Australian
National
University

Validated RIS: A within subject test of incentive compatibility in the random (lottery) incentive scheme

by

Evan M. Calford

and

Anujit Chakraborty

ANU Working Papers in Economics and Econometrics
705

October 2025

JEL Codes: C90, D81, D90

ISBN: 0 86831 705 5

Validated RIS: A within subject test of incentive compatibility in the random (lottery) incentive scheme *

Evan M. Calford† and Anujit Chakraborty‡

October 23, 2025

Abstract

We propose the Validated Random Incentive Scheme (VRIS)—an enhancement of the standard Random Incentive Scheme (RIS)—that enables internal validation of RIS's incentive compatibility (IC) without altering an experiment's core design. The key innovation is that researchers can observe each subject under both RIS and single-choice incentives, enabling direct within-person comparisons that form the basis for testing IC. We derive a set of testable predictions from assumptions about the structure of random noise in decision making. In an online experiment involving three risky decisions and a perception task, we use the VRIS design to compare behavior under the RIS with that under single-choice incentives. We find that the RIS distorts the marginal distribution of choices in two of the four decisions and the joint distributions in five of six decision-pairs. In contrast, our novel VRIS directly measures all four marginal distributions under single-choice incentives and enables recovery of undistorted pairwise distributions in all but one case.

Keywords: Experimental Economics, Random Incentive Scheme, Incentive Compatibility, Stochastic Choice
JEL Codes: C90, D81, D90

*We thank Alex Brown, Yoram Halevy, PJ Healy, and Chen Li for helpful discussions and feedback.

†John Mitchell Fellow, Research School of Economics, Australian National University, Canberra, Australia. Evan.Calford@anu.edu.au

‡Department of Economics, University of California, Davis; chakraborty@ucdavis.edu.

1 Introduction

Suppose you wish to run an economics experiment where each subject makes multiple decisions. A common approach to encourage truthful revelation of preferences across all decisions is to use the Random Incentive Scheme (RIS), in which one decision is randomly selected and paid for. A critical question is: how can you validate that subject responses are not being distorted by the additional risk introduced by the RIS, i.e, that the RIS satisfies incentive compatibility (IC)?

Running additional "single-choice" treatments – where each subject faces only one incentivized decision, which is paid with certainty – is the standard way to test whether the RIS satisfies incentive compatibility. These treatments let researchers compare responses under RIS to those under single-choice incentives, providing a benchmark for whether RIS preserves truthful revelation. But such treatments are expensive: each new question of interest requires its own separate sample, making it costly to validate IC across a large set of decisions.

Even more importantly, single-choice treatments only allow researchers to compare marginal distributions – the distribution of responses to each individual question under RIS versus single-choice incentives. Agreement in marginal distributions does not, however, imply agreement in the correlation between questions. In other words, even if RIS reproduces the distribution of choices for Question A and for Question B separately, it may still distort the correlation between A and B without currently available techniques being able to identify it. Thus, the IC of joint choices cannot be tested, even if the IC of every single question, considered alone, cannot be rejected.

This creates a trilemma: researchers must either blindly accept the risk of incentive distortions under RIS *without validation*, or add costly single-choice treatments that only validate IC for individual questions but not for pairs, or abandon the many merits of the RIS in favor of single-choice designs.

In this paper, we make three contributions that essentially resolve this trilemma. First, we introduce the Validated RIS (VRIS), which allows the experimenter to observe the same subject make multiple choices under the RIS *and* make one of those choices under the single-choice incentive, within the same experiment. By systematically varying, between subjects, which decision is incentivized under the single-choice incentive, the experimenter can ensure that all relevant decisions are observed under both the RIS and the single-choice benchmark. Comparing RIS choices to the single-choice benchmark allows for within-subject comparisons that can validate whether the RIS is incentive compatible. Moreover, VRIS also permits testing whether the *joint distribution* of responses to a pair of questions is invariant to which of the two is measured under single-choice incentives, thereby allowing the IC of pairs of questions to be validated as well – this is why we call our method the Validated RIS.

Second, our conceptual framework formally establishes how every experimental test of IC, including ours, is in fact a joint test of both incentive compatibility and assumptions on the stochastic component of choice. We expand the framework of

Azrieli et al. (2018) to incorporate stochastic choice, making the underlying assumptions in different IC tests explicit and providing a new perspective on previous IC tests. Our exercise shows that the VRIS allows for stronger tests of IC than was possible under previous experiments.

Third, by allowing the experimenter to observe the same subject make multiple choices under the RIS, *and* make one of those choices under the single-choice incentive, within the same experiment, the VRIS expands the information recoverable from the experiment. For example, even when the RIS part of the VRIS fails IC on a decision Q, and distorts the observed distribution of behavior relative to the benchmark of single-choice behavior, the VRIS still provides a direct observation of the undistorted behavior on Q. More importantly, we identify a simple, testable condition under which the *undistorted joint distribution* over decisions Q and Q' can be recovered via the VRIS—even if IC failures distort the joint distribution within the RIS.

We implement the VRIS in an online experiment on Prolific with 1,003 subjects, and compare this to a separate sample of 490 subjects who faced only standard single-choice incentives. We find that choices made under the single-choice component of the VRIS-data are statistically indistinguishable from those made under the much more expensive, between-subject single-choice incentive.[1] This validates the use of VRIS as a cost-effective and information-rich alternative: researchers can use it to test for incentive compatibility without sacrificing data quality or requiring large, between-subject samples.

How VRIS works: Consider any experiment with two or more decisions that uses the RIS and call this the *original experiment*. The VRIS version of the same experiment consists of two parts: Part A, an exact replica of the original experiment, and Part B, a single question repeated from Part A. We adopt the convention that the VRIS consists of n total questions, implying that the original RIS experiment contained $n-1$ questions. Before the experiment begins, subjects are informed that they will be paid based on either Part A or Part B decision(s), with equal probability. They are also told that if Part A is selected for payment, one of the $n-1$ decisions will be randomly chosen to determine their earnings (i.e. they are paid via a standard RIS).

Once Part A is completed but before Part B begins, subjects are informed which part, A or B, will determine their payment (Figure 1). If a subject is to be paid for Part A, then her payment is determined via RIS and disclosed before she proceeds to Part B. The subject then responds to a single question in Part B, knowing that response is unincentivized. Conversely, if a subject is to be paid for Part B, she is told that her decisions in Part A will not affect her payment. Instead, the subject's

[1]There is some debate in the literature as to whether stand alone single-choice data is a "better" measure of preferences than the type of single-choice data that can be observed using the VRIS; see Brown and Healy (2018) for a discussion. Our result suggests that, at least for our implementation of the VRIS, this debate is not empirically relevant.

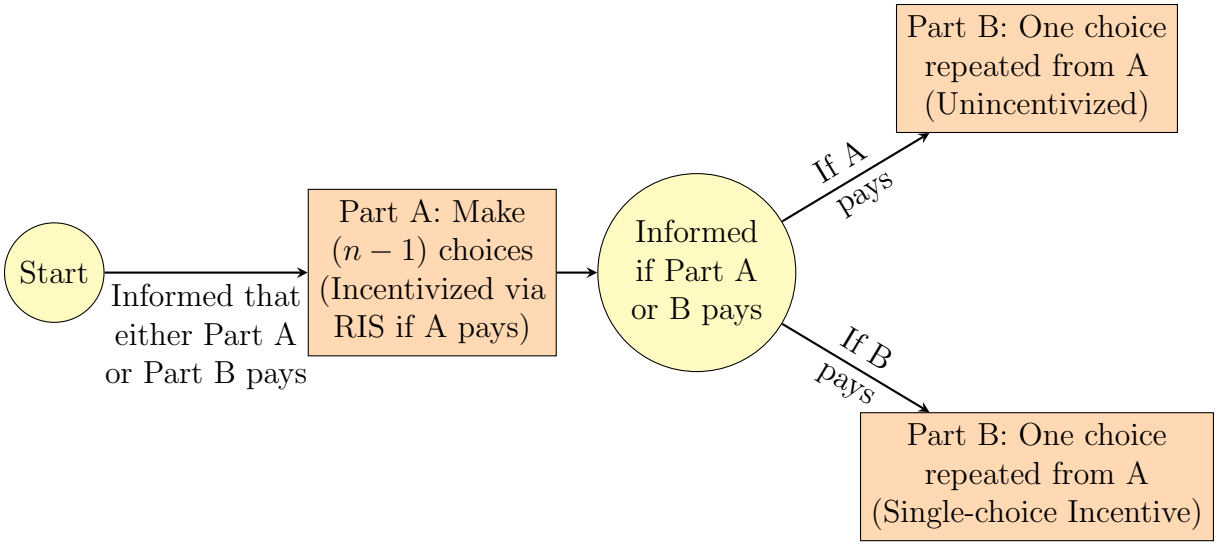

Figure 1: Flowchart of the VRIS Experimental Design. After Part A concludes but before Part B starts, subjects are randomly assigned to one of two branches. Subjects assigned to the *top branch* are informed that they would be paid based on their Part A decisions, and their Part B decision is *unincentivized*. Subjects assigned to the *bottom branch* are told that their Part B decision is the *only* one that pays. The incentive compatibility of the original experiment is assessed by comparing RIS responses from Part A with the single-choice incentivized responses from Part B among subjects in the bottom branch.

response in Part B alone will determine her earnings. Knowing this, the subject then completes Part B, which contains a *single choice* that is a repetition of a question from Part A.

Even though subjects know about parts A and B, and the the payment mechanism, they encounter decision problems one at a time and do not know in advance what decision problems will follow. They are also not told that the decision in Part B is identical to a decision from Part A. These features prevent subjects from anticipating future decisions or recognizing the repetition in advance, ensuring that such factors cannot influence their choices.

Our VRIS Experiment: In an online experiment run on Prolific with 1003 subjects, we apply the VRIS to study three binary lottery choice tasks (*Cer1*, *CIA1*, *MIA1*) and one perception task (*Percep1*). **Cer1** presents a simple binary choice between a sure payoff and a risky lottery. **CIA1** mixes the two alternatives from *Cer1* with the same third lottery and offers a choice between the resulting compound lotteries. **MIA1** reduces both the compound lotteries from *CIA1*, and offers a choice between the resulting reduced lotteries. **Percep1** offers a deterministic payoff, contingent on correctly identifying which of two visual displays contains more black dots.

Name	Option 1	Option 2
Cer1	$l_1 = (1, \$5)$	$l_2 = (0.2, \$0; 0.8, \$7)$
CIA1	$(0.75, l_3; 0.25, l_1)$ where $l_3 = (1, \$0)$	$(0.75, l_3; 0.25, l_2)$
MIA1	$(0.75, \$0; 0.25, \$5)$	$(0.8, \$0; 0.2, \$7)$
Percep1	\$6 if left figure has more black dots than right figure, \$1 otherwise	\$6 if right figure has more black dots than left figure, \$1 otherwise

Table 1: Summary of the four target questions in our VRIS experiment.

For any fixed subject, one of these four tasks is randomly chosen to be *repeated* in Part B of the experiment.[2] The remaining three tasks are each repeated within Part A, and provide an experimental control for the effect of repetition on choice. Thus, for any decision we observe repeated choices under two distinct scenarios: for some subjects, the repetition occurs in Part B under single-choice incentives (or, with no incentives); for everyone else, that repetition occurs within Part A under RIS incentives. This structure allows us to control for the effects of repetition when isolating the impact of incentives in Part A versus Part B.

Four tests of Incentive Compatibility: We define and characterize four distinct tests of Incentive Compatibility of the RIS in Section 3, as shown in Table 2. Each test is a joint test of the incentive compatibility of latent preferences (also called latent IC) and an assumption on stochastic perturbations, where observed behavior is treated as a perturbation of the underlying latent preference.

Test	**Assumption added to latent IC**	**Implication(s)**
Weak Stochastic IC	The distribution of perturbations on question Q is identical across Parts A and B.	(1) The distribution of choices on Q is identical across Parts A and B.
(Q, Q') Partial Stochastic IC	The distribution of perturbations on Q is identical across Parts A and B, conditional on every possible perturbation on Q'.	(2) The joint distribution of choices over (Q, Q') is invariant to whether Q is measured in Part A or B.
(Q, Q') Pairwise Stochastic IC	Both (Q, Q') and (Q', Q) Partial Stochastic IC hold.	(3) The joint distribution of choices over (Q, Q') is invariant to whether Q or Q' is measured in Part A or B.
Deterministic IC	Choices on question Q are observed without any noise or perturbations.	(4) Every subject makes identical choices on Q in Parts A and B.

Table 2: Four tests of Incentive Compatibility (IC) of the RIS.

Weak Stochastic IC is the most permissive test, requiring only that perturbations average out equally across Parts A and B of one question. At the other extreme, De-

[2]Our experiment includes an additional 10 binary choice tasks, as outlined in Table 4, although these questions are only displayed in Part A and never in Part B of the experiment.

terministic IC assumes choices are noise-free in *all questions*, making it the strongest but most restrictive test. If both Q and Q satisfy the Deterministic IC assumptions individually, then it also follows that RIS preserves the correlation between their choices.

Partial and Pairwise Stochastic IC lie in between: they assess whether RIS distorts the correlation between two choices. These conditions are also central for identifying the joint distribution of behavior across questions. Ideally, we would like to observe the hypothetical joint distribution of (Q, Q'), where both choices were made under single-choice incentives— we call this the *target distribution*. While this target distribution is unobservable by definition, it serves as a natural benchmark, representing how choices would look if each decision were made in isolation. In Section 3.3, we show that VRIS can recover an unbiased estimate of this target distribution whenever at least one of (Q, Q') or (Q', Q) satisfies Partial Stochastic IC.

Illustrative example: The usefulness of the four tests in Table 2 is best illustrated by zooming in on a familiar pair of questions: *Cer1* and *MIA1*. Recall that *MIA1* is constructed by mixing the two lotteries from *Cer1* with a common third lottery. The Independence Axiom predicts that subjects should make the same choice in both questions (i.e., choose the riskier option in both or choose the safer option in both).

If we look only at Part A of the VRIS—the data that would also be available in a standard RIS—we find that 59% of subjects (95% CI: [56%, 62%]) behave consistently across *Cer1* and *MIA1*, suggesting they satisfy the Independence Axiom. But here's the catch: behavior on *MIA1* differs significantly across Parts A and B. In fact, *MIA1* fails Weak SIC, and the pair (*MIA1*, *Cer1*) fails Partial SIC. That means that the 59% estimate is biased, because it's based on data that fails both tests of incentive compatibility.

By contrast, we cannot reject that *Cer1* satisfies Weak SIC nor that (*Cer1*, *MIA1*) *does* satisfy Partial SIC. Thus, not only is the behavior on *Cer1* similar across Parts A and B, it is also similar when conditioning for *MIA1* choices. This allows us to construct an unbiased estimate by combining Part A data from *Cer1* with Part B data from *MIA1*. Doing so, we find that 69% of subjects (95% CI: [61%, 77%]) make consistent choices—significantly higher than the naive Part A estimate of 59% ($p = 0.03$). In Section 3.3 we provide the theoretical justification for this recovery method, and in Section 6 we describe the algorithm that implements it in general VRIS applications.

1.1 Road map

Section 2 discusses the prior literature on incentive compatibility in the RIS. Section 3 outlines the theory of incentive compatibility in the presence of stochastic choice, generates testable conditions to establish IC using VRIS data, and outlines when underlying preferences can be recovered from VRIS data. Proofs are gathered in the

Appendix. Section 4 describes our implementation of the VRIS and Section 5 presents the experimental results. Section 6 presents an algorithm that can be used to recover underlying preferences using data generated by the VRIS. Section 7 discusses the interpretation and utility of the data produced by the VRIS, and Section 8 concludes.

A reader who is primarily interested in understanding how to implement the VRIS can focus their attention on Section 4, Section 5 and Section 6.

2 Literature review

The literature on incentive compatibility in experiments has developed along two interconnected pathways: a theoretical pathway and an empirical pathway.

Theoretical pathway. Early theoretical work (Holt, 1986; Karni and Safra, 1987) showed that the incentive compatibility of RIS depends critically on axiomatic assumptions imposed on preferences, like the reduction of compound lotteries and the independence axiom. For example, several common behavioral models such as rank-dependent utility violate both independence and IC (Karni and Safra, 1987), while models like Loomes and Sugden's disappointment aversion can preserve IC even without independence (Loomes and Sugden, 1986; Starmer and Sugden, 1991) because they do not impose reduction. Azrieli, Chambers, and Healy (2018) unified this literature, showing that *statewise monotonicity* is a sufficient condition for IC – we extend their framework to the richer case of stochastic choice.

Empirical pathway. Empirical work, evolving in parallel, typically assumes that when a subject faces a single incentivized decision (*single-choice*), their choice truthfully reveals preferences. In this context, IC is reinterpreted as the empirical equivalence between behavior under a single-choice mechanism and under the payment mechanism under consideration. An Early contribution, Starmer and Sugden (1991), exemplifies a classic *between-subject* design: different groups of subjects face different payment mechanisms, and behavior in the single-choice treatment is compared to that under the RIS.[3] Starmer and Sugden (1991) and other studies—including Camerer (1989), Beattie and Loomes (1997), Cubitt et al. (1998), Aoyama and Hanaki (2024), and Baltussen et al. (2012)—find behavior consistent with IC. In contrast, Freeman and Mayraz (2019), Harrison and Swarthout (2014), Harrison et al. (2015), Aoyama and Hanaki (2021), and Baillon et al. (2022) find evidence against IC. Yet other papers, including Cox et al. (2015) and Freeman et al. (2019), present evidence that supports IC for some treatments or questions but also report evidence against IC for other treatments or questions. A common feature of this literature is its reliance on

[3]We refer to "single-choice" experiments as any design in which subjects face multiple decisions but are informed ex ante that only one prespecified decision will be paid. Starmer and Sugden (1991) call this a real-choice experiment.

smaller sample sizes and relatively low-powered between-subject tests. We catalogue these sample sizes in Appendix B.2.

Among these studies, the design in Camerer (1989) most closely anticipates our approach. Subjects made thirteen binary choices and, after learning which one was paid, were given the opportunity to revise their earlier decision. Only two of eighty subjects chose to change their minds—a remarkably high consistency compared to a 68.4% rate observed when a question was simply repeated within the original set. This highlights an important methodological point: presenting a subject with their previous choice and asking whether they would like to revise it (as in Camerer's design) encourages consistency by framing the task as self-comparison, unlike the neutral repetition used in the VRIS. Recent work, including Kneeland (2015), Breig and Feldman (2024), and Nielsen and Rehbeck (2022), explores related revision designs where subjects can revise decisions before payment uncertainty is resolved, while Cheung and Ellis (2025) use variation in payment timing to implement a between-subject analog of the VRIS, though their design does not generate within-subject data.[4]

Framing and presentation effects. Brown and Healy (2018) highlight two key factors that determine whether the RIS satisfies incentive compatibility. First, *framing effects* arise when the presence of additional decisions alters preferences, even when incentives are unchanged.[5] For instance, if subjects face both $D_1 = \{x, y\}$ and $D_2 = \{x^+, y^-\}$—where x^+ dominates x and y^- is dominated by y—the context provided by D_2 can shift how x and y are perceived. To disentangle such effects from genuine failures of incentive compatibility, Brown and Healy (2018) propose a "Framed Control" single-choice incentive: subjects see the full set of questions as in the RIS but know in advance which single decision determines their payment. A similar approach is used in the K-list treatment of Freeman and Mayraz (2019). This approach isolates IC failures from framing effects. Second, Brown and Healy (2018) show that *choice presentation* also matters. When choices are displayed as a list, the RIS is more likely to violate IC, with behavior in the list being less risk averse (Freeman and Mayraz, 2019), as subjects tend to treat the list as one compound decision. Presenting each question separately prevents this and yields more reliable preference revelation. Our VRIS design incorporates both these insights—framed control and separated presentation—to provide a cleaner test of incentive compatibility.

[4]In our notation, the Cheung and Ellis (2025) design assigns one question to Part A and another to Part B, so each question is asked only once per subject.

[5]Baltussen et al. (2012) present evidence of a related phenomena: carry-over effets from the outcomes of previous tasks.

3 Incentive Compatibility with Stochastic Choice

Let X be a finite set of prizes and $\mathcal{L}(X)$ be the set of all lotteries with support in X.[6] Subjects in an experiment are randomly assigned to a fixed set of questions, $\mathcal{Q}$, where each question $Q \in \mathcal{Q}$ is a menu: $Q \subset X$.

Let $\mathcal{I} = \{1, 2, ..., n\}$. A subject's question sequence is represented by a mapping $d : \mathcal{I} \to \mathcal{Q}$, where $d(i)$ is the i-th question shown. The mapping may assign the same question to multiple indices, allowing for repetitions. For the VRIS, we will use the convention that $d(n)$ is the decision encountered last, in Part B, and we write D_i to denote $d(i)$ and let D represent the vector $(D_1, \ldots, D_n)$.

Message: The data from each subject is called a *message*, which is a vector, ordered across the choices in D. Let $M = \otimes_i D_i$ be the finite set of all such messages. For any message $m \in M$, we will use m_i to refer to the choice made on the i-th decision problem.

Preference-types: Subjects can differ in their *preference-types*. Each *preference-type* $s \in S$ is characterized by a strict preference ordering $\succ_s$ over items in X, that is complete, transitive, and anti-symmetric. There is an unobserved probability distribution $w(\cdot)$ over the subject types such that $w(s) \geq 0$ and $\sum_s w(s) = 1$. Each w uniquely defines a *subject pool*.

Latent Preference: We associate the response vector or message $\mu^s \in M = \otimes_i D_i$ with subject-type s's *latent* preference, if $\mu_i^s \succ_s x$ for all $x \in D_i / \mu_i^s$.

Latent message: The decisions $D_1, .., D_{n-1}, D_n$ are incentivized through a mechanism φ, which maps each message to a lottery: $\varphi : M \to \mathcal{L}(2^X)$. For example, the RIS maps the message m to the lottery $\varphi^{RIS}(m) = (m_1, \frac{1}{n}; \ldots; m_n, \frac{1}{n})$. Given mechanism φ, the subjects choose their optimal message to maximize their utility over the mechanism-delivered outcomes.[7] Let $m^{s,\varphi}$ be the optimal message chosen by type s given mechanism φ. Henceforth, we refer to $m^{s,\varphi}$ as the *latent message*. The latent message maximizes subject utility, conditional on the mechanism, but it need not coincide with latent preferences.[8]

The structure of our model, to this point, follows Azrieli et al. (2018) closely. Azrieli et al. provide preference-based conditions under which the RIS is latent incentive compatible (i.e. the latent message coincides with latent preferences, see Definition 4). Instead, we are interested in how the latent incentive compatibility of the VRIS and RIS translates into testable restrictions on the observed data. Given

[6] Prizes are not restricted in their domain. For example, prizes might be monetary amounts, lotteries, pieces of fruit, or any combination thereof.

[7] For RIS, this can be micro-founded by assuming that $\succ_s$ has a unique extended preference $>_s$ over $\Delta(X)$, that mimics $\succ_s$ over degenerate lotteries. See Azrieli et al. (2018) for details.

[8] For a trivial example, consider a single binary choice question that is incentivized with the mechanism that awards subjects the prize that they did *not* choose. The latent message reports the least preferred item.

that observed data is typically stochastic, rather than deterministic (see Agranov and Ortoleva (2017) and references therein), we extend the deterministic model of Azrieli et al. (2018) by allowing for perturbations of latent messages.

Latent vs Perturbed messages: Fix some mechanism φ. For each subject-type $s \in S$, consider the state-space $(\Omega = M \times M, 2^{\Omega}, P^s)$. For each pair $m, \hat{m} \in M$, the state $(m, \hat{m})$ is understood to be the state where the latent message is m but the analyst observes $\hat{m}$, a potentially *perturbed message*. There are multiple, related, feasible interpretations of why the latent and perturbed messages might differ: perhaps the subject was unable to understand the question and reported some randomly chosen $\hat{m}$, or perhaps the subject intended to report m but trembled and reported $\hat{m}$ by mistake instead, or perhaps the subject was (rationally, maybe) inattentive and reported $\hat{m}$ despite truly preferring m in the counterfactual where she had been fully attentive.

For $x, y \in D_i$, define the event $\mathcal{E}_{i,x\to y} = \{\omega \in \Omega : m_i = x, \widehat{m}_i = y\}$ as the collection of states where a subject's latent message on question D_i was x, but the trembled message was y. In our framework, deterministic choice is the special case where the latent message and trembled message are always identical. Formally,

Definition 1 (Deterministic Choice)**.** $P^s(\mathcal{E}_{i,x\to x}) = 1$ for all $i \in \mathcal{I}$, for all $x \in D_i$, and all $s \in S$ such that $\mu^s_{D_i} = x$.

For the remainder of this section, we focus on the case where the payment mechanism is VRIS.

3.1 VRIS

This subsection develops the theoretical framework for testing incentive compatibility under the VRIS payment mechanism. Our goal is to formalize how behavioral assumptions about preferences and perturbations translate into observable, testable implications. We proceed in stages. First, we define the experimental data structure, introducing the notion of a sub-dataset to capture groups of subjects who faced the same ordered sequence of decision problems. Next, we define Latent Incentive Compatibility, which describes when the VRIS preserves subjects' underlying preferences. We then relax deterministic choice by introducing probabilistic trembles, leading to the conditions of Equal Marginal Probability and Q'-Exchangeability, which ensure that the marginal and joint distributions of choices remain invariant across incentive regimes. Together, these concepts provide the foundation for our later tests of stochastic incentive compatibility.

We begin by defining the basic data structure of the experiment. Each subject is assigned an *ordered set of decision problems*. Many subjects share the same order, and the data from all subjects with that order form what we call a *sub-dataset*.

Definition 2. Let $D = \{D_i\}_{i\in\mathcal{I}}$ denote an ordered sequence of n decision problems drawn from the set of all questions $\mathcal{Q}$. The data collected from all subjects assigned to the same ordered sequence D constitute a *sub-dataset*, also denoted by D.

The next notation allows us to describe precisely how a given question may be repeated in the VRIS experiment, across Parts A and B, and how those repetitions can be used to test incentive compatibility (IC).

Definition 3. For sub-dataset D, let $\mathcal{I}_n^D = \{i : D_i = D_n\}$ be the indices for all decision problems that are identical to its Part B decision problem (D_n).

Next, we introduce the concept of *Latent Incentive Compatibility*, which represents the idealized benchmark in which subjects' latent message coincides with her latent preference. Since latent preference must be identical across identical decisions, in the VRIS context, this also means that the latent message is invariant to both parts (part A or part B) of the VRIS payment structure.

Definition 4. The VRIS mechanism is latent incentive compatible on question $Q \in \mathcal{Q}$ for subject-type $s \in S$ if $m_i^s = \mu_n^s$ for all $i \in \mathcal{I}_n^D$, for all D with $D_n = Q$.

Note that we define incentive compatibility at the level of individual questions, even though in our framework, the subject makes a choice over the set of all possible messages. This allows incentive compatibility to hold for some decision problems but not for others.

We can now state our first result, outlining the testable implications of incentive compatibility for deterministic choices with any sub-dataset.

Proposition 1. Assume that φ^{VRIS} is latent incentive compatible on $Q \in \mathcal{Q}$. If choice is deterministic (Definition 1), then $\hat{m}_i^s = \hat{m}_n^s$ for all $i \in \mathcal{I}_n^D$ with $D_n = Q$ and for all subjects s.

This result also provides our first testable implication: under deterministic behavior, VRIS yields perfectly consistent within-subject choices across repeated questions. Of course, such strict consistency is rarely observed in practice—choice data are inherently noisy. Hence, we next seek assumptions on stochastic perturbations that preserve the same observable behavior across incentive regimes.

Our first condition, *Equal Marginal Probability*, requires that the likelihood of a perturbation for a given question is stable across repetitions of that question, regardless of whether it appears under single-choice or RIS incentives.

Definition 5. Equal Marginal Probability (EMP): $Q \in \mathcal{Q}$ satisfies Equal Marginal Probabilities if for all $i \in \mathcal{I}_n^D$ with $D_n = Q$, for all $x, y \in D_n$, and for all preference-types s with $m_{D_n}^s = x$, $P^s(\mathcal{E}_{i,x\to y}) = P^s(\mathcal{E}_{n,x\to y})$.

To analyze how incentives might shape the correlation of choices across questions, we extend EMP to a joint condition, Q'-*Exchangeability*. It requires EMP on Q to hold conditional on every possible tremble on another question Q', ruling out systematic dependence between the two.

Definition 6. Q'-Exchangeability: Fix $Q' \neq Q$ such that $D_j = Q', D_n = Q$. Q satisfies Q'-Exchangeability if for all $i \in \mathcal{I}_n^D$ with $i \neq n$, for all $x, y_1, y_2 \in D_n$, $a, b \in D_j$ and for all preference-types $s \in S$ with $m^s_{D_j} = a, m^s_{D_n} = x$

$$P^s(\mathcal{E}_{j,a\to b} \cap \mathcal{E}_{i,x\to y_1} \cap \mathcal{E}_{n,x\to y_2}) = P^s(\mathcal{E}_{j,a\to b} \cap \mathcal{E}_{i,x\to y_2} \cap \mathcal{E}_{n,x\to y_1}).$$

It is perhaps easiest to understand Q'-Exchangeability by first considering the weaker condition

$$P^s(\mathcal{E}_{i,x\to y_1} \cap \mathcal{E}_{n,x\to y_2}) = P^s(\mathcal{E}_{i,x\to y_2} \cap \mathcal{E}_{n,x\to y_1})$$

which states that the probability of observing any particular pattern of trembles across two repeated problems, i and n, is symmetric across the repetitions. This symmetry reflects the core intuition behind EMP (see Proposition 2), and in fact, it is equivalent to EMP when $|D_n| = 2$. Q'-Exchangeability strengthens this requirement by insisting that this symmetry holds not just in aggregate, but for every possible realization of trembles on question j. Formally, this is captured by taking intersections with the question-j tremble $\mathcal{E}_{j,a\to b}$ on both sides. This additional conditioning rules out the possibility that the likelihood of a perturbation in decision i is systematically influenced by the perturbation that occurred in decision j, and this property is crucial to identifying the joint distributions of behavior across multiple questions. Intuitively, Q'-Exchangeability prevents the RIS from distorting the correlation structure of responses across questions.

Proposition 2. If Q satisfies Q'-Exchangeability for some $Q' \neq Q$ then it also satisfies Equal Marginal Probability.

In the next subsection we derive testable conditions that are implied by Q'-Exchangeability and EMP. Each testable condition is a joint test of latent incentive compatibility and the properties of the perturbations.

3.2 The testable implications of incentive compatibility

The analyst observes only trembled messages– the reported choices– while the underlying latent messages and the perturbation process that maps them to observed choices remain hidden. Given this limitation, our goal is to formulate conditions on observable data that are implied by our assumptions on the latent stochastic perturbation process and latent IC. Let $\rho^w(\hat{m}_i = x)$ be the expected proportion of reports equal to x in problem i within subject pool w, where the expectation is taken over the distribution of perturbations. Similarly, $\rho^w(\hat{m}_i = x, \hat{m}_j = y)$ is the expected proportion of reports that equal x in problem i and y in problem j.

Our main theorem, Theorem 1, provides necessary and sufficient conditions on the data for a given set of assumptions on the trembles and latent IC to hold. In doing so, it offers a rigorous foundation for testing when the incentive structure distorts, or when decision noise masks, the underlying preferences.

Theorem 1. Assume that φ^{VRIS} is latent incentive compatible on $Q \in \mathcal{Q}$. Then, the following properties hold for all D with $D_n = Q$:

a) (Weak Stochastic IC) Q satisfies Equal Marginal Probability if and only if $\rho^w(\hat{m}_i = x) = \rho^w(\hat{m}_n = x)$ for all $i \in \mathcal{I}_n^D$, for all $x \in D_n$, and for all subject pools w.

b) (Partial Stochastic IC) If Q satisfies Q'-Exchangeability then $\rho^w(\hat{m}_j = a, \hat{m}_i = x) = \rho^w(\hat{m}_j = a, \hat{m}_n = x)$ for all $i \in \mathcal{I}_n^D - \{n\}$, for all $x \in D_n = Q, a \in D_j = Q'$, and for all subject pools w. If $|D_n| = 2$ then the converse also holds.

c) (Pairwise Stochastic IC) Assume additionally that φ^{VRIS} is latent incentive compatible on $Q' \neq Q$. Now consider sub-datasets D and D' such that $D_n = D'_{j'} = Q$ and $D_j = D'_n = Q'$. If Q satisfies Q'-Exchangeability and Q' satisfies Q-Exchangeability then $\rho_D^w(\hat{m}_j = a, \hat{m}_i = x) = \rho_{D'}^w(\hat{m}_{i'} = a, \hat{m}_{j'} = x) = \rho_D^w(\hat{m}_j = a, \hat{m}_n = x) = \rho_{D'}^w(\hat{m}_{n'} = a, \hat{m}_{j'} = x)$ for all $i \in \mathcal{I}_n^D - \{n\}$ and all $i' \in \mathcal{I}'_n - \{n\}$, for all $x \in D_n, a \in D_j$, and for all subject pools w. If $|D_n| = |D'_n| = 2$ then the converse also holds.

Weak Stochastic IC is, perhaps, the most familiar condition. The stipulation that $\rho^w(\hat{m}_i = x) = \rho^w(\hat{m}_n = x)$ requires that the average behavior on question Q is identical, irrespective of whether it shows up under RIS (in Part A) or under single-choice incentives (Part B). This is precisely the test of incentive compatibility that the prior literature (Section 2) has conducted using between-subject designs, without explicitly specifying it as the joint hypothesis of latent IC *and* Equal Marginal Probability.

Partial Stochastic IC presents a stronger condition, one that requires equality of joint probability distributions: the joint distribution of questions Q, Q' must stay identical irrespective of if Q is observed under RIS (part A) or single-choice (part B) incentives. When Q only has two alternatives, this is equivalent to Weak Stochastic IC holding conditional on each possible Q' choice. We will often write (Q, Q') Partial SIC to clarify that we are referring to the case where Q satisfies Q'-Exchangeability.

Any experiment that seeks to uncover the joint distribution of choices over two questions Q, Q' using RIS, such as in tests of the independence axiom, must inherently assume that this joint distribution would remain unchanged if either one, or both, of those decisions were made under single-choice incentives. In doing so, such experiments *implicitly* assume both (Q, Q') Partial SIC and (Q', Q) Partial SIC. That is, they must assume Pairwise SIC. This highlights the central importance of Pairwise SIC in such experiments.

3.3 Recovering joint preferences

While the previous section developed *testable implications* of incentive compatibility, our focus now shifts to *recovery*—that is, whether data observed under the VRIS can

be used to reconstruct the counterfactual distribution of behavior that would have been observed if incentive compatibility were satisfied. Conceptually, testing and recovery are related but distinct: testing asks whether two observable distributions are statistically indistinguishable, whereas recovery asks whether one can approximate an unobservable distribution with an observable distribution. Our *target distribution*, which we seek to recover, is the counterfactual joint distribution where both questions are observed in Part B of the VRIS for all subjects. This target distribution is clearly unobservable, but it provides a conceptual benchmark where the joint distribution of behavior is unaffected by concerns about RIS' incentive compatibility.

In practice, recovery problems rarely involve exact equalities. Even if incentive compatibility cannot be rejected, stochastic perturbations and sampling noise imply that the relevant conditions hold only approximately. Consequently, we will work with inequalities—expressing that two distributions are *close enough*—rather than exact equalities.

Target and observable joint distributions. To study how incentive structure shapes the correlation between responses to a pair of questions, we focus on the distribution of preferences across two questions, (Q, Q'). Define the 4×1 vector

$$\Delta_{BB} = \begin{pmatrix} \rho^{\omega}(\hat{m}_{Q_B} = 0, \hat{m}_{Q'_B} = 0) \\ \rho^{\omega}(\hat{m}_{Q_B} = 0, \hat{m}_{Q'_B} = 1) \\ \rho^{\omega}(\hat{m}_{Q_B} = 1, \hat{m}_{Q'_B} = 0) \\ \rho^{\omega}(\hat{m}_{Q_B} = 1, \hat{m}_{Q'_B} = 1) \end{pmatrix}$$

as the unobserved target distribution of preferences had both questions been measured under Part B incentives. Define $\mathcal{S}_4 = \{x \in \mathbb{R}^4 : x_i \geq 0, \sum x_i = 1\}$.
The VRIS design provides the following three observable 4×1 vectors,

$$\Delta_{AA} = \begin{pmatrix} \rho^{\omega}(\hat{m}_{Q_A} = 0, \hat{m}_{Q'_A} = 0) \\ \vdots \\ \rho^{\omega}(\hat{m}_{Q_A} = 1, \hat{m}_{Q'_A} = 1) \end{pmatrix}, \Delta_{AB} = \begin{pmatrix} \rho^{\omega}(\hat{m}_{Q_A} = 0, \hat{m}_{Q'_B} = 0) \\ \vdots \\ \rho^{\omega}(\hat{m}_{Q_A} = 1, \hat{m}_{Q'_B} = 1) \end{pmatrix}, \Delta_{BA} = \begin{pmatrix} \rho^{\omega}(\hat{m}_{Q_B} = 0, \hat{m}_{Q'_A} = 0) \\ \vdots \\ \rho^{\omega}(\hat{m}_{Q_B} = 1, \hat{m}_{Q'_A} = 1) \end{pmatrix} \in \mathcal{S}_4$$

which can be used to recover the target distribution Δ_{BB}.

Comparison based on approximate equality. Instead of requiring equality of two distributions, we ask whether they are *close enough*—in the sense that the differences between them, measured by the Total Variation Distance (TVD), are statistically or economically negligible. Formally, for $X, Y, Z, W \in A, B$, we say that Δ_{XY} is *approximately equal to* Δ_{ZW} whenever their total variation distance (TVD) is small enough, i.e,

$$\text{TVD}(\Delta_{XY}, \Delta_{ZW}) := \frac{1}{2} \sum_{i,j \in \{0,1\}} \left| \rho^{\omega}(\hat{m}_{Q_X} = i, \hat{m}_{Q'_Y} = j) - \rho^{\omega}(\hat{m}_{Q_Z} = i, \hat{m}_{Q'_W} = j) \right| < \varepsilon$$

for some small $\varepsilon > 0$.

Assumptions relating observed and latent behavior. We assume that the VRIS is latent incentive compatible (Definition 4), and thus the observed joint distributions in Parts A and B can be viewed as noisy transformations of the latent preference distribution $\Delta^* \in \mathcal{S}_4$. Formally, for $X \in A, B$ there exist stochastic matrices G_{XX} describing the perturbation processes in Parts A and B such that:

$$\Delta_{AA} = G_{AA}\Delta^*, \qquad \Delta_{BB} = G_{BB}\Delta^*.$$

where the 4×4 stochastic matrix G_{XX} captures the perturbation of Q, Q' within part X. For example, G_{XX}'s $(1,4)$ entry is $P^s(\mathcal{E}_{Q_X,1\to 0} \cap \mathcal{E}_{Q'_X,1\to 0})$.

Additionally, we shall assume in Theorem 2 that the incentives in Part A make responses weakly noisier than those in Part B. This implies that there exists a stochastic 'garbling' matrix G that maps B-type perturbations into A-type perturbations.

$$G_{AA} = G \times G_{BB} \quad \Rightarrow \quad \Delta_{AA} = G\Delta_{BB}.$$

In this sense, Δ_{AA} is a *Blackwell less informative* signal of latent preferences than Δ_{BB}.

Recovery: In a general experiment, the main objective of the experimenter is to design the environment so the observable data is as close as possible to the counterfactual data that would be observed if each question were fully incentivized in isolation from every other question. Formally, instead of requiring perfect equality, we will express the approximate equality as the Total variation distance being bounded:

$$\mathrm{TVD}(G\Delta_{BB}, \Delta_{BB}) \leq \varepsilon$$

Intuitively this approximate equality can be interpreted in a few different ways; for example, it could mean that the two distributions are close enough so that the experimenter does not care about their distance, or that their equality cannot be rejected via statistical tests. There are two conceptually distinct cases in which this can occur.

1. **Case 1 (Uniform Recovery):** G is approximately equal to the identity matrix, such that $\mathrm{TVD}(G\Delta_{BB}, \Delta_{BB}) \leq \varepsilon$ for every $\Delta_{BB} \in \mathcal{S}_4$. In this case, recovery is robust to the underlying latent preference distribution.

2. **Case 2 (Δ_{BB}-dependent Recovery):** The inequality holds only for some Δ_{BB}—specifically those close to an eigenvector of G with eigenvalue 1. In this case, recovery depends on a fortunate coincidence between the data-generating process and the perturbation structure, and will fail generically.

Hence, in the limit as $\varepsilon \to 0$, uniform recovery dominates: for a generic population (with non-atomic latent distributions), the probability that Δ_{BB} happens to satisfy the second case is vanishingly small.

Uniform recovery and testable implications. The following theorem establishes how uniform recovery connects the unobservable target distribution Δ_{BB} to the three observable ones—Δ_{AA}, Δ_{AB}, and Δ_{BA}—through a set of testable conditions.

Theorem 2. Let $G \in \mathbb{R}^{4\times 4}$ be a stochastic matrix such that $\Delta_{AA} = G\Delta_{BB}$.

a) If

$$\sup_{\Delta_{BB}\in\mathcal{S}_4} \mathrm{TVD}(\Delta_{AA}, \Delta_{BB}) \leq \varepsilon,$$

then

$$\sup_{\Delta_{BB}\in\mathcal{S}_4} \mathrm{TVD}(\Delta_{AA}, \Delta_{BA}) \leq \varepsilon, \tag{1}$$

$$\sup_{\Delta_{BB}\in\mathcal{S}_4} \mathrm{TVD}(\Delta_{AA}, \Delta_{AB}) \leq \varepsilon. \tag{2}$$

Moreover, (1)–(2) together imply

$$\sup_{\Delta_{BB}\in\mathcal{S}_4} \mathrm{TVD}(\Delta_{AA}, \Delta_{BB}) \leq 2\varepsilon.$$

b) If

$$\sup_{\Delta_{BB}\in\mathcal{S}_4} \mathrm{TVD}(\Delta_{AA}, \Delta_{BA}) \leq \varepsilon,$$

then

$$\sup_{\Delta_{BB}\in\mathcal{S}_4} \mathrm{TVD}(\Delta_{AB}, \Delta_{BB}) \leq \varepsilon.$$

Theorem 2 formalizes the practical value of our stochastic IC conditions. Suppose $\varepsilon > 0$ represents the researcher's tolerance level for when two distributions are considered statistically or economically indistinguishable. Part (a) addresses the case where the observed distribution under standard RIS incentives (Δ_{AA}) is "close enough" to the target distribution Δ_{BB}. While Δ_{BB} itself is unobservable, the theorem shows that this condition is equivalent to two testable implications: Equation 1 and Equation 2. Empirically, failing to reject these two conditions corresponds to failing to reject (Q, Q') Partial SIC and (Q', Q) Partial SIC, respectively. Moreover, the result works in both directions: if neither condition is rejected then the deviation between Δ_{AA} and Δ_{BB} must be bounded by at most 2ε.

Part (b) implies that whenever (Q, Q') Partial SIC holds, the cross-incentive distribution Δ_{AB} must approximate the counterfactual target Δ_{BB} within the same tolerance. Hence, even if the standard RIS distribution (Δ_{AA}) fails to replicate the target, as long as Partial SIC holds, VRIS recovers a close empirical analogue.

We provide a simple intuition for how the proof of Theorem 2 works (see Appendix A for the full proof). Consider Δ_{BB} at one of the vertices of the simplex, say

at $\Delta_{BB} = (1, 0, 0, 0)$. In Table 3, we show the implied distributions of $\Delta_{AA}, \Delta_{AB}, \Delta_{BA}$ as functions of the entries of G, which then characterizes the TVD bounds:

$$\text{TVD}\,(\Delta_{AA}, \Delta_{BB}) \leq \varepsilon \iff |G_{11} - 1| + |G_{21}| + |G_{31}| + |G_{41}| \leq 2\varepsilon \quad (3)$$

$$\text{TVD}\,(\Delta_{AB}, \Delta_{BB}) \leq \varepsilon \iff |(G_{11} + G_{21}) - 1| + |G_{31} + G_{41}| \leq 2\varepsilon \quad (4)$$

$$\text{TVD}\,(\Delta_{AA}, \Delta_{BA}) \leq \varepsilon \iff 2|G_{31}| + 2|G_{41}| \leq 2\varepsilon \quad (5)$$

Using $1 = G_{11}+G_{21}+G_{31}+G_{41}$, because G is a stochastic matrix, it is straightforward to verify, on the right hand sides, that (3) $\Rightarrow$ (5) $\Rightarrow$ (4). Hence the corresponding total variation distances bound each other as stated in Theorem 2. A similar exercise shows that the same chain of implications holds for the TVDs at each of the other three vertices of the Δ_{BB} simplex. Since the supremum of a convex operator like TVD must be attained at one of the vertices, and since the TVD relationships hold at all four vertices of the simplex, the proposition follows.

Δ_{BB}	Δ_{AA}	Δ_{AB}	Δ_{BA}
1	G_{11}	$G_{11} + G_{21}$	$G_{11} + G_{31}$
0	G_{21}	0	$G_{21} + G_{41}$
0	G_{31}	$G_{31} + G_{41}$	0
0	G_{41}	0	0

Table 3: A degenerate Δ_{BB} is transformed under G into distributions over Δ_{AA}, Δ_{AB}, and Δ_{BA}.

4 The experimental design

We ran an online experiment using VRIS incentives in June 2023 with a sample of 1003 US residents recruited via Prolific. Subjects were paid $4 for completing the experiment, plus any earnings generated from the VRIS payment mechanism. On average, total compensation per subject was approximately $8 and the median time to completion was 20 minutes.

Each subject made 14 distinct choices in our experimental design, listed as C_1 to C_{14} in Table 4. C_1 to C_{11} were repeated, and thus subjects made a total of 25 choices. We refer to the repeated versions as R_1 to R_{11}. In the first Part, Part A, subjects made 24 binary choices and in the second Part, Part B, subjects made a single binary choice, chosen randomly from one of the *target questions* (R_1 to R_4 in Table 4).

ID (Repeat ID)	Name	Option 1	Option 2
C_1 (R_1)	Cer1	$l_1 = (1, \$5)$	$l_2 = (0.2, \$0; 0.8, \$7)$
C_2 (R_2)	CIA1	$(0.75, l_3; 0.25, l_1)$ where $l_3 = (1, \$0)$	$(0.75, l_3; 0.25, l_2)$
C_3 (R_3)	MIA1	$(0.75, \$0; 0.25, \$5)$	$(0.8, \$0; 0.2, \$7)$
C_4 (R_4)	Percep1	\$6 if left figure has more black dots than right figure, \$1 otherwise	\$6 if right figure has more black dots than left figure, \$1 otherwise
C_5 (R_5)	Mon1	$l_4 = (1, \$4)$	$(1, \$5)$
C_6 (R_6)	FOSD1	$(0.01, \$1.25; 0.19, \$4.50; 0.8, \$9.75)$	$(0.21, \$1.25; 0.18, \$4.50; 0.61, \$9.75)$
C_7 (R_7)	PORU1	$(0.8, l_5; 0.2, l_6)$ where $l_5 = (0.8, \$0; 0.2, \$10)$ and $l_6 = (0.2, \$0; 0.8, \$10)$	$(0.68, \$0; 0.32, \$10)$
C_8 (R_8)	Cer2	$l_7 = (1, \$7)$	$l_8 = (0.2, \$0; 0.8, \$9)$
C_9 (R_9)	CIA2	$(0.5, l_9; 0.5, l_7)$ where $l_9 = (0.16, \$0; 0.68, \$7; 0.16, \$9)$	$(0.5, l_9; 0.5, l_8)$
C_{10} (R_{10})	MIA2	$(0.08, \$0; 0.84, \$7; 0.08, \$9)$	$(0.18, \$0; 0.34, \$7; 0.48, \$9)$
C_{11} (R_{11})	Percep2	\$6 if left figure has more black dots than right figure, \$1 otherwise	\$6 if right figure has more black dots than left figure, \$1 otherwise
C_{12}	Mon2	$(1, \$4)$	$(1, \$4.05)$
C_{13}	FOSD2	$(0.2, l_1; 0.8, l_1)$	$(0.8, l_4; 0.2, l_4)$
C_{14}	PORU2	$(0.5, l_9; 0.5, l_8)$	$(0.18, \$0; 0.34, \$7; 0.48, \$9)$

Table 4: List of all 14 choices. The first 4 choices (highlighted) were the **target questions**: one of these was randomly selected, its repetition was omitted from Part A, and instead shown in Part B. The first 11 questions were repeated and hence have both original and repeat IDs.

4.1 Main focus: The target questions

The study focuses on the four target questions (C_1 to C_4), which we refer to as *Cer1*, *CIA1*, *MIA1* and *Percep1*. Each subject was randomly assigned to one of four treatments. Under each treatment, the repetition of exactly one target question was omitted from Part A and instead presented in Part B (see Figure 2).

The *Percep1* question is not a lottery choice, but is a test of cognitive effort. In this task, subjects were presented with two 10-by-10 dot matrices side-by-side, where each dot was either black or white. Subjects were instructed to select the matrix with the most black dots, receiving \$6 for a correct answer and \$1 for an incorrect one. All subjects saw the same configuration of dots within the matrix, but the configuration of dots was changed between C_4 and its repetition R_4. One set of the dot matrices used for the *Percep1* question are displayed in Figure 3.

An important feature of the *Percep1* question is that preferences of all subjects are known to the experimenter: all subjects prefer choosing the matrix with the most black dots. In the notation of Section 3, there exists a unique preference type, s, such that all subjects have the same preference for this question. Given this, we know that

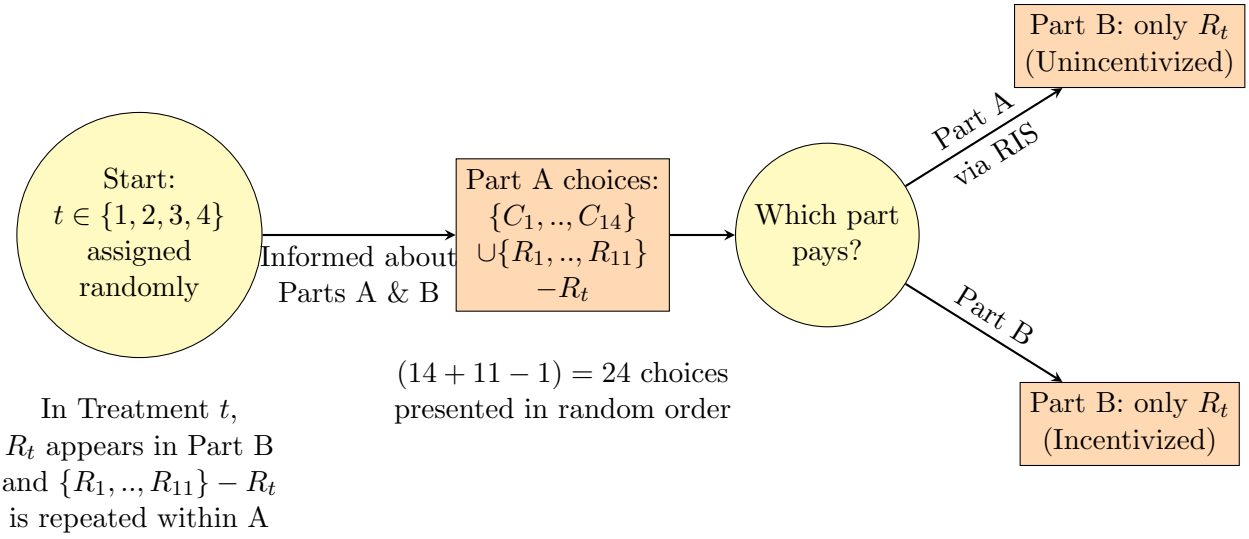

Figure 2: Flowchart of our VRIS Experimental Design. In treatment t, choice R_t is omitted from Part A and becomes the single-choice in Part B.

Figure 3: Two dot matrices used for the *Percep1* question. The left matrix contains 45 black dots, and the right contains 47 black dots.

latent IC is satisfied by construction for this question, and that IC tests involving the *Percep1* question are therefore pure tests of the perturbation function for that question.

The remaining ten non-target questions are detailed in Table 4. Included in the non-target questions is a second set of lotteries, *Cer2*, *CIA2*, and *MIA2*, which were constructed in a similar fashion to the target lotteries. These questions are not the focus of our current study, but can be used to construct subject-level explanatory variables, as we demonstrate in Section 7.2, to analyze the failure of IC.

Timeline of a session: Upon entry into the experiment, subjects were randomized into one of four treatments with each treatment being identified by the target question that the subject was to be shown in Part B. Before Part A began, subjects were informed that there are 2 parts, Part A and Part B, and one of those would be randomly chosen for payment with equal chance. They also knew that Part A had

multiple questions, and if Part A was chosen for payment, then one of the questions from Part A would be randomly chosen to be paid.

Next, instructions for Part A were given, and subjects were tested on their understanding of the choice-objects and how the incentive system works. Subjects had two attempts to correctly answer all comprehension questions, those that did not were excluded from the study.[9] A second randomization, determining which part (A or B) was to be paid, occurred at the end of Part A for each subject.

Before Part B began, subjects were told whether Part A or Part B would pay *and* that Part B had a single question. If Part A was chosen for payment, then each of the 24 questions was selected for payment with a probability of $\frac{1}{24}$. If Part B was chosen for payment then the subjects were paid their choice in the single Part B question.[10] Thus, before making the Part B choice, half of the subjects knew that this single-choice was the only choice that mattered for their payment, and the remaining half knew that this single-choice had no bearing on their payment. All payments were completed within 72 hours of the study ending.

4.2 The robustness treatment

Consider the Part B choices made under single-choice incentives by half of our original sample. Would their choices be different had they never completed Part A and, instead, only faced Part B? To investigate, we ran a robustness treatment in March 2025 with a sample of 490 US residents recruited via Prolific.[11] Subjects were paid $2 for completing the experiment, plus any earnings generated from the single incentivized task. On average, total compensation per subject was approximately $5.30 and the median time to complete was 10 minutes.[12]

The robustness treatment was designed to be as similar as possible to the original experiment, with the only changes to the instructions being amendments to remove references to Part A of the experiment. Each subject responded to one question, which is randomly chosen from the four target questions.

[9] This complied with Prolific requirements for excluding subjects. Excluded subjects are not included in the count of the 1003 subjects who completed the experiment.

[10] Contrast this with Agranov et al. (2023), who also include a Part with a single question (their *One* treatment) but don't reveal which Part will be paid until after all Parts are completed. Agranov et al. (2023) use the single question to infer latent preferences, given an assumption of incentive compatibility.

[11] We targeted a sample of 500 subjects to reflect the design of our original experiment where we expected 500 subjects to be incentivized for Part B. After removing eight subjects who abandoned or otherwise did not complete the experiment, and removing both records of a single subject who completed the experiment twice, we obtained our final sample of 490.

[12] The completion payment was set to half of the amount of the original experiment, after piloting established that the average time taken was half that of the original experiment. The earnings from the incentivized task are slightly lower than in the original experiment because the four target questions exhibited slightly lower earnings than the non-target questions.

5 Results

This section presents the results of our experiment. We begin, in Section 5.1, by testing the incentive compatibility conditions established in Theorem 1. In Section 5.2 we investigate the data from subjects who were not incentivized during Part B. We find a surprising association: two of our target questions exhibit the same behavior for fully incentivized and unincentivized subjects, and these are the same two questions that satisfy Weak Stochastic IC. In Section 5.3 we analyze the data from the robustness treatment. Section 6 provides further analysis and discussion including a demonstration that the data from the VRIS can, in some cases, be used to recover pairwise joint distributions when the Part A RIS data fails Stochastic IC.

Throughout this section we provide both p-values and Anderson's (2008) sharpened q-values, which are constructed using the technique of Benjamini et al. (2006), to correct for the multiple hypothesis testing problem that is introduced by considering four target questions simultaneously. Sharpened q-values control the False Discovery Rate (FDR) which is an alternative to using the Bonferroni correction to control for the Family Wise Error Rate (FWER).[13]

Intuitively, controlling for the FWER would be appropriate if we were intent on rejecting the RIS as a whole whenever we reject the RIS for any individual question; in this case, a false rejection of IC for any question would be particularly problematic. Instead, we take the position that the RIS may satisfy IC for some questions, or in some environments, but not in others. Thus, we are interested in controlling the False Discovery Rate, or the expected rate of false rejections of IC, to be sufficiently small.[14] One implication of controlling for the FDR is that q-values are not independent across tests, and in some cases this can cause q-values to be lower than the corresponding p-values.[15]

There is also a practical reason that we provide sharpened q-values rather than Bonferroni corrections: Readers who prefer controlling for the FWER can easily calculate Bonferroni corrections, while the calculations for the sharpened q-values require machine computation.

[13]Consider a family of hypothesis tests, and suppose that V is the number of false rejections of the null hypotheses and T is the total number of rejections. Then the FWER is given by $\Pr(V > 0)$ and the FDR is given by $E[R]$ where $R = \frac{V}{T}$ whenever $T > 0$ and $R = 0$ if $T = 0$. The FWER and FDR are equal whenever all of the null hypothesis being considered are true, but $E[R] < \Pr(V > 0)$ if some false hypotheses are correctly rejected. This implies that FDR corrections are less extreme than FWER corrections, and the sharpened q-values are more powerful than FWER corrections. The sharpened q-values presented here are only valid if the underlying p-values are not negatively associated.

[14]As is standard, we use a threshold for significance of $\alpha = 0.05$ throughout. We correct for the FDR for each of our different tests of incentive compatibility independently.

[15]If one null hypothesis has very strong evidence against it (increasing T), then we can be slightly more permissive when rejecting the other hypothesis while still controlling the overall expectation $E[R]$.

5.1 Tests of Incentive Compatibility

This section focuses on tests of incentive compatibility across our four target questions. For each question, $Q \in \{\text{Cer1}, \text{CIA1}, \text{MIA1}\}$, each subject made choices in two of four possible scenarios:

1. The first time the question was encountered in Part A (Q_A)
2. The second time the question was encountered in Part A (Q_R, for repeat)
3. When the question was encountered and incentivized in Part B (Q_B)
4. When the question was encountered but unincentivized in Part B (Q_U).

For the *Percep1* question we use a slightly different nomenclature given subjects were shown two distinct sets of dot matrices; in this case we denote with $Percep1_A$ the set of dot matrices that all subjects saw in Part A, and we denote the other set of dot matrices by $Percep1_R$, $Percep1_B$ or $Percep1_U$ when the subject saw this set in either Part A, incentivized in Part B, or unincentivized in Part B, respectively.[16] We will often write $Q_A = x$, as shorthand for $m_{Q_A} = x$, to denote that the message in Q_A was x. Given this, and adapting the notation from Section 3, $\rho(Q_A = x)$ denotes the expected proportion of subjects sending a message of x in Q_A, and $\hat{\rho}(Q_A = x)$ denotes the sample realization observed in the experiment.

We shall use the convention that $Percep1_A = 1$ indicates a correct response to *Percep1*, and $Q_A = 1$ denotes a choice of the less risky option for the three lottery questions. Conversely, a message of 0 indicates an incorrect response to *Percep1* or a risky response to a lottery question.

5.1.1 Testing for Deterministic IC

Deterministic Incentive Compatibility (DIC) requires that $Q_A = Q_B$ for every subject, a restriction that can be refuted when subjects make different choices in Parts A versus B. While it is tempting to claim that DIC is approximately satisfied if, for example, 99 out of 100 subjects satisfy $Q_A = Q_B$, there is no principled threshold for when approximate DIC should be considered acceptable. We recommend using the testing criterion for Stochastic IC established in Section 3 instead.

Nevertheless, Column 4 of Table 5 reports the frequency and confidence intervals of $\rho(Q_A = Q_B)$ for each target question. DIC is not particularly close to holding for any of the questions.

Stochastic IC (SIC) is the natural generalization of DIC for data that has been generated by subjects who exhibit stochastic behavior. Following Theorem 1, we consider three key conditions: Weak SIC, Partial SIC, and Pairwise SIC.

[16] This notation implies that some subjects saw $Percep1_R$ before $Percep1_A$. There were no effects of the order in which the matrices were seen on subject decisions, but the matrices in $Percep1_A$ appear to have been slightly easier than the other set of matrices. See Section 7.1 for details.

5.1.2 Weak SIC

Weak SIC, as implied by Latent IC plus equiprobable trembles across all repetitions, requires that choices are, on average, the same across repetitions of a question. Therefore, our testable hypothesis for Weak SIC is:

$$\rho(Q_R = 1) = \rho(Q_A = 1) = \rho(Q_B = 1). \tag{6}$$

which leads to three testable conditions.[17] We present multiple ways to test Equation 6, and indicate our preferred option below.

The test of $\rho(Q_R = 1) = \rho(Q_B = 1)$ only leverages between subject variation, and we employ a z-test for equality of proportions. This is the same test that is available to previous researchers who used a between subject design (see Section 2) and is useful for comparing the VRIS to previous results.

There are two methods for testing $\rho(Q_A = 1) = \rho(Q_B = 1)$. The first uses a within-subject McNemar test, which leverages the within-subject design features of the VRIS, but limits the sample size only to subjects who were incentivized in Part B. Alternatively, one could run a test for the equality of proportions with partially overlapping samples as proposed in Derrick et al. (2015), and reviewed in Derrick and White (2022), that uses all of the data from question A.[18] The z-statistic for the overlapping samples test is given by

$$z = \frac{\hat{\rho}(Q_A = 1) - \hat{\rho}(Q_B = 1)}{\sigma_A^2 + \sigma_B^2 - 2\phi\sigma_A\sigma_B}$$

where σ_Q is the standard deviation associated with question Q, and ϕ is Pearson's correlation coefficient in the within subject sample.

Our recommendation: We recommend the overlapping samples test for Weak SIC. Intuitively, it can be thought of as performing a between subject test on the full data sample and then correcting the z-statistic for the observed within-subject correlation for subjects who saw the same question in both Parts. In Section B.1 we report simulations that confirm that this overlapping sample test outperforms McNemar's test, and that it is also more powerful than the between subject test comparing question R and question B.

We present the non-overlapping sample tests in Columns 5-7 of Table 5, and summarize the test statistics for the overlapping sample tests in Column 8. It is immediately apparent, with all tests generating the same conclusions, that we can

[17] However, testing all three equalities and then rejecting the joint hypothesis if any of the three individual hypotheses are rejected would lead to a test that rejects the null hypothesis of equality too frequently (approximately 11% of the time at a 5% level of significance, see Appendix B.1 for a discussion and simulations).

[18] To avoid confusion, the test that we are referring to as *the* Derrick et al. (2015) test is referred to as z_8 by Derrick et al. (2015).

reject Weak SIC (Equation 6) for both *Percep1* and *MIA1*, but cannot reject the same for the other two questions.[19]

	(1) $\hat{\rho}(Q_A = 1)$	(2) $\hat{\rho}(Q_R = 1)$	(3) $\hat{\rho}(Q_B = 1)$	(4) $\hat{\rho}(Q_A = Q_B)$	(5) R vs B	(6) A vs B	(7) A vs R	(8) Overlap test
Cer1	0.82	0.82	0.84	0.83	0.67	0.84	0.91	0.69
	$N = 1003$	$N = 742$	$N = 140$	[0.77,0.89]	(0.79)	(0.84)	(0.84)	(0.53)
CIA1	0.73	0.74	0.69	0.75	0.22	0.72	0.65	0.29
	$N = 1003$	$N = 747$	$N = 122$	[0.67,0.82]	(0.35)	(0.79)	(0.79)	(0.24)
MIA1	0.56	0.54	0.75	0.71	<0.001	0.01	0.35	<0.001
	$N = 1003$	$N = 747$	$N = 126$	[0.63,0.79]	(<0.01)	(0.01)	(0.54)	(<0.01)
Percep1	0.72	0.61	0.84	0.71	<0.001	0.047	<0.001	<0.01
	$N = 1003$	$N = 773$	$N = 129$	[0.64,0.79]	(<0.01)	(0.08)	(<0.01)	(<0.01)

Table 5: For each question, columns 1–3 present the raw frequencies in A, R, B. Column 4 reports the frequency of identical choice in A vs B, alongside its 95% confidence interval in square brackets — this serves as a test of Deterministic IC. Columns 5–7 report results from non-overlapping samples test of Weak SIC (Equation 6): The R vs B test is conducted between subject, using a z-test for equality of proportions. The A vs R test and the A vs B test are conducted within subject using McNemar's exact test. Column 8 adds the overlapping sample tests of Equation 6, formed by testing the full data from question A and against the data from Part B (Derrick et al., 2015). Anderson's (2008) sharpened q-values, that adjust p-values for multiple hypotheses testing, are displayed in parentheses; q-values for columns 5 through 7 are calculated independently of q-values for column 8. $p, q \geq 0.05$ are noted in gray.

For the *MIA1* question, the rejection of Weak SIC could be caused be a failure of either the latent IC assumption or the Equal Marginal Probability assumption. However, for the *Percep1* question the rejection of Weak SIC must be caused by the failure of EMP: by construction, preferences are fixed across Part A and Part B for the *Percep1* question given that subjects always prefer to select the image with the most black dots. EMP fails, in this case, because subjects spend longer studying the images (i.e. counting the dots) in Part B when they know that a correct decision will earn \$5 more than an incorrect decision with certainty.[20]

Result 1. Weak SIC is not rejected for *Cer1* and *CIA1*, but is rejected for *Percep1* and *MIA1*.

[19]The interpretation of the overlapping sample test for the three lottery questions is straightforward, given that we cannot reject that the data from the A and R questions is drawn from the same data generating process. The interpretation is a little less clear for the *Percep1* question, given that the A and R data is distinct (and subjects saw a different set of dot matrices in R, see footnote 16). Nevertheless, the result of the overlapping samples test is in concordance with the results found without pooling the A and R questions.

[20]The median time spent on the *Percep1* question was 20 seconds in Part A and 57 seconds in Part B ($p < 0.001$).

5.1.3 Partial and Pairwise SIC

(Q, Q') Partial SIC is implied by Latent IC plus decision problem Q satisfying Q'-Exchangeability, and requires that the joint distribution of Question Q and Question Q' is not altered when using the Part A response or Part B response to Question Q. Our testable hypothesis for (Q, Q') Partial SIC is, therefore,

$$\rho(Q_A = x, Q'_A = x') = \rho(Q_B = x, Q'_A = x') \tag{7}$$

for all $x, x' \in \{0, 1\}$. We present the p-values for these tests in Table 6, where each pairwise set of questions is tested using a Fisher Exact test. Note that Partial SIC is not symmetric across questions: we can reject (*MIA1*,*Cer1*) Partial SIC ($p < 0.001$), but cannot reject (*Cer1*,*MIA1*) Partial SIC ($p = 0.19$).

		Question Q'			
		Cer1	CIA1	MIA1	Percep1
Question Q	Cer1		0.72	0.19	0.85
			(0.40)	(0.14)	(0.40)
	CIA1	0.06		0.67	0.39
		(0.06)		(0.40)	(0.28)
	MIA1	< 0.01	< 0.001		< 0.01
		(< 0.01)	(< 0.01)		(< 0.01)
	Percep1	< 0.001	< 0.001	< 0.001	
		(< 0.01)	(< 0.01)	(< 0.01)	

Table 6: Tests of (Q, Q') Partial Stochastic IC, testing $\hat{\rho}(Q_A, Q'_A) = \hat{\rho}(Q_B, Q'_A)$ using Fisher exact tests. For the *Percep1* question Percep1_R is used in place of Percep1_A. Anderson's (2008) sharpened q-values, that adjust p-values for Multiple Hypotheses testing, are displayed in parentheses.

We cannot reject (*Cer1*,Q') Partial SIC or (*CIA1*,Q') Partial SIC, at the 5% level, for any Q'. On the other hand, *MIA1* and *Percep1*, fail to satisfy (*MIA1*,Q') Partial SIC and (*Percep1*,Q') Partial SIC for all Q'.[21]

Pairwise Stochastic SIC, for a pair of questions Q and Q', can be rejected if either (Q, Q') Partial SIC or (Q', Q) Partial SIC is rejected. It is immediate, from Table 6, that the only pair of questions for which we do not reject Pairwise SIC is the *Cer1* and *CIA1* pair.

Result 2. Partial SIC is not rejected for (*Cer1*,Q') and (*CIA1*,Q') for any Q'; it is rejected for (*MIA1*,Q') and (*Percep1*,Q') for all Q'.

[21] Recall that Proposition 2 establishes that if a question satisfies Q'-Exchangeability, for any Q', then the question also satisfies Equal Marginal Probabilities. It is, therefore, natural that we find a close empirical relationship between Partial SIC and Weak SIC. The two questions that fail Weak SIC also fail Partial SIC.

Result 3. Pairwise SIC is not rejected for *(Cer1,CIA1)*; it is rejected for all other pairs (Q, Q').

5.2 Unincentivized data

The VRIS design allows for a natural test of the effects of incentives on behavior: comparing the aggregate frequency of responses in incentivized Part B to that in unincentivized Part B.[22] Table 7 presents the results of these tests, and also includes the data from the repetition of each question in Part A for comparison. For the *Cer1* and *CIA1* questions we cannot reject the null hypothesis that incentivized and unincentivized behavior is the same. For the *MIA1* and *Percep1* questions incentivized behavior is significantly different from unincentivized behavior.[23] This leads to the following interesting observation:

Result 4. Across all four target questions: responses are indistinguishable across single-choice and RIS choices *if and only if* they are indistinguishable across incentivized and unincentivized choices.

An additional insight our data provide concerns how we evaluate the effect of incentivization itself. A standard test for evaluating the effect of incentives on behavior is to compare unincentivized behavior to behavior incentivized by an RIS. Our findings, however, reinforce the critique of Beattie and Loomes (1997): this test can yield false inferences. For instance, applying that conventional test to the *MIA1* question would fail to reject the null that unincentivized and RIS-incentivized behaviors are identical (R vs U, $p = 0.06$). Yet when we use the appropriate comparison—between incentivized single-choice behavior and unincentivized behavior (B vs U)—we clearly reject the null that incentives have no effect on responses ($p < 0.001$).

5.3 Robustness S-treatment

Besides VRIS, the other way of validating the data collected under RIS is multiple standalone single-choice treatments with additional subjects in a between-subject design. In these treatments, subjects never face any RIS and instead complete a single decision and are paid for that question. We ran four such robustness treatments, one for each of the target questions, to see if the single-choice Part B data under VRIS is statistically different from the between-subject single-choice data (which we denote as the S treatment). We present this data in Table 7. For all four of our target questions, we fail to reject the null hypothesis that behavior in Part B of the original experiment

[22] Even in the absence of monetary incentives, subjects may still respond honestly—out of reciprocity for the participation payment, or due to intrinsic honesty. Additionally, there is no instrumental reason or scope to integrate decisions across questions.

[23] Notably, the unincentivized behavior in *MIA1* and *Percep1* questions is indistinguishable from a 50-50 split between options ($p = 0.29$ for *MIA1* and $p = 0.77$ for *Percep1*).

	$\hat{\rho}(Q_R = 1)$	$\hat{\rho}(Q_B = 1)$	$\hat{\rho}(Q_U = 1)$	$\hat{\rho}(Q_S = 1)$	R vs B	R vs U	B vs U	B vs S
Cer1	0.82	0.84	0.82	0.80	0.67	0.95	0.71	0.47
	$N = 742$	$N = 140$	$N = 121$	$N = 126$	(0.61)	(0.66)	(0.61)	(0.42)
CIA1	0.74	0.69	0.71	0.60	0.22	0.43	0.72	0.15
	$N = 747$	$N = 122$	$N = 134$	$N = 120$	(0.23)	(0.48)	(0.61)	(0.33)
MIA1	0.54	0.75	0.45	0.64	$< .001$	.06	$< .001$	.06
	$N = 747$	$N = 126$	$N = 130$	$N = 118$	($< .01$)	(0.08)	($< .01$)	(0.33)
Percep1	0.61	0.84	0.49	0.79	$< .001$	.01	$< .001$	.22
	$N = 773$	$N = 129$	$N = 101$	$N = 126$	($< .01$)	($< .01$)	($< .01$)	(0.33)

Table 7: Comparison of response rates across the four incentive regimes: repeated incentivized (R), Part B incentivized (B), Part B unincentivized (U), and single-question incentivized (S). All tests are two-sample between-subject tests of proportions. Anderson's (2008) sharpened q-values, adjusted for multiple hypothesis testing, are reported in parentheses with columns 6-8 being adjusted independently of column 9. Insignificant results ($p \geq 0.05$) are shown in gray.

is identical to the behavior in the robustness treatments. The uncorrected p-value for the *MIA1* question is close to the standard threshold for significance ($p = 0.06$), but the sharpened q-value is much further from the threshold ($q = 0.33$).

Result 5. For all four of our target questions, we cannot reject the null hypothesis that behavior is the same in Part B of the VRIS and in the single-choice Robustness treatment.

6 Estimating pairwise joint distributions

We now apply the recovery framework from Section 3.3. For two questions Q, Q', let

$$\Delta_{XY} = \begin{pmatrix} \rho(\hat{m}_{Q_X} = 0, \hat{m}_{Q'_Y} = 0) \\ \rho(\hat{m}_{Q_X} = 0, \hat{m}_{Q'_Y} = 1) \\ \rho(\hat{m}_{Q_X} = 1, \hat{m}_{Q'_Y} = 0) \\ \rho(\hat{m}_{Q_X} = 1, \hat{m}_{Q'_Y} = 1) \end{pmatrix}$$

denote the vector of joint expected choice frequencies when Q is measured in Part X and Q' in Part Y, and let $\hat{\Delta}_{XY}$ denote its empirical realization observed in the VRIS.

A common goal in experiments is to estimate the joint distribution of choices across multiple decision problems. For two questions Q and Q', the ideal—but unobservable—target is the joint distribution Δ_{BB}, where both choices are elicited under single-choice incentives. In practice, we only observe three distinct joint distributions from distinct subsets of the sample: (1) $\hat{\Delta}_{BA}$ — from subjects who faced Q in Part B (available only in VRIS), (2) $\hat{\Delta}_{AB}$ — from subjects who faced Q' in Part B (available only in VRIS), and (3) $\hat{\Delta}_{AA}$ — from the Part A responses of all subjects (available in RIS and VRIS).

Theorem 2 provides a blueprint for how these observables can be used to recover the unobservable target. Informally, the theorem establishes that approximate equality between Δ_{AA} and Δ_{BA} (or between Δ_{AA} and Δ_{AB}) implies that Δ_{AB} (or Δ_{BA}) approximate the target joint distribution. Formally:

$$\Delta_{AA} \approx \Delta_{BA} \Rightarrow^* \Delta_{AB} \approx \color{gray}{\Delta_{BB}} \tag{8}$$

$$\Delta_{AA} \approx \Delta_{AB} \Rightarrow^* \Delta_{BA} \approx \color{gray}{\Delta_{BB}} \tag{9}$$

$$\Delta_{AA} \approx \color{gray}{\Delta_{BB}} \iff^* \begin{cases} \Delta_{AA} \approx \Delta_{BA}, \\ \Delta_{AA} \approx \Delta_{AB}. \end{cases} \tag{10}$$

where, $\approx$ denotes that the total variation distance between two distributions (Section 3.3) is smaller than some threshold $\varepsilon > 0$, and $\Rightarrow^*$ is used to indicate that the implication holds for sufficiently small ε. We write the latent unobservable target, Δ_{BB}, in gray to separate it from the directly observable distributions. Only the conditions on the left hand side of Equation 8 and Equation 9 (which are the same equations as the right hand side of Equation 10) are testable because both quantities involved are observable, and our recovery method builds on these testable conditions.

Algorithmic implementation. Algorithm 1 operationalizes the logic above. For the unobservable target distribution Δ_{BB} to be approximately recovered from observable data, at least one of two necessary conditions must hold: either $\hat{\Delta}_{AA} \approx \hat{\Delta}_{BA}$ or $\hat{\Delta}_{AA} \approx \hat{\Delta}_{AB}$. The algorithm thus begins by testing these conditions in Steps 1.1 and 1.2. In Step 2, if exactly one of the two condition holds, the corresponding recovery rule is applied, using $\hat{\Delta}_{AB}$ or $\hat{\Delta}_{BA}$ as the estimate, respectively. If both conditions hold, the algorithm uses $\hat{\Delta}_{AA}$ as the estimate, leveraging its larger sample size. If both conditions are rejected, the algorithm halts - reflecting that Δ_{BB} cannot be estimated by the VRIS data.

Intuitively, whenever $\hat{\Delta}_{AA} \approx \hat{\Delta}_{BA}$, we can substitute Part A data for Q with Part B data without materially changing its joint distribution with Q'. This equivalence holds in the *uniform recovery* case (see Section 3.3), where the similarity reflects genuine equality in behavior across incentive conditions rather than a spurious equality arising from a coincidental alignment between the data-generating process and the perturbation structure. As $\varepsilon \to 0$, the probability of such spurious equality vanishes, allowing us to focus on the uniform recovery case.

Design conditions for uniform recovery. In practical terms whether an experiment achieves genuine equality in behavior across incentive conditions (uniform recovery) depends partly on experimental implementation. Achieving this requires creating experimental conditions such as:

(a) Designing the experimental interface and instructions to encourage subjects to *isolate* decisions, i.e, treat each decision as the only decision they would face,

(b) Offering *sufficient incentives* to encourage attentive responses to all decisions, including decisions embedded within the Part A RIS,

(c) *Filtering for attentive subjects* through screening questions or attention checks.[24]

Brown and Healy (2018) highlight, for example, that subjects are more likely to isolate decisions when each question is presented on a different page compared to placing multiple questions on a single page.

VRIS vs RIS recovery. Our testing-based recovery approach highlights the value of VRIS: even when one of Equation 8 or Equation 9 fails, VRIS can still recover the ideal unobservable joint distribution of behavior. In contrast, RIS—which only observes Δ_{AA}— requires both Equation 8 and Equation 9 to hold, even though the RIS data is not sufficient to test either. When both testable conditions fail, neither RIS nor VRIS can succeed.

Next, we apply this algorithm to our target questions, and in particular to each of the three question pairs that test canonical axioms of decision theory. For each axiom, consistency implies identical choices across a specific pair of questions:

- *Independence (IND)*: Implies identical choices in *Cer1* vs. *MIA1*
- *Reduction of Compound Lotteries (ROCL)*: Implies identical choices in *CIA1* vs. *MIA1*
- *Compound Independence (C-IND)*: Implies identical choices in *Cer1* vs. *CIA1*

Table 8 summarizes the outcome of applying Algorithm 1 to each axiom test. The first three columns list the axiom and associated question pair. The next three columns (*Algo-steps*) report which of the two Step 1 tests were satisfied (i.e., where the null of equality could not be rejected), followed by the recovery rule used in Step 2. A ✓ indicates the null could not be rejected, and ✗ indicates rejection. The final columns compare behavioral estimates from the original RIS data (i.e. $\hat{\Delta}_{AA}$) and the corrected VRIS data (i.e. the output of Algorithm 1).

For instance, for the Independence Axiom (*Cer1* vs. *MIA1*), Step 1.2 held but 1.1 did not, so we used $\hat{\Delta}_{BA}$ in Step 2 to recover the joint distribution. This correction increased the estimate of adherence with the Indpendence axiom from 59% of subjects to 69% of subjects ($p = 0.03$). That is, the RIS provides a biased estimate of adherence to the Independence Axiom which can be corrected using the VRIS. For ROCL (*CIA1* vs. *MIA1*), the same procedure applied, but the overall estimate changed only marginally (0.61 to 0.63, $p = 0.52$). However, using RIS data, subjects who violated ROCL exhibited a significant bias toward the safer lottery in *CIA1* ($p < 0.001$), but

[24]Specifically, the benefit arises from screening for subjects who are attentive to all aspects of the experiment, irrespective of the strength of incentives. This suggests that unincentivized attention checks may be preferred.

Algorithm 1 Algorithm for recovering unobservable Δ_{BB} using three data sources. Comments are in gray and start with "//".

Input: Questions Q and Q' from VRIS data (both Part A and Part B)
Initialize: $\delta_{ba} \leftarrow null$, $\delta_{ab} \leftarrow null$

Step 1: Test for Partial SIC

if H_0: $\Delta_{AA} = \Delta_{AB}$ is not rejected **then** //**Step 1.1:** Verify if $\hat{\Delta}_{AA} \approx \hat{\Delta}_{AB}$
  $\delta_{ba} \leftarrow \hat{\Delta}_{BA}$ // Implies $\hat{\Delta}_{BA} \approx \Delta_{BB}$
end if

if H_0: $\Delta_{AA} = \Delta_{BA}$ is not rejected **then** //**Step 1.2:** Verify if $\hat{\Delta}_{AA} \approx \hat{\Delta}_{BA}$
  $\delta_{ab} \leftarrow \hat{\Delta}_{AB}$ // Implies $\hat{\Delta}_{AB} \approx \Delta_{BB}$
end if

Step 2 (Recovery): Which data recovers Δ_{BB}

if $\delta_{ba} = null$ and $\delta_{ab} = null$ **then**
  Exit: Recovery impossible // No observable distribution equals Δ_{BB}
else if $\delta_{ba} \neq null$ and $\delta_{ab} \neq null$ **then**
  Use $\hat{\Delta}_{AA}$ to recover Δ_{BB} // Use the largest feasible sample
else if $\delta_{ba} \neq null$ **then**
  Use $\delta_{ba} = \hat{\Delta}_{BA}$ to recover Δ_{BB} // Use data only from subjects who saw Q in part B
else if $\rho_{ab} \neq null$ **then**
  Use $\delta_{ab} = \hat{\Delta}_{AB}$ to recover Δ_{BB}// Use data only from subjects who saw Q' in part B
end if

			Algorithm steps			$\hat{\rho}(Q = Q')$		$\hat{\rho}(Q = 1) - \hat{\rho}(Q' = 1)$	
Axiom	Q	Q'	1.1	1.2	2. Recovery	RIS	VRIS	RIS	VRIS
IND	Cer1	MIA1	✗	✓	$\hat{\rho}(Q_A, Q'_B)$	0.59 [0.56, 0.62]	0.69 [0.61, 0.77]	0.26 [0.22, 0.30]	0.12 [0.02, 0.22]
ROCL	CIA1	MIA1	✗	✓	$\hat{\rho}(Q_A, Q'_B)$	0.61 [0.57, 0.64]	0.63 [0.55, 0.72]	0.17 [0.13, 0.21]	-0.03 [-0.15, 0.08]
C-IND	CIA1	Cer1	✓	✓	$\hat{\rho}(Q_A, Q'_A)$	0.77 [0.74, 0.80]	n.c. n.c.	-0.09 [-0.12, -0.06]	n.c. n.c.

Table 8: **Applying the algorithm to our data:** The first three columns specify the axiom and the pair of questions being tested. The next three columns show the outcomes of the statistical tests corresponding to Steps 1.1 and 1.2 in Algorithm 1, and the resulting recovery method from Step 2. A ✓ indicates that the corresponding condition held; a ✗ indicates it did not. The next four columns show the proportion of subjects who satisfy each axiom, measured using either raw RIS data (Part A only) or the corrected VRIS data (combining Parts A and B via the algorithm).

this asymmetry vanished after correcting for stochastic IC violation using the VRIS ($p = 0.66$). For C-IND (*Cer1* vs. *CIA1*), both questions show identical responses across incentives, so VRIS produced no change—consistent with the theory.

7 Discussion

7.1 Framing and order effects

Brown and Healy (2018) emphasize that tests of incentive compatibility should be conducted within a consistent frame, as preferences can be influenced by the frame in which they are observed. Thus, to test whether the presence (or absence of) a particular payment mechanism alters preferences, subjects should be presented with the same set of questions, regardless of the mechanism. The VRIS design satisfies this condition as it displays the same set of questions to all subjects, by construction. However, there a residual concern remains: Subjects who repeat a decision in Part B will have seen strictly more questions than those who repeated the same question in Part A (incentivized by the RIS). Could this additional exposure influence their choices and account for any observed differences in responses? Evidence from the Robustness treatment (Section 5.3) already rules this out.

However, the Robustness treatment imposes additional cost and complexity, and so may not be included in all implementations of VRIS. For this reason, we also propose and conduct a second approach to testing for framing or order effects that uses only data from the core VRIS design. Each subject in our implementation of the VRIS faces ten distinct pairs of repeated questions in Part A, allowing us to test whether responses to a given question remain consistent between its first and second appearance.[25] Failure to reject equality of responses supports the claim that exposure to intervening questions does not meaningfully shift preferences.

For the nine preference questions that are identical across repetitions, we use a McNemar exact test. We cannot reject the null hypothesis of equality of responses across repetitions, at a 5% significance level, for any of the nine questions.[26]

The remaining two questions, *Percep1* and *Percep2*, involve perceptual judgments with objectively correct answers. Each displays a pair of dot matrices differing by two dots (45 vs. 47 for *Percep1*; 55 vs. 57 for *Percep2*), with dot positions permuted across repetitions. Using χ^2 tests, we fail to reject the null hypothesis of no association

[25] Subjects also answer an 11th repeated question, but only one of the two appearances of this question occurs in Part A.

[26] All nine questions have a FDR adjusted q-value of $q = 1$. The power for these within-subject McNemar tests varies with the expected level of consistency across repetitions. If 80% of subjects are expected to report consistent preferences across repetitions, a figure which is broadly consistent with the data in our lottery questions then we have over 70% power to detect an odds ratio of 1.5 for our Cer1, CIA1 and MIA1 questions (where our sample size is approximately 750 for each question) and over 80% power for the remaining questions (where our sample size is 1003).

between the likelihood of correct answer and order, for either question (*Percep1*: $p = 0.26$; *Percep2*: $p = 0.94$).

In sum, our direct tests find no evidence of order or framing effects in the sense of Brown and Healy (2018), using only within-subject data from the VRIS. These findings reinforce the results from the Robustness treatment and support the internal validity of our design.

7.2 Statewise Monotonicity and consistency across incentives

Many of the papers cited in Section 2 use the term *isolation* when discussing incentive compatibility of the RIS. A subject exhibits isolation in the context of the RIS, when, she treats each decision problem independently of the other decision problems; in other words, the subject treats each decision in isolation.[27]

Within Azrieli et al. (2018, 2019)'s framework, isolation can be given a more precise interpretation. The idea that each decision within the RIS is evaluated independently corresponds closely to the *Compound Independence Axiom (CIA)* introduced by Segal (1990), or equivalently to *Statewise Monotonicity (SM)* axiom discussed in Azrieli et al. (2018, 2019). When a subject's latent preferences satisfy CIA or SM, her utility-maximizing choice in any one decision is independent of the presence of other decisions within the RIS - in other words, CIA or SM guarantees Latent IC. We utilize this conceptual bridge between isolation and the testable behavioral conditions of CIA/ SM in the subsequent discussion.

Building on this interpretation, we note that observed consistency between Parts A and B of the experiment reflects two distinct forces. First, it depends on whether the subject's latent preferences satisfy *Latent Incentive Compatibility*—that is, whether the utility-maximizing latent message is identical across incentive conditions. Second, it depends on the subject's individual-level *perturbation rate*: how often the observed (possibly noisy) message coincides with the latent message. Hence, AB-consistency is jointly determined by the structure of latent preferences (via CIA/SM) and by the reliability with which preferences are expressed in choice data.

To empirically quantify these two mechanisms, we proceed as follows. We construct an individual-level proxy for adherence to the *SM/ CIA* axiom using responses to *Cer2* and *CIA2*. The SM proxy equals one if and only if $\hat{m}_{\text{Cer2}_A} = \hat{m}_{\text{CIA2}_A}$ and $\hat{m}_{\text{Cer2}_R} = \hat{m}_{\text{CIA2}_R}$. This variable (noisily) captures whether a subject's choices in these structurally related questions remain stable across repetitions and incentive conditions, as would be implied by SM.

To account for subject-level heterogeneity in perturbation rates - i.e., how noisy each subject's choices are *for that task* - we construct a task-specific consistency measure. For each question type $X \in \text{Cer}, \text{MIA}, \text{CIA}$, the task-specific indicator equals

[27] In contrast, a subject exhibits integration when she treats the entire RIS experiment as a single decision problem.

one if $\hat{m}_{\text{X2}_A} = \hat{m}_{\text{X2}_R}$. This captures within-question consistency across repetitions, independent of Latent IC.

Finally, to ensure that our results are not merely driven by general cross-question consistency unrelated to SM, we construct a placebo proxy by substituting *MIA2* for *CIA2*: it equals one if and only if $\hat{m}_{\text{Cer2}_A} = \hat{m}_{\text{MIA2}_A}$ and $\hat{m}_{\text{Cer2}_R} = \hat{m}_{\text{MIA2}_R}$. Because SM makes no predictions about the relationship between these two questions, any predictive power of this placebo variable would suggest confounding factors unrelated to isolation or CIA.[28]

Next, we regress AB-consistency, defined as an indicator for identical responses in one of the three questions (*Cer1*, *CIA1*, or *MIA1*) across Parts A and B (incentivized), on the SM proxy, placebo proxy, and task-specific consistency indicator. Note that while the dependent variable is based on *Cer1*, *CIA1*, and *MIA1*, the explanatory variables rely on *Cer2*, *CIA2*, and *MIA2*, ensuring that our regressors and outcome are based on distinct tasks and thus not mechanically correlated.

Table 10 reports results from a linear probability regression.[29] In both specifications, the SM proxy significantly predicts consistency between Parts A and B, with a coefficient between 11 and 14 percentage points. Subjects whose choices satisfy SM are therefore more likely to behave consistently across incentive conditions. By contrast, the placebo proxy has no predictive power, and the task-specific indicator has an additional positive effect of 15 percentage points.

These findings provide direct evidence that adherence to CIA/SM—a behavioral manifestation of isolation—predicts the degree to which individual behavior remains stable across incentive regimes. In other words, violations of isolation appear to stem not only from random perturbations or noise, but also from systematic departures from statewise monotonicity in subjects' latent preferences. This link between CIA/SM and consistency across incentives highlights how the theoretical structure of isolation maps onto observable patterns of stochastic incentive compatibility.

As an additional placebo test we repeat the same exercise for the *Percep1* question, in Table 10. We should not expect either the SM proxy (a measure of latent IC) nor the placebo proxy to affect AB consistency for the *Percep1*, given that latent IC should be trivially satisfied for all subjects. The results in Table 10 are consistent with this conjecture.

8 Conclusion

This paper proposed and implemented the Validated Random Incentive Scheme (VRIS) - a practical enhancement of the standard Random Incentive Scheme (RIS) - that allows internal validation of incentive compatibility within the same experimental

[28] The SM proxy and placebo proxy are discordant for approximately 30% of our sample, indicating sufficient independent variation for estimation.

[29] The marginal effects calculated from a logit regression are similar to the coefficients found using the linear probability model.

Table 9: Linear regression: Comparison of AB Consistency Regressions Across Samples

	Three Lottery Questions		*Percep1* question	
	(1)	(2)	(1)	(2)
	AB consistent	AB consistent	AB consistent	AB consistent
Statewise Mon. Indicator	0.14**	0.11*	-0.02	-0.06
	(0.05)	(0.05)	(0.10)	(0.10)
Placebo Indicator	0.05	0.01	0.06	0.07
	(0.05)	(0.05)	(0.09)	(0.09)
Task-Specific Consistency		0.15*		0.11
		(0.06)		(0.09)
Constant	0.63***	0.55***	0.69***	0.64***
	(0.04)	(0.06)	(0.09)	(0.10)
N	388	388	129	129

Table 10: Linear probability model of AB consistency for the three lottery questions and *Percep1*. Standard errors are shown in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

design. The key idea is conceptually simple but empirically powerful: by observing each participant's choices under both RIS and single-choice incentives, one can test whether the RIS elicits the preferences it intends to reveal.

Our conceptual framework shows that the validity of RIS depends jointly on two behavioral forces: latent incentive compatibility (whether latent preferences are stable across incentive regimes) and perturbation rates across incentives (how observed choices deviate from latent preferences within and across questions). Our theory leverages the VRIS design to define stronger tests for latent incentive compatibility than was previously possible.

Empirically, using data from an online experiment involving three risky decisions and a perception task, we implement our tests and find that the standard RIS systematically distorts behavior in a subset of decisions. The distortions affect not only marginal choice frequencies in two out of four tasks, but also the joint distributions across decision pairs. We show how the VRIS framework can be used to successfully recover underlying joint choice distributions in such cases.

Beyond this specific implementation, the VRIS introduces a broader methodological contribution. It offers a template for self-validating experimental designs - those that can test their own incentive assumptions without external validation. We hope that this approach can be applied to a wide range of experimental settings, from risk and time preferences to strategic and social-choice environments.

A Proofs

Proof of Proposition 1. First, note that uniqueness of preference ensures that $\mu_i^s = \mu_j^s$ for all $i, j \in \mathcal{I}_n^D$ and that latent incentive compatibility implies that $m_i^{s^*} = \mu_i^s$ for all $i \in \mathcal{I}_n^D$ and for all s.

Definition 1 implies that $m_i^{s^*} = \hat{m}_i^s$ for all $i \in \mathcal{I}_n^D$ and for all s. Therefore, $\hat{m}_i^s = m_i^{s^*} = \mu_i^s = \mu_n^s = m_n^{s^*} = \hat{m}_n^s$ for all $i \in \mathcal{I}_n^D$ and all s. □

Proof of Proposition 2.

$$\begin{aligned} P^s(\mathcal{E}_{i_1,x\to y}) &= P^s(\mathcal{E}_{i_1,x\to y} \cap \mathcal{E}_{i_2,x\to y}) + \sum_{z\neq y} P^s(\mathcal{E}_{i_1,x\to y} \cap \mathcal{E}_{i_2,x\to z}) \\ &= P^s(\mathcal{E}_{i_2,x\to y} \cap \mathcal{E}_{i_1,x\to y}) + \sum_{z\neq y} P^s(\mathcal{E}_{i_2,x\to y} \cap \mathcal{E}_{i_1,x\to z}) \\ &= P^s(\mathcal{E}_{i_2,x\to y}) \end{aligned}$$

where we have used the fact that Q'-Exchangeability conditional on $\mathcal{E}_{j,a\to b}$ for all b implies unconditional exchangeability when moving from line 1 to line 2. □

Proof of Theorem 1. For all parts of the proof, recognize that latent incentive compatibility and unique preferences imply that the optimal choice in both D_i and D_n with $i \in \mathcal{I}_n^D$ is $m_n^{s*} = \mu_n^s = \mu_i^s = m_i^{s*}$ for all subjects s.

(a) For any w,

$$\begin{aligned} \rho^w(\hat{m}_i = x) &= \sum_{\{s:m_i^{s*}=x\}} w(s)P^s(\mathcal{E}^s_{i,x\to x}) + \sum_{y\in D_n-\{x\}} \sum_{\{s:m_i^{s*}=y\}} w(s)P^s(\mathcal{E}^s_{i,y\to x}) \\ &= \sum_{\{s:m_i^{s*}=x\}} w(s)P^s(\mathcal{E}^s_{n,x\to x}) + \sum_{y\in D_n-\{x\}} \sum_{\{s:m_i^{s*}=y\}} w(s)P^s(\mathcal{E}^s_{n,y\to x}) \\ &= \sum_{\{s:m_n^{s*}=x\}} w(s)P^s(\mathcal{E}^s_{n,x\to x}) + \sum_{y\in D_n-\{x\}} \sum_{\{s:m_n^{s*}=y\}} w(s)P^s(\mathcal{E}^s_{n,y\to x}) \\ &= \rho^w(\hat{m}_n = x) \end{aligned}$$

To prove the opposite direction, fix any pair of alternatives x, y with $x, y \in D_n$, and fix any preference-type s' with $\mu_i^{s'} = x$. Choose a subject pool w such that it assigns weight 1 to the preference-type s', i.e, all subjects are of type s'. Then, for this subject pool,

$$\begin{aligned} P^{s'}(\mathcal{E}_{i,x\to y}) &= \rho^w(\hat{m}_i = y) \\ &= \rho^w(\hat{m}_n = y) \\ &= P^{s'}(\mathcal{E}_{n,x\to y}) \end{aligned}$$

Any subject pool can then be constructed as a linear combination of these single-type subject pools.

b) We prove in two steps.

Step 1: Show Q'-Exchangeability is necessary and sufficient for the following property $\mathcal{P}$

$$P^s(\mathcal{E}_{j,a\to b} \cap \mathcal{E}_{i,x\to y}) = P^s(\mathcal{E}_{j,a\to b} \cap \mathcal{E}_{n,x\to y}) \quad ,\forall i \in \mathcal{I}_n^D - \{n\}, \forall x, y \in D_n$$

Fix any i and x, y satisfying the criteria above. Sufficiency of Q'-Exchangeability follows from

$$\begin{aligned} P^s(\mathcal{E}_{j,a\to b} \cap \mathcal{E}_{i,x\to y}) &= P^s(\mathcal{E}_{j,a\to b} \cap \mathcal{E}_{i,x\to y} \cap \mathcal{E}_{n,x\to y}) + \sum_{z\neq y} P^s(\mathcal{E}_{j,a\to b} \cap \mathcal{E}_{i,x\to y} \cap \mathcal{E}_{n,x\to z}) \\ &= P^s(\mathcal{E}_{j,a\to b} \cap \mathcal{E}_{i,x\to y} \cap \mathcal{E}_{n,x\to y}) + \sum_{z\neq y} P^s(\mathcal{E}_{j,a\to b} \cap \mathcal{E}_{i,x\to z} \cap \mathcal{E}_{n,x\to y}) \\ &= P^s(\mathcal{E}_{j,a\to b} \cap \mathcal{E}_{n,x\to y}) \end{aligned}$$

When $|D_n| = 2$, necessity is also implied by the steps above.

Step 2: We now show that the property $\mathcal{P}$ from Step 1 is necessary and sufficient for

$$\rho^w(\hat{m}_j = b, \hat{m}_i = y) = \rho^w(\hat{m}_j = b, \hat{m}_n = y), \forall i \in \mathcal{I}_n^D - \{n\}, \forall y \in D_n, b \in D_j, \forall w$$

$$\begin{aligned} \rho^w(\hat{m}_j = b, \hat{m}_i = y) &= \sum w(s) P^s(\mathcal{E}^s_{j,\mu_j^s\to b} \cap \mathcal{E}^s_{i,\mu_n^s\to y}) \\ &= \sum w(s) P^s(\mathcal{E}^s_{j,\mu_j^s\to b} \cap \mathcal{E}^s_{n,\mu_n^s\to y}) \\ &= \rho^w(\hat{m}_j = b, \hat{m}_n = y) \end{aligned}$$

To prove the opposite direction, fix any pair of alternatives x, y with $x, y \in D_n$ and similarly $a, b \in D_j$, and fix any preference-type s' with $\mu_j^{s'} = a, \mu_i^{s'} = x$. Choose a subject pool w such that it assigns weight 1 to the preference-type s', i.e, all subjects are of type s'. Then, for this subject pool,

$$\begin{aligned} P^{s'}(\mathcal{E}_{j,a\to b} \cap \mathcal{E}_{i,x\to y}) &= \rho^w(\hat{m}_j = b, \hat{m}_i = y) \\ &= \rho^w(\hat{m}_j = b, \hat{m}_n = y) \\ &= P^{s'}(\mathcal{E}_{j,a\to b} \cap \mathcal{E}_{n,x\to y}) \end{aligned}$$

Any subject pool can then be constructed as a linear combination of these single type subject pools.

c) Partial Stochastic IC immediately implies that $\rho^\omega_D(\hat{m}_j = a, \hat{m}_i = x) = \rho^\omega_D(\hat{m}_j = a, \hat{m}_n = x)$ and $\rho^\omega_{D'}(\hat{m}_{i'} = a, \hat{m}_{j'} = x) = \rho^\omega_{D'}(\hat{m}_{n'} = a, \hat{m}_{j'} = x)$. Note that $D_i = D'_{j'}$ and $D_j = D'_{i'}$, so that $\rho^\omega_D(\hat{m}_j = a, \hat{m}_i = x) = \rho^\omega_{D'}(\hat{m}_{i'} = a, \hat{m}_{j'} = x)$.

□

Proof. Proof of Theorem 2

Step 1: From $\sup_{\Delta_{BB}\in\mathcal{S}_4} \mathrm{TVD}\left(\Delta_{AA},\Delta_{BB}\right)\leq\varepsilon$ to constraint on G.

We want the total variation distance between Δ_{AA} and Δ_{BB} to be bounded:

$$\forall\Delta_{BB}\in\mathcal{S}_4,\quad \mathrm{TVD}(\Delta_{AA},\Delta_{BB})=\frac{1}{2}\sum_{i=1}^{4}|\Delta_{AA_i}-\Delta_{BB_i}|\leq\varepsilon$$

Define the difference matrix $D=G-I$, where I is the identity matrix. Then:

$$\Delta_{AA}-\Delta_{BB}=(G-I)\Delta_{BB}=D\Delta_{BB}$$

and the total variation condition becomes $\sup_{\Delta_{BB}\in\mathcal{S}_4}\|D\Delta_{BB}\|_1\leq 2\varepsilon$. Since this supremum over the simplex is achieved at the extreme points e_j (the unit basis vectors), the condition is equivalent to:

$$\|De_j\|_1=\sum_{i=1}^{4}|G_{ij}-\delta_{ij}|\leq 2\varepsilon\quad\text{for all } j=1,2,3,4$$

That is, for each column j of the perturbation matrix, we require:

$$\sum_{i=1}^{4}|G_{ij}-\delta_{ij}|\leq 2\varepsilon$$

This gives 4 linear inequalities:

$$\begin{cases}|G_{11}-1|+|G_{21}|+|G_{31}|+|G_{41}| &\leq 2\varepsilon\\ |G_{12}|+|G_{22}-1|+|G_{32}|+|G_{42}| &\leq 2\varepsilon\\ |G_{13}|+|G_{23}|+|G_{33}-1|+|G_{43}| &\leq 2\varepsilon\\ |G_{14}|+|G_{24}|+|G_{34}|+|G_{44}-1| &\leq 2\varepsilon\end{cases}\tag{11}$$

Step 2: From $\sup_{\Delta_{BB}\in\mathcal{S}_4} \mathrm{TVD}\left(\Delta_{AB},\Delta_{BB}\right)\leq\varepsilon$ to constraint on G.

$$\begin{aligned}\Pr(Q_A=0,Q'_B=0)&=\sum_{q,q'\in\{0,1\}}\Pr(Q_A=0,Q'_A=q,Q_B=q',Q'_B=0)\\ &=\sum_{q,q'\in\{0,1\}}\Pr(Q_A=0,Q'_A=q|Q_B=q',Q'_B=0)\Pr(Q_B=q',Q'_B=0)\\ &=(G_{11}+G_{21})p_1+(G_{13}+G_{23})p_3\end{aligned}$$

Thus,

$$\Pr(Q_A=0,Q'_B=0)-\Pr(Q_B=0,Q'_B=0)=(G_{11}+G_{21})p_1+(G_{13}+G_{23})p_3-p_1.$$

The other three differences can be derived similarly:

$$\begin{aligned}\Pr(Q_A=0,Q'_B=1)-\Pr(Q_B=0,Q'_B=1)&=(G_{12}+G_{22})p_2+(G_{14}+G_{24})p_4-p_2,\\ \Pr(Q_A=1,Q'_B=0)-\Pr(Q_B=1,Q'_B=0)&=(G_{33}+G_{43})p_3+(G_{31}+G_{41})p_1-p_3,\\ \Pr(Q_A=1,Q'_B=1)-\Pr(Q_B=1,Q'_B=1)&=(G_{34}+G_{44})p_4+(G_{32}+G_{42})p_2-p_4.\end{aligned}$$

Now, the total variation distance between the joint distributions Δ_{AB} and Δ_{BB} is given by:

$$\text{TVD} = \frac{1}{2} \sum_{q,q' \in \{0,1\}} \left| \Pr(Q_A = q, Q'_B = q') - \Pr(Q_B = q, Q'_B = q') \right|.$$

This is a linear function of $p = (p_1, p_2, p_3, p_4)$, so it achieves its maximum over the unit simplex at one of its vertices. Therefore, it suffices to evaluate the TVD at each vertex.

This generate 4 linear inequalities:

$$\begin{cases} |(G_{11} + G_{21}) - 1| + |G_{31} + G_{41}| & \leq 2\varepsilon \\ |(G_{12} + G_{22}) - 1| + |G_{32} + G_{42}| & \leq 2\varepsilon \\ |(G_{33} + G_{43}) - 1| + |G_{13} + G_{23}| & \leq 2\varepsilon \\ |(G_{34} + G_{44}) - 1| + |G_{14} + G_{24}| & \leq 2\varepsilon \end{cases} \tag{12}$$

Note that the inequalities in (11) imply those in (12), and therefore:

$$\sup_{\Delta_{BB} \in \mathcal{S}_4} \text{TVD}((Q_A, Q'_B), (Q_B, Q'_B)) \leq \sup_{\Delta_{BB} \in \mathcal{S}_4} \text{TVD}((Q_A, Q'_A), (Q_B, Q'_B))$$

Step 3: From $\sup_{\Delta_{BB} \in \mathcal{S}_4} \text{TVD}\left(\Delta_{AA}, \Delta_{BA}\right) \leq \varepsilon$ to constraint on G.
As before, we start by noting

$$\begin{aligned} \Pr(Q_A = 0, Q'_A = 0) &= G_{11}p_1 + G_{12}p_2 + G_{13}p_3 + G_{14}p_4, \\ \Pr(Q_B = 0, Q'_A = 0) &= G_{11}p_1 + G_{31}p_1 + G_{12}p_2 + G_{32}p_2, \\ \Rightarrow ((Q_A = 0, Q'_A = 0) - (Q_B = 0, Q'_A = 0) &= G_{13}p_3 + G_{14}p_4 - G_{31}p_1 - G_{32}p_2. \end{aligned}$$

Similarly

$$\begin{aligned} (Q_A = 0, Q'_A = 1) - (Q_B = 0, Q'_A = 1) &= G_{23}p_3 + G_{24}p_4 - G_{41}p_1 - G_{42}p_2 \\ (Q_A = 1, Q'_A = 0) - (Q_B = 1, Q'_A = 0) &= G_{31}p_1 + G_{32}p_2 - G_{13}p_3 - G_{14}p_4 \\ (Q_A = 1, Q'_A = 1) - (Q_B = 1, Q'_A = 1) &= G_{41}p_1 + G_{42}p_2 - G_{23}p_3 - G_{24}p_4 \end{aligned}$$

At $p_1 = 1$, all other $p_i = 0$, the TVD become:

$$\begin{aligned} \text{TVD} &= \frac{1}{2}\left(|-G_{31}| + |-G_{41}| + |G_{31}| + |G_{41}|\right) \\ &= |G_{31}| + |G_{41}| \end{aligned}$$

Overall, with the 4 vertices we get 4 linear inequalities:

$$|G_{31}| + |G_{41}| \leq \varepsilon, \quad |G_{32}| + |G_{42}| \leq \varepsilon, \quad |G_{13}| + |G_{23}| \leq \varepsilon, \quad |G_{14}| + |G_{24}| \leq \varepsilon \tag{13}$$

Note that the inequalities in (13) imply those in (12), and that (11) implies those in (13).

Finally, note that $\mathrm{TVD}\left(\Delta_{AA}, \Delta_{AB}\right) \leq \varepsilon$ is characterized by

$$|G_{21}| + |G_{41}| \leq \varepsilon, \quad |G_{12}| + |G_{32}| \leq \varepsilon, \quad |G_{23}| + |G_{43}| \leq \varepsilon, \quad |G_{14}| + |G_{34}| \leq \varepsilon$$

which together with the equations of (13) imply the equations of (11) and hence

$$\sup_{\Delta_{BB} \in \mathcal{S}_4} \mathrm{TVD}(\Delta_{AA}, \Delta_{BB}) \leq 2\varepsilon, \tag{14}$$

□

Remark 1. Neither $\sup_{\Delta_{BB} \in \mathcal{S}_4} \mathrm{TVD}(\Delta_{AB}, \Delta_{BB}) \leq \varepsilon$ or $\sup_{\Delta_{BA} \in \mathcal{S}_4} \mathrm{TVD}(\Delta_{AA}, \Delta_{BA}) \leq \varepsilon$ individually implies (14). Even $\sup_{\Delta_{BB} \in \mathcal{S}_4} \mathrm{TVD}(\Delta_{AB}, \Delta_{BB}) \leq \varepsilon$ and $\sup_{\Delta_{BB} \in \mathcal{S}_4} \mathrm{TVD}(\Delta_{BA}, \Delta_{BB}) \leq \varepsilon$ together do not imply (14).

Proof. Take

$$G_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \mathbf{1} & \mathbf{1} & 0 & 0 \\ 0 & 0 & \mathbf{1} & \mathbf{1} \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad G_2 = \begin{bmatrix} 0.85 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.85 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.85 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.85 \end{bmatrix}$$

G_1 retains the same Q response between A and B perfectly, but flips the Q' response. It is easy to verify that $\sup_{\Delta_{BB} \in \mathcal{S}_4} \mathrm{TVD}(\Delta_{AB}, \Delta_{BB}) = 0$, whereas $\sup_{\Delta_{BB} \in \mathcal{S}_4} \mathrm{TVD}(\Delta_{AA}, \Delta_{BB}) > 0$.

For $\varepsilon = .11 > .10$ and G_2, $\sup_{\Delta_{BB} \in \mathcal{S}_4} \mathrm{TVD}(\Delta_{AB}, \Delta_{BB}) \leq \varepsilon$ and $\sup_{\Delta_{BB} \in \mathcal{S}_4} \mathrm{TVD}(\Delta_{BA}, \Delta_{BB}) \leq \varepsilon$ hold but (14) does not. □

References

Agranov, M., Healy, P. J., and Nielsen, K. (2023). Stable randomization. *The Economic Journal*, 133:2553–2579. → pages 20

Agranov, M. and Ortoleva, P. (2017). Stochastic choice and preferences for randomization. *Journal of Political Economy*, 125(1):40–68. → pages 10

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry preschool and Early Training Projects. *Journal of the American Statistical Association*, 103:1481–1495. → pages 21, 24, 25, 27

Aoyama, T. and Hanaki, N. (2021). Preference for randomization and validity of random incentive system under ambiguity: an experiment. The Institute of Social and Economic Research discussion paper No. 1140. → pages 7, 47

Aoyama, T. and Hanaki, N. (2024). Experimental evaluation of random incentive system under ambiguity. The Institute of Social and Economic Research discussion paper No. 1236. → pages 7, 47

Azrieli, Y., Chambers, C. P., and Healy, P. J. (2018). Incentives in experiments: A theoretical analysis. *Journal of Political Economy*, 126:14720–1503. → pages 3, 9, 10, 32

Azrieli, Y., Chambers, C. P., and Healy, P. J. (2019). Incentives in experiments with objective lotteries. *Experimental Economics*. → pages 32

Baillon, A., Halevy, Y., and Li, C. (2022). Randomize at your own risk: On the observability of ambiguity aversion. *Econometrica*, 90:1085–1107. → pages 7, 47

Baltussen, G., Post, T. G., van den Assem, M. J., and Wakker, P. P. (2012). Random incentive systems in a dynamic choice experiment. *Experimental Economics*, 15:418–443. → pages 7, 8, 47

Beattie, J. and Loomes, G. (1997). The impact of incentives upon risky choice experiments. *Journal of Risk and Uncertainty*, 14(2):155–168. → pages 7, 26, 47

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93:491–507. → pages 21

Breig, Z. and Feldman, P. (2024). Revealing risky mistakes through revisions. *Journal of Risk and Uncertainty*, 68:227–254. → pages 8

Brown, A. L. and Healy, P. J. (2018). Separated decisions. *European Economic Review*, 101:20–34. → pages 3, 8, 29, 31, 32, 47

Camerer, C. F. (1989). An experimental test of several generalized utility theories. *Journal of Risk and Uncertainty*, 2:61–104. → pages 7, 8, 47

Cheung, P. and Ellis, K. (2025). Non-isolation, reversals, and social preference. Mimeo. → pages 8

Cox, J. C., Sadiraj, V., and Schmidt, U. (2015). Paradoxes and mechanisms for choice under risk. *Experimental Economics*, 18:215–250. → pages 7, 47

Cubitt, R. P., Starmer, C., and Sugden, R. (1998). On the validity of the random lottery incentive system. *Experimental Economics*, 1(2):115–131. → pages 7, 47

Derrick, B., Dobson-Mckittrick, A., Toher, D., and White, P. (2015). Test statistics for comparing two proportions with partially overlapping samples. *Journal of Applied Quantitative Methods*, 10:1–14. → pages 23, 24, 44, 46

Derrick, B. and White, P. (2022). Review of the partially overlapping samples framework: Paired observations and independent observations in two samples. *The Quantitative Methods for Psychology*, 18:55–64. → pages 23

Freeman, D. J., Halevy, Y., and Kneeland, T. (2019). Eliciting risk preferences using choice lists. *Quantitative Economics*, 10:217–237. → pages 7, 47

Freeman, D. J. and Mayraz, G. (2019). Why choice lists increase risk taking. *Experimental Economics*, 22:131–154. → pages 7, 8, 47

Harrison, G. W., Martinez-Correa, J., and Swarthout, J. T. (2015). Reduction of compound lotteries with objective probabilities: Theory and evidence. *Journal of Economic Behavior and Organization*, 119:32–55. → pages 7

Harrison, G. W. and Swarthout, J. T. (2014). Experiment payment protocols and the Bipolar Behaviorist. *Theory and Decision*, 77:423–438. → pages 7, 47

Holt, C. A. (1986). Preference reversals and the independence axiom. *The American Economic Review*, 76(3):508–515. → pages 7

Karni, E. and Safra, Z. (1987). "Preference reversal" and the observability of preferences by experimental methods. *Econometrica*, 55(3):675–685. → pages 7

Kneeland, T. (2015). Identifying higher order rationality. *Econometrica*, 83(5):2605–2079. → pages 8

Loomes, G. and Sugden, R. (1986). Disappointment and dynamic consistency in choice under uncertainty. *The Review of Economic Studies*, 53(2):271–282. → pages 7

Nielsen, K. and Rehbeck, J. (2022). When choices are mistakes. *American Economic Review*, 112:2237–2268. → pages 8

Segal, U. (1990). Two-stage lotteries without the reduction axiom. *Econometrica*, 58(2):349–377. → pages 32

Starmer, C. and Sugden, R. (1991). Does the random-lottery incentive system elicit true preferences? An experimental investigation. *The American Economic Review*, 81(4):971–978. → pages 7

B Additional data and results (Supplemental Appendix starts here)

In this Appendix we present some additional data and analysis. This section begins by presenting, in Table A1, Table A2, Table A3, and Table A4, the full data for each of our four target questions. Section B.1 presents simulations designed to understand the properties of various statistical tests that can be used to test for Weak and Strong SIC.

		$\hat{m}_R$				$\hat{m}_B$		
		0	**1**			**0**	**1**	
$\hat{m}_A$	**0**	95	40	135	**0**	10	11	21
		(0.13)	(0.05)	(0.18)		(0.07)	(0.08)	(0.15)
	1	38	569	607	**1**	13	106	119
		(0.05)	(0.77)	(0.82)		(0.09)	(0.75)	(0.85)
		113	609	742		23	117	140
		(0.18)	(0.82)			(0.16)	(0.84)	

Table A1: Contingency tables for the *Cer* questions, with cell probabilities in parentheses.

		$\hat{m}_R$				$\hat{m}_B$		
		0	**1**			**0**	**1**	
$\hat{m}_A$	**0**	111	89	200	**0**	21	14	35
		(0.15)	(0.12)	(0.27)		(0.17)	(0.11)	(0.29)
	1	82	465	547	**1**	17	70	87
		(0.11)	(0.62)	(0.73)		(0.14)	(0.57)	(0.71)
		193	554	747		38	84	122
		(0.26)	(0.74)			(0.31)	(0.68)	

Table A2: Contingency tables for the *CIA* questions, with cell probabilities in parentheses.

Table A5 contains the vRIS pairwise distribution corrections for the *Percep1* question. It is difficult to place a meaningful interpretation on these numbers, given that there is no ex-ante reason to expect any particular pattern of behavior to arise in the joint distribution of the *Percep1* question with the lottery questions. Nevertheless, we include the results here to illustrate the effectiveness of the correction procedure. We are not able to provide corrections for the the *Percep1* and *MIA1* question as we reject both *Percep1*×*MIA1* SIC and *MIA1*×*Percep1* SIC.

		$\hat{m}_R$				$\hat{m}_B$		
		0	**1**			**0**	**1**	
$\hat{m}_A$	**0**	252	77	329	**0**	22	27	49
		(0.34)	(0.10)	(0.44)		(0.18)	(0.21)	(0.39)
	1	90	328	418	**1**	10	67	77
		(0.12)	(0.44)	(0.56)		(0.08)	(0.53)	(0.61)
		342	405	747		32	94	126
		(0.46)	(0.54)			(0.25)	(0.75)	

Table A3: Contingency tables for the *MIA* questions, with cell probabilities in parentheses.

		$\hat{m}_R$				$\hat{m}_B$		
		0	**1**			**0**	**1**	
$\hat{m}_A$	**0**	96	113	209	**0**	8	25	33
		(0.12)	(0.14)	(0.27)		(0.06)	(0.19)	(0.26)
	1	204	360	564	**1**	12	84	96
		(0.26)	(0.47)	(0.73)		(0.09)	(0.65)	(0.74)
		300	473	773		20	109	129
		(0.39)	(0.61)			(0.16)	(0.85)	

Table A4: Contingency tables for the *Percep* questions, with cell probabilities in parentheses.

B.1 Simulation of statistical properties

Table A6 presents the results from a series of simulations designed to illustrate the size and power of a range of plausible tests Weak Stochastic IC.[30] We consider five different tests. The first three are simple pairwise tests of $\rho^w(\hat{m}_A = 1) = \rho^w(\hat{m}_R = 1)$, $\rho^w(\hat{m}_A = 1) = \rho^w(\hat{m}_B = 1)$ and $\rho^w(\hat{m}_R = 1) = \rho^w(\hat{m}_B = 1)$. The fourth test is a combined test that rejects the null hypothesis if any of the three simple pairwise tests do, and the fifth test is the test of Derrick et al. (2015) discussed in the main text.

We use four different simulated datasets, and simulate each data set 10,000 times. Each dataset consisted of 1000 simulated subjects, with each subject having a 75% chance of responding to questions A and B, a 12.5% chance of responding to questions A and B, and a 12.5% chance of responding to questions A and U, reflecting the design of our experimental implementation of the vRIS.[31] To generate the data for each question, we simulated each data set as consisting of three types of subjects with weights w_1, w_2 and w_3, respectively. Type 1 subjects always choose option 1, type 2

[30] These simulations are not exhaustive, but instead illustrate plausible scenarios for the data.

[31] We use the notation B to describe the Part B decision when the question is incentivized, and U for the case where the Part B question is unincentivized. We do not use the U data in the simulations.

Q	Q'	$\Pr(Q = Q')$ RIS	$\Pr(Q = Q')$ vRIS	$\Pr(Q = 1) - \Pr(Q' = 1)$ RIS	$\Pr(Q = 1) - \Pr(Q' = 1)$ vRIS
Cer1	Percep1	0.59	0.73	0.21	0.01
		[0.55, 0.62]	[0.65, 0.81]	[0.17, 0.25]	[−0.09, 0.11]
CIA1	Percep1	0.55	0.64	0.11	-0.11
		[0.51, 0.58]	[0.56, 0.73]	[0.07, 0.16]	[−0.22, 0.00]
MIA1	Percep1	0.50	n.a.	-0.05	n.a.
		[0.47, 0.54]	n.a.	[−0.10, 0.01]	n.a.

Table A5: The first two columns show the proportion of subjects that either (1) answer the perception question correctly and choose the safer option in the lottery task or (2) answer the perception question incorrectly and choose the risky option in the lottery task. The second two columns indicate, among subjects who do not satisfy either (1) or (2), whether subjects are more likely to select the risky lottery or to answer the perception question incorrectly. 95% confidence intervals are shown in square brackets. n.a. indicates not available.

subjects always choose option 0, and type 3 subjects choose option 1 with probability p. Type 3 subjects have their decision drawn independently across repetitions unless stated otherwise.

The data generating process for Sim 1 is constructed to approximate the data from the *CIA* questions. It is generated using $w_1 = 0.4$, $w_2 = 0.1$, $w_3 = 0.5$, and $p = 0.6$. Sim 1 satisfies Weak Stochastic IC. The data generating process for Sim 2 approximates the data from the *Cer* questions, with the data for the A and B questions generated using $w_1 = 0.45$, $w_2 = 0.05$, $w_3 = 0.5$ and $p = 0.8$. The data for the R question was then generated to have the same marginal distributions, but the action of the w_3 type is correlated with their action in the A question (but independent of the B question). Given that the marginal distributions are the same for all questions the Sim 2 data satisfies Weak SIC.

The data generating process for Sim 3 approximates the data from the *MIA* questions, with the data for the A and R questions generated using $w_1 = \frac{25}{36}$, $w_2 = \frac{1}{4}$, $w_3 = \frac{1}{18}$, and $p = 0.55$. The data for the B question was generated by modifying this so that type 2 subjects chose option 1 with probability $p_2 = 0.4$ and type 3 subjects chose option 1 with probability $p_3 = 0.7$. The data generating process for Sim 4 uses the same structure as Sim 3 for questions A and R, but instead modifies the data generating process such that $p_3 = 0.8$ for question B. Each of these last two simulations violates both Weak SIC by changing the marginal distribution of behavior in Part B.

Table A6 shows the rate of rejecting the null hypotheses at the standard significance level of 5%, which implies that a test is Type 1 error robust if the test rejects the null hypothesis 5% of the time when the null is true. The first thing to note from Table A6 is that the Combined test for Weak SIC, which combines all three pairwise

Test	Data required	Test type	Sim 1 Weak IC Satisfied	Sim 2 Weak IC Satisfied	Sim 3 Weak IC Violated	Sim 4 Weak IC Violated
A vs R	Within	McNemar	0.040	0.037	0.042	0.042
A vs B	Within	McNemar	0.030	0.027	0.909	0.742
R vs B	Between	Proportions	0.048	0.050	0.916	0.594
Combined	Combined	Joint test	0.111	0.106	0.976	0.855
Derrick et al. (2015)	Combined	z-test	0.047	0.049	0.969	0.730

Table A6: Simulated rate of rejecting the null hypothesis, at a 5% significance level, for five statistical tests across four data sets, with each data set being simulated 10000 times. Sim 1 used a data set that approximates the data from the *CIA* questions. Sim 2 used a data set that approximates the data from the *Cer* questions. Sim 3 used a data set that approximates the data from the *MIA* questions. Sim 4 used a dataset with the same distribution as Sim 3 for the A and R questions, but a different correlation structure across Parts A and B.

tests, rejects the null hypothesis too often, approximately 11% of the time, in Sim 1 and Sim 2. The two McNemar exact tests, the pairwise within subject tests, are also not Type 1 error robust as they do not reject the null often enough in this case. Of the remaining two tests for Weak SIC, the Derrick et al. (2015) test has a higher power (probability of rejecting the null when the null is false) in both Sim 3 and Sim 4 than the between subject test of R vs B. Our simulations here complement the results of the simulations contained in Derrick et al. (2015), and we agree with their conclusion to use their test.

B.2 Sample sizes in prior experiments

Between subject binary choice statistical tests are notoriously low powered. Our between subject tests do, however, have substantially better power than the tests for IC in most of the previous literature. Our sample size in the Part B vs Part S tests, of approximately 125 subjects in each treatment, allows us to detect a minimum effect size of approximately 10 to 17 percentage points at 80% power (we can detect smaller effect sizes in the *Cer* question which has more extreme data than the *MIA* question which has data closer to the middle of the unit interval). The comparison of sample sizes across studies in Table A7 ranks us as one of the highest powered studies. Many of the previous studies have large aggregate sample sizes (often many hundreds of subjects), but are also testing incentive compatibility across many different questions which reduces the sample size per question. This reflects a fundamental tension in testing for the incentive compatibility of the RIS: testing for IC across multiple questions generates a richer and more nuanced data set, at the cost of the statistical power of each individual test.

Study	**N per Treatment/ Question**	**Comments**
Current Study	~ 125	Within and between subject treatments
Brown and Healy (2018)	60	
Baillon et al. (2022)	84-89	Main study
Freeman et al. (2019)	20-49	Pools similar treatments
Cox et al. (2015)	40-46	
Cubitt et al. (1998)	38-62	
Beattie and Loomes (1997)	50	
Camerer (1989)	80	Within-subject, pooled across questions
Freeman and Mayraz (2019)	120	Similar power to current paper
Harrison and Swarthout (2014)	75	single-choice treatment
Baltussen et al. (2012)	88-97	Dynamic task
Aoyama and Hanaki (2021)	95-97	
Aoyama and Hanaki (2024)	~ 200	Within subject, two week delay

Table A7: Comparison of sample sizes in previous studies. All comparisons are between subject unless noted.