

Refund Amount, Cancellation Timing, and the Decision to Cancel a Practical Driving Test: A Behavioural Study

TVPZ9, UCL Interaction Centre, University College London, United Kingdom

CONSTANTIN REZLESCU, UCL Interaction Centre, University College London, United Kingdom

Practical driving test capacity in Great Britain is heavily constrained, and every booked slot that is neither used nor released early represents capacity lost to another learner. Current Driver and Vehicle Standards Agency (DVSA) policy applies an all-or-nothing refund rule: candidates who give at least ten working days' notice receive a full refund, while those inside that window forfeit the entire fee, and so have no financial reason to declare a cancellation rather than simply not attend. This study tested experimentally how the size of a refund and the timing of the cancellation decision affect stated willingness to release an unwanted test slot. Two hundred UK-based adults imagined having booked and paid £62 for their first practical test and being uncertain about their readiness. Refund was manipulated between participants at four levels (£0, £15, £31, £62) and held constant within each participant; timing was manipulated across two decisions, a fixed five-day anchor followed by a randomly assigned interval of 1, 3, 7 or 10 days. Refund amount produced a large increase in stated cancellation likelihood, rated on a seven-point scale from 1 (*very unlikely*) to 7 (*very likely*), rising from $M = 1.74$ with no refund to $M = 3.62$ with a full refund, $F(3, 196) = 10.95$, $p < .001$, $\eta^2 = .144$, 95% CI [.056, .225], $d = 1.17$, 95% CI [0.75, 1.60]; the same gradient reappeared when participants rated the identical refund a second time at a different notice period, and survived demographic adjustment. All four conditions nonetheless sat at or below the scale midpoint, so no refund level produced majority willingness to release the slot. No test detected departure from a straight line in pounds, though with four levels this is weak evidence for linearity rather than evidence against a threshold; the contrast between any refund and none was itself substantial ($d = 0.71$), leaving open whether magnitude or mere presence is doing the work. Timing had no reliable effect. Because the fixed anchor yields four contrasts running in opposite directions, these were analysed separately rather than pooled; three were null and the largest movement away from the test showed a decrease in willingness to cancel, opposite to prediction, with the appropriate between-groups test marginal ($p = .052$). Equivalence testing excluded timing effects larger than one scale point ($p = .025$) but not smaller ones. Refund generosity rather than notice period appears to be the operative lever on stated intentions. A central caveat limits the policy reading: the task offered no option to keep the slot and not attend, so it cannot speak directly to whether a refund converts silent non-attendance into early release, which is the behaviour the policy would target.

1 Introduction

1.1 The problem: a constrained testing system and the cost of unused slots

For several years, demand for practical driving tests in Great Britain has outstripped the capacity of the system that delivers them. As of May 2026, the Driver and Vehicle Standards Agency (DVSA) reported a national median waiting time of 9.7 weeks between booking a car practical test and taking it, alongside its longer-standing “10% availability” measure, which stood at 21.8 weeks [4]. These headline figures conceal substantial regional variation: the median wait ranged from roughly 7 weeks in the South West to almost 14 weeks in Scotland, and individual centres in London and the West Midlands recorded median waits above 20 weeks [4]. The agency has not been idle. It delivered close to a quarter of a million additional car tests

Authors' Contact Information: TVPZ9, UCL Interaction Centre, University College London, London, United Kingdom; Constantin Rezlescu, UCL Interaction Centre, University College London, London, United Kingdom.

between June 2025 and May 2026, and raised examiner numbers to their highest level since 2019. Waits in many centres are still long enough to have drawn scrutiny from the National Audit Office [13].

When the supply of test slots is this heavily constrained, the efficient use of each slot becomes a matter of real consequence. Every appointment that is paid for but not converted into a completed test, whether through a late cancellation that cannot be re-sold or through outright non-attendance, represents capacity that could have gone to another learner in the queue. Unused slots do not simply disappear; they lengthen waiting times for others, reduce the number of tests the system delivers relative to its theoretical capacity, and erode public confidence in a service that is already under pressure. The challenge facing DVSA is therefore not only to increase the absolute number of tests available, but to ensure that booked slots reliably convert into completed tests, or, failing that, are released early enough to be offered to someone else.

The rules that govern cancellations are a potentially powerful lever on this conversion. Under current DVSA policy, a candidate who cancels or moves a car test receives a full refund of the fee only if they give at least 10 full working days' notice; inside that window, they forfeit the entire fee [6]. The fee itself ranges from £62 for a standard weekday test to £75 for evenings and weekends [6]. This is an all-or-nothing, cliff-edged structure: the marginal financial consequence of cancelling a single day too late is the loss of the whole fee. Viewed through a behavioural lens, such a structure may have the perverse effect of discouraging precisely the cancellations that would free slots for reallocation. A candidate who has already crossed the refund deadline has no financial incentive to tell DVSA that they will not attend, because the fee is lost whether they formally cancel or simply fail to appear. The slot is then far less likely to be re-offered in time for another learner to use it. The present research asks whether the design of the refund, both its size and its timing, changes people's willingness to cancel an appointment they no longer intend to use.

1.2 Theoretical Background

Four bodies of work bear on the questions asked here. The first three converge on the same prediction about refunds, and this convergence should be stated plainly at the outset, because it constrains what the present design can establish. Sunk-cost reasoning, loss aversion and mental accounting all imply that a larger refund makes cancellation more likely. They differ in the mechanism they propose, not in the direction or ordering of the effect they predict. Section 5.3 works through what each would say about the pattern actually observed, and shows that they come apart only in the interval this design leaves thinnest. A design that varies only the refund amount therefore tests a prediction the three accounts share, and cannot adjudicate between them. They are set out together on that basis, and the study is presented throughout as a test of the shared prediction rather than as a competition between mechanisms.

The sunk-cost account begins from the observation that people take account of unrecoverable prior investment in decisions where only future consequences should matter [3, 16]. A candidate who has paid £62 and secured a scarce appointment has invested both money and effort, and abandoning the slot means writing off that investment. Escalation of commitment predicts that the larger the investment appears, the more reluctant the person will be to walk away, and a refund reduces the portion of the investment that must be written off. Behaviourally this should show up as a candidate keeping an appointment they privately expect to fail, because attending at least converts the outlay into something, whereas cancelling

with no refund converts it into nothing. The prediction for this study is therefore ordinal: the more of the fee that comes back, the weaker that pull should be.

Loss aversion and reference dependence supply a more precise account of why the size of the refund should matter. Under prospect theory, outcomes are evaluated as gains and losses relative to a reference point, and losses are weighted more heavily than gains of equivalent magnitude [11, 20]. Forfeiting the fee is naturally coded as a loss from the status quo of holding a paid-for appointment, an asymmetry closely related to the endowment effect and to status quo bias [9, 10]. Because the disutility of that loss is disproportionate to its monetary size, reducing it should lower the barrier to cancelling by more than the returned amount alone would suggest. The same framework supplies the reference point. With a full refund available, cancellation carries no financial loss at all. Inside the forfeiture window the entire fee is lost whether the candidate cancels or simply fails to attend, which means the two courses of action are financially identical and the choice between them turns on non-financial considerations alone. The behavioural consequence that motivates this study is that a candidate in that position has no monetary reason to notify anyone, and notification is what releases the slot.

Mental accounting extends this by describing how a specific outlay is tracked and closed [17, 18]. The test fee plausibly occupies a dedicated account which cancelling without a refund closes at a loss. A partial refund allows part of the outlay to be recovered and the account to be partly settled, which offers a natural reading of why intermediate refunds might produce intermediate willingness rather than an all-or-nothing response. Mental accounting is also the account most compatible with a threshold rather than a smooth function, since the psychologically significant event may be recovering something rather than recovering a particular amount. This distinction is taken up in the analysis of the shape of the refund effect.

The fourth body of work concerns time rather than money, and yields an independent prediction about the timing of the decision. Hyperbolic discounting and present bias imply that immediate costs and benefits are weighted disproportionately relative to delayed ones [1, 5, 12]. As the test approaches, its costs and benefits become more immediate and more salient, while the effort of cancelling remains constant, which predicts declining willingness to release the slot as the date draws near. A second and partly independent route to the same prediction runs through procrastination, inertia and defaults. Cancelling requires an active step, and not cancelling is the default [8, 15, 19]; present-biased agents systematically defer effortful actions with delayed benefits [14], and deadlines are known to shape whether such actions are completed at all [2]. Both routes predict that greater notice increases the probability of an early cancellation, giving the timing hypothesis two theoretical grounds rather than one.

1.3 Empirical Background

Three applied literatures bear on the present study, and each carries a specific implication for its design or predictions.

The first concerns missed appointments in public services. In two randomised trials with a UK hospital, Hallsworth and colleagues [7] found that adding the cost of a missed appointment to text-message reminders significantly reduced non-attendance. That result matters here on two counts. It shows that non-attendance responds to how financial consequences are made salient, which supports the premise that the refund schedule, and not only its monetary value, is a behavioural lever. More importantly, it suggests that people often fail to attend through inertia and inattention rather than settled intent. If the same is true of driving

tests, then a modest incentive to cancel formally may have a disproportionate effect, because it is acting on candidates who have not decided to waste the slot so much as failed to release it. This motivates the choice to study the decision to cancel rather than the decision to attend, and to present participants with a scenario in which they are explicitly undecided.

The second is the commercial treatment of cancellations. Airlines, hotels and event promoters routinely face customers who book capacity and do not use it, and manage this with schedules ranging from wholly non-refundable fares to graduated partial refunds. Xie and Gerstner [21] showed that refunding cancelled reservations can be advantageous to the firm, because recovered capacity can be re-sold. The logic transfers to a public testing body: a slot released early can be re-offered to a waiting learner, whereas one lost to a last-minute absence usually cannot. Two things it does not settle define the gap this study addresses. It does not establish how large a refund must be, since commercial schedules are set to maximise revenue rather than utilisation, and it does not establish how the influence of a refund varies with notice, since commercial cancellation windows are set by operational convenience rather than estimated from behaviour. The present design therefore varies refund amount and notice period explicitly, at the fee and threshold DVSA actually uses.

The third concerns the status of scenario evidence itself, and shapes how the findings should be read rather than what was predicted. Vignette methods are widely used in behavioural and health economics because they permit controlled variation of factors, such as a refund amount or a deadline, that cannot readily be randomised in the field, and influential reviews of intertemporal choice rest substantially on such designs [5]. Their limitation is equally well established: stated and actual behaviour diverge, particularly where present bias and inertia exert their strongest influence at the moment of action rather than in prospect [14]. This has a direct consequence for the present study. It means the design is better suited to establishing the direction and relative magnitude of the refund and timing effects than their absolute size, and it means a null result for timing must be interpreted with the possibility in view that the manipulation conveyed temporal distance without making it felt. Both points are returned to in the discussion.

The driving test case sharpens what these literatures leave open, in two ways. The first is that the appointment is rationed rather than merely scheduled. In the health setting studied by Hallsworth and colleagues [7] a missed appointment wastes clinician time and the intervention aims to reduce missing altogether. Here the appointment is a unit of scarce national capacity with a visible queue behind it, so an early cancellation can be the socially preferable outcome rather than a failure. The behavioural target is therefore not attendance. It is the timely release of an appointment the candidate no longer intends to use. That inverts the usual framing, and it is why the scenario here asks about cancelling rather than about turning up.

The second is that current policy already embeds a behavioural model, and it is that model the study tests. The ten working day threshold treats notice period as the variable that matters and treats the refund as a reward for giving notice. That is a coherent design if candidates decide early and the barrier to cancelling is inattention. It is the wrong design if candidates remain undecided until close to the date and the barrier is the financial loss, because the rule then removes the incentive at exactly the point where the decision is actually being made. Since the two factors are confounded in current policy, where more notice and a larger refund always arrive together, observational data cannot separate them. Only an experiment that varies them independently can, which is the gap this study addresses.

A third consideration concerns what a partial refund does that a full refund does not. The commercial literature [21] treats refunds instrumentally, as the price of recovering saleable capacity, and firms set them to maximise revenue. A public body faces a different objective function: it is not trying to maximise fee income but to maximise tests delivered, and forfeited fees are at best a weak proxy for that. This matters for interpretation because it means the commercial finding that refunds can be profitable does not transfer directly. For DVSA the question is not whether a refund pays for itself in recovered revenue. It is whether the refund converts wasted capacity into delivered tests. That is an empirical question, and one this design can inform but not settle.

What is absent from all three literatures is direct evidence on how refund amount and notice period operate, and interact, in a setting where the fee is modest, the emotional stakes are high, the appointment is scarce, and the social value of early release is considerable. Refund amount and notice period have never been varied against each other in such a setting, so their relative importance is unknown. That comparison is what this study provides.

1.4 The Current Study

The rationale for this project follows from the mismatch between the current cliff-edged refund policy and what the behavioural literature predicts about human decision-making. Forfeiting the whole fee is coded as a salient loss [11, 20], and the sunk investment grows more compelling as the test approaches [3, 14]. Put those together and the current rule looks badly aimed. It offers nothing to a candidate inside the window, so there is no financial reason to cancel at exactly the point when releasing the slot would help most. A graduated refund schedule, one that returns part of the fee even at short notice, might convert what would otherwise be silent non-attendance into early, formal cancellation, allowing slots to be reallocated to waiting learners. Two questions follow, both empirical. How large does a refund need to be before stated behaviour shifts? And does its influence weaken as the test draws nearer? Answering these questions is a necessary first step toward evaluating whether a change to DVSA’s refund structure could improve slot utilisation without simply subsidising cancellations that would have happened anyway.

The study addresses them with an online scenario-based design in which participants read a single, richly specified vignette and rate how they would behave. The design is set out in full in Section 4; in outline, refund amount varies between participants and the notice period varies within them, so that each factor can be examined without the other moving.

1.5 Research Questions and Hypotheses

The study addresses three questions. Three questions follow from the review: whether refund amount affects stated willingness to cancel, whether the timing of the decision does, and whether the two interact. They are stated as directional hypotheses.

H1 (Refund amount). Greater refund amounts will be associated with a higher stated likelihood of cancelling. Because a larger refund reduces the financial loss absorbed on cancellation (loss aversion and reference dependence) and partially recovers the sunk investment (sunk-cost and mental-accounting accounts), participants offered a full refund are expected to report the

greatest willingness to cancel, followed by the half and quarter refunds, with the no-refund condition producing the lowest willingness.

H2 (Timing). A longer interval before the test will be associated with a higher stated likelihood of cancelling; as the decision point moves closer to the test date, willingness to cancel will decline. This follows from the increasing salience of the sunk investment as the event nears, present-biased focus on the imminent test, and the tendency to defer the effortful act of formally cancelling. The opposite prediction is at least as defensible, and the data speak to it. As the date approaches, a candidate’s uncertainty about their own readiness resolves: a day or two before the test most people know whether they intend to sit it, so the probability of a decisive cancellation might reasonably be expected to rise rather than fall. H2 as stated therefore rests on the assumption that inertia dominates information, and that assumption is itself testable rather than obvious.

H3 (Interaction). The positive effect of refund amount on willingness to cancel will be attenuated at short notice relative to longer notice. As the test approaches and non-financial pressures to keep the slot intensify, the marginal influence of a given refund on the decision is expected to weaken. This interaction hypothesis is treated as exploratory.

Supporting these primary hypotheses, an exploratory within-person comparison will examine whether participants report a different likelihood of cancelling at their assigned second-part timing than at the fixed 5-day anchor, holding their refund constant, providing an individual-level test of the timing effect. Confirming H1 and H2 would establish that both the size of a refund and the timing of the cancellation window shape the willingness to release a scarce test slot, and would provide the behavioural foundation for evaluating whether a graduated, partially refundable cancellation policy could improve the utilisation of driving-test capacity.

2 Method

2.1 Design

The study used a 4×4 mixed factorial design. Refund amount was manipulated between participants at four levels corresponding to the proportion of the standard £62 weekday car test fee returned on cancellation: no refund (£0), a partial refund of £15 (approximately one quarter), a partial refund of £31 (one half), and a full refund of £62. Cancellation timing was manipulated across two successive decisions. In Part 1 all participants decided at a fixed anchor of five days before the test. In Part 2 the interval was randomly assigned to 1, 3, 7, or 10 days.

Refund amount was held constant within each participant across both parts. This was deliberate: holding the financial consequence fixed while varying only temporal distance isolates the effect of timing and permits a within-person contrast in which each participant serves as their own control. The dependent variable in both parts was a single-item rating of stated cancellation likelihood from 1 (*very unlikely*) to 7 (*very likely*).

Because Part 1 is fixed at five days, the four Part 2 intervals do not form a single ordered manipulation relative to a common baseline. They constitute four distinct contrasts against the same anchor, two in which the decision moves closer to the test (5 to 1 days, 5 to 3 days) and two in which it moves further

away (5 to 7 days, 5 to 10 days). These run in opposite directions, so the within-person changes are not pooled into a single average. They are reported and interpreted separately throughout.

Allocation to both refund and timing conditions used balanced randomisation without replacement, which distributes participants in blocks equal to the sum of the allocation ratios and yields approximately equal cell sizes. The two randomisations operated independently.

2.2 Participants

Two hundred UK-based adults ($N = 200$) completed the study, recruited through the Testable Minds academic participant panel and remunerated at the panel's standard rates. Eligibility required participants to be aged 18 or over and resident in the United Kingdom.

Mean age was 35.0 years ($SD = 9.2$, range 18–50). One hundred and two participants (51.0%) reported male sex at birth and 98 (49.0%) female. On driving status, 153 (76.5%) held a full car licence, 37 (18.5%) held a provisional licence and had not yet passed the practical test, and 10 (5.0%) held no licence. Among the 37 provisional holders, 14 had passed the theory test and 12 had a practical test booked; of those 12, 11 had booked it themselves, 11 had paid the fee themselves, and 9 reported it would be their first attempt. Participants were geographically dispersed, with London ($n = 30$), Manchester ($n = 10$) and Newcastle upon Tyne ($n = 7$) most frequently reported.

All 200 participants who began the study completed it. There was no attrition after consent and no participant was excluded.

That pattern warrants comment, since perfect completion and zero exclusions can signal weak quality control rather than good data. The study carried no attention check and no manipulation check on whether participants registered their assigned refund; that is a design shortcoming, listed among the limitations. The platform does record event-level timestamps, which give indirect evidence of engagement. Median time on the Part 1 screen, which held the full scenario, was 49.7 seconds (IQR 36.9–71.3), none under five seconds. Part 2, a shortened restatement, took a median of 16.5 seconds. Median session time was 136 seconds (IQR 106–198). These figures are consistent with engagement but are no substitute for an attention check.

Ethical approval was granted by the UCL Research Ethics Committee (PaLS LREC EP-2024-0203-148). Participants gave informed consent through the eight-item form described below, were told participation was voluntary and could be discontinued at any point before submission, and were informed that responses were anonymous and could not be withdrawn afterwards because no identifying information was held. No personal data were collected beyond age, sex at birth, city and licence status. Participants were debriefed and given contact details for the researcher and the Principal Investigator.

2.3 Materials and Procedure

The study was built on the Gorilla Experiment Builder and administered as a single online session of approximately four minutes. Participants viewed an information screen, then completed an eight-item consent form covering comprehension of the information sheet, voluntary participation, anonymity and the irrevocability of submitted responses, age eligibility, intended uses of the data including sharing with DVSA, archiving, data processing under UK data protection law, and voluntary agreement. Progression required agreement with all eight statements.

Participants then answered background questions on age, sex at birth and current city, followed by a categorical question on driving licence status. Licence status was asked with three response options rather than as a yes-or-no question, because UK licensing distinguishes provisional from full entitlement and a yes-or-no item collapses that distinction, a point raised by the industry partners. Conditional follow-ups were displayed contingent on the response: provisional holders were asked about theory test completion and whether a practical test was booked, and if booked, whether they had arranged and paid for it themselves and whether it would be their first attempt; full licence holders were asked the year they obtained their licence.

In Part 1, participants imagined preparing for their first practical test. The scenario established that they had passed the theory test, booked the practical test themselves, and paid the £62 fee from their own funds; that suitable slots had been difficult to find, so they had taken the first available; and that they had continued taking lessons since booking. It then established genuine ambivalence about readiness, describing fluctuating confidence, concern that sitting the test prematurely risked failure and further delay, and the hope of feeling more prepared closer to the date. Participants were told they were five days before the test and could cancel quickly and easily online, and were informed of their assigned refund. The scenario presented considerations on both sides: cancelling would release the slot to another learner who might be better prepared, and unused slots waste scarce capacity; but cancelling would mean surrendering a hard-won appointment, with no guarantee of a suitable replacement date, potential knock-on delay to work, study, travel or family plans, and the possibility of feeling ready on the day. Participants then gave their rating.

Because the wording carries the manipulation, the operative passages are reproduced in full. The scenario opened by establishing ownership of the decision and the scarcity of the appointment:

You are preparing for your first practical driving test. You passed your theory test, booked the practical test yourself, and paid the fee of £62 out of your own pocket. When you booked, suitable slots were difficult to find, so you took the first available opportunity. You have continued having lessons since then.

Three details in it are doing work. The participant both booked and paid, which makes the fee a personal sunk cost; the slot is described as hard to obtain, supplying the scarcity that makes an unused appointment costly to others; and lessons are ongoing, so the dilemma cannot be resolved by assuming preparation has stopped.

The second paragraph established genuine ambivalence, which is what prevents the task from reducing to a simple arithmetic comparison:

As the test date approaches, you are not sure how you feel. Some days you feel like you are making progress and might be ready in time. Other days, you doubt yourself and worry that sitting the test now risks a failure, and failing could set you back further. You go back and forth. You have not made a decision yet, and part of you hopes that you will feel more confident as the date gets closer.

Holding this uncertainty constant was deliberate. Had the scenario specified that the participant felt unprepared, the refund would merely have priced a decision already made. The cost, taken up in the discussion, is that it may also suppress the mechanism through which timing would ordinarily operate.

The final paragraph delivered the manipulation and set out the considerations on both sides:

You are 5 days before the driving test. You have the option to cancel, quickly and easily online. If you cancel now, you get [refund]. Your slot becomes available for another learner driver who may be better prepared. Practical test slots are in short supply, and an unused slot is a wasted one. However, cancelling means giving up the slot you worked hard to secure, and finding another suitable date soon may not be easy. A further delay could also set back your plans for work, study, travel, or family. And you might cancel now, only to wake up on the test day feeling ready after all.

The bracketed term was replaced with the participant’s assigned refund. Two arguments favour cancelling and three favour keeping the slot, a balance intended to stop the scenario reading as an invitation to cancel, which would inflate the outcome measure uniformly across conditions.

In Part 2, participants were told the refund remained the same but the decision now occurred at a different point before the test. They were shown their assigned interval, reminded of the identical refund, told they remained undecided, and asked the same question. A debrief screen followed.

The outcome measure was a single seven-point item, “How likely are you to cancel your practical test now?”, anchored at 1 (*very unlikely*) and 7 (*very likely*). A single item was chosen over a multi-item scale because the construct is narrow and behavioural rather than a latent disposition, and because each participant had to make the judgement twice: repeating a battery would have introduced fatigue and demand characteristics into the second measurement, which is exactly where the timing manipulation had to act. The cost, discussed in the limitations, is that no internal-consistency estimate is available and the measure cannot separate strength of intention from confidence in it.

Consent was obtained through eight separate statements rather than a single tick, covering comprehension of the information sheet, voluntary participation, the right to withdraw during the study, anonymity and the consequent irrevocability of submitted responses, age eligibility, the intended uses of the data including sharing with DVSA, archiving for future research and teaching, and processing under UK data protection law. Agreement with all eight was required to proceed, so no partial consent was possible.

2.4 Analytic Approach

Analyses proceeded in four pre-specified stages.

First, the between-participants effect of refund amount on the Part 1 rating, where timing was constant for all participants. This used one-way analysis of variance with Tukey-corrected pairwise comparisons, a Kruskal–Wallis test appropriate to the ordinal outcome, and a proportional-odds ordinal logistic regression.

Second, the functional form of the refund effect. Because a positive relationship between refund size and stated willingness to cancel is close to a manipulation check, the more informative question is the shape of that relationship. A linear fit on refund value in pounds was compared against quadratic and logarithmic alternatives by Akaike information criterion; departure from linearity was tested by comparing the linear model against a saturated model treating refund as a factor; the increment per pound was computed between adjacent levels; and a single contrast compared any refund against none.

Third, the effect of timing. The four within-person contrasts against the five-day anchor were analysed separately for the reason given in the design section, then grouped by direction of movement, and then compared between those two groups. Because Part 2 always follows Part 1, any general drift between the

two responses affects all participants equally, so the between-groups comparison of change scores is the appropriate test of whether direction of movement matters. The between-participants comparison across the four Part 2 intervals and a linear mixed-effects model with random intercepts for participant were also fitted.

Fourth, robustness: the refund model refitted with age, sex and licence status as covariates, and the refund effect examined separately within licence subgroups.

Because the timing hypothesis was not supported, statistical equivalence was assessed directly: a smallest effect size of interest was specified in advance and two one-sided tests were conducted against it.

Throughout, effect sizes and 95% confidence intervals are reported alongside test statistics rather than p values alone. An alpha of .05 was adopted.

3 Results

3.1 Randomisation and Data Quality

Balanced randomisation produced closely equal cells. Refund conditions contained 50 (no refund), 51 (£15), 49 (£31) and 50 (£62) participants; timing conditions contained 51 (1 day), 49 (3 days), 49 (7 days) and 51 (10 days). The two factors were allocated independently and no participant was lost after allocation.

Data completeness was high. Every participant provided a valid rating in both parts and there were no missing responses on any measure. Conditional routing operated as designed: the theory and booking items were presented to exactly the 37 provisional holders, and the licence year item to exactly the 153 full licence holders. Two of the 153 licence year responses were implausible as four-digit years (14 and 96, most plausibly two-digit abbreviations or a mis-entered age); the remaining 151 fell between 1993 and 2026. Licence year was not used in any confirmatory analysis, so these two entries were retained and flagged rather than corrected.

Sample characteristics are summarised in Table 1. Two of them bear on interpretation. The sample skews towards people who have already completed the process the scenario describes: 153 held a full licence, against 37 provisional holders and 10 with none, and only 12 had a test booked at the time. For most participants the task therefore involved recall or projection rather than a live decision. This is revisited in the robustness checks, where the refund effect is examined separately by licence status, and again in the limitations. Second, the conditional routing means the sub-questions were answered only by those for whom they applied, so the denominators differ by item and should be read against the 37 provisional holders rather than the full sample.

3.2 Effect of Refund Amount

Refund amount had a substantial effect on stated cancellation likelihood. Descriptive statistics with confidence intervals appear in Table 2.

The omnibus test was significant, $F(3, 196) = 10.95$, $p < .001$, $\eta^2 = .144$, 95% CI [.056, .225]. The non-parametric equivalent agreed, $H(3) = 32.77$, $p < .001$. Tukey-corrected comparisons showed the no-refund condition differing from the £31 partial refund (mean difference 1.14, 95% CI [0.26, 2.02], $p = .005$) and from the full refund (1.88, 95% CI [1.01, 2.75], $p < .001$), and the £15 partial refund differing from the full refund (1.19, 95% CI [0.32, 2.06], $p = .003$). Adjacent levels did not differ reliably: £0 versus £15 (0.69, 95%

Table 1. Participant characteristics ($N = 200$).

Characteristic	n	%
<i>Sex at birth</i>		
Male	102	51.0
Female	98	49.0
<i>Driving licence status</i>		
Full car licence	153	76.5
Provisional (learner) licence	37	18.5
No licence	10	5.0
<i>Provisional holders only ($n = 37$)</i>		
Passed the theory test	14	37.8
Practical test currently booked	12	32.4
<i>Those with a test booked ($n = 12$)</i>		
Booked it themselves	11	91.7
Paid the fee themselves	11	91.7
Would be their first attempt	9	75.0

Note. Mean age 35.0 years ($SD = 9.2$, range 18–50). Participants were geographically dispersed, most frequently reporting London ($n = 30$), Manchester ($n = 10$) and Newcastle upon Tyne ($n = 7$). Percentages for the conditional items are of the relevant sub-sample rather than of $N = 200$.

Table 2. Cancellation likelihood (1–7) by refund condition, with 95% confidence intervals.

Refund condition	n	Part 1: M [95% CI]	Part 2: M [95% CI]
No refund (£0)	50	1.74 [1.39, 2.09]	1.54 [1.21, 1.87]
Partial refund (£15)	51	2.43 [2.00, 2.86]	2.39 [1.95, 2.83]
Partial refund (£31)	49	2.88 [2.33, 3.42]	2.57 [2.02, 3.12]
Full refund (£62)	50	3.62 [3.10, 4.14]	3.72 [3.17, 4.27]

CI $[-0.18, 1.56]$, $p = .170$), £15 versus £31 (0.45, 95% CI $[-0.43, 1.32]$, $p = .550$) and £31 versus £62 (0.74, 95% CI $[-0.14, 1.62]$, $p = .129$). The standardised difference between extremes was large, $d = 1.17$, 95% CI $[0.75, 1.60]$.

The proportional-odds model gave the same conclusion, $b = 0.033$ per pound, $z = 5.59$, $p < .001$. Each additional £10 of refund was associated with 1.39 times higher odds of a higher rating, and moving from no refund to a full refund with 7.55 times higher odds.

The same gradient appeared again in Part 2, where the refund was unchanged but the notice period differed, with means rising from 1.54 to 3.72, $F(3, 196) = 13.73$, $p < .001$. This is a re-measurement rather than an independent replication: the same participants judged the same refund against the same vignette about a minute later, and the two ratings correlate at $r = .82$. It shows the gradient is stable within a session, not that it reproduces in a fresh sample. Figure 1 displays both.

The Part 2 column of Table 2 averages over the four notice periods, which hides variation, so Table 3 gives the full breakdown. Cell sizes range from 8 to 16, so it is included for transparency rather than inference. The refund gradient appears within every notice period; the ordering of notice periods does not.

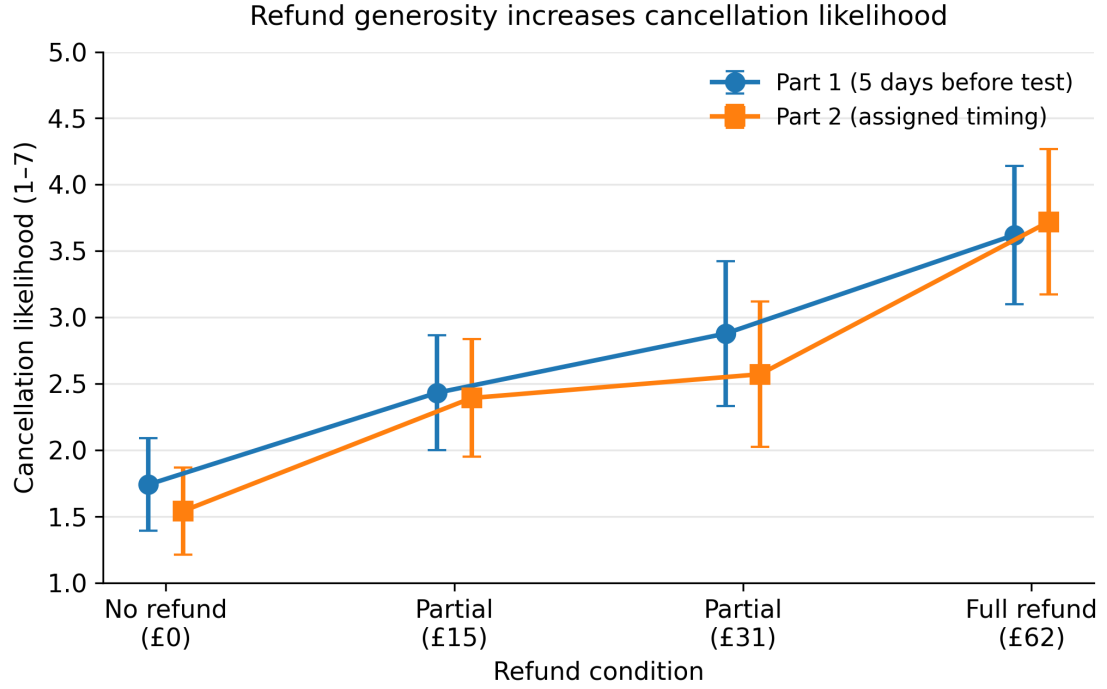


Fig. 1. Mean cancellation likelihood by refund condition for both decisions. Error bars are 95% confidence intervals.

Table 3. Mean Part 2 cancellation likelihood for every refund by notice period cell, with cell sizes in parentheses.

Refund condition	1 day	3 days	7 days	10 days
No refund (£0)	1.00 (11)	1.62 (8)	2.07 (15)	1.38 (16)
Partial refund (£15)	2.83 (12)	3.00 (16)	1.79 (14)	1.67 (9)
Partial refund (£31)	2.64 (14)	2.79 (14)	1.50 (8)	2.92 (13)
Full refund (£62)	3.43 (14)	4.27 (11)	4.08 (12)	3.23 (13)

Note. Refund was fixed for each participant across both parts; notice period was assigned independently for Part 2. Because the two randomisations were independent, cell sizes are unequal and no cell exceeds 16 participants.

3.3 Shape of the Refund Effect

That willingness to cancel increases with the size of the refund is a weak claim on its own, since presenting a larger sum and asking about cancellation invites an ordered response. The substantive question is the form the increase takes, and specifically whether the marginal pound buys less at the top of the range than at the bottom, or whether some threshold must be crossed before any movement occurs.

The increment per pound was largest for the first step and smaller thereafter: 0.046 scale points per pound from £0 to £15, 0.028 from £15 to £31, and 0.024 from £31 to £62. This pattern is consistent with diminishing marginal returns, but it is not statistically supported. A quadratic term did not improve fit ($p = .396$), the logarithmic specification fitted worse than the linear one (AIC 780.7 versus 777.3), and a

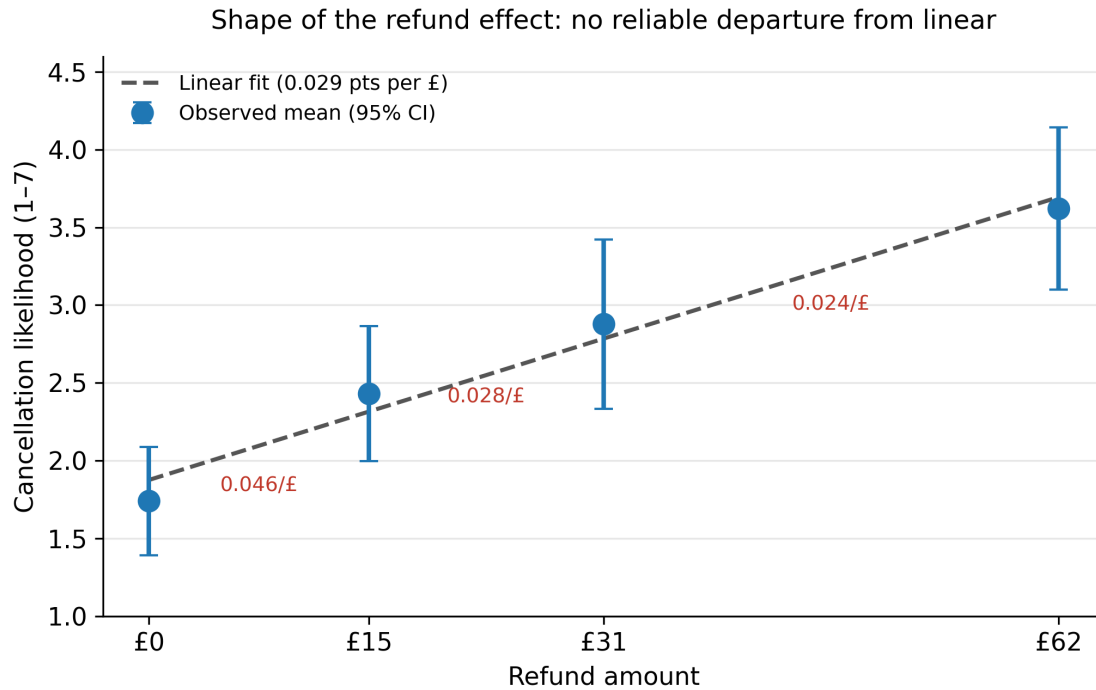


Fig. 2. Observed means by refund level against the fitted straight line, with the increment per pound annotated for each step. The per-pound return falls across the range, but every observed mean lies close to the line and no test of curvature reaches significance.

saturated model treating refund as a factor did not improve on the linear model, $F(2, 196) = 0.40$, $p = .668$. On the available evidence the relationship is indistinguishable from a straight line in pounds, and there is no reliable evidence of a threshold, a plateau, or a curve.

Figure 2 makes the tension visible. The annotated increments fall steadily from left to right, which is the pattern diminishing returns would produce, yet each observed mean sits close to the fitted line and the confidence intervals overlap it throughout. The apparent curvature is therefore within the range that sampling variation alone would generate at this sample size, and should not be reported as a finding.

One contrast is nevertheless worth isolating. Comparing any refund against none, participants offered nothing rated 1.74 while those offered any refund at all rated 2.97, $t(198) = 4.36$, $p < .001$, $d = 0.71$, 95% CI [0.38, 1.04]. Since the four adjacent-level comparisons above were individually non-significant while this single step is substantial, the data are equally consistent with a design in which the presence of a refund matters more than its precise magnitude. Distinguishing a linear-in-pounds account from a presence-versus-absence account would require levels spaced more finely at the bottom of the range.

Figure 3 shows the distribution of individual ratings by condition, indicating that the shift involves the whole distribution rather than a minority of extreme responses.

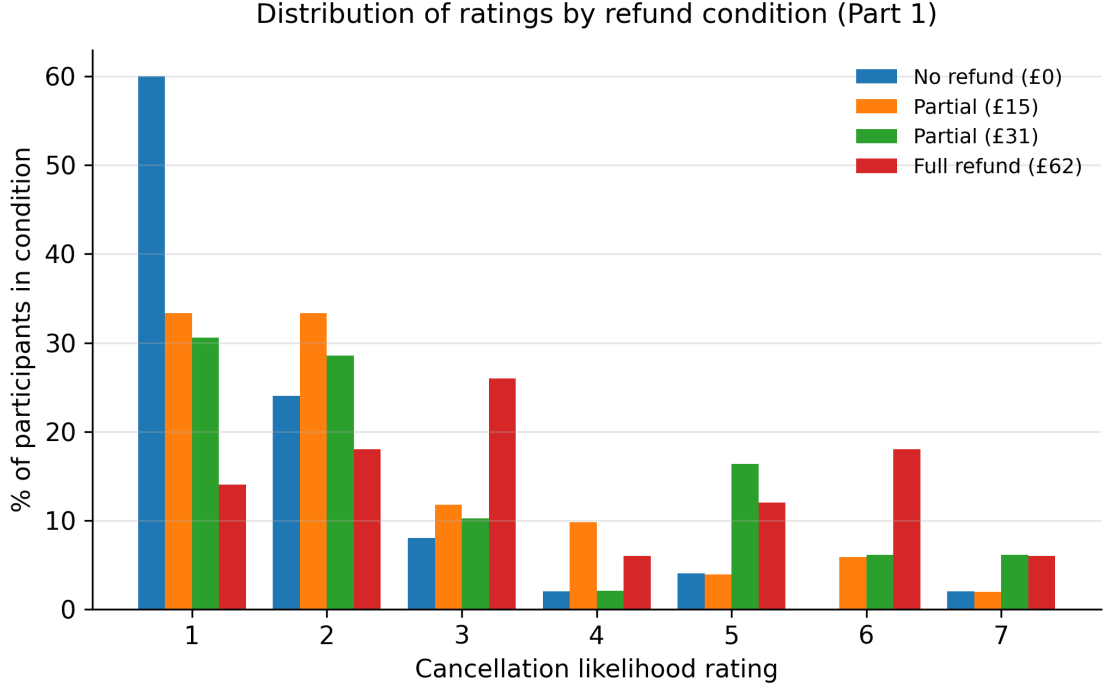


Fig. 3. Distribution of cancellation likelihood ratings by refund condition (Part 1), as a percentage of participants within each condition.

3.4 Effect of Cancellation Timing

Every participant decided at five days in Part 1 and at their assigned interval in Part 2. This yields four separate comparisons against a common anchor: two in which the decision moved closer to the test (5 to 1 days; 5 to 3 days, a shift of two) and two in which it moved further away (5 to 7 days; 5 to 10 days). Because these run in opposite directions, averaging them would cancel genuine effects against one another, so they are reported separately in Table 4.

Three of the four contrasts were null. The exception was the largest movement away from the test, 5 to 10 days, where willingness to cancel fell by 0.33 scale points, $d_z = -0.38$. This contrast survives Bonferroni correction for the four tests ($\alpha = .0125$). Grouping by direction, participants who moved closer to the test showed no change ($+0.04$, 95% CI $[-0.19, +0.27]$, $p = .735$), while those who moved further away showed a decrease (-0.26 , 95% CI $[-0.45, -0.07]$, $p = .009$).

Figure 4 shows why the contrasts cannot be combined. The two comparisons in which the decision moved closer to the test straddle zero in opposite directions, while both comparisons in which it moved further away sit below zero. A single averaged change score would cancel the first pair against the second and return a value near zero, which is what the pooled analysis produced and what the disaggregated analysis shows to be misleading.

Table 4. Within-person change in cancellation likelihood from the five-day anchor to the assigned interval, reported separately by contrast. Positive values indicate greater willingness to cancel at the assigned interval than at five days.

Contrast	Movement	<i>n</i>	Mean change	95% CI	<i>p</i>
5 to 1 day	closer	51	−0.10	[−0.46, +0.26]	.597
5 to 3 days	closer	49	+0.18	[−0.10, +0.47]	.211
5 to 7 days	further	49	−0.18	[−0.48, +0.12]	.237
5 to 10 days	further	51	−0.33	[−0.58, −0.09]	.010
Grouped: closer (1, 3 d)	closer	100	+0.04	[−0.19, +0.27]	.735
Grouped: further (7, 10 d)	further	100	−0.26	[−0.45, −0.07]	.009

Note. The between-groups comparison of the two grouped rows, which removes drift common to both, gave a difference of 0.30 scale points, 95% CI [−0.001, +0.601], $t(198) = 1.96$, $p = .052$, $d = 0.28$.



Fig. 4. The four within-person contrasts against the five-day anchor, grouped by whether the decision moved closer to the test or further away. Error bars are 95% confidence intervals. Pooling contrasts that run in opposite directions would average these against one another.

These within-group tests are confounded with any general drift between the two responses, since Part 2 always came second. Across all participants the mean change was -0.11 ($p = .156$), indicating a slight overall downward drift. The appropriate test is therefore the comparison between the two direction groups, which removes drift common to both. That comparison did not reach significance: participants who moved closer changed by 0.30 scale points more than those who moved further away, 95% CI [−0.001, +0.601],

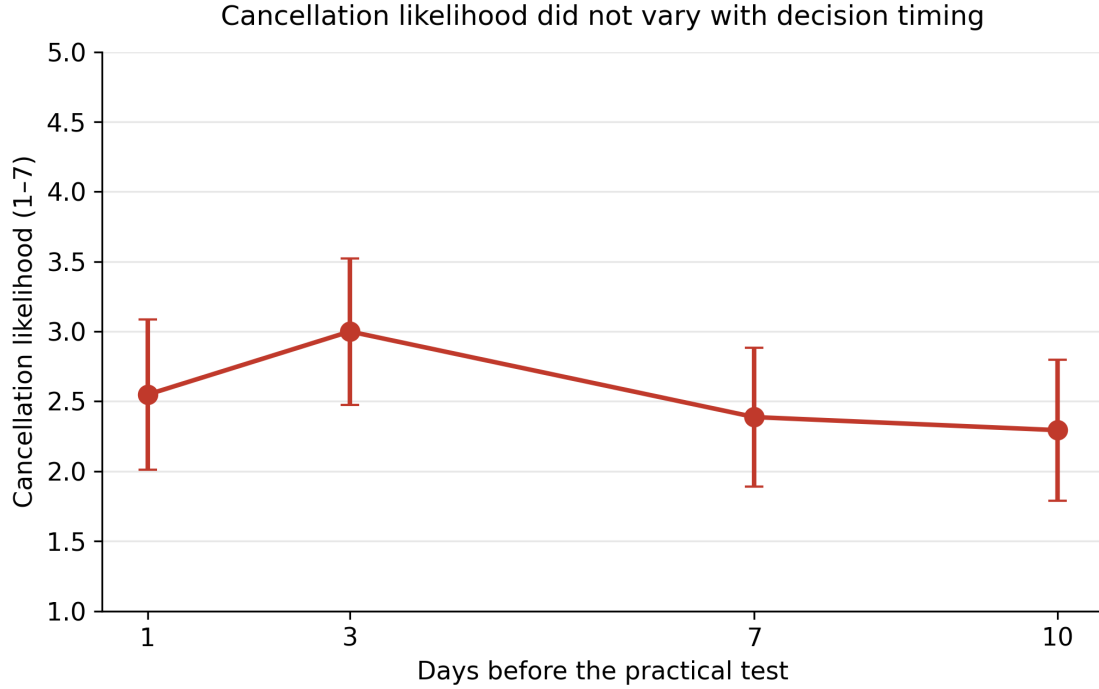


Fig. 5. Mean cancellation likelihood by days remaining before the practical test (Part 2). Error bars are 95% confidence intervals.

$t(198) = 1.96$, $p = .052$, $d = 0.28$. Regressing individual change on the number of days moved gave a slope of -0.039 per day, 95% CI $[-0.082, +0.004]$, $p = .079$.

The between-participants comparison across the four Part 2 intervals was also null. Means were 2.55 [2.01, 3.09] at 1 day, 3.00 [2.48, 3.52] at 3 days, 2.39 [1.89, 2.89] at 7 days and 2.29 [1.79, 2.80] at 10 days, $F(3, 196) = 1.40$, $p = .245$, $\eta^2 = .021$, 95% CI $[.000, .062]$. There was no monotonic association with days remaining, $\rho = -.095$, $p = .181$, and the linear slope was small, $b = -0.051$ per day, 95% CI $[-0.124, +0.023]$. Figure 5 displays the pattern.

The Kruskal–Wallis test on the four Part 2 intervals reached nominal significance, $H(3) = 8.14$, $p = .043$, at odds with the analysis of variance. This is attributable to the three-day condition, which had an atypically low proportion of floor responses (18.4% rated 1, against 47.1%, 46.9% and 54.9% in the one-, seven- and ten-day conditions). The comparisons involving the shortest interval were null (1 versus 7 days $p = .818$; 1 versus 10 days $p = .478$; 1 versus 3 days $p = .058$), and no comparison survived Bonferroni correction for the six tests ($\alpha = .0083$). The pattern is non-monotonic, confined to a single cell, and opposite in direction to the hypothesis, and is most parsimoniously attributed to sampling variation.

Taken together, the disaggregated analysis shows that the pooled null concealed heterogeneity between contrasts, but the direction of the reliable component is opposite to that predicted: moving further from the test reduced rather than increased willingness to cancel. Because this emerged from disaggregation rather

Table 5. Linear mixed-effects model of cancellation likelihood (both decisions stacked; random intercept for participant).

Predictor	<i>b</i>	<i>SE</i>	<i>z</i>	<i>p</i>
Intercept	1.790	0.242	7.40	< .001
Days remaining	−0.005	0.032	−0.14	.885
Refund amount (per £)	0.038	0.007	5.62	< .001
Days × Refund	−0.001	0.001	−1.50	.133

Note. Random intercept variance = 2.29. $N = 200$ participants, 400 observations.

than prior specification, and because the appropriate between-groups test does not reach significance, it is treated as exploratory.

3.5 Equivalence Testing of the Timing Effect

Observed power is not reported, for the reason given in the analysis section. Statistical equivalence was assessed directly instead. A smallest effect size of interest of one scale point on the seven-point measure was specified, on the grounds that a shift of less than one point on a single item is difficult to interpret substantively and would not support a change to notice periods.

The difference between the extreme intervals, 1 day against 10 days, was 0.26 scale points, 95% CI $[-0.49, +1.00]$. Two one-sided tests against the one-point bound were significant, $p = .025$, so effects larger than one scale point can be rejected. Against a tighter bound of half a scale point the equivalence test was not significant, $p = .258$, so effects in the range of half a point cannot be excluded.

The conclusion is therefore bounded rather than absolute. Any effect of notice period in this paradigm is smaller than one scale point, but the data do not establish that it is zero.

3.6 Within-Person Consistency and the Combined Model

Ratings across the two decisions correlated at $r = .82$. That figure needs care, because 112 participants (56.0%) rated 1 or 2 in both parts, and identical floor responses inflate the correlation without indicating psychological stability. Excluding those who sat at the floor twice, the correlation falls to $r = .54$ ($n = 88$), and among participants above the floor in both parts it is $r = .64$ ($n = 67$). A substantial part of the apparent stability is therefore a measurement artifact of the compressed scale.

The mixed-effects model on the stacked data is reported in Table 5. Refund amount was a significant predictor while days remaining was not, and the interaction was non-significant. Substantial between-participant variance was present, consistent with stable individual differences in baseline willingness to cancel.

3.7 Robustness Checks

The refund effect was unchanged by demographic covariates. In a model including age, sex and licence status, refund remained a strong predictor, $b = 0.030$ per pound, $t = 5.72$, $p < .001$. None of the covariates reached significance: age ($b = -0.023$ per year, $p = .079$), sex ($b = 0.396$ for male, $p = .101$) and licence status (both $p > .40$).

The effect also held within licence subgroups. Among the 153 full licence holders the association was $\rho = .369$, $p < .001$; among the 37 provisional holders, who most closely resemble the population of policy interest, it was $\rho = .487$, $p = .002$. Subgroup means were comparable (2.63 and 2.86), suggesting that responses from experienced drivers imagining the scenario did not differ markedly from those of current learners.

3.8 Summary in Relation to Hypotheses

H1, that greater refunds would increase stated cancellation likelihood, was supported. The increase was monotonic, reappeared on re-measurement within the same session, and was robust to covariates and subgroup splits, though its functional form is indistinguishable from linear and the presence-versus-magnitude question remains open.

H2, that a longer interval before the test would increase cancellation likelihood, was not supported. Effects larger than one scale point can be excluded. An effect in the opposite direction emerged on disaggregation, falling just short of significance, and is treated as exploratory.

H3, that the refund effect would be attenuated at short notice, was not supported.

4 Discussion

4.1 Principal Findings

Refund amount had a clear effect on stated willingness to release a booked driving test slot, and the effect was large by the standards of vignette research. Cancellation timing did not, and the notice period did not moderate the influence of the refund. The design cannot separate the three theoretical accounts that motivated the refund prediction, since all three predict the same direction; what it can do is characterise the form of the relationship and bound the size of the timing effect.

4.2 Interpretation of the Refund Effect

The direction of the refund result is close to a manipulation check. Show someone a larger sum and ask how likely they are to cancel, and an ordered response is more or less what you would expect. A study finding no such ordering would point to a failed manipulation rather than an interesting result. The contribution therefore lies in the magnitude and the shape rather than the sign.

On magnitude, the difference between no refund and a full refund was 1.88 scale points on a seven-point measure, $d = 1.17$. Expressed in odds, a full refund raised the odds of a higher rating more than sevenfold relative to none. That the gradient reappeared on re-measurement and survived adjustment for demographics and subgroup splits suggests it is not fragile, though the re-measurement was within the same session and is not independent evidence.

On shape the study cannot yet separate two accounts. The observed increments per pound decline from the bottom of the range to the top, which would suggest diminishing marginal returns and, in policy terms, that a modest partial refund captures much of the available benefit. But no test supports that reading: a quadratic term added nothing, a logarithmic specification fitted worse than a linear one, and a saturated model did not improve on the straight line. At the same time, none of the adjacent-level comparisons was individually significant, while the single contrast between any refund and none was substantial, $d = 0.71$.

The data are therefore consistent both with a smooth linear function of the amount and with a step at the point where a refund exists at all. Deciding between these matters for policy, because the first implies that the size of the refund is what needs setting and the second implies that the mere existence of a partial refund does most of the work. Resolving it requires levels spaced more finely between zero and a quarter of the fee, which the present four-level design cannot deliver.

Both accounts are compatible with the theory reviewed earlier. Loss aversion, sunk-cost reasoning and mental accounting all imply that reducing the loss to be absorbed lowers the barrier to cancelling, and mental accounting in particular offers a natural reading of a threshold effect, since a partial refund allows the original outlay to be partly recovered and the account partly closed. The study was not designed to adjudicate between these mechanisms and does not do so.

4.3 Interpretation of the Timing Result

Disaggregating the four contrasts, for the reason given in the results, reveals heterogeneity the pooled test concealed: the largest movement away from the test produced a reliable decrease in willingness to cancel, with no change among those whose decision moved closer.

The direction of that pattern is the opposite of what was predicted. The hypothesis, grounded in present bias and in the tendency to defer effortful administrative action, was that greater notice would increase willingness to cancel. What appears instead is that greater notice slightly reduced it. Because the appropriate test, comparing the two direction groups so that drift common to both is removed, did not reach significance ($p = .052$), and because this emerged from disaggregation rather than prior specification, the finding should not be over-interpreted. It is reported as exploratory and as a candidate for direct test.

If it is real, one interpretation is that a distant deadline invites deferral of the cancellation decision itself rather than of the test: with ten days remaining there is no pressure to resolve anything now, so the status quo of holding the slot persists. It also matches the alternative prediction set out alongside H2, that approaching the date resolves uncertainty and so makes a decisive cancellation more likely.

The policy reading is concrete, and DVSA would have an interest in it. The reliable movement was confined to the largest step away from the test, and willingness to cancel was no lower at one day than at five. That suggests a refund would do most work inside roughly five days rather than further out: candidates ten days ahead are least inclined to resolve the question, those close to the date readiest to act. Since the current rule offers a full refund only outside ten working days, it is generous precisely where candidates are least disposed to use it and silent precisely where they are most disposed to. If the effect replicates, the implication is that extending some refund into the final week, rather than lengthening the notice period, is where the capacity gain lies. The evidential basis is thin, so this is a hypothesis for direct test rather than a recommendation. That reading is consistent with the inertia literature but inverts the usual prediction about deadline proximity, and would need confirming.

The equivalence test permits a more disciplined statement than a bare null. Effects larger than one scale point can be excluded; effects of around half a point cannot. Any influence of notice period in this paradigm is therefore small relative to the influence of refund amount, without being demonstrably zero.

Several design features may have limited the manipulation. Temporal distance was conveyed as a number of days rather than felt, and present bias is triggered by experienced proximity. Holding ambivalence constant may have suppressed the very mechanism timing works through. Cancellation was described as

quick and easy in both parts, so the procedural friction that builds near a deadline never appeared. And the floor discussed above left little room to move ratings down, which is the direction H2 required.

4.4 Relating the Findings to the Theoretical Accounts

Because the three accounts introduced earlier predict the same direction, the design cannot rank them, and no such ranking is claimed. Setting out what each would say about the pattern observed is still useful, because it points to the follow-up study that would separate them.

On a sunk-cost reading the relevant quantity is the residual loss, which falls smoothly from £62 to zero, so the account predicts a smooth function of the amount and sits comfortably with the linear fit. It struggles with the size of the step between no refund and any refund, since £15 still leaves £47 unrecovered and should feel closer to total forfeiture than it evidently did.

Loss aversion makes the same directional prediction but locates the action in how the outcome is coded. Under a value function steep near the reference point and flattening as losses grow, the first pound recovered weighs more than the last, which is what the observed increments hint at. It therefore anticipates the diminishing pattern the data suggested but did not confirm, and would be tested by spacing levels finely near zero.

Of the three, mental accounting sits most comfortably with the presence-versus-absence reading. If what matters psychologically is closing the account rather than the sum recovered, the distinction is between recovering nothing and recovering something, and the amount is secondary. That predicts a step near zero and then a shallow slope, which these data can neither confirm nor exclude.

The accounts therefore diverge exactly where this design is thinnest, between no refund and a quarter of the fee. Levels at, say, £0, £3, £6, £15 and £31 would separate them, and would answer how small a refund can be while still moving behaviour.

For the timing result, none of the accounts is straightforwardly supported. Both routes to the timing prediction, present bias and the tendency to defer effortful action, imply that willingness to cancel should fall as the test approaches. The observed pattern, in which the only reliable movement was a decrease among those whose decision moved further away, is if anything the reverse. One reading consistent with the inertia literature is that a distant deadline removes the pressure to resolve anything now, so the status quo of holding the slot persists; on that view the operative variable is not proximity to the event but whether the decision feels due. That reading is post hoc and is offered as a hypothesis for direct test rather than as an interpretation the present data establish.

4.5 Policy Implications

Current policy makes the refund all-or-nothing at a ten working day threshold. The findings suggest this is poorly matched to the behaviour it needs to influence. Willingness to cancel rises with the refund. So a candidate inside the current window who no longer wants to sit the test has the weakest possible reason to say so: the fee is gone either way, whether they cancel formally or just fail to turn up. The slot is then unlikely to be released in time for reallocation. A graduated partial refund inside the current cliff edge would be expected to increase early release, and because even the smallest refund tested shifted intentions, the benefit may not require returning the entire fee.

Two objections require serious treatment before this can be recommended.

The first is revenue. A partial refund inside the current window returns money the agency currently retains. If a proportion of forfeited fees is being used to fund test capacity, then refunding part of it reduces the resource available for the very capacity the intervention aims to protect. The cost is bounded and calculable in a way the benefit is not. Forfeited fee income can be computed directly. The value of an earlier-released slot depends on whether anyone takes it up. This asymmetry means the policy should be evaluated on realised reallocation rates rather than on intentions, and it is a further reason to prefer the smallest effective refund over a generous one.

The second, and more serious, is speculative booking. Making cancellation cheaper also makes it cheaper to hold a slot you are unsure about. Candidates might respond by booking earlier, or more tentatively, simply to secure a place while they make up their minds. That would raise the number of provisionally held slots and worsen the congestion the policy is meant to relieve. The present study cannot speak to this, because participants were presented with a slot already booked and paid for; the design says nothing about how refund terms affect the decision to book in the first place. This is a genuine limitation of the inference, not merely a caveat.

There are nonetheless reasons to expect a positive net effect. The counterfactual is unusually weak, since a slot lost to late non-release is wasted entirely while a released one has some chance of reallocation. The refund need only apply inside the window where the incentive is currently zero, leaving the ten-day rule and its deterrent against casual booking intact outside it. A partial rather than full refund keeps a real cost to cancelling, limiting the speculative-booking risk. And the whole thing can be trialled in a subset of centres, with booking volumes monitored, before any national rollout. What would settle the question is not further vignette work but a field trial measuring both slot release and booking behaviour together, since the two effects run in opposite directions and only their net matters.

4.6 What the Dependent Variable Does and Does Not Capture

One limitation needs its own section, because the policy argument turns on it. The case for a graduated refund rests on a specific behavioural claim: that a candidate inside the forfeiture window has no reason to cancel formally rather than simply not turn up, and that returning part of the fee would convert silent non-attendance into early release. The task never put that choice to participants. The item asked how likely they were to cancel, with the implied alternative of keeping the slot and sitting the test. There was no option to keep the slot and not attend.

This changes how the central numbers should be read. Under no refund, a mean of 1.74 most plausibly means participants would keep the appointment and try, not that they would waste it. The behaviour the policy targets, holding a slot and then failing to appear without telling anyone, was never available to be chosen and so cannot have been measured. The low no-refund ratings are as consistent with deciding to attend as with deciding to let the slot go unused, and the design cannot separate the two.

The refund finding therefore needs stating narrowly. It shows that refund size shifts stated willingness to cancel relative to attending. It does not show that refund size shifts the split between cancelling and silently not attending, which is the margin that determines whether capacity is recovered. Testing that would need a three-way response, cancel now, attend as planned, or keep the slot and probably not attend, with the refund attached only to the first. This study establishes that the financial lever moves intentions at all, which is a precondition for the policy argument rather than a demonstration of it.

4.7 Absolute Levels of Willingness to Cancel

The absolute level of the ratings is as important for policy as the size of the treatment effects. Every refund condition fell at or below the scale midpoint of 4. Under no refund the mean was 1.74, and only 8% of participants chose 4 or above. Even with the whole fee returned the mean was 3.62, statistically indistinguishable from the midpoint ($t(49) = -1.43$, $p = .160$), with 42% at 4 or above and 36% at 5 or above. The corresponding Part 2 figure was 3.72, with 48% at 4 or above.

This cuts two ways. Optimistically, the shift from 8% to 42% choosing the upper half of the scale is substantial, and in a system where recovering a small proportion of slots would be a meaningful gain, moving two fifths of candidates into the region where cancellation is seriously entertained may be enough to matter. Less comfortably, no condition came close to majority willingness. A full refund, the most generous intervention available, still left most participants disinclined to cancel. Whatever holds people to the appointment is not principally the money, and a refund schedule alone will not solve the utilisation problem.

Some of this is the scenario, which described an undecided candidate rather than one who knew they were unprepared; some is reluctance to abandon a scarce appointment; some may be scale avoidance. Either way the refund is a lever with a real but bounded effect, and the case for it rests on recovering a modest fraction of slots rather than on transforming behaviour.

4.8 Limitations

The outcome was a stated intention in a hypothetical vignette rather than observed behaviour. Intentions and actions come apart once real money and real inconvenience are at stake. The effect on actual cancellation rates could be larger or smaller than what is reported here.

The sample was drawn from an online panel and consisted predominantly of full licence holders. Only 37 participants held provisional licences and only 12 had a test booked, so most were imagining or recalling rather than reporting a live decision. The refund effect was at least as strong within the provisional subgroup, which is reassuring, but replication with candidates holding current bookings would materially strengthen the inference.

The dependent variable was a single item, which minimises fatigue but provides no internal-consistency estimate and cannot separate willingness from confidence in that willingness.

The design fixed the fee at the standard weekday rate and the scenario at a first test booked and paid for by the participant. Effects may differ where a third party paid, where the test is a retake, or at the higher evening and weekend rate, all of which alter the ownership and sunk-cost considerations at stake.

The timing manipulation was structurally asymmetric: Part 1 was fixed at five days, so the design provides four unequal contrasts rather than a balanced ordering, and no contrast beyond ten days was tested. Refund was manipulated between participants while timing was manipulated within, so the two factors are not tested with equal sensitivity.

Finally, the design cannot separate the three theoretical accounts behind the refund prediction, and does not attempt to.

The compression of the outcome scale, discussed above, has one further consequence for the timing null. A manipulation predicted to *raise* willingness to cancel has room to operate; one predicted to *lower* it,

which is what H2 required at short notice, has very little. The floor therefore works against the timing hypothesis specifically, and the equivalence test bounds the effect without correcting for the compression.

The vignette also fixes parameters that vary in the population: the standard weekday fee, a first attempt, and a participant who both booked and paid. Those choices maximise the force of the sunk cost, making this a strong but unrepresentative test. Candidates whose parents paid, who are retaking, or who booked at the higher evening rate face different trade-offs, and nothing here speaks to them.

Finally, the sample came from a general online panel rather than from candidates with a live booking, and the two are unlikely to be equivalent where it matters. Holding an appointment you are unsure about is affectively different from imagining it, and the stated-versus-revealed preference literature suggests the gap is widest exactly where present bias and inertia bite hardest. The subgroup analysis reassures partially, since the effect was if anything stronger among provisional holders, but 37 participants is a thin basis and only 12 had a test booked.

4.9 Future Directions

Four extensions follow. The most valuable is a field trial varying the refund schedule inside the current forfeiture window, with early slot release and booking volume measured together so the net capacity effect can be established rather than assumed. Second, the presence-versus-magnitude question should be settled with levels spaced finely at the bottom of the range, since the policy implication differs sharply between the two accounts. Third, the timing question deserves a stronger and better-balanced manipulation, measuring intentions prospectively from candidates at genuinely different distances from their own booked test, and including intervals beyond ten days. Fourth, the exploratory finding that greater notice slightly reduced willingness to cancel should be tested directly rather than inferred from disaggregation.

All four are feasible with existing infrastructure. DVSA records booking, cancellation and attendance per appointment, so a graduated refund trialled in some centres could be evaluated against matched centres without new data collection. The comparison that matters is not whether cancellations rise, which they almost certainly would, but whether delivered tests rise with them.

The finer-grained refund study is the cheapest of the four and would resolve the most consequential ambiguity. If the presence of a refund is doing the work, the policy is inexpensive and the design question becomes where to set the threshold. If the amount is doing the work, the policy is a schedule and its cost scales with its effect. The present data are compatible with both, and the difference between them is the difference between a policy that costs the agency very little and one that costs it a proportion of every refunded fee.

4.10 Conclusion

Willingness to release a scarce driving test appointment was strongly related to how much of the fee would be returned and, at most weakly, to how many days remained before the test. Refund generosity rather than notice period is the operative lever, though the study cannot yet establish whether what matters is the size of the refund or simply its existence. Against a constrained testing system in which every unreleased slot lengthens the queue, the findings support trialling a graduated partial refund in place of the present all-or-nothing rule, with booking behaviour monitored alongside slot release so that the net effect on capacity can be established rather than presumed.

Acknowledgements

Conducted in partnership with the Department for Transport and the Driver and Vehicle Standards Agency.

Use of Generative AI

This assessment falls under UCL’s Category 2, in which generative AI tools may be used in an assistive role. A generative AI assistant (Anthropic’s Claude) was used in that capacity during the preparation of this dissertation: to help structure and draft sections of the text, to refine wording and sentence structure, to write and check the analysis scripts used to produce the reported statistics, and to generate the figures from the analysis output. The research question, study design, scenario wording, data collection and the interpretation of the findings are the author’s own, and all statistical output was inspected and verified by the author against the raw data. The author takes full responsibility for the content of this dissertation.

References

- [1] George Ainslie. 1975. Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin* 82, 4 (1975), 463–496.
- [2] Dan Ariely and Klaus Wertenbroch. 2002. Procrastination, deadlines, and performance: Self-control by precommitment. *Psychological Science* 13, 3 (2002), 219–224.
- [3] Hal R. Arkes and Catherine Blumer. 1985. The psychology of sunk cost. *Organizational Behavior and Human Decision Processes* 35, 1 (1985), 124–140.
- [4] Driver and Vehicle Standards Agency. 2026. Giving you a clearer picture of driving test waiting times. GOV.UK. <https://despatch.blog.gov.uk/2026/06/18/giving-you-a-clearer-picture-of-driving-test-waiting-times/>.
- [5] Shane Frederick, George Loewenstein, and Ted O’Donoghue. 2002. Time discounting and time preference: A critical review. *Journal of Economic Literature* 40, 2 (2002), 351–401.
- [6] GOV.UK. 2026. Change your driving test appointment. Driver and Vehicle Standards Agency. <https://www.gov.uk/change-driving-test>.
- [7] Michael Hallsworth, Dan Berry, Michael Sanders, Anna Sallis, Dominic King, Ivo Vlaev, and Ara Darzi. 2015. Stating appointment costs in SMS reminders reduces missed hospital appointments: Findings from two randomised controlled trials. *PLoS ONE* 10, 9 (2015), e0137306.
- [8] Eric J. Johnson and Daniel Goldstein. 2003. Do defaults save lives? *Science* 302, 5649 (2003), 1338–1339.
- [9] Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler. 1990. Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy* 98, 6 (1990), 1325–1348.
- [10] Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler. 1991. Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives* 5, 1 (1991), 193–206.
- [11] Daniel Kahneman and Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47, 2 (1979), 263–291.
- [12] David Laibson. 1997. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 112, 2 (1997), 443–478.
- [13] National Audit Office. 2025. *Investigation into car driving test waiting times*. Technical Report. National Audit Office. <https://www.nao.org.uk/reports/investigation-into-car-driving-test-waiting-times/>.
- [14] Ted O’Donoghue and Matthew Rabin. 1999. Doing it now or later. *American Economic Review* 89, 1 (1999), 103–124.
- [15] William Samuelson and Richard Zeckhauser. 1988. Status quo bias in decision making. *Journal of Risk and Uncertainty* 1, 1 (1988), 7–59.
- [16] Richard Thaler. 1980. Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization* 1, 1 (1980), 39–60.
- [17] Richard Thaler. 1985. Mental accounting and consumer choice. *Marketing Science* 4, 3 (1985), 199–214.
- [18] Richard H. Thaler. 1999. Mental accounting matters. *Journal of Behavioral Decision Making* 12, 3 (1999), 183–206.
- [19] Richard H. Thaler and Cass R. Sunstein. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press.
- [20] Amos Tversky and Daniel Kahneman. 1991. Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics* 106, 4 (1991), 1039–1061.
- [21] Jinhong Xie and Eitan Gerstner. 2007. Service escape: Profiting from customer cancellations. *Marketing Science* 26, 1 (2007), 18–30.