79(2), 399-421.
- Gathergood, J., Mahoney, N., Stewart, N., & Weber, J. (2019). How do individuals repay their debt? The balance-matching heuristic. *American Economic Review*, 109(3), 844-875.
- Gil, P., Holz, J., List, J. A., Simon, A., & Zentner, A. (2024). Toward an understanding of tax amnesty take-up: Evidence from a natural field experiment. *Journal of Public Economics*, 239, 105245.
- Gross, T., Notowidigdo, M. J., & Wang, J. (2014). Liquidity constraints and consumer bankruptcy: Evidence from tax rebates. *The Review of Economics and Statistics*, 96(3), 431-443.
- Guggenmos, R. D., Piercey, M. D., & Agoglia, C. P. (2018). Custom contrast testing: Current trends and a new approach. *The Accounting Review*, 93(5), 223-244.
- Hallsworth, M. (2014). The use of field experiments to increase tax compliance. *Oxford Review of Economic Policy*, 30(4), 658-679.
- Hallsworth, M., List, J. A., Metcalfe, R. D., & Vlaev, I. (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics*, 148, 14-31.
- Hayes, T. R., & Petrov, A. A. (2016). Pupil diameter tracks the exploration-exploitation trade-off during analogical reasoning and explains individual differences in fluid intelligence. *Journal of Cognitive Neuroscience*, 28(2), 308-318.

- Hecht, G., Rotaru, K., Schulz, A. K. D., Towry, K. L., & Webb, A. (2025). Decoding effort: Toward a measure — and a new understanding — of effort intensity in accounting research. *Management Accounting Research*, 66, 100926.
- Hirsh, J. B., Mar, R. A., & Peterson, J. B. (2012). Psychological entropy: a framework for understanding uncertainty-related anxiety. *Psychological Review*, 119(2), 304.
- HM Revenue and Customs. (2021). *Timely payment: Summary of responses*. HM Revenue and Customs.
- HM Revenue and Customs. (2026). *Measuring tax gaps 2026 edition: Tax gap estimates for 2024 to 2025*. <https://www.gov.uk/government/statistics/measuring-tax-gaps>
- Holz, J. E., List, J. A., Zentner, A., Cardoza, M., & Zentner, J. E. (2023). The \$100 million nudge: Increasing tax compliance of firms using a natural field experiment. *Journal of Public Economics*, 218, 104779.
- Hurwitz, H., & McGowan, J. R. (2024). The efficiency of Interactive Voice Response systems and tax compliance: a field experiment by the California Franchise Tax Board. *Journal of the American Taxation Association*, 46(2), 63-80.
- Hyde, J., Ryan, K. M., & Waters, A. M. (2019). Psychophysiological markers of fear and anxiety. *Current Psychiatry Reports*, 21, Article 56.
- Internal Revenue Service. (2024). *The tax gap: Tax year 2022 projections*. <https://www.irs.gov/statistics/irs-the-tax-gap>
- Keys, B. J., & Wang, J. (2019). Minimum payments and debt paydown in consumer credit cards. *Journal of Financial Economics*, 131(3), 528-548.
- Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, 21(2), 153-174.
- Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, 127(2), 267-286.
- Loughran, T. A., Paternoster, R., Piquero, A. R., & Pogarsky, G. (2011). On ambiguity in perceptions of risk: Implications for criminal decision making and deterrence. *Criminology*, 49(4), 1029-1061.
- Lynch, E. J., & Andiola, L. M. (2019). If eyes are the window to our soul, what role does eye-tracking play in accounting research? *Behavioral Research in Accounting*, 31(2), 107-133.
- Mathôt, S. (2018). Pupillometry: Psychology, physiology, and function. *Journal of Cognition*, 1(1), Article 16.
- Mertens, S., Herberz, M., Hahnel, U. J., & Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences*, 119(1), Article e2107346118.
- Midgette, G., Loughran, T. A., & Tahamont, S. (2021). The impact of ambiguity-induced error in offender decision-making: Evidence from the field. *Journal of Research in Crime and Delinquency*, 58(6), 635-665.

- Nagin, D. S. (2013). Deterrence: A review of the evidence by a criminologist for economists. *Annual Review of Economics*, 5, 83-105.
- Nagin, D. S. (2017). Deterrence. In E. Luna (Ed.), *Reforming criminal justice, Vol. 4: Punishment, incarceration, and release* (pp. 19-36). Academy for Justice.
- OECD. (2014). *Measures of tax compliance outcomes: A practical guide*. OECD Publishing.
- Perez-Truglia, R., & Troiano, U. (2018). Shaming tax delinquents. *Journal of Public Economics*, 167, 120-137.
- Reinganum, J. F., & Wilde, L. L. (1988). A note on enforcement uncertainty and taxpayer compliance. *The Quarterly Journal of Economics*, 103(4), 793-798.
- Robitaille, N., House, J., & Mazar, N. (2021). Effectiveness of Planning Prompts on Organizations' Likelihood to File Their Overdue Taxes: A Multi-Wave Field Experiment. *Management Science*, 67(7), 4327-4340.
- Rose, A. M., Rose, J. M., Rotaru, K., Sanderson, K. A., & Thibodeau, J. C. (2022). Effects of uncertainty visualization on attention, arousal, and judgment. *Behavioral Research in Accounting*, 34(1), 113-139.
- Rotaru, K., Rose, J. M., Ganbold, O., Rose, A., Loveridge, J., & Stamatelatos, A. (2025). Neuroaccounting: Integrating Neurophysiological Techniques into Accounting and AIS Research. *Journal of Information Systems*. Advance online publication.
- Salmon, T. C., & Shniderman, A. (2019). Ambiguity in criminal punishment. *Journal of Economic Behavior & Organization*, 163, 361-376.
- Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2004). Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. *Risk Analysis*, 24(2), 311-322.
- Snow, A., & Warren, R. S., Jr. (2005). Ambiguity about audit probability, tax compliance, and taxpayer welfare. *Economic Inquiry*, 43(4), 865-871.
- Snow, A., & Warren, R. S., Jr. (2007). Audit uncertainty, Bayesian updating, and tax evasion. *Public Finance Review*, 35(5), 555-571.
- Sommer, J., Musolino, J., & Hemmer, P. (2024). Updating, evidence evaluation, and operator availability: A theoretical framework for understanding belief. *Psychological Review*, 131(2), 373-401.
- Tan, F., & Yim, A. (2014). Can strategic uncertainty help deter tax evasion? An experiment on auditing rules. *Journal of Economic Psychology*, 40, 161-174.
- Vossler, C. A., & Gilpatric, S. M. (2018). Endogenous audits, uncertainty, and taxpayer assistance services: Theory and experiments. *Journal of Public Economics*, 165, 217-229.
- Wang, H.-J., & White, M. J. (2000). An optimal personal bankruptcy procedure and proposed reforms. *The Journal of Legal Studies*, 29(1), 255-286.

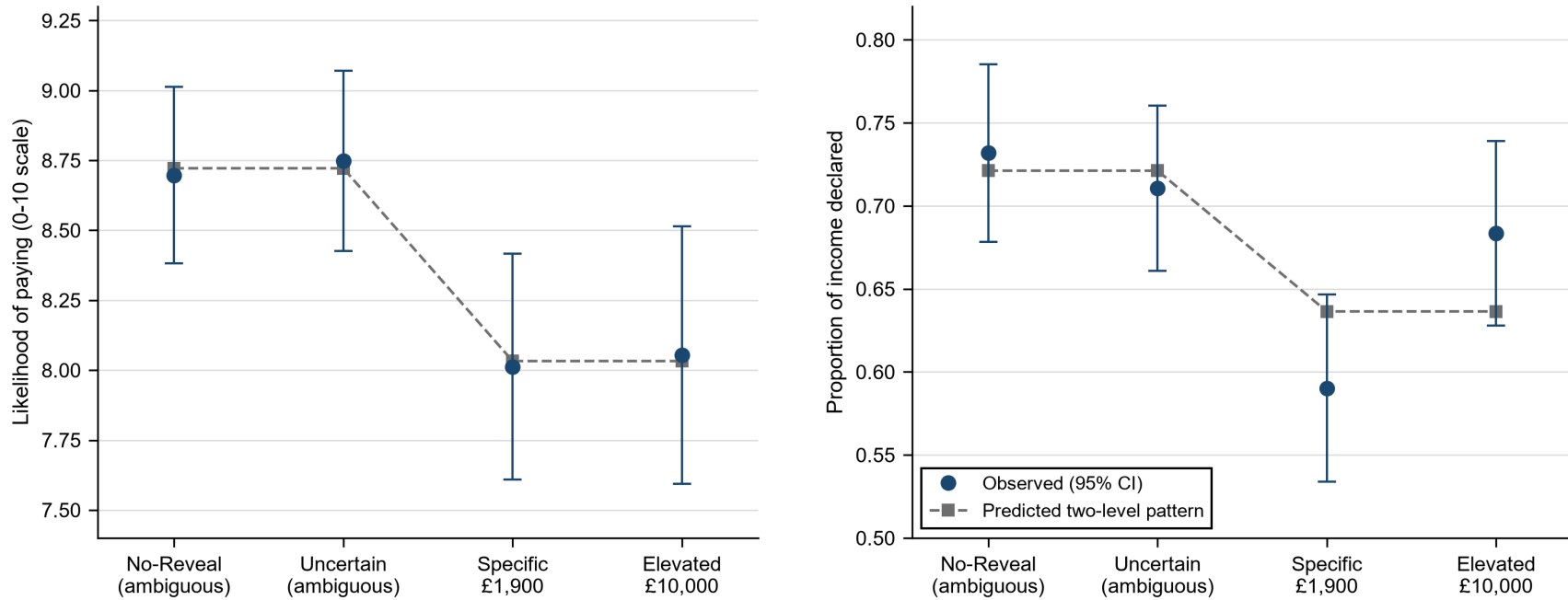
- Wang, J. T. Y., Spezio, M., & Camerer, C. F. (2010). Pinocchio's pupil: using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. *American Economic Review*, 100(3), 984-1007.
- Widianto, R., Rieger, M., Adrison, V., & Murshed, M. (2024). Nudging taxpayers via instant messaging to submit their annual tax return: Evidence from a field experiment in Indonesia. *Bulletin of Indonesian Economic Studies*. Advance online publication.
- Yang, W., Zhao, J., & Zhou, H. (2024). Information nudges and tax compliance: Evidence from a field experiment in China. *Journal of Economic Behavior & Organization*, 228, 106779.
- Yogama, E. A., Gray, D. J., & Rablen, M. D. (2024). Nudging for prompt tax penalty payment: evidence from a field experiment in Indonesia. *Journal of Economic Behavior & Organization*, 224, 548-579.
- Yoon, S.-S., Yoo, S.-W., & Kim, J. (2011). Ambiguity, audit errors, and tax compliance. *Asia-Pacific Journal of Accounting & Economics*, 18(1), 11-26.

Fig. 1. The four experiments, their compliance margins, and the hypothesis each carries.



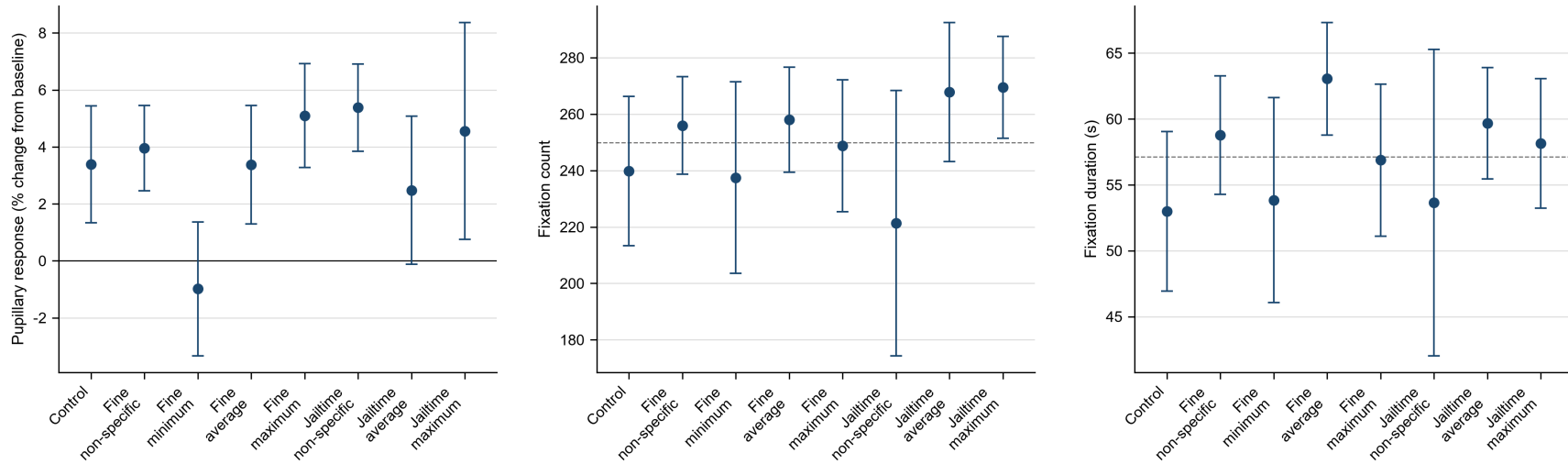
Notes. The four experiments are shown left to right in the order in which they are reported, with the setting, the outcome measure and the analysis sample for each. The band above each experiment marks the behavioral margin it addresses, payment or reporting. The label below each experiment gives the hypothesis it carries. Study A, the belief-elicitation experiment reported in Online Appendix A, sits outside the sequence as a dashed box, because it supplies a premise for the experiments that follow rather than testing a hypothesis of its own. The lower left panel shows which experiment separates which of the three explanations set out in Section 2.2; the lower right panel states the four hypotheses in full.

Fig. 2. The same ambiguity contrast on two compliance margins: Experiment 2 (payment) and Experiment 3 (reporting).



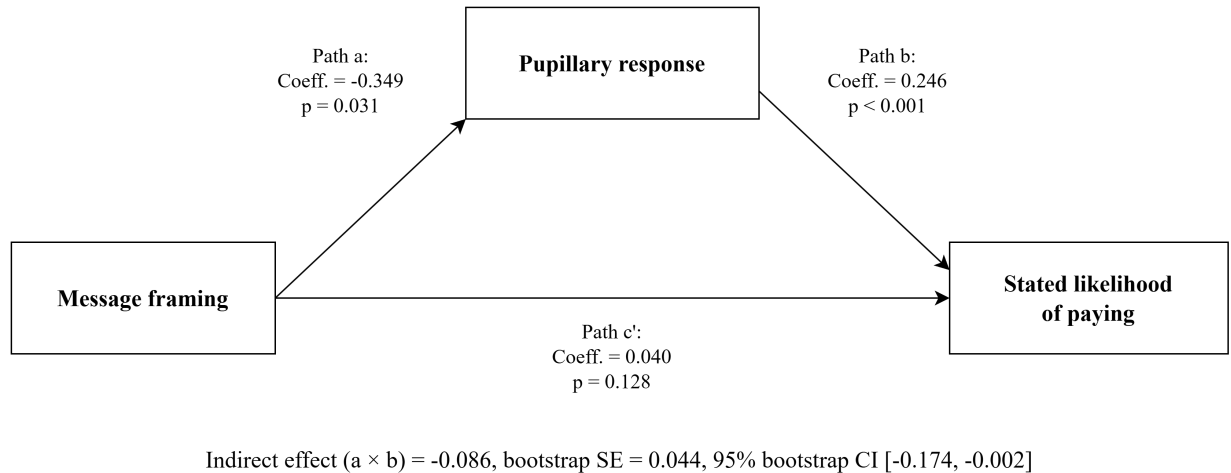
Notes. Panel (a) gives the Experiment 2 condition means for the stated likelihood of paying on the eleven-point scale, with 95 percent confidence intervals: No-Reveal 8.697, Uncertain 8.748, Specific (£1,900) 8.013 and Elevated (£10,000) 8.054, with 155, 147, 157 and 149 participants respectively. Panel (b) gives the Experiment 3 mean declarations as a proportion of the endowment, with 95 percent confidence intervals: No-Reveal 0.732, Uncertain 0.711, Specific 0.590 and Elevated 0.684, with 206, 207, 209 and 205 participants respectively. Each panel plots the observed means against the two-level pattern the ambiguity contrast predicts. The contrast leaves 0.5 percent of the between-cells variance unexplained on the hypothetical payment margin and 39.6 percent on the incentivized reporting margin. The vertical axis of each panel is truncated so that the confidence intervals are legible; the full range is 0 to 10 in panel (a) and 0 to 1 in panel (b).

Fig. 3. Experiment 4: enforcement messages differ in pupillary response but not in visual attention.



Notes. Each panel gives the condition means for the eight enforcement messages with 95 percent confidence intervals, on its own natural scale. Pupillary response is the baseline-corrected pupil percentage change and differs across the messages, $F(7, 231) = 3.65$, $p < 0.001$. Neither attention measure differs: fixation count, $F(7, 228) = 1.01$, $p = 0.42$; fixation duration, $F(7, 228) = 1.37$, $p = 0.22$. Fixation measures are available for 236 of the 239 participants. Fixation duration is plotted in seconds.

Fig. 4. Experiment 4: the indirect effect of the enforcement message on the stated likelihood of paying through pupillary response.



Notes. The model is PROCESS Model 4. The message-framing variable is the treatment identifier entered as a single continuous predictor taking the values 1 to 8 across the eight message conditions; pupillary response, the baseline-corrected pupil percentage change, is the mediator; and the stated likelihood of paying is the dependent variable. No covariates are included. The interval on the indirect effect is a 95 percent percentile bootstrap confidence interval computed on 5,000 resamples; an interval that excludes zero indicates a significant indirect effect. N = 239.

Table 1

Experiment 1: payment within 23 days of receiving the letter.

	(I) Baseline	(II) Covariate-adjusted
Interest Today	0.000 (0.005)	0.001 (0.005)
Interest Tomorrow	0.008 (0.005)	0.008 (0.005)
Less Effort	-0.003 (0.005)	-0.010 (0.005)
More Effort	0.002 (0.005)	0.003 (0.005)
Costs Further Action	0.044*** (0.005)	0.046*** (0.006)
Costs Peace	-0.008 (0.005)	-0.007 (0.005)
Interest & Effort	0.006 (0.005)	0.009 (0.005)
Interest & Action	0.017*** (0.005)	0.019*** (0.005)
Contact Now	0.022*** (0.005)	0.024*** (0.005)
Contact Now Information	0.008 (0.005)	0.010 (0.005)
Contact Later Information	0.029*** (0.005)	0.031*** (0.005)
Age		0.004*** (0.000)
Male		-0.032*** (0.003)
Debt size		-4.59e-06*** (0.000)
Self-employed		0.121*** (0.004)
Accountant		-0.004 (0.003)
Experienced		-0.048*** (0.002)
Constant	0.421*** (0.016)	0.420*** (0.031)
Pseudo R ²	0.00	0.01
N	204,936	204,367

Notes. Marginal effects from a logistic regression; the dependent variable is an indicator for making a payment within 23 days of receiving the letter. The omitted category is the control letter, whose payment rate is 42.1 percent. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Specifications by recent-debt subgroup, the 48-day and 70-day windows, and the balance table are reported in Online Appendix C. Of the 204,945 taxpayers randomized, specification (I) estimates on 204,936 and specification (II), which adds covariates, on 204,367.

Table 2

Experiment 2: likelihood of paying the outstanding tax obligation, by cell.*Panel A: Descriptive statistics*

Debt	Fine type	Information	Mean (SD)	n
£500	Minimum	No-Reveal	8.80 (2.00)	147
£500	Minimum	Revealed	8.63 (2.45)	139
£500	Maximum	No-Reveal	8.76 (1.93)	138
£500	Maximum	Revealed	8.28 (2.80)	149
£2,500	Minimum	No-Reveal	8.55 (2.01)	154
£2,500	Minimum	Revealed	8.21 (2.52)	155
£2,500	Maximum	No-Reveal	8.70 (1.99)	155
£2,500	Maximum	Revealed	8.01 (2.56)	157
Marginal: debt	£500		8.61 (2.33)	573
	£2,500		8.37 (2.30)	621
Marginal: fine type	Minimum		8.54 (2.26)	595
	Maximum		8.43 (2.37)	599
Marginal: information	No-Reveal		8.70 (1.98)	594
	Revealed		8.27 (2.59)	600
Overall			8.48 (2.32)	1,194

Panel B: Three-way analysis of variance (Type II)

Source	SS	df	F	p
Debt	18.32	1	3.44	0.064
Fine type	3.58	1	0.67	0.413
Information	53.19	1	10.00	0.002
Debt x Fine type	2.12	1	0.40	0.528
Debt x Information	2.61	1	0.49	0.484
Fine type x Information	8.34	1	1.57	0.211
Debt x Fine type x Information	0.01	1	0.00	0.967
Residual	6311.77	1186		

Notes. The dependent variable is the participant's stated likelihood of paying the outstanding tax obligation on an eleven-point scale (0 = certain not to pay, 10 = certain to pay). Standard deviations in parentheses. "No-Reveal" denotes conditions in which the prior belief was elicited and no fine was disclosed; "Revealed" denotes conditions in which the prior belief was elicited and the actual fine was then disclosed. Covariate-adjusted regressions (age, gender, household income, risk propensity, leadership and management experience, debt size, and elicited fine belief), estimated on the 1,180 participants with complete covariate data, are reported in Online Appendix D; inferences are unchanged.

Table 3

Experiment 2: the four maximum-fine conditions (£2,500 debt).*Panel A: Cell means*

Condition	Maximum fine information	Mean (SD)	n
No-Reveal	mentioned, never resolved	8.697 (1.988)	155
Uncertain	set case-by-case, may increase	8.748 (1.979)	147
Specific	£1,900	8.013 (2.557)	157
Elevated	£10,000	8.054 (2.842)	149
Pooled ambiguity-preserved		8.722 (1.980)	302
Pooled resolved		8.033 (2.695)	306

Panel B: Contrast decomposition

Source of variance	SS	df	F	p
Ambiguity contrast {1, 1, -1, -1}	72.17	1	12.84	0.0002
Residual between-cells variance	0.35	2	0.03	0.970
Total between-cells variance	72.52	3		
Error	3395.98	604		
Contrast variance residual, q²				0.5%

Notes. Standard deviations in parentheses. The contrast is evaluated following Guggenmos et al. (2018). The reported p value for the contrast is one-tailed, consistent with the directional prediction in H2; the residual between-cells test is two-tailed. q² is the proportion of systematic between-cells variance left unexplained by the contrast. The Specific and Elevated conditions differ by 0.041 scale points, 95 percent confidence interval [-0.568, 0.650]; no equivalence is claimed.

Table 4

Experiment 3: income declared, by condition.*Panel A: Descriptive statistics*

Condition	Maximum fine information	Mean (SD)	n	Full declaration (%)
No-Reveal	mentioned, never resolved	0.732 (0.389)	206	62.6
Uncertain	set case-by-case, may increase	0.711 (0.364)	207	51.2
Specific	£1,900	0.590 (0.414)	209	43.1
Elevated	£10,000	0.684 (0.403)	205	56.6
Pooled ambiguity-preserved		0.721 (0.376)	413	
Pooled resolved		0.636 (0.411)	414	
Overall		0.678	827	53.3

Panel B: One-way analysis of variance

Source	df	F	p
Condition	3, 823	5.25	0.001

Panel C: Contrast decomposition

Source of variance	SS	df	F	p
Ambiguity contrast {1, 1, -1, -1}	1.469	1	9.52	0.001
Residual between-cells variance	0.962	2	3.12	0.045
Total between-cells variance	2.431	3		
Contrast variance residual, q^2				39.6%

Notes. The dependent variable is the number of tokens declared divided by 10,000, so that it runs from 0 to 1 and higher values indicate greater compliance; standard deviations in parentheses. The primary sample is 827 participants of the 887 who completed the task, after the attention and instructional checks. The contrast p value is one-tailed, consistent with the directional prediction in H2; the residual between-cells test is two-tailed. The effect size for the ambiguity contrast is $d = 0.21$ against the pooled standard deviation and 0.22 against the root mean square error. The Specific and Elevated conditions are compared with a confidence interval in Section 5.3 and no equivalence is claimed. Robustness to the exclusion rule, functional form and covariates is reported in Online Appendix E.

Table 5

Experiment 4: compliance, pupillary response and visual attention, by condition.*Panel A: Descriptive statistics*

Condition	n	Pupillary response (%)	Fixation count	Fixation duration (s)
Control (costs further action)	41	3.383 (6.495)	239.80 (83.64)	52.99 (19.17)
Fine (non-specific)	41	3.955 (4.729)	255.98 (54.75)	58.77 (14.27)
Jailtime (non-specific)	19	5.376 (3.180)	221.32 (97.50)	53.65 (24.13)
Fine (minimum)	32	-0.984 (6.525)	237.50 (94.00)	53.84 (21.58)
Fine (average)	38	3.379 (6.329)	258.03 (56.48)	63.05 (12.96)
Jailtime (average)	15	2.476 (4.695)	267.80 (44.41)	59.67 (7.62)
Fine (maximum)	39	5.099 (5.621)	248.74 (72.02)	56.88 (17.82)
Jailtime (maximum)	14	4.558 (6.584)	269.45 (26.93)	58.14 (7.32)

Omnibus F-tests: pupillary response $F(7, 231) = 3.65$, $p < 0.001$, $N = 239$; fixation count $F(7, 228) = 1.01$, $p = 0.42$, $n = 236$; fixation duration $F(7, 228) = 1.37$, $p = 0.22$, $n = 236$.

Panel B: Stated likelihood of paying, regressed on condition

Condition	Coefficient (SE)	t	p
Fine (non-specific)	0.073 (0.373)	0.20	0.845
Jailtime (non-specific)	0.569 (0.469)	1.21	0.226
Fine (minimum)	-0.849 (0.398)	-2.13	0.034
Fine (average)	0.253 (0.380)	0.67	0.507
Jailtime (average)	0.730 (0.509)	1.43	0.153
Fine (maximum)	0.848 (0.378)	2.25	0.026
Jailtime (maximum)	-0.179 (0.523)	-0.34	0.732

Intercept 8.537; $R^2 = 0.087$; $N = 239$.

Panel C: Pupillary response, regressed on condition

Condition	Coefficient (SE)	t	p
Fine (non-specific)	0.571 (1.270)	0.45	0.653
Jailtime (non-specific)	1.993 (1.596)	1.25	0.213
Fine (minimum)	-4.367 (1.356)	-3.22	0.001
Fine (average)	-0.004 (1.295)	-0.00	0.997
Jailtime (average)	-0.908 (1.735)	-0.52	0.601
Fine (maximum)	1.715 (1.286)	1.33	0.184
Jailtime (maximum)	1.174 (1.780)	0.66	0.510

Intercept 3.383; $R^2 = 0.100$; $N = 239$.

Notes. The dependent variable in Panel B is the participant's stated likelihood of paying on an eleven-point scale (0 = certain not to pay, 10 = certain to pay). Standard deviations in Panel A and standard errors in Panels B and C are in parentheses. The pupillary measure is the baseline-corrected pupil percentage change, relative to the mean pupil diameter over the full 60-second baseline period, which was recorded while the participant viewed a fixation cross, before any treatment message. Fixation duration is reported in seconds; fixation measures are available for 236 of the 239 participants. The omitted category in Panels B and C is the control letter. Background characteristics, covariate models and the indirect-effect analyses are reported in Online Appendix F.

Online Appendix

How resolving enforcement ambiguity reduces tax compliance

Contents

Part	Content
Online Appendix A	Study A: the belief-elicitation experiment, in full
Online Appendix B	Formal model, reframed as illustrative
Online Appendix C	Experiment 1 supplements
Online Appendix D	Experiment 2 supplements
Online Appendix E	Experiment 3 supplements
Online Appendix F	Experiment 4 supplements
Online Appendix G	All experimental instruments

Online Appendix A: Study A, the Belief-Elicitation Experiment

Study A was reported as Experiment 2 in the previous submission. It is relocated here in full because three of its findings are premises the main text relies on, while its own results do not bear directly on the hypotheses tested in Experiments 1 to 4.

A.1 Design and sample

7,064 UK taxpayers were recruited via Prolific in 2021 (£1 compensation; UK residents aged 21-65 with experience of entrepreneurship or self-employment; 60.0% female, 38.7% male, 0.7% non-binary, 0.7% undisclosed; mean age 36.8). Each participant was randomly assigned to one of seven enforcement messages and then stated the likelihood of paying the money owed to the government, in the item reproduced at Appendix G.1, on an 11-point scale (0 = certain not to pay, 10 = certain to pay); Experiment 2 later used the same scale with the object of payment modified to the outstanding tax obligation (§4.1). Perceived punishment severity and perceived realism were also measured on 11-point scales.

The design is a **message-exposure** design: each participant saw one message, no prior beliefs were elicited, and nothing was resolved against a stated baseline. This is the structural feature that governs how Study A relates to Experiment 2, and it is set out at Appendix D.3.

Condition	Message added to the standard letter	n	Likelihood of paying	Severity	Realism
Costs Further Action (<i>baseline</i>)	If you pay this debt now, you can stop us taking any further action.	1,010	8.53 (1.93)	6.52 (1.92)	6.88 (2.19)
Fine, non-specific	...you may be subject to a fine.	1,005	8.59 (1.93)	6.41 (1.90)	7.60 (2.11)
Jailtime, non-specific	...you may be subject to prosecution that results in time in prison.	1,010	8.73 (1.77)	6.99 (1.95)	6.37 (2.44)
Fine, average	...we could fine you 5% of your current outstanding tax obligation.	1,008	8.51 (1.91)	6.10 (2.25)	7.02 (2.33)
Jailtime, average	...you could face 6 months in prison.	1,014	8.77 (1.80)	6.91 (2.16)	5.83 (2.69)
Fine, maximum	...you could face a fine of up to 100% of the outstanding obligation and the possibility that the government seizes your personal assets.	1,007	8.75 (1.71)	7.72 (1.76)	7.00 (2.38)

Condition	Message added to the standard letter	n	Likelihood of paying	Severity	Realism
Jailtime, maximum	...you could face up to 7 years in prison.	1,010	8.66 (1.78)	6.89 (2.29)	5.60 (2.76)

Standard deviations in parentheses. Groups are balanced on age, gender and risk propensity.

A.2 Treatment effects

Ordinary least squares on the stated likelihood of paying, with Costs Further Action as the omitted baseline.

Condition	(I)	(II) + demographics	(III) + severity, realism	(IV) full
Fine, non-specific	0.058	0.055	-0.063	-0.059
Jailtime, non-specific	0.203*	0.204*	0.208**	0.209**
Fine, average	-0.019	-0.021	0.051	0.036
Jailtime, average	0.237**	0.251**	0.364***	0.372***
Fine, maximum	0.220**	0.223**	-0.050	-0.042
Jailtime, maximum	0.130	0.132	0.310***	0.304***
Punishment severity			0.214***	0.202***
Realism			0.198***	0.192***
Constant	8.530***	8.991***	5.777***	6.197***
R ²	0.003	0.047	0.177	0.204
N	7,064	7,017	7,020	7,017

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in the source tables.

Three points matter for the main text.

Absolute effects are small. The whole range of condition means is 8.51 to 8.77 on an 11-point scale, and treatment alone explains 0.3% of the variance. Study A establishes what taxpayers *infer* from enforcement language rather than a large behavioral effect, which is part of why the paper does not rest on it.

Severity alone did not raise compliance. The most extreme threat, up to seven years' imprisonment, did not significantly raise the stated likelihood of paying (0.130, $p > 0.05$), while the message stating a maximum *fine* did (0.220, $p < 0.01$). The realism ratings show why this is not a question of magnitude: the jailtime-maximum message was rated *least* realistic of all seven conditions (5.60, against 7.60 for the non-specific fine). The effect of a stated penalty is bounded by its credibility. A seven-year custodial threat may strike taxpayers as implausible, and on this evidence that is what participants themselves reported.

The fine-maximum effect runs through perceived severity and realism. Adding those two measures to the model removes the fine-maximum coefficient entirely (0.220 to -0.050) while leaving the jailtime coefficients intact or increasing them. The parallel mediation analysis of the previous submission's Table 8 is reproduced below: PROCESS Model 4, with each message in turn against Costs Further Action as the independent variable (message framing), punishment severity and realism as parallel mediators, the stated likelihood of paying as the dependent variable, and age, gender and risk propensity as covariates; the intervals on the indirect effects are bootstrap confidence intervals as reported there. Path labels use the likelihood-of-paying label of this version.

Fine non-specific against Costs Further Action:

Path	B	SE	t	p	95% CI
a1: message framing → punishment severity	-0.099	0.085	-1.159	0.247	[-0.266, 0.068]
a2: message framing → realism	0.733	0.096	7.663	< 0.001	[0.546, 0.921]
b1: punishment severity → likelihood of paying	0.262	0.021	12.469	< 0.001	[0.221, 0.304]
b2: realism → likelihood of paying	0.301	0.019	16.065	< 0.001	[0.264, 0.338]
c': direct effect	-0.140	0.074	-1.886	0.060	[-0.285, 0.006]
indirect a1 × b1	-0.026	0.023			[-0.070, 0.018]
indirect a2 × b2	0.221	0.033			[0.159, 0.288]
a1 × b1 against a2 × b2	-0.247	0.033			[-0.313, -0.185]
age	0.002	0.003	0.456	0.648	[-0.005, 0.008]
gender	0.303	0.076	3.993	< 0.001	[0.154, 0.452]
risk propensity	-0.114	0.016	-7.159	< 0.001	[-0.145, -0.082]

Jailtime non-specific against Costs Further Action:

Path	B	SE	t	p	95% CI
a1: message framing → punishment severity	0.236	0.043	5.482	< 0.001	[0.151, 0.320]
a2: message framing → realism	-0.253	0.052	-4.894	< 0.001	[-0.354, -0.152]
b1: punishment severity → likelihood of paying	0.251	0.021	12.054	< 0.001	[0.210, 0.291]
b2: realism → likelihood of paying	0.203	0.017	11.747	< 0.001	[0.169, 0.237]
c': direct effect	0.094	0.037	2.528	0.012	[0.021, 0.166]
indirect a1 × b1	0.059	0.013			[0.036, 0.086]
indirect a2 × b2	-0.051	0.012			[-0.075, -0.030]
a1 × b1 against a2 × b2	0.111	0.013			[0.085, 0.137]
age	0.001	0.003	0.149	0.882	[-0.006, 0.007]
gender	0.286	0.076	3.773	< 0.001	[0.137, 0.435]
risk propensity	-0.111	0.016	-6.992	< 0.001	[-0.142, -0.080]

Fine average against Costs Further Action:

Path	B	SE	t	p	95% CI
a1: message framing → punishment severity	-0.140	0.031	-4.512	< 0.001	[-0.201, -0.079]

Path	B	SE	t	p	95% CI
a2: message framing → realism	0.045	0.034	1.353	0.176	[-0.020, 0.111]
b1: punishment severity → likelihood of paying	0.220	0.020	11.250	< 0.001	[0.181, 0.258]
b2: realism → likelihood of paying	0.256	0.018	14.141	< 0.001	[0.220, 0.291]
c': direct effect	0.013	0.025	0.498	0.618	[-0.037, 0.062]
indirect a1 × b1	-0.031	0.007			[-0.046, -0.017]
indirect a2 × b2	0.012	0.009			[-0.005, 0.029]
a1 × b1 against a2 × b2	-0.042	0.009			[-0.060, -0.024]
age	0.0001	0.003	0.027	0.978	[-0.006, 0.007]
gender	0.217	0.077	2.813	0.005	[0.066, 0.368]
risk propensity	-0.100	0.017	-6.050	< 0.001	[-0.133, -0.068]

Jailtime average against Costs Further Action:

Path	B	SE	t	p	95% CI
a1: message framing → punishment severity	0.100	0.023	4.413	< 0.001	[0.056, 0.144]
a2: message framing → realism	-0.261	0.027	-9.575	< 0.001	[-0.314, -0.207]
b1: punishment severity → likelihood of paying	0.265	0.020	13.307	< 0.001	[0.226, 0.303]
b2: realism → likelihood of paying	0.169	0.017	10.207	< 0.001	[0.136, 0.201]
c': direct effect	0.080	0.019	4.200	< 0.001	[0.043, 0.118]
indirect a1 × b1	0.026	0.007			[0.014, 0.040]
indirect a2 × b2	-0.044	0.007			[-0.057, -0.032]
a1 × b1 against a2 × b2	0.070	0.007			[0.056, 0.085]
age	0.007	0.003	2.135	0.033	[0.001, 0.013]
gender	0.270	0.076	3.532	< 0.001	[0.120, 0.420]
risk propensity	-0.095	0.016	-5.941	< 0.001	[-0.126, -0.063]

Fine maximum against Costs Further Action:

Path	B	SE	t	p	95% CI
a1: message framing → punishment severity	0.240	0.016	14.592	< 0.001	[0.208, 0.272]
a2: message framing → realism	0.025	0.020	1.207	0.228	[-0.015, 0.065]
b1: punishment severity → likelihood of paying	0.283	0.021	13.334	< 0.001	[0.242, 0.325]
b2: realism → likelihood of paying	0.190	0.017	11.063	< 0.001	[0.156, 0.223]
c': direct effect	-0.028	0.015	-1.845	0.065	[-0.057, 0.002]
indirect a1 × b1	0.068	0.008			[0.052, 0.085]
indirect a2 × b2	0.005	0.004			[-0.003, 0.013]
a1 × b1 against a2 × b2	0.063	0.009			[0.047, 0.081]
age	0.003	0.003	0.912	0.362	[-0.003, 0.009]
gender	0.271	0.073	3.730	< 0.001	[0.129, 0.413]
risk propensity	-0.098	0.016	-6.306	< 0.001	[-0.128, -0.067]

Jailtime maximum against Costs Further Action:

Path	B	SE	t	p	95% CI
a1: message framing → punishment severity	0.063	0.016	4.023	< 0.001	[0.032, 0.094]
a2: message framing → realism	-0.214	0.019	-11.557	< 0.001	[-0.250, -0.177]
b1: punishment severity → likelihood of paying	0.200	0.020	10.082	< 0.001	[0.161, 0.238]
b2: realism → likelihood of paying	0.167	0.017	9.985	< 0.001	[0.134, 0.200]
c': direct effect	0.045	0.013	3.433	< 0.001	[0.019, 0.071]
indirect a1 × b1	0.013	0.004			[0.006, 0.020]
indirect a2 × b2	-0.036	0.005			[-0.047, -0.026]
a1 × b1 against a2 × b2	0.048	0.005			[0.039, 0.059]
age	0.002	0.003	0.679	0.497	[-0.004, 0.009]
gender	0.196	0.078	2.520	0.012	[0.043, 0.348]
risk propensity	-0.115	0.016	-7.087	< 0.001	[-0.146, -0.083]

The pattern the main text relies on is carried by the fine-maximum comparison: its effect on the stated likelihood of paying runs through perceived severity (indirect effect 0.068, interval excluding zero) with no remaining direct effect, whereas the jailtime comparisons retain direct effects alongside negative indirect effects through realism.

A.3 What Study A contributes to this paper

Three premises, and only these three, are carried forward:

1. **When enforcement consequences are left unspecified, taxpayers fill the gap with fines rather than imprisonment.** This is why Experiments 2 to 4 manipulate fine information rather than imprisonment information, and it is the reason the precursor study is retained at all rather than dropped.
2. **Responses to penalty information mix belief updating with reactions that updating cannot explain,** which motivates measuring prior beliefs directly in Experiments 2 and 3.
3. **Stated-penalty effects are bounded by credibility,** as Appendix A.2 shows.

Study A's own treatment comparisons are not evidence for H2 or H3, because it never resolves an ambiguity against a measured prior. Its relationship to Experiment 2 is set out at Appendix D.3.

A.4 The distribution of responses

The full distribution of responses is reported here, because the concentration of responses at the top of the scale bears on the economic significance of the condition differences.

Responses are heavily concentrated at the top of the 11-point scale. Of 7,064 participants, **238 (3.4%) answered strictly below the midpoint of 5** and **500 (7.1%) answered at the midpoint or below**, while **3,486 (49.3%) chose the maximum of 10**. The full distribution:

Response	0	1	2	3	4	5	6	7	8	9	10
n	27	34	40	52	85	262	326	681	1,051	1,020	3,486

The below-midpoint share varies little across conditions, from 2.3% (Jailtime, maximum) to 4.4% (Costs Further Action).

Two consequences follow, and we accept both. First, skepticism about the economic significance of Study A's condition differences is well founded. With half the sample at the ceiling and 96.6% at or above the midpoint, the seven condition means can differ only within a narrow band, and that is what is observed (8.51 to 8.77, with message condition explaining 0.3% of the variance). Second, this compression is a property of a hypothetical stated-likelihood measure with nothing at stake, and it is one of the reasons the paper's claims do not rest on Study A. The behavioral evidence comes from Experiment 1, which measures whether real debts were actually paid, and from Experiment 3, where declarations are incentivized with real money.

Online Appendix B: An Illustrative Model of the Payment Decision

This appendix formalizes the decision problem faced by the taxpayers in Experiment 1. It is offered as an aid to intuition, not as the source of the paper's hypotheses. The previous version's Prediction 1 was drawn from comparative statics on parameters the experiments never manipulated (the interest rate, the disutility of present payment, the moral cost), and that a prediction of that kind cannot be tested by the designs reported here. The hypotheses developed in §2.3 to §2.5 refer only to constructs that are manipulated or directly measured. What follows is therefore presented as illustration, and Appendix B.3 sets out plainly what the model does *not* capture.

B.1 Setup

The model formalizes the decision problem of the taxpayer population studied by Hallsworth et al. (2017). It describes a situation in which (a) taxpayers who owe money and are liquidity-constrained, earning less than a threshold level of income, have no margin on which an enforcement letter can act, whereas (b) taxpayers who are not liquidity-constrained trade off an explicit financial liability against a moral cost of the kind described by Levitt and List (2007).¹

Consistent with present-bias preferences, the agent experiences a higher disutility from paying tax today ($\beta > 1$) than from the prospect of future payment. The agent chooses whether to pay now ($i = 0$) or later ($i = 1$) so as to maximize

$$U = u_i(Y, M, t, p, C) = Y - \beta t \text{ if } i = 0 \text{ and } Y \geq \beta t = \max[Y + (1 + r)t - M - g(p, C | \\ I) - (1 + \alpha)t, 0] \text{ if } i = 1 \text{ or } Y < \beta t$$

¹ Two literatures stand behind the two branches. Wang and White (2000) set the debtor's repayment decision in a second period in which wealth, and so the ability to pay, has been realized. Gross et al. (2014) find that the 2001 and 2008 tax rebates raised consumer bankruptcy filings, the rebate relieving a constraint that had previously prevented filing, which is the evidence that liquidity constraints bind household financial decisions. Neither work is drawn on beyond these two points.

The model begins on the self-assessment day. The agent earns stochastic labour income Y , drawn i.i.d. from $f(y)$, and reports it to the authority, which converts it by a deterministic rule into a liability $t \geq 0$. A no-borrowing condition applies: an agent who experiences a negative income shock ($Y < \beta t$) has no choice but $i = 1$. Otherwise the timing is a choice.

Deferring payment causes interest to accrue geometrically at rate $\alpha \in (0, 1)$, while the agent can reinvest the retained funds at rate $r \in (0, 1)$. The agent bears a moral cost $M > 0$ from not paying immediately, induced by the letter. Finally the agent faces an enforcement cost C , imposed with probability p .

Both p and C are unknown to the agent. The agent forms beliefs about them from an information set I , and the function $g(p, C | I)$ maps the probability and cost of enforcement into utility space conditional on that information. g is assumed monotonic in both arguments.

B.2 The payment condition

Paying now is preferred whenever

$$Y - \beta t > Y + (1 + r)t - M - g(p, C | I) - (1 + \alpha)t$$

which holds as long as

$$\beta + r - \alpha < M/t + g(p, C | I)/t$$

The condition is easier to satisfy as the moral cost rises, as the expected cost of enforcement rises, and as the penalty rate rises; it is harder to satisfy as the prevailing interest rate rises, as the liability grows, and as the disutility of paying today grows.

Only one of these terms is manipulated in this paper. Experiment 1 varies the enforcement message, which is a change in I and therefore acts on $g(\cdot | I)$. The remaining comparative statics are stated here for completeness and are **not** offered as predictions; none of

the four experiments manipulates the interest rate, the liability, or the disutility of present payment, and the moral cost is neither manipulated nor measured.

B.3 What the model does not capture

The model treats the taxpayer's informational position as a matter of *where* beliefs sit. Because g is monotonic in p and C and depends on I only through the location of those beliefs, the model implies that behavior follows the direction of the news: an enforcement message that raises the perceived penalty raises compliance, and one that lowers it lowers compliance.

That is the statistical-updating explanation which §2.2 sets up as a rival and which §4.3 rejects. In Experiment 2, participants whose measured priors were *below* the disclosed maximum received objectively bad news, news that should have raised compliance under belief updating, and their stated likelihood of paying fell instead. A model in which the information set enters only through the mean of the belief distribution cannot represent that result.

Ambiguity aversion is the claim that g responds to the **spread** or vagueness of the belief distribution and not only to its central tendency: an unresolved maximum is aversive *because* it is unresolved, independently of where the taxpayer's point estimate lies. Formalizing that requires non-expected-utility machinery which this model does not contain. Two treatments in the literature do contain it, and either would be the natural formal home for the mechanism the paper documents:

- **Snow and Warren (2005)** model ambiguity over the audit *probability* under ambiguity preferences, and show that greater ambiguity raises compliance for ambiguity-averse taxpayers and lowers it for ambiguity-loving ones. Their skepticism about ambiguity as a policy instrument turns on the heterogeneity of those preferences;

our field evidence speaks to the average effect of a communication choice rather than to screening on preference type.

- **Calford and DeAngelo (2023)** model an enforcement agency's decision to reveal or shroud its resource allocation when offenders hold non-expected-utility uncertainty preferences, which is the disclosure problem discussed in §7 stated formally.

We set out the simpler model here because it makes the field decision problem legible, and we state its limitation explicitly because the paper's central result lies in the place where the model is silent.

Online Appendix C: Experiment 1 Supplements

Source material: the previous submission's Tables 1-5 and its Appendices 3 and 7. Because the field data were analyzed within HMRC under the authority's data-access arrangements, this part reproduces the previously reported analyses; it does not re-estimate them.

C.1 Treatment messages, group sizes and balance

All letters carried the amount owed and payment instructions. Treatment letters added one short, bolded message. The Control group received the standard letter with no added message.

Group	Added message	n	Total debt	Mean debt	Mean age	% male
Control	<i>(none)</i>	17,208	£43,066,664.71	£2,503.29	49.57	71.34
Interest Today	We are charging you interest daily. You will save money if you pay now.	17,061	£42,772,379.82	£2,507.61	49.58	72.00
Interest Tomorrow	We are charging you interest daily. You will lose money if you pay later.	16,866	£41,703,233.28	£2,473.35	49.58	70.70
Less Effort	Paying now means less hassle because we won't contact you about this again.	17,172	£43,911,399.60	£2,557.75	48.99	71.44
More Effort	Paying later means more hassle because we will contact you about this again.	16,960	£42,966,785.52	£2,533.87	49.36	71.63
Costs Further Action	If you pay this debt now, you can stop us taking any further action.	17,048	£43,138,652.55	£2,531.02	49.26	70.46
Costs Peace	If you pay this debt now, you won't have to think about it anymore.	17,162	£42,752,837.97	£2,491.57	49.35	71.92
Interest & Effort	We are charging you interest daily. You will save money if you pay today. Paying now means less hassle because we won't contact you about this again.	17,134	£42,621,810.33	£2,487.99	49.45	71.96
Interest & Action	We are charging you interest daily. You will save money if you pay today. If you pay this debt now, you can stop us taking any further action.	16,858	£42,738,043.43	£2,535.63	49.18	71.58
Contact Now	Please contact us as soon as you receive this letter.	16,994	£42,323,999.15	£2,491.55	49.20	71.63
Contact Now Information	Please contact us as soon as you receive this letter. We are open Monday to Saturday 8am to 8pm, Sundays 8am to 5pm - or you can pay online any time.	17,134	£43,657,948.21	£2,548.48	49.25	71.54
Contact Later Information	Please choose a time tomorrow to contact us. We are open Monday to Saturday 8am to 8pm, Sundays 8am to 5pm - or you can pay online any time.	17,348	£43,663,133.99	£2,517.33	49.29	71.66

The three groups discussed in the main text are shown in bold. Mean debt, mean age and gender composition are closely balanced across the twelve groups. Regressions of treatment-

group membership on background characteristics, reported in the previous submission's Appendix 3, found no systematic predictors of assignment.

The group sizes reported above sum to the 204,945 taxpayers randomized. Two of them correct the corresponding table of the previous submission, which gave Contact Now as 16,960 and Contact Now Information as 17,124. The values reported here, 16,994 and 17,134, are those of the date-of-issue table in Appendix C.2, and each of the twelve group totals in that table reconciles with its seven daily issue counts, which the background-characteristics table did not permit us to check. The previous submission's text stated a total of 204,936.

C.2 Dates of letter issue

Letters were issued over seven working days, from Tuesday 5 March to Wednesday 13 March 2013, with each group's letters spread across all seven days so that day of issue is balanced across conditions.

Group	5 Mar	6 Mar	7 Mar	8 Mar	11 Mar	12 Mar	13 Mar	Total
Control	2,913	2,808	2,651	2,579	2,575	2,482	1,200	17,208
Interest Today	2,985	2,784	2,646	2,584	2,483	2,420	1,159	17,061
Interest Tomorrow	2,947	2,704	2,679	2,489	2,379	2,441	1,227	16,866
Less Effort	3,070	2,809	2,697	2,531	2,436	2,469	1,160	17,172
More Effort	3,012	2,741	2,609	2,568	2,431	2,401	1,198	16,960
Costs Further Action	2,971	2,852	2,656	2,516	2,493	2,465	1,095	17,048
Costs Peace	3,033	2,711	2,684	2,599	2,492	2,462	1,181	17,162
Interest & Effort	3,072	2,767	2,679	2,641	2,392	2,413	1,170	17,134
Interest & Action	2,924	2,775	2,607	2,552	2,370	2,481	1,149	16,858
Contact Now	2,957	2,755	2,659	2,583	2,431	2,440	1,169	16,994
Contact Now Information	2,978	2,740	2,735	2,533	2,452	2,453	1,243	17,134
Contact Later Information	3,008	2,727	2,718	2,669	2,485	2,455	1,286	17,348

Note on the issue date. The previous submission's §3.2 stated that letters were sent to taxpayers with standalone debts "on February 1, 2013". Read against this table, 1 February 2013 is the date as at which the outstanding debts were identified, the reminder-statement date, and not the date of issue. The letters themselves went out between 5 and 13 March 2013. The main text now states this explicitly, because the earlier phrasing invites the reading that the letters were issued on 1 February.

C.3 Regression tables

The following tables from the previous submission are reproduced here without re-estimation.

Entries are marginal effects from logistic regressions on payment, with standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table C.3a. Payment within 23 days (previous Table 3): treatment indicators alone (I), with covariates (II), and separately for taxpayers with and without recent debt (III, IV).

	(I) Full sample	(II) Full sample	(III) Recent debtors	(IV) No recent debt
Interest Today	0.000 (0.005)	0.001 (0.005)	-0.013* (0.007)	0.025** (0.009)
Interest Tomorrow	0.008 (0.005)	0.008 (0.005)	-0.002 (0.007)	0.024** (0.009)
Less Effort	-0.003 (0.005)	-0.010 (0.005)	-0.005 (0.007)	0.001 (0.009)
More Effort	0.002 (0.005)	0.003 (0.005)	-0.004 (0.007)	0.012 (0.009)
Costs Further Action	0.044*** (0.005)	0.046*** (0.006)	0.038*** (0.007)	0.057*** (0.009)
Costs Peace	-0.008 (0.005)	-0.007 (0.005)	-0.009 (0.007)	-0.006 (0.009)
Interest & Effort	0.006 (0.005)	0.009 (0.005)	-0.003 (0.007)	0.022* (0.009)
Interest & Action	0.017*** (0.005)	0.019*** (0.005)	0.014* (0.007)	0.024** (0.009)
Contact Now	0.022*** (0.005)	0.024*** (0.005)	0.022*** (0.007)	0.021* (0.009)
Contact Now Information	0.008 (0.005)	0.010 (0.005)	0.013* (0.007)	0.021* (0.009)
Contact Later Information	0.029*** (0.005)	0.031*** (0.005)	0.030*** (0.007)	0.027** (0.009)
Age		0.004*** (0.000)		
Male		-0.032*** (0.003)		
Debt size		-4.59e-06*** (0.000)		
Self-employed		0.121*** (0.004)		
Accountant		-0.004 (0.003)		
Experienced		-0.048*** (0.002)		
Constant	0.421*** (0.016)	0.420*** (0.031)	0.405*** (0.019)	0.452*** (0.026)
Pseudo R ²	0.00	0.01	0.00	0.00
N	204,936	204,367	133,665	71,271

Table C.3b. Payment by background characteristics (previous Table 4): treatment effects estimated separately by gender, median age and use of an accountant, the source of the heterogeneity results summarized in §3.3.

	(I) Male	(II) Female	(III) Below median age	(IV) Above median age	(V) Accountant	(VI) No accountant
Interest Today	-0.007 (0.006)	0.015 (0.011)	-0.017* (0.008)	0.018* (0.008)	-0.008 (0.006)	0.025* (0.011)
Interest Tomorrow	0.004 (0.006)	0.014 (0.010)	0.001 (0.008)	0.015* (0.008)	0.007 (0.006)	0.008 (0.011)
Less Effort	-0.008 (0.006)	0.010 (0.010)	-0.009 (0.008)	0.004 (0.008)	-0.002 (0.006)	-0.007 (0.011)
More Effort	-0.002 (0.006)	0.009 (0.011)	-0.009 (0.008)	0.012 (0.008)	0.002 (0.006)	0.001 (0.011)
Costs Further Action	0.042*** (0.006)	0.049*** (0.010)	0.039*** (0.008)	0.051*** (0.008)	0.046*** (0.008)	0.039*** (0.011)

	(I) Male	(II) Female	(III) Below median age	(IV) Above median age	(V) Accountant	(VI) No accountant
Costs Peace	-0.012* (0.006)	-0.001 (0.011)	-0.021** (0.008)	0.005 (0.008)	-0.008 (0.006)	-0.009 (0.011)
Interest & Effort	-0.003 (0.006)	0.021 (0.011)	0.006 (0.008)	0.006 (0.008)	0.003 (0.006)	0.013 (0.011)
Interest & Action	0.012 (0.006)	0.029** (0.011)	0.011 (0.008)	0.022** (0.008)	0.016** (0.006)	0.018 (0.011)
Contact Now	0.019** (0.006)	0.034** (0.011)	0.016* (0.008)	0.030*** (0.008)	0.023*** (0.006)	0.019 (0.011)
Contact Now Information	0.004 (0.006)	0.016 (0.011)	0.007 (0.008)	0.009 (0.008)	0.015* (0.008)	-0.012 (0.011)
Contact Later Information	0.026*** (0.006)	0.034** (0.010)	0.019* (0.008)	0.039*** (0.008)	0.032*** (0.006)	0.020 (0.011)
Constant	0.410***	0.446***	0.456***	0.386***	0.420***	0.424***
Pseudo R ²	0.00	0.00	0.00	0.00	0.00	0.00
N	146,484	54,320	102,686	102,250	153,007	51,929

Table C.3c. Payment by debt quartile (previous Table 5): the Costs Further Action

effect is present in every quartile of the debt distribution while other messages are heterogeneous.

	(I) First quartile	(II) Second quartile	(III) Third quartile	(IV) Fourth quartile
Interest Today	0.003 (0.011)	0.014 (0.011)	-0.010 (0.011)	-0.006 (0.010)
Interest Tomorrow	0.016 (0.011)	0.016 (0.011)	0.003 (0.011)	-0.006 (0.010)
Less Effort	-0.008 (0.011)	0.002 (0.011)	0.007 (0.011)	-0.015 (0.010)
More Effort	-0.001 (0.011)	0.006 (0.011)	-0.001 (0.011)	0.001 (0.011)
Costs Further Action	0.049*** (0.011)	0.064*** (0.011)	0.035*** (0.011)	0.029** (0.011)
Costs Peace	-0.004 (0.011)	0.003 (0.011)	-0.016 (0.011)	-0.017 (0.010)
Interest & Effort	0.023* (0.011)	0.020 (0.011)	-0.008 (0.011)	-0.013 (0.010)
Interest & Action	0.036** (0.011)	0.021 (0.011)	0.012 (0.011)	-0.004 (0.010)
Contact Now	0.037** (0.011)	0.019 (0.011)	0.024* (0.011)	0.009 (0.011)
Contact Now Information	0.006 (0.011)	0.014 (0.011)	0.003 (0.011)	0.007 (0.011)
Contact Later Information	0.028** (0.011)	0.035** (0.011)	0.021* (0.011)	0.030** (0.011)
Constant	0.460	0.422	0.420	0.380
Pseudo R ²	0.00	0.00	0.00	0.00
N	51,224	51,223	51,222	51,268

Table C.3d. Predictors of treatment-group membership (previous Appendix 3), the

randomization check: each column regresses a background characteristic on the treatment indicators. No systematic pattern of assignment appears.

	(I) Male	(II) Age	(III) Initial debt	(IV) Self-employed	(V) Accountant
Interest Today	0.007	-0.002	4.321	-0.001	-0.002
Interest Tomorrow	-0.007	0.004	-29.939	-0.003	-0.004
Less Effort	0.001	-0.381**	54.453	0.001	0.003
More Effort	0.003	-0.212	30.573	-0.003	-0.009
Costs Further Action	-0.009	-0.314	27.723	0.001	0.003
Costs Peace	0.006	-0.226	-11.724	0.001	0.001
Interest & Effort	0.002	-0.390**	32.336	-0.059*	-0.010*
Interest & Action	0.006	-0.125	-15.301	0.002	0.000

	(I) Male	(II) Age	(III) Initial debt	(IV) Self-employed	(V) Accountant
Contact Now	0.002	-0.325**	46.167	-0.001	-0.003
Contact Now Information	0.001	-0.381**	-11.741	-0.003	-0.008
Contact Later Information	0.003	-0.286*	14.039	-0.002	-0.002
N	204,936	204,413	204,889	204,936	204,936

Table C.3e. Payment within 48 and 70 days (previous Appendix 7), the longer outcome

windows, with covariates.

	(I) Payment within 48 days	(II) Payment within 70 days
Interest Today	-0.006 (0.005)	-0.007 (0.005)
Interest Tomorrow	-0.013* (0.005)	-0.013* (0.005)
Less Effort	-0.013* (0.005)	-0.011* (0.005)
More Effort	-0.005 (0.005)	-0.002 (0.005)
Costs Further Action	0.010 (0.005)	0.004 (0.005)
Costs Peace	-0.009 (0.005)	-0.006 (0.005)
Interest & Effort	-0.009 (0.005)	-0.006 (0.005)
Interest & Action	0.001 (0.005)	0.001 (0.005)
Contact Now	-0.001 (0.005)	-0.001 (0.005)
Contact Now Information	-0.013* (0.005)	-0.019*** (0.005)
Contact Later Information	0.001 (0.005)	0.001 (0.005)
Age	0.003*** (0.000)	0.002*** (0.000)
Male	-0.022*** (0.002)	-0.019*** (0.002)
Debt size	-1.36e-06*** (0.000)	-3.42e-07*** (0.000)
Self-employed	0.146*** (0.004)	0.123*** (0.004)
Accountant	0.005 (0.002)	0.003 (0.002)
Experienced	0.001 (0.002)	0.011*** (0.002)
Constant	0.651	0.710
Pseudo R ²	0.00	0.00
N	204,367	204,367

Note on the longer windows. The previous submission stated that the

enforcement-uncertainty effect persisted while other message effects diminished after 48 days, and inferred that ambiguity “operates through more durable psychological mechanisms”. We report the 48-day and 70-day windows descriptively and do not repeat that inference. Persistence of a one-shot letter effect requires an identified mechanism, habit formation or a change in the taxpayer’s capital stock of knowledge, and neither was instrumented here; see Brandon et al. (2026) on what establishing nudge persistence requires.

C.4 Design correspondence with the experiments of Hallsworth et al. (2017)

Hallsworth et al. (2017) report two large-scale natural field experiments conducted in the same Self Assessment debt-recovery process, in which social-norm and public-services messages were

inserted into the authority’s standard reminder letters and payment within the first 23 days was the outcome. Experiment 1 is a separate randomization conducted later in the same process. The correspondence, and the differences, are as follows.

Feature	Hallsworth et al. (2017)	Experiment 1
Population	UK Self Assessment debtors	UK Self Assessment debtors
Vehicle	One message added to the authority’s standard reminder letter	One message added to the authority’s standard reminder letter
Trials and issue	Two experiments, conducted in 2011 and 2012; the first issued its letters in August 2011	One experiment; letters issued 5 to 13 March 2013 (Appendix C.2) on debts identified at 1 February 2013
Treatments	Three social-norm and two public-services messages (first experiment); social-norm variants (second experiment)	Twelve groups varying enforcement, financial-cost, administrative-effort and contact messages (Appendix C.1)
Outcome	Payment within 23 days	Payment within 23 days
Scale	Around 100,000 taxpayers randomized in the first experiment; more than 200,000 across the two	204,945 taxpayers randomized (Appendix C.1)

The two designs share the population, the letter vehicle and the outcome window; they differ in the class of message content (social information as against enforcement communication) and in the cohort of debts treated. Experiment 1 is therefore a later randomization in the same recovery process, on a distinct cohort of debts, and not a re-analysis of the data of the experiments reported in that paper.

Online Appendix D: Experiment 2 Supplements

All statistics are computed from the archived Experiment 2 datasets. The covariate-adjusted models in Appendix D.1 and Appendix D.2 are the original analyses, and their coefficients agree with the original estimation log and with the previous submission's regression table.

D.1 Covariate-adjusted models

The main text reports the design as an analysis of variance. This section reports the equivalent regressions with covariates, which is how the previous version presented them.

Effect of revealing the actual fine on the stated likelihood of paying:

Specification	b	SE	p	n
Information only	-0.425	0.134	0.001	1,194
+ age, gender, risk propensity, debt size, elicited fine belief	-0.436	0.133	0.001	1,193
+ household income, leadership, management experience	-0.426	0.133	0.001	1,180

Within the maximum-fine conditions:

Specification	b	SE	p	n
Information only	-0.586	0.192	0.002	599
+ first covariate block	-0.621	0.191	0.001	599
+ second covariate block	-0.623	0.193	0.001	596

Within the minimum-fine conditions:

Specification	b	SE	p	n
Information only	-0.260	0.185	0.162	595
+ first covariate block	-0.260	0.184	0.159	594
+ second covariate block	-0.232	0.183	0.206	584

Inferences are unchanged by covariate adjustment throughout: the disclosure effect is significant overall and within the maximum-fine conditions, and is not significant within the minimum-fine conditions, in every specification.

Note on sample sizes, and on the figure of 1,180. The previous submission's regression table reported $N = 1,180$ with subsamples of 584 and 596. Those are **complete-case samples for the fullest covariate specification**, not counts of participants recruited. The first collection period comprised 1,195 taxpayers, of whom 1,194 provided the dependent variable; every cell mean and marginal reported in §4 is

computed on those 1,194. Of them, 1,193 have complete data on the first covariate block and 1,180 on the full covariate set; 584 and 584 plus 12 further cases, that is 596, are the same quantity within the minimum-fine and maximum-fine conditions respectively. This also identifies the two sample-classification rows at the foot of that table (“High/Low Belief” and “Minimum/Maximum”): they record which subgroup each column was estimated on, whether the column used both fine types or only one, and whether it used both belief groups or only one; they are not estimated coefficients. The previous submission did not explain those rows; that is what they record.

D.2 Individual-level belief analyses

Every participant stated a prior belief about the fine before any disclosure. For participants in a disclosure condition we can therefore classify the disclosure as good news or bad news for that individual, by comparing the belief with the value actually disclosed. Following the original coding, a participant is classified as having **over-estimated** if the stated belief exceeded the disclosed value and **under-estimated** if it fell below. One participant, in the minimum-fine No-Reveal condition, stated a belief equal to the disclosed value and belongs to neither group.

Unadjusted cell means, by fine type and belief group:

Fine type	Belief group	No-Reveal	Revealed	Difference	Welch p
Maximum	Under-estimated (bad news)	8.734 (n = 199)	8.036 (n = 220)	-0.697	0.003
Maximum	Over-estimated (good news)	8.713 (n = 94)	8.407 (n = 86)	-0.306	0.360
Minimum	Under-estimated	8.569 (n = 209)	8.454 (n = 196)	-0.115	0.610
Minimum	Over-estimated	8.923 (n = 91)	8.327 (n = 98)	-0.597	0.070

The same models with the full covariate block:

Fine type	Belief group	b (revealed)	p	n
Maximum	Under-estimated	-0.673	0.004	416
Maximum	Over-estimated	-0.205	0.557	180
Minimum	Under-estimated	0.019	0.932	397
Minimum	Over-estimated	-0.673	0.051	186

The result the main text relies on is the first row of each table. Among participants for whom the disclosed maximum was worse than they had privately expected (the subgroup for

which statistical updating makes an unambiguous directional prediction that compliance should rise), the stated likelihood of paying **fell**, significantly, in both specifications. The belief $\times$ information interaction is not significant within either fine type (maximum: $F(1, 595) = 0.87, p = 0.352$; minimum: $F(1, 591) = 1.51, p = 0.220$), so we do not claim that the effect differs reliably between belief groups; the claim rests on the direction of movement in the subgroup where updating is most sharply falsifiable.

Two claims from the previous submission are not supported and are not repeated. First, that submission reported that participants whose priors exceeded the disclosed *minimum* significantly reduced the stated likelihood of paying, labeled willingness to pay in that version ($\beta = -0.673, p = 0.051$). That coefficient is reproduced unchanged in the bottom row of the table above, but $p = 0.051$ is not significant at the conventional threshold, and the unadjusted equivalent is weaker still ($-0.597, p = 0.070$). The associated interpretation, that minimum-fine disclosures operate through rational updating while maximum-fine disclosures operate through ambiguity resolution, two distinct pathways, rested on comparing a significant coefficient with a non-significant one and is withdrawn. Second, that submission described the reduction as occurring “regardless of prior beliefs”. The analysis above shows a sharper and better-supported pattern, and the main text now states it in that form.

D.3 Reconciliation with Study A

Study A (Online Appendix A) and Experiment 2 appear at first to point in opposite directions. In Study A, a message stating a **maximum fine** produced the highest stated likelihood of paying of any condition, higher than the ambiguous “Costs Further Action” baseline. In Experiment 2,

resolving the maximum reduced the stated likelihood of paying relative to a preserved-ambiguity baseline. The two are not in conflict, for the same structural reason set out in §6.4 for Experiment 4.

Study A is a **message-exposure** design. Each participant saw one enforcement message; no prior belief was elicited, and nothing was resolved relative to a stated baseline. Its comparisons therefore concern message *content*: what a letter says, measured against a different letter. Experiment 2 is a **resolution** design. Every participant states a prior belief about the fine, and the manipulation is whether that belief is then met with a specific figure or left unresolved. Its comparisons concern what a disclosure *removes*.

A message that names a large fine can therefore raise compliance relative to a letter that names nothing in particular (Study A's finding), while resolving an unspecified maximum can lower compliance relative to leaving it unspecified (Experiment 2's finding). The comparison classes differ.

Study A also supplies the boundary that Experiment 2 does not test. Its jailtime conditions show that severity alone did not increase compliance: the most extreme threat did not raise the stated likelihood of paying relative to the baseline, and Study A attributes this to the low perceived realism of extreme penalties. The effect of a stated penalty is bounded by its credibility. That premise is one reason the paper manipulates fine information rather than imprisonment information throughout Experiments 2 to 4, and it is consistent with the observation that the seven-year imprisonment conditions read as unrealistic.

D.4 The second collection period: the £500-debt supplementary conditions

As described in §4.1, Experiment 2 was conducted in two periods of data collection. The first period comprised the factorial conditions and the two further £2,500-debt conditions of the four-

cell comparison. In the second, later period, four further conditions were collected at the £500 debt level: an Uncertain condition, an Elevated condition revealing the £10,000 maximum, and a second pair of revealed values, a £1,200 maximum and a £600 minimum, complementing the £1,600 and £700 values of the main design.

The two periods account for the experiment's fourteen conditions as follows. Across all fourteen, 2,050 taxpayers provided the dependent variable. The first period accounts for 1,490 of them: 1,194 in the eight factorial conditions, and 296 in the two further £2,500-debt conditions (Uncertain, $n = 147$; Elevated, $n = 149$). The second period accounts for the remaining 560, in the four conditions tabulated below. The 1,180 of the previous submission is a subset of the 1,194, being those with complete data on the full covariate set (Appendix D.1).

The £500-debt four-cell comparison reported in §4.3 combines cells from the two periods by construction: the No-Reveal and revealed-£1,600 cells belong to the factorial design of the first period, and the Uncertain and Elevated cells to the second period. Cell-level statistics, standard deviations in parentheses:

Cell	n	Stated likelihood of paying	Collection period
No-Reveal	138	8.761 (1.932)	first
Uncertain	138	8.543 (2.069)	second
Specific (£1,600 revealed)	149	8.275 (2.804)	first
Elevated (£10,000 revealed)	139	8.036 (3.147)	second
Specific (£1,200 revealed)	142	8.127 (2.668)	second
Minimum (£600 revealed)	141	8.312 (2.426)	second
Minimum (£700 revealed)	139	8.633 (2.447)	first

The four-cell pattern replicates the §4.3 comparison at the second debt level: the pooled ambiguity-preserved cells average 8.652 ($n = 276$) against 8.160 in the pooled resolved cells ($n = 288$), a difference of 0.492 (ambiguity contrast $F(1, 560) = 5.36$, one-tailed $p = 0.011$, $d = 0.19$), with the omnibus F-test across the four cells not significant ($F(3, 560) = 2.14$, $p = 0.094$). The four magnitude comparisons among revealed values are not significant (£1,200 against £1,600, Welch $p = 0.644$; £600 against £700, $p = 0.271$; £1,600 against £10,000, $p = 0.498$; £1,200

against £10,000, $p = 0.795$). Together with the £1,900 against £10,000 comparison reported in §4.3 (Welch $p = 0.895$), the absence of a magnitude effect among revealed values rests on five comparisons spanning £600 to £10,000 across two debt levels rather than on one.

Online Appendix E: Experiment 3 Supplements

All statistics in this part are computed from the archived Experiment 3 export.

E.1 Sensitivity of the primary result to the exclusion rule

The primary sample retains the 827 participants who passed both the attention check and the instructional manipulation check. The table below re-estimates the primary ambiguity contrast under three exclusion rules.

Sample	N	Ambiguous	Resolved	Difference	Contrast F	One-tailed p	d	q ²
All valid completes, no exclusions	887	0.720	0.632	0.088	F(1, 883) = 11.23	0.0004	0.22	40.9%
Primary: attention check + IMC	827	0.721	0.636	0.085	F(1, 823) = 9.52	0.0011	0.21	39.6%
Strict: + consequence comprehension	806	0.721	0.643	0.078	F(1, 802) = 7.85	0.0026	0.20	39.8%

The effect size d is standardized by the pooled standard deviation, the convention of §5;

under the root mean square error of the four-cell model the primary-sample value is 0.22. The contrast is significant under all three rules and the effect size is stable. The difference narrows slightly as the rule tightens, because the additional exclusions remove proportionally more participants from the resolved conditions.

E.2 Pairwise comparisons and equivalence tests

Comparison	Difference	95% CI	Welch p
Uncertain against No-Reveal	-0.021	[-0.094, 0.052]	0.567
Elevated against Specific	0.093	[0.014, 0.172]	0.021
Elevated against pooled ambiguity-preserved	-0.038	[-0.104, 0.029]	0.265

Two one-sided tests are reported below **for completeness only**. No equivalence claim is made

anywhere in the paper. The Specific-versus-Elevated comparison is examined in the main text as a secondary boundary question and is reported there as a difference with a confidence interval, which is the appropriate form given that the two conditions do differ reliably.

Smallest effect size of interest	Comparison	Difference	TOST p	Conclusion
±0.05	No-Reveal against Uncertain	0.021	0.219	not shown equivalent
±0.05	Specific against Elevated	-0.093	0.859	not shown equivalent
±0.10	No-Reveal against Uncertain	0.021	0.017	within the bound

Smallest effect size of interest	Comparison	Difference	TOST p	Conclusion
±0.10	Specific against Elevated	-0.093	0.434	not shown equivalent

The two ambiguity-preserved conditions fall within a ±0.10 bound of one another; the two resolved conditions do not fall within either bound. This is consistent with the main text, which pools the ambiguity-preserved cells and declines to treat the resolved cells as interchangeable.

E.3 Robustness to functional form

The dependent variable is a proportion bounded at 0 and 1 with mass at both ends, so the primary linear analysis is re-estimated under alternative specifications.

Specification	Estimate on the ambiguity indicator	SE	p
OLS	0.0847	0.0274	0.0021
OLS with HC3 robust standard errors	0.0847	0.0274	0.0020
Fractional logit (HC3)	0.3903	0.1267	0.0021
Mann-Whitney U, one-tailed	U = 94,266		0.0027
Kruskal-Wallis across the four cells	H(3) = 15.47		0.0015

The binary margin behaves the same way: 56.9% of participants in the ambiguity-preserved conditions declared their full income against 49.8% in the resolved conditions ($\chi^2(1) = 3.96$, $p = 0.047$).

E.4 Covariate-adjusted models

Model	Estimate on the ambiguity indicator	SE	p	n
Unadjusted	0.0847	0.0274	0.0020	827
+ age, risk preference	0.0811	0.0271	0.0028	827
+ self-employment experience, Self-Assessment filing	0.0831	0.0270	0.0021	827
+ household income	0.0886	0.0275	0.0013	827

All models use HC3 robust standard errors. Inferences are unchanged.

E.5 Manipulation checks and mechanism items

Recognition of whether a specific maximum fine had been stated, by condition:

Condition	Correctly recognized
No-Reveal	95.6%
Uncertain	93.7%
Specific £1,900	68.4%
Elevated £10,000	42.4%

Mechanism items, ambiguity-preserved versus resolved conditions:

Item	Ambiguous	Resolved	Welch t	p
Felt uncertainty about the penalty	6.26	5.49	3.78	0.0002
Perceived realism of the information	7.23	6.62	3.56	0.0004
Worry about the consequences	7.01	6.75	1.24	0.215
Perceived penalty severity	7.60	7.42	1.19	0.235
Perceived audit likelihood	46.29	48.48	-1.23	0.217

Perceived audit likelihood does not differ across the four conditions ($F(3, 823) = 1.45, p = 0.227$), confirming that the manipulation moved beliefs about penalty ambiguity and not beliefs about the probability of being audited.

Perceived realism as a check on the stimuli. The realism item asks, “In your opinion, how realistic was the penalty information mentioned in the notice?” It is a check on the stimuli rather than a mechanism, and is reported separately here for that reason. Cell means:

Condition	Perceived realism	(SD)	n
No-Reveal	7.40	(2.36)	206
Uncertain	7.06	(2.24)	207
Specific £1,900	6.45	(2.62)	209
Elevated £10,000	6.80	(2.63)	205

Three features of this pattern indicate that the difference in declarations does not follow the credibility of the stimuli. First, perceived realism is only weakly related to the amount declared: $r = 0.10$ across all conditions, and 0.02, 0.03, 0.19 and 0.07 within the four conditions in the order above. Second, the Elevated notice named the least plausible amount and was nevertheless rated no less realistic than the Specific notice (6.80 against 6.45, Welch $p = 0.176$), which reverses the ordering a credibility explanation requires. Third, the item is measured after the treatment and after the declaration, so it records a consequence of the manipulation rather than a characteristic that could confound it, and adjusting the ambiguity contrast for it would be conditioning on a post-treatment variable. For completeness, such an adjustment leaves the contrast coefficient at 0.038 against 0.042 unadjusted, absorbing the tenth of the effect that travels through the item. No mediation through perceived realism is claimed.

E.6 Recognition-restricted analysis

Reported in the main text at §5.4 and repeated here in full. This is a post-hoc split on a post-treatment variable and is presented as interpretation rather than as a causal test. Restricting the sample to participants who correctly recognized whether a specific maximum had been stated:

Condition	Mean	n
No-Reveal	0.738	197
Uncertain	0.725	194
Specific £1,900	0.585	143
Elevated £10,000	0.665	87

N = 621; ambiguity-preserved 0.732 versus resolved 0.615, difference 0.116; contrast $F(1, 617) = 10.44$, one-tailed $p = 0.0006$, $d = 0.30$; residual between-cells $F(2, 617) = 2.47$, $p = 0.086$, $\eta^2 = 32.1\%$. Where the £10,000 disclosure registered as a specific figure, the Elevated cell behaved as a resolved condition and the residual between-cells variance ceased to be significant.

E.7 Mediation through felt uncertainty

Path	Estimate	p
a: ambiguity → felt uncertainty	0.774	0.0002
b: felt uncertainty → declared income	-0.0010	0.836
c': direct effect	0.0855	0.002

Indirect effect $a \times b = -0.0008$, 95% bootstrap confidence interval $[-0.0081, 0.0071]$ (5,000 resamples). The interval includes zero. The manipulation moves felt uncertainty strongly, but felt uncertainty does not predict declarations, so there is no indirect effect to claim. The mechanism items are presented in the main text as convergent evidence that the manipulation moved the intended construct, and the process question is assigned to Experiment 4's real-time physiological measures.

E.8 Instrument

The Experiment 3 instrument is reproduced in Online Appendix G.

Online Appendix F: Experiment 4 Supplements

All statistics below are computed from the merged two-wave dataset ($N = 239$). Treatment labels are standardized across every table in this appendix and in the main text.

F.1 Design, waves, and the composition of the eight conditions

Experiment 4 pools two laboratory waves. The condition set is **not** identical across them, and the composition is set out here in full because it bears on how the omnibus F-test should be read.

Condition	Wave 1	Wave 2	Total
Control (costs further action)	18	23	41
Fine (non-specific)	18	23	41
Fine (minimum)		32	32
Fine (average)	17	21	38
Fine (maximum)	15	24	39
Jailtime (non-specific)	19		19
Jailtime (average)	15		15
Jailtime (maximum)	14		14
Total	116	123	239

Four conditions run in both waves. The three Jailtime conditions appear only in the first wave and the Fine (minimum) condition only in the second. Treatment is therefore partially confounded with wave, and the eight-cell omnibus F-test compares some cells that were never fielded concurrently.

This composition does not alter the central result, and we checked it directly.

Restricting to the four conditions present in both waves, there is no wave main effect on pupillary response (3.35 vs 4.40; $F(1, 151) = 1.12$, $p = 0.291$) and no treatment $\times$ wave interaction ($F(3, 151) = 0.97$, $p = 0.408$). In addition, the one significant pupillary contrast, the reduction in pupillary response under the minimum-fine message, can be estimated **within the second wave alone**, where both it and the Control condition were fielded. Doing so removes the confound entirely and yields a larger estimate:

Estimate of the Fine (minimum) effect on pupillary response	Difference	p
Pooled across waves, against the Control baseline (main text)	-4.367	0.001
Within the second wave only	-5.297	0.005

We report the pooled estimate in the main text because it uses the full sample, and record the within-wave estimate here because it is the cleaner comparison and it points the same way.

F.2 Background characteristics

Condition	n	Mean age	% female	Risk propensity
Control (costs further action)	41	21.95	46.3	4.37
Fine (non-specific)	41	22.88	63.4	4.46
Fine (minimum)	32	21.28	50.0	5.06
Fine (average)	38	23.42	65.8	5.29
Fine (maximum)	39	22.15	59.0	4.18
Jailtime (non-specific)	19	23.11	79.0	4.26
Jailtime (average)	15	22.73	66.7	4.33
Jailtime (maximum)	14	22.64	64.3	4.50

Conditions do not differ on age ($F(7, 231) = 1.35$, $p = 0.227$), risk propensity ($F(7, 231) = 0.94$, $p = 0.476$), or gender ($\chi^2(7) = 8.48$, $p = 0.292$).

F.3 Compliance by condition

Stated likelihood of paying, 0-10 scale. Omnibus $F(7, 231) = 3.13$, $p = 0.004$.

Condition	n	Mean (SD)	vs Control	p
Control (costs further action)	41	8.54 (1.45)		
Fine (non-specific)	41	8.61 (1.36)	0.073	0.845
Fine (minimum)	32	7.69 (2.48)	-0.849	0.034
Fine (average)	38	8.79 (2.02)	0.253	0.507
Fine (maximum)	39	9.39 (1.18)	0.848	0.026
Jailtime (non-specific)	19	9.11 (1.05)	0.569	0.226
Jailtime (average)	15	9.27 (1.16)	0.730	0.153
Jailtime (maximum)	14	8.36 (2.27)	-0.179	0.732

$R^2 = 0.087$, $N = 239$. Only the minimum-fine and maximum-fine messages differ from the control letter. §6.4 of the main text discusses why this pattern does not conflict with Experiment 2.

The previous submission reported this outcome in four specifications, which are reproduced here: (I) message-condition indicators only; (II) with age, gender and risk propensity; (III) the fine and jailtime message groups pooled against the control letter; (IV) the pooled groups with the same covariates. Standard errors in parentheses.

	(I)	(II)	(III)	(IV)
Fine (non-specific)	0.073 (0.373)	0.147 (0.375)		
Fine (minimum)	-0.849 (0.398)	-0.794 (0.399)		
Fine (average)	0.253 (0.380)	0.416 (0.388)		

	(I)	(II)	(III)	(IV)
Fine (maximum)	0.848 (0.378)	0.871 (0.377)		
Jailtime (non-specific)	0.569 (0.469)	0.673 (0.475)		
Jailtime (average)	0.730 (0.509)	0.799 (0.511)		
Jailtime (maximum)	-0.179 (0.523)	-0.105 (0.523)		
Fine group (pooled)			0.123 (0.307)	0.191 (0.309)
Jailtime group (pooled)			0.401 (0.371)	0.456 (0.375)
Age		-0.021 (0.033)		-0.007 (0.033)
Male		0.271 (0.228)		0.202 (0.233)
Risk propensity		-0.086 (0.046)		-0.099 (0.047)
Constant	8.537 (0.264)	9.229 (0.797)	8.537 (0.272)	9.020 (0.811)
R ²	0.087	0.105	0.006	0.026
N	239	239	239	239

Table note. Male is an indicator equal to one for male participants. In the corresponding table of the previous submission the sex variable was entered in its raw two-level coding, so the row there labeled Male carried the female direction and the constants differed accordingly; the treatment coefficients are unaffected. Inferences are unchanged by covariate adjustment: the minimum-fine and maximum-fine messages differ from the control letter in both specifications, and the pooled message groups do not.

F.4 Pupillary and attention measures by condition

Cell means, with standard deviations in parentheses.

Condition	Pupil % change	Fixation count	Fixation duration (s)
Control (costs further action)	3.38 (6.50)	239.8 (83.64)	52.99 (19.17)
Fine (non-specific)	3.96 (4.73)	256.0 (54.75)	58.77 (14.27)
Fine (minimum)	-0.98 (6.53)	237.5 (94.00)	53.84 (21.58)
Fine (average)	3.38 (6.33)	258.0 (56.48)	63.05 (12.96)
Fine (maximum)	5.10 (5.62)	248.7 (72.02)	56.88 (17.82)
Jailtime (non-specific)	5.38 (3.18)	221.3 (97.50)	53.65 (24.13)
Jailtime (average)	2.48 (4.70)	267.8 (44.41)	59.67 (7.62)
Jailtime (maximum)	4.56 (6.58)	269.5 (26.93)	58.14 (7.32)

Fixation duration is the total duration of fixations within the area of interest, recorded in milliseconds by the eye-tracker and reported here in seconds. Omnibus F-tests: pupillary response $F(7, 231) = 3.65$, $p < 0.001$ ($N = 239$); fixation count $F(7, 228) = 1.01$, $p = 0.423$ ($N = 236$); fixation duration $F(7, 228) = 1.37$, $p = 0.220$ ($N = 236$). This is the dissociation reported in §6.3 and plotted in Figure 3.

F.5 Pupillary response regressed on condition

Ordinary least squares, Control as the omitted baseline (intercept 3.383).

Condition	Coefficient	SE	t	p	95% CI
Fine (non-specific)	0.571	1.270	0.45	0.653	[-1.931, 3.074]
Fine (minimum)	-4.367	1.356	-3.22	0.001	[-7.040, -1.695]
Fine (average)	-0.004	1.295	-0.00	0.997	[-2.556, 2.547]
Fine (maximum)	1.715	1.286	1.33	0.184	[-0.819, 4.250]
Jailtime (non-specific)	1.993	1.596	1.25	0.213	[-1.151, 5.138]
Jailtime (average)	-0.908	1.735	-0.52	0.601	[-4.327, 2.511]
Jailtime (maximum)	1.174	1.780	0.66	0.510	[-2.333, 4.682]

$R^2 = 0.100$, $N = 239$. Only the minimum-fine message differs significantly from the control letter. The cells with the largest positive point estimates, non-specific jailtime and the maximum fine, are directionally elevated but not significant, and are described as such throughout.

Corrections to the previous submission's Experiment 4 tables. Re-deriving every Experiment 4 statistic from the raw data identified four transcription errors in the previously submitted tables, corrected here and in the main text. (a) In the pupillary regression, the Fine (maximum) and Jailtime (maximum) rows were transposed: 1.715 belongs to Fine (maximum) and 1.174 to Jailtime (maximum). (b) In the pupillary descriptives, the standard deviation printed for Jailtime (maximum) repeated the value of an adjacent row; the cell computes as 6.58. (c) In the mediation table, the t statistic of the direct effect was printed as 1.258 and computes as 1.528. (d) The pupillary omnibus F-test was reported as $p = 0.04$; it is $p < 0.001$.

F.6 Indirect-effect analyses

Reports in full the mediation summarized in §6.3 and shown in Figure 4 of the main text, in the specification of the previous submission.

The model is a single-mediator model (PROCESS Model 4). The message-framing variable, the treatment identifier entered as one continuous predictor taking the values 1 to 8 across the eight message conditions, is the independent variable; the baseline-corrected pupillary response is the mediator; and the stated likelihood of paying is the dependent variable. No

covariates are included, and the confidence interval is a 95 percent percentile bootstrap interval computed on 5,000 resamples.

Path	Estimate	SE	t	p	95% CI
a: message framing → pupillary response	-0.349	0.161	-2.168	0.031	[-0.666, -0.032]
b: pupillary response → stated likelihood of paying	0.246	0.011	23.511	< 0.001	[0.225, 0.267]
c': direct effect of message framing	0.040	0.026	1.528	0.128	[-0.012, 0.092]

The indirect effect of message framing on the stated likelihood of paying through pupillary response is -0.086 (bootstrap SE 0.044), with a 95 percent percentile bootstrap confidence interval of [-0.174, -0.002], which does not include zero.

F.7 Apparatus and pre-processing

Eye movements and pupil diameter were recorded with a Tobii TX300 eye-tracking system at 300 Hz. The letter from the tax authority constituted the area of interest, identical in size across treatments; the pupillary measure was computed from the samples recorded while the participant's gaze was fixated on it, and the fixation measures were computed within the same area of interest. Pupillary response is expressed as a percentage change relative to baseline, the baseline being the mean pupil diameter over the full 60-second baseline period rather than over a subinterval of it, following Hayes and Petrov (2016); that period was recorded at the start of the session while the participant viewed a fixation cross, before the participant saw any treatment message, and the measure was processed through the procedure documented in Hecht et al. (2025). The measure analyzed throughout is the baseline-corrected pupil percentage change. Fixation duration and fixation count index visual attention (Lynch & Andiola, 2019).

Of 250 participants, 11 were excluded: six for failing the instructional manipulation check and five for not meeting the pupillometry data-quality threshold, which required at least 85 percent of pupil samples to be present during the presentation of the enforcement letter, leaving 239 for analysis. Fixation measures are available for 236 of these.

Online Appendix G: Experimental Instruments

Sources. Each instrument is reproduced from the archived final instrument document of the study concerned and has been checked against the archived Qualtrics export where one is held. The selection of items follows the abridged form filed with the previous submission; the wording follows the administered instruments.

Numbering. The source documents use the previous numbering. Study A was Experiment Two, the present Experiment 2 was Experiment Three, and Experiment 4 was Experiment Four; the present Experiment 3 was built as “Experiment 5B”. The headings below use the numbering of the revised manuscript and give the former label in parentheses.

Conventions. Participant-facing text is reproduced verbatim. Text in square brackets is a researcher annotation describing routing, randomization or screen structure and was never shown to participants. Response scales were presented as 11 radio buttons; only the anchors are reproduced here, in the wording participants saw.

One label has been changed, deliberately. The construct label “Willingness to pay”, used as a researcher heading in the previously filed instruments document, is replaced throughout by the label of the measure as reported, the likelihood of paying. This is a change to the researcher’s heading only; every participant-facing item is reproduced in the administered wording.

Corrections to the previously filed instruments document. Checking each instrument against the archived original instruments and the archived Qualtrics exports identified three respects in which the previously filed document did not reproduce the administered instruments, and they are corrected here. (a) The dependent-variable item in Study A and in Experiment 4 asks “How likely are you to pay the money to the government?”; the previous filing printed

Experiment 2's wording, "the outstanding tax obligation", for all three studies. The two wordings and their source are reported in the manuscript (§4.1 and §6.2). (b) The previous filing showed an instructional-manipulation-check screen within Study A; the archived Study A instrument and its export contain no such item, and it is not reproduced. (c) In Experiment 4, the instructional manipulation check and the demographic items were administered within the laboratory session rather than as screens of the survey reproduced in Appendix G.4, and Study A's consent statement and scenario are restored to the administered wording.

G.1 Study A (previous Experiment Two)

Between-participants message-exposure design; each participant saw one of seven enforcement messages.

Screen 1: Consent.

EXPLANATORY STATEMENT

You are invited to take part in this research study.

This is a short survey about taxes and personal finance. Please be assured that all your answers will be strictly confidential. The survey should take around 3 minutes. You may leave the study at any point.

If you agree to take part in this research study, please click the "Yes, I consent" option below to proceed.

_____ Yes, I consent. _____ No, I do not consent.

Screen 2: Scenario.

Imagine that you own a cash constrained business. You have previously declared the total amount of taxes you owe to the government. However, you have not yet fulfilled your tax obligation. Consequently, the government sent you a letter

reminding you of the amount you owe. Because the government did not receive a response from you to this letter, it sends a follow up letter with the following statement:

Screen 2, continued: the manipulation. Each participant saw one of the following seven statements. Every statement opens with the same clause, so the messages differ only in the consequence they name.

Condition	Statement
Costs Further Action	‘We told you recently that you have an outstanding tax obligation with the government. If you pay this debt now, you can stop us taking any further action.’
Fine, non-specific	‘...If you do not pay your taxes, you may be subject to a fine.’
Jailtime, non-specific	‘...If you do not pay your taxes, you may be subject to prosecution that results in time in prison.’
Fine, average	‘...If you do not pay your taxes, we could fine you 5% of your current outstanding tax obligation.’
Jailtime, average	‘...If you do not pay your taxes, you could face 6 months in prison.’
Fine, maximum	‘...If you do not pay your taxes, you could face a fine of up to 100% of the outstanding tax obligation and the possibility that the government seizes your personal assets (e.g. property).’
Jailtime, maximum	‘...If you do not pay your taxes, you could face up to 7 years in prison.’

The elision marks the shared opening clause, which was presented in full in every condition.

Screen 3: Likelihood of paying (*dependent variable*).

Referring to the letter above, please answer the following questions.

How likely are you to pay the money to the government?

Please give your answer out of 10, where 0 means you are certain not to pay and 10 means you are certain to pay.

Anchors: 0 (certain not to pay) ... 10 (certain to pay).

Screen 4: Perceived punishment severity and perceived realism.

How severe do you think the punishment would be if you do not respond to the letter?

Please give your answer out of 10, where 0 means no action and 10 means the maximum penalty possible.

Anchors: 0 (no action) ... 10 (maximum penalty possible).

In your opinion, how realistic was the punishment mentioned in the letter?

Anchors: 0 (not at all realistic) ... 10 (very realistic).

Screen 5: Risk propensity.

How willing are you to take risks, in general?

Anchors: 0 ... 10.

Screen 6: Demographics.

What is your gender? Male / Female / Other (please specify) _____

What is your age? _____

G.2 Experiment 2 (previous Experiment Three)

Participants were first randomly assigned to the £2,500 or the £500 outstanding-balance condition. Within each balance condition they were assigned to a fine *type* (minimum or maximum) and then to an information condition (No-Updating: prior belief elicited, nothing revealed; Updating: prior belief elicited, then the actual fine revealed). That fine *type* and fine amount were both manipulated, which was not evident from the previous version's description, is visible directly in the screens below: the *type* is carried by the wording of the notice ("a minimum fine of at least" versus "a maximum fine of at most"), while the *amount* appears only on the later reveal screen, and only in the Updating conditions.

Screen 1: Consent.

EXPLANATORY STATEMENT

You are invited to take part in this research study.

This is a brief survey about taxes and personal finance. Your responses will remain completely confidential. The survey is designed to take approximately 2 minutes, and you are free to withdraw from the study at any time. Please note that a simple attention check question will appear in the survey. If this question is answered incorrectly, the survey will end, and reimbursement will not be provided.

If you consent to take part in this study, please select the “Yes, I consent” option below to begin.

_____ Yes, I consent. _____ No, I do not consent.

Screen 2: Scenario. Shown with the participant’s assigned balance.

Imagine you own a cash constrained business with an outstanding tax balance of £2,500.

[The £500 condition was identical with “£500” substituted.]

Screen 3: Attention check.

Please recall the information from the previous page:

What is the outstanding tax balance that your cash constrained business owes to the government?

£500 / £1,000 / £2,500 / £5,000

Screen 4: The notice.

As noted earlier, you own a cash-constrained business with an outstanding tax balance of £2,500. You have not yet paid this amount to the government. As a result, the government sent you a reminder letter about the outstanding balance. Since you did not respond to this letter, the government has now issued a follow-up letter with the following statement:

Participants in the **minimum**-fine conditions then saw:

‘We recently informed you that you have an outstanding tax obligation to the government. If you fail to pay your taxes, you may face a minimum fine of at least £_____ GBP’

Participants in the **maximum**-fine conditions saw:

‘We recently informed you that you have an outstanding tax obligation to the government. If you fail to pay your taxes, you may face a maximum fine of at most £_____ GBP’

The amount was left blank for every participant at this stage. This is what makes the design a test of ambiguity rather than of magnitude: the notice establishes that a minimum or a maximum exists without stating its value.

Screen 5: Prior belief (*elicited before any disclosure, from every participant*).

What is your belief about the maximum fine that the UK tax authority can impose on you? (The amount that is presented as £_____ GBP above)

Write your belief about the maximum fine in £ (GBP) below: _____

[In the minimum-fine conditions “maximum” was replaced by “minimum” throughout the item.]

Screen 6: The information manipulation and the dependent variable. All participants saw the lead-in “Based on the context presented earlier:”. In the **No-Updating** conditions the dependent-variable item followed directly; in the **Updating** conditions one additional sentence preceded it on the same screen:

Condition	Sentence shown
Minimum revealed, £2,500 balance	Assume the actual minimum fine for not paying your tax obligation is £800.
Maximum revealed, £2,500 balance	Assume the actual maximum fine for not paying your tax obligation is £1,900.
Minimum revealed, £500 balance	Assume the actual minimum fine for not paying your tax obligation is £700.

Condition	Sentence shown
Maximum revealed, £500 balance	Assume the actual maximum fine for not paying your tax obligation is £1,600.

Six further conditions were administered separately: two at the £2,500 balance (the Elevated and Uncertain cells of the four-cell comparison reported in §4.3) and four at the £500 balance (the second collection period's conditions, reported in §4.3 and tabulated in Appendix D.4):

Condition	Sentence shown
Elevated maximum, £2,500 balance	Assume the actual maximum fine for not paying your tax obligation is £10,000.
Uncertain maximum, £2,500 balance	Assume the actual maximum fine is set by the regulator based on the severity of non-compliance. The maximum fine amount may increase depending on factors such as payment delay duration, compliance history, and other circumstances determined by the regulatory authority.
Elevated maximum, £500 balance	Assume the actual maximum fine for not paying your tax obligation is £10,000.
Uncertain maximum, £500 balance	Assume the actual maximum fine is set by the regulator based on the severity of non-compliance. The maximum fine amount may increase depending on factors such as payment delay duration, compliance history, and other circumstances determined by the regulatory authority.
Second maximum revealed, £500 balance	Assume the actual maximum fine for not paying your tax obligation is £1,200.
Second minimum revealed, £500 balance	Assume the actual minimum fine for not paying your tax obligation is £600.

The £1,900 and £800 values were calibrated against taxpayers' own beliefs in a separate elicitation study (N = 199, May 2024), in which each taxpayer estimated the minimum and the maximum penalty for one of the two debt levels; the mean estimated maximum for a £2,500 debt was £1,960.55 (n = 99). £1,600 and £700 are the corresponding values for the £500 debt.

The dependent-variable item, presented on the same screen as the information manipulation:

How likely are you to pay the outstanding tax obligation to the government? (0
= Certain not to pay, 10 = Certain to pay)

[The anchors were displayed within the item as shown; the response was one of eleven radio buttons.]

Screen 7: Covariates and demographics (*presented to all participants*).

How willing are you to take risks, in general? [0 ... 10]

Do you currently hold, or have you previously held, a position that involves supervisory duties or leadership responsibilities? Yes / No

Have you ever held a position with management responsibilities (e.g., managing projects, budgets, or team members)? Yes / No / Prefer not to say

What is your gender? Male / Female / Other (please specify) _____

What is your age? _____

What was your total household income before taxes during the past 12 months? (Please select one): twelve bands from “Less than £10,000” to “More than £150,000”.

[Note on a labeling error in the source document.] In the previously filed instruments document, the reveal screen for the £500-balance minimum-fine Updating condition is headed “[Updating Minimum Fine Condition (£2,500 Balance)]” while the value it reveals is £700, which is the £500-balance value. The heading is a copy-and-paste error in that document; the revealed value is correct, and Appendix G.2 above labels the condition by the balance it actually belongs to.

G.3 Experiment 3 (*built as Experiment 5B; incentivized income reporting*)

Sixteen Qualtrics screens. The four conditions differ **only** in whether and how the maximum fine is resolved; every other screen is identical across conditions. Declarations were incentive-compatible: 10,000 tokens = £1.00 of real bonus, paid through Prolific.

Screen 1: Consent.

EXPLANATORY STATEMENT

You are invited to take part in this research study.

This is a brief survey about taxes and personal finance (about 5 minutes). It includes a decision task with a real bonus payment: depending on your decisions and on chance, you can earn a bonus of up to £1.00 on top of your guaranteed participation payment. Your guaranteed participation payment cannot be reduced. Your responses will remain completely confidential, and you are free to withdraw at any time. The survey includes a short attention check and questions about the scenario. Please read each question carefully.

If you consent to take part in this study, please select the “Yes, I consent” option below to begin.

_____ Yes, I consent. _____ No, I do not consent.

Screen 2: Scenario and payoff rules. Identical in all four conditions. The audit probability was a real 25% and was never stated to participants.

PLEASE READ THE FOLLOWING CAREFULLY. The information below determines the decisions you will make and the bonus you can earn.

Imagine you own a cash-constrained business. For this task, your business earned 10,000 tokens of additional taxable profit this year, on top of your other income. The tokens determine a real bonus payment: 10,000 tokens = £1.00.

You must decide how much of the 10,000 tokens to declare on your tax return. The tax rate in this task is 40% of the amount you declare.

Declaring more income means paying more tax now. Declaring less income means paying less tax now, which can increase your bonus if your return is not checked.

Some tax returns are checked by the tax authority. If your return is checked and you have under-declared your income, the under-declared income will be identified. You will then pay the additional tax due plus a penalty, as described in the notice you will receive. Any tax and penalty are deducted only from your bonus. Your guaranteed Prolific payment cannot be reduced, and your bonus cannot go below £0.00.

The bonus rules described here are real and will be applied exactly as stated. At the end of the survey you will see whether your return was checked and your final bonus.

The 40% rate is anchored to the UK higher-rate income-tax band; Class 4 National Insurance is omitted as a simplification.

Screen 3: Scenario checks. [Used for analysis exclusions only; all participants were paid regardless of their answers.]

In this task, how much additional taxable profit did your business earn?

500 tokens / 1,000 tokens / **10,000 tokens** / 25,000 tokens

In this task, what happens if your return is checked and you have under-declared your income?

Nothing happens / You only receive a warning / **You pay the additional tax due plus any penalty described in the notice, deducted from your bonus** / Your business is automatically closed

Correct responses in bold. These are `attn_check` and `conseq_check`; the primary sample requires `attn_check` together with the instructional manipulation check at screen 13.

Screen 4: The notice and prior belief. Presented identically to all four conditions.

As noted earlier, you own a cash-constrained business that earned 10,000 tokens of additional taxable profit this year, and you are deciding how much of this income to declare on your tax return. The government has issued the following notice:

‘We remind you that you must declare your full income to the government. If you under-declare your income, you may be subject to a maximum fine of at most _____ tokens.’

What is your belief about the maximum fine that the tax authority can impose on you? (The amount presented as _____ tokens above)

Write your belief about the maximum fine in tokens below: _____

Screen 5: The manipulation. All four conditions saw the lead-in “Based on the context presented earlier:”. This is the only screen that differs between conditions.

Condition	Text following the lead-in
No-Reveal (<i>ambiguity-preserved baseline</i>)	Nothing further. Participants were routed directly to the declaration decision.
Specific	Assume the actual maximum fine for under-declaring your income is 1,900 tokens.
Elevated	Assume the actual maximum fine for under-declaring your income is 10,000 tokens.
Uncertain	Assume the actual maximum fine is set by the regulator based on the severity of non-compliance. The maximum fine amount may increase depending on factors such as the amount under-declared, your compliance history, and other circumstances determined by the regulatory authority.

The four conditions port the four maximum-fine cells of Experiment 2 onto the reporting margin, with tokens substituted for pounds and “under-declaring your income” for “not paying your tax obligation”.

[How the ambiguous cells were paid. For No-Reveal and Uncertain the applicable maximum was drawn at random from {1,900; 4,600; 7,300; 10,000} tokens with equal

probability, **after** the declaration and independently of it. The applied fine was the drawn maximum multiplied by the under-declared share, so the fine did increase with the amount under-declared, consistent with the notice wording, while its ceiling was unpredictable. The endpoints of the pool are the two revealed maxima and its mean is their midpoint, so the ambiguous cells are constructed to be neutral in expectation relative to the resolved cells rather than more or less severe. Because the disclosed floor is £0, a participant who reasons about the worst case in the ambiguous cells faces a bounded risk, which biases the design *against* the ambiguity effect.]

Screen 6: Declaration decision (*primary dependent variable, incentivized*).

Reminder: The tax rate is 40% of the amount you declare. Some tax returns are checked. If your return is checked and you have under-declared, you will pay the additional tax due plus the penalty described in the notice, deducted from your bonus. Your bonus cannot go below £0.00.

Referring to the notice, please answer the following question.

Your business earned 10,000 tokens of additional taxable profit this year. How much of this income will you declare on your tax return?

Please enter an amount between 0 and 10,000 tokens: _____

[Integer 0-10,000, forced response. Note the wording “will you declare”: this is a real decision with money attached, not a stated intention. The analysis divides this by 10,000.]

Screen 7: Projective declaration (*secondary*).

Now, thinking about a typical self-employed business owner in this situation, how much of the 10,000 tokens do you think they would declare on their tax return?

Please enter an amount between 0 and 10,000 tokens: _____

[Deliberately placed after the participant's own incentivized decision so that it cannot anchor it.]

Screen 8: Mechanism measures. [All four items post-decision, in randomized order, each 0-10.]

How uncertain do you feel about the penalty you could face for under-declaring your income? [0 = not at all uncertain, 10 = extremely uncertain]

How worried are you about the possible consequences of under-declaring your income? [0 = not at all worried, 10 = extremely worried]

How severe do you think the penalty would be if you under-declare your income and your return is checked? [0 = no action, 10 = maximum penalty possible]

In your opinion, how realistic was the penalty information mentioned in the notice? [0 = not at all realistic, 10 = completely realistic]

Screen 9: Perceived audit likelihood. [Asked after the declaration but before the outcome was revealed, so that the realized audit result could not contaminate it.]

Before you see whether your return was checked: what do you think the chance was that your return would be checked?

Please enter a number from 0 to 100, where 0 means no chance and 100 means certain to be checked: _____

Screen 10: Payoff calculation. [A JavaScript payoff engine attached to this screen read the declaration, drew the audit at 25%, drew the maximum fine for No-Reveal and Uncertain, and computed the bonus before the next screen loaded.]

Your declaration has been recorded.

Click Next to see whether your return was checked and to see your final bonus.

Screen 11: Outcome. One of three variants was shown, by display logic on the realized audit draw.

Your return was checked. You had under-declared your income, so the additional tax due and a penalty were deducted from your bonus, as described in the notice. Your final bonus for this task is £[amount].

Your return was checked. You declared your full income, so no penalty applies. Your final bonus for this task is £[amount].

Your return was not checked. Your final bonus for this task is £[amount].

[All variants:] This bonus will be paid to your Prolific account after the study has been reviewed.

Screen 12: Manipulation check. [For analysis only; not a payment condition.]

What did the government notice (and the information that followed) say about the maximum fine for under-declaring income?

It stated an exact maximum fine amount. (*correct for Specific and Elevated*)

It said the maximum fine is set case-by-case and could increase. (*correct for Uncertain*)

It mentioned a maximum fine but did not provide the actual amount. (*correct for No-Reveal*)

What exact maximum fine did the information state? If no exact amount was stated, select the last option.

1,900 tokens / 5,000 tokens / 10,000 tokens / It did not state an exact amount

Screen 13: Instructional manipulation check.

Note, answer ‘yellow’ to the next question that asks the colour of the sky.

What colour is the sky? Blue / White / Yellow / Orange / Purple

Screen 14: Risk propensity.

How willing are you to take risks, in general?

Please give your answer out of 10, where 0 means not at all willing and 10 means very willing.

Screen 15: Covariates and demographics.

Do you currently hold, or have you previously held, a position that involves supervisory duties or leadership responsibilities? Yes / No

Have you ever held a position with management responsibilities (e.g., managing projects, budgets, or team members)? Yes / No

Have you ever filed a Self-Assessment tax return in the UK? Yes / No

Have you ever had income from self-employment or running a business? Yes / No

What is your gender? Male / Female / Other (please specify) _____

What is your age? _____

What was your total household income before taxes during the past 12 months? (Please select one): twelve bands from “Less than £10,000” to “More than £150,000”.

Screen 16: Completion.

Thank you. Your responses have been recorded.

Your bonus of £[amount] will be paid to your Prolific account after the study has been reviewed (normally within a few days).

Please click the button below to return to Prolific and register your submission.

G.4 Experiment 4 (previous Experiment Four; pupillometry and eye-tracking)

Between-participants message-exposure design, run in the laboratory across two waves. The instrument presents Study A's scenario and messages with one condition added.

Screen 1: Participant identifier.

Please type in your Participant ID (NOT your Student ID) that was provided to you at the beginning of this session: _____

Screen 2: Scenario.

Imagine that you own a cash constrained business. You have previously declared the total amount of taxes you owe to the government. However, you have not yet fulfilled your tax obligation. Consequently, the government sent you a letter reminding you of the amount you owe. Because the government did not receive a response from you to this letter, it sends a follow up letter with the following statement:

Screen 2, continued: the manipulation. Each participant saw one of eight statements: the seven of Study A, reproduced word for word at Appendix G.1, plus a Fine-minimum condition.

Condition	Statement
Fine, minimum (added)	'We told you recently that you have an outstanding tax obligation with the government. If you do not pay your taxes, you may be subject to a \$100 fine.' ²

The other seven statements are identical to Study A's. Which conditions ran in which wave, and why the eight conditions have unequal cell sizes, is set out at Appendix F.1: the Jailtime conditions ran only in wave 1 and Fine-minimum only in wave 2, with four conditions in both.

² The dollar amount reflects the Australian setting of the laboratory experiment (§6.2); the studies run with UK taxpayers are denominated in pounds.

Screen 3: Likelihood of paying (*dependent variable*).

Referring to the letter above, please answer the following questions.

How likely are you to pay the money to the government?

Please give your answer out of 10, where 0 means you are certain not to pay
and 10 means you are certain to pay.

Anchors: 0 (certain not to pay) ... 10 (certain to pay).

Screen 4: Risk propensity.

How willing are you to take risks, in general? [0 ... 10]

[The instructional manipulation check and the demographic items were administered
within the same laboratory session, outside the survey reproduced above.]

Pupillary and eye-tracking measures were recorded while participants read the message
screen; the apparatus, sampling and pre-processing are described at Appendix F.7.