Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

journal homepage: www.elsevier.com/locate/jbee

Order-constrained inference to supplement experimental data analytics in behavioral economics: A motivational case study[☆]

Jonas Ludwig[a,*], Daniel R. Cavagnaro[b], Michel Regenwetter[c]

[a] Coller School of Management, Tel Aviv University, Tel Aviv 6997801, Israel
[b] College of Business and Economics, California State University, Fullerton, USA
[c] Department of Psychology, University of Illinois at Urbana-Champaign, USA

ARTICLE INFO

Keywords:
Order-constrained inference
Bayesian statistics
Regression analysis

ABSTRACT

A common approach to theory testing in behavioral and experimental economics relies on null hypothesis significance testing via (generalized) linear regression models. Here, we showcase order-constrained inference as an alternative route to theory testing. Order-constrained inference can improve the precision and nuance of behavioral decision analytics. For example, the method can be leveraged to quantify the evidence in support of, or against, a given hypothesis. It also offers advanced model selection tools for quantitative competition among multiple theories. To illustrate our case for order-constrained methods, we re-analyze data from a pre-registered experiment on incentives, cognitive reflection, and dishonest behavior. Building on this publicly available dataset, we further highlight the advantages of Bayesian order-constrained inference. We discuss how the method can deliver more convincing and more nuanced evidence than frequentist null hypothesis significance testing, pointing to new research avenues for supplementing and expanding on experimental designs in behavioral economics.

1. Introduction

A common approach to theory testing in behavioral and experimental economics relies on null hypothesis significance testing via (generalized) linear regression models. This article discusses an alternative that "removes the shackles of regression analysis" (Regenwetter & Cavagnaro, 2019) and thereby facilitates more nuanced theory testing: order-constrained likelihood-based inference. Experimental data analytics can benefit greatly from order-constrained inference, for instance, when this method is leveraged to quantify the evidence in support of or against a given hypothesis, or to facilitate quantitative competition among several models. To illustrate our case for order-constrained inference, we re-analyze publicly available data from a pre-registered experiment on dishonest behavior (Ludwig & Achtziger, 2021, see also https://osf.io/3va2w). Using this example, we advocate for order-constrained inference as a tool for researchers to better tailor their analytical procedure to the theory under investigation. This allows them to eschew arbitrary auxiliary assumptions on the theoretical level whose only purpose is to legitimize the statistical model underlying conventional analyses. In the following, we further highlight

[☆] **Author Note:** Ludwig contributed to this project with the project idea, translation of verbal hypotheses into order constraints, data analytics using QTEST, and nearly all of the writing of the manuscript. The first author also served as the team expert on the substantive research area. Cavagnaro and Regenwetter trained the first author in order-constrained modeling and inference at a summer school, helped translate verbal hypotheses into mathematical models, ran the final the QTEST analyses on powerful computers, helped with the interpretation of the results, and edited various versions of the manuscript. All authors have read and approved the final manuscript.We gratefully acknowledge National Science Foundation grants SES # 10–62,045 (PI: M. Regenwetter), SES # 14–59,699 (PI: M. Regenwetter), and SES # 20–49,896 (PI: M. Regenwetter, Co-PI: D. Cavagnaro). The development of QTEST was further supported by a grant from the Alexander von Humboldt Foundation (Co-PIs: J. Stevens and M. Regenwetter). The first author is supported by a grant from the John Templeton Foundation. We further thank Sylvia E, Emily Neu Line, Alexandra Ortmann, Carrie Dale Shaffer-Morrison, Ori Weisel, and the participants of the 2nd *Advances in Behavioral Decision Analytics* Summer School held in Champaign in June 2022 for valuable discussions and feedback. Data and materials used in this manuscript are available on the OSF, see https://osf.io/3va2w.This work has been presented at the 65th Conference of Experimental Psychologists (TeaP, Trier, Germany). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of funding agencies or the authors' home institutions. Jonas Ludwig is now at Technische Universität Berlin (Email: j.ludwig.1@tu-berlin.de).

[*] Corresponding author.
E-mail address: jonasl@tauex.tau.ac.il (J. Ludwig).

https://doi.org/10.1016/j.socec.2023.102116
Received 6 December 2022; Received in revised form 1 August 2023; Accepted 4 October 2023

the advantages of Bayesian order-constrained inference and show how, in an experimental setting, it can deliver more convincing and more nuanced evidence than frequentist null hypothesis significance testing. This also opens new avenues of research for supplementing and expanding experimental designs in behavioral economics.

2. The case: cognitive misers on the web

Ludwig and Achtziger (2021) investigated the role of unsolicited online search behavior as a potential source of distortion in the assessment of cognitive performance in a web experiment. They administered two versions of the Cognitive Reflection Test (CRT; Frederick, 2005; Toplak et al., 2014). The original version of this brief, entertaining task is extremely popular and generally well-known, both in the academic judgment and decision making community and in the general population (see Haigh, 2016; Kahneman, 2012; Stieger & Reips, 2016). Correct responses can be found easily by a simple online search within a matter of seconds. On the other hand, a slightly modified variant introduced by Ludwig and Achtziger (2021), while requiring essentially the same arithmetical steps to find a solution, used different wordings that made it impossible to find the correct solution by a quick web search shortcut. Ludwig and Achtziger (2021) counted the number of times participants changed browser tabs while working on their four-item versions of the CRT (potentially to search for the correct answers). Ludwig and Achtziger also offered a piece-rate incentive for correct responding to half of the participants working on either CRT version (the other half received a fixed-rate bonus payment). This design aimed to assess the impact of online searches on task performance, as well as study performance and online searches (hence, potentially, cheating behavior) under performance-based incentivization.

In the following, we first describe Ludwig and Achtziger's (2021) analytical procedures for testing their primary research questions. We highlight limitations of the chosen analytical approach and identify questions that the original publication left unanswered. We then turn to order-constrained likelihood-based inference (Silvapulle & Sen, 2005; Regenwetter & Cavagnaro, 2019) as an alternative analytical approach to address the same research questions more directly, with greater nuance, and with fewer, and better motivated, auxiliary assumptions. In the style of a tutorial, we walk the reader through the formulation of order-constrained hypotheses based on the scientific predictions in the original publication. We provide a step-by-step description of data pre-processing required to submit Ludwig and Achtziger's (2021) dataset to QTest (Regenwetter et al., 2014; Zwilling et al., 2019), a public domain software for order-constrained inference. We then walk through a detailed and fully reproducible report of the QTest analysis. We explain how these more custom-tailored data analytics form a useful alternative to the conventional route of analyzing experimental data by means of regression. Finally, we discuss future directions.

2.1. Regression analysis by Ludwig and Achtziger (2021)

Ludwig and Achtziger (2021) were interested in three main questions. First, do participants cheat when solving the CRT online? Second, if so, how does cheating affect assessments of CRT performance? And third, does a piece-rate incentive increase performance and/or cheating? In addition, given the extant literature on gender differences in CRT performance (e.g., Brañas-Garza et al., 2019) and dishonest behavior (e.g., Abeler et al., 2019; Gerlach et al., 2019; Leib et al., 2021), the authors investigated heterogeneity across genders.

To address these questions, Ludwig and Achtziger (2021, see p. 5) performed a stepwise proportional odds logistic regression analysis (McCullagh, 1980; Moffatt, 2016). Essentially, they regressed CRT performance (i.e., a score between 0 and 4) on the experimental treatments (original vs. new test version; fixed vs. piece-rate incentive) and gender. In line with the prediction that cheating takes place and boosts performance, the number of browser tab changes was significantly related with performance in the original CRT (for which answers were readily available on the web) but not in the modified variant (for which responses were unavailable online). Ludwig and Achtziger interpreted this finding to mean that participants clicked different browser tabs to search for the correct answers, and thus, that tab clicks represented cheating behavior.

Most importantly for the aim of this re-analysis, Ludwig and Achtziger further suggested that the piece-rate incentive in their study merely increased cheating, rather than improving CRT performance by virtue of boosting cognitive reflection (see also Brañas-Garza et al., 2019; Yechiam & Zeif, 2022, on incentives and CRT performance). This argument rested on the descriptive finding that piece-rate incentives, relative to a fixed bonus payment, appeared to increase performance for the original CRT, but not for the new CRT (see also Fig. 1 in Ludwig & Achtziger, 2021).

Consistent with this descriptive finding, the regression models indicated a pattern of interactions between the predictors: The interaction term for CRT version and incentive treatment was statistically significant, indicating that participants responded differently to the incentive depending on the CRT version. However, when adding the interaction term for CRT version and cheating (i.e., the click count), the former interaction was no longer statistically significant (see Table 3 in Ludwig & Achtziger, 2021). According to the authors, this finding suggested that performance improvements in the original CRT that were observed under piece-rate incentives were likely due to more cheating. On the other hand, they "did not find clear evidence, but only a tendency, that cheating was more prevalent" (p. 4) in the piece-rate incentive treatment when analyzing cheating as a separate dependent variable. Consequently, Ludwig and Achtziger (2021, p. 1) reservedly concluded that "performance-based payment improved CRT performance, but probably through cheating."

2.2. Limitations and open questions

Although it seems plausible that cheating plays a causal role in explaining performance differences between the incentive treatments, the experimental design limits the strength of conclusion that can be drawn from Ludwig and Achtziger's analysis. One reason for this is that cheating is a measured variable, not an experimentally manipulated one. More generally, the design leaves room for multiple alternative explanations for the performance improvement. We investigate several such explanations.

For instance, if one were skeptical as to the informational value of the cheating indicator, one might fancy an alternative explanation for the performance difference that leaves out any reference to dishonest behavior. It is theoretically possible, for example, that the changes to the items' wording in the new version (while retaining the basic structural properties of the task), introduced idiosyncrasies that rendered the new CRT slightly more difficult.[1] Possibly, a rather small boost in cognitive reflection, motivated by the piece-rate incentive, may have sufficed to significantly increase performance in the original task. However, at overall higher difficulty of the new CRT, the same incentive-born increment of cognitive reflection might no longer suffice to translate into an observable performance improvement. In this sense, it remained an open question whether the performance differences observed in the original CRT could be attributed to just cheating, rather than increased effort, or to a combination of both.

[1] This example is to illustrate a theoretical possibility. Ludwig and Achtziger advocated that the new version was by no means more difficult than the original CRT (e.g., see their Fig. 1). Notwithstanding, it is virtually impossible to keep *all* aspects of the CRT items constant while changing their wording. So, one cannot rule out that the adaptation introduced some arbitrary changes that relate to how strongly participants engaged in cognitive reflection while solving the task.

Moreover, like all regression modeling, Ludwig and Achtziger's (2021) analytical approach relies on the validity of various auxiliary assumptions (Regenwetter & Cavagnaro, 2019; but see also Alós-Ferrer et al., 2021). Regenwetter and Cavagnaro (2019) argued that extraneous assumptions like (log-) linearity or even the assumption of a functional relationship may well bias the results of empirical analysis. This line of reasoning implies that adding or relaxing constraints just to convert a theory into conventionally testable predictions can have serious ramifications for the generalizability and replicability of empirical findings (see Regenwetter, 2020; Regenwetter & Cavagnaro, 2019; Zwilling et al., 2019, for more detail).

A noteworthy feature of the analytical approach in the original publication is that it does not consider the two behavioral outcomes of interest, CRT performance and dishonesty, simultaneously as dependent variables within one single analytical step. We aim to determine whether the performance increment observed for the piece-rate incentive relied on increased cheating, rather than more effort, or both. To that goal, we predict certain patterns of performance and online searches to hold jointly. In the next section, we present an analytical approach that facilitates such a more nuanced view. We also review model selection tools for quantitative competitions among theories.

3. Order-constrained modeling and inference

3.1. *The basic idea behind order-constrained inference*

Order-constrained inference is a procedure for studying hypotheses that are characterized by order constraints, such as $P_1 \leq P_2 \leq P_3, P_2 \leq P_5, P_1 \leq P_4$ (for five binomial probabilities $P_1, P_2, \ldots, P_5$). Both frequentist and Bayesian approaches are available. Order-constrained inference in a frequentist framework allows one to test a conjunction of constraints like the above, directly, as the null hypothesis, thereby making the scientific hypothesis refutable. Contrast this with, say regression analysis, where the null hypothesis is the absence of an effect. In a Bayesian framework, order constrained-inference allows one to select between different order-constrained hypotheses based on the evidence provided by data. In each case, a given conjunction of constraints, like those above, forms one single hypothesis.

Order-constrained inference has several major advantages over conventional regression. These mostly have to do with stating formal models that tightly align with a scientific hypothesis. The approach also supports these models through advanced quantitative data analytics: First, the approach lets the scholar specify the relationship between variables, including the possibility of not even requiring a functional relationship between certain variables at all. Second, it eschews the standard a-theoretical distributional assumptions embedded in regression analyses. Third, it makes it possible to consider multiple behavioral outcomes jointly. Fourth, by aligning its statistical model closely with the scientific hypothesis, order-constrained modeling allows us to carry out more nuanced and concise theory testing. Fifth, the associated data-analytics provide highly nuanced frequentist and Bayesian inference. Finally, the Bayesian analytics facilitate full-fledged quantitative model selection in that they move beyond heuristic model-complexity measures that typically limit themselves to counting parameters and degrees of freedom. The cost of entry is that, essentially, the data-generating process must be (adequately approximated by) a product of binomial processes.

Many scholars have advocated for order-constrained methods as a tool to improve theory testing, and several software packages are available to date to conduct order-constrained modeling and inference for a variety of statistical models (Gu et al., 2019; Heck & Davis-Stober, 2019; Hoijtink et al., 2019; Klugkist et al., 2005; Mulder et al., 2021; Regenwetter et al., 2014; Sarafoglou et al., 2023a; Zwilling et al., 2019). For instance, the R package Bain was developed specifically for testing order-constrained hypotheses in structural equation models (Gu et al., 2019). In this article, we rely on QTest (Regenwetter et al., 2014; Zwilling et al., 2019) to evaluate hypotheses about equality of probabilities and order constraints in binomial models.

3.2. *The application of order-constrained inference to the Ludwig and Achtziger (2021) dataset*

In this section, we leverage these unique qualities to supplement and expand on the experimental evidence obtained by Ludwig and Achtziger (2021). We re-analyze their data (as available at https://osf.io/3va2w) with order-constrained methods to address a key question for which the authors, in the original paper, were reluctant to draw strong conclusions: Does CRT performance increase under piece-rate incentives by virtue of more effort and cognitive reflection, or alternatively, by cheating, or by a combination of the two?

To answer this question, we formulate three models that explain higher performance under incentives as a result of (a) increased effort, (b) more cheating, or (c) both. The models specify the expected behavioral patterns of both CRT performance and online searches (dishonest behavior). Importantly, one should only consider the models as supported if the predictions on both behavioral outcomes hold jointly. We seek convincing evidence for or against conjunctions of constraints on both behavioral outcomes (see also Davis-Stober & Regenwetter, 2019).

The upcoming tutorial is organized into five steps. First, we look at the distinct predictions of the three competing explanatory models in terms of performance and online searches. We formulate hypotheses in the form of order-constrained probabilities for the behaviors of interest in each experimental treatment. Second, we describe the data structure required for order-constrained inference with QTest (Regenwetter et al., 2014; Zwilling et al., 2019), as well as how we recoded and pre-processed the original data. The third step then walks the reader through the data input in QTest, as well as the software settings and options for configuration. Step 4 presents and interprets the results. Finally, the fifth step sketches further modeling possibilities when taking gender into account as an additional predictor of performance and cheating. We conclude with a discussion of the re-analysis.

3.2.1. *Hypotheses: order-constrained probabilities*

We consider two behavioral outcomes: high performance and cheating, in each of the four treatments resulting from crossing the experimental factors CRT version (original vs. new) and incentive (fixed vs. piece-rate). We use two different cutoffs on the number of correctly solved items to define high performance. We treat behavior as cheating if the participant changed tabs during the CRT at least once. In the data processing and results sections below, we provide more detail on how we recode the data and how we check the robustness of our results under different cutoffs for high performance.

Our hypotheses address the probabilities of being categorized as a high performer or as a cheater, given one of the four treatments. Specifically, we describe below three competing explanations for higher performance under incentivization (Models 1- 3): incentives improve performance (1) through boosting cognitive reflection, (2) through increased cheating, or (3) through both cognitive reflection and cheating.

We convert verbal predictions into a stream of rank-ordered, (in-) equality-constrained probabilities to form models that are amenable to order-constrained inference. We denote the probability that a randomly sampled respondent performs highly as P. We write the probability that a randomly sampled respondent cheats as C. The subscript denotes the test version (O for original, N for new) and the incentive treatment ($\$$ indicates the presence of piece-rate incentives). Table 1 summarizes seven sets of order constraints that encapsulate verbal hypotheses as described below. We show below how we combine constraint sets to form competing models.

First, let us consider some basic predictions for the probabilities of high performance and of cheating. These predictions capture the general

Table 1
Summary of constraint sets and models.

Constraint set	Order constraints	Model					
		0	1*	1	2*	2	3
1	$0 \leq P_N \leq P_O \leq 1$ and $0 \leq P_{N\$} \leq P_{O\$} \leq 1$.	○	○	○	○	○	○
2	$0 \leq C_N \leq C_O \leq 1$ and $0 \leq C_{N\$} \leq C_{O\$} \leq 1$.	○	○	○	○	○	○
3	$0 \leq P_N \leq P_{N\$} \leq 1$ and $0 \leq P_O \leq P_{O\$} \leq 1$.		○	○			○
4*	$0 \leq C_N = C_{N\$} \leq 1$ and $0 \leq C_O = C_{O\$} \leq 1$.		○				
4	$\mid C_N - C_{N\$} \mid \leq 0.05$ and $\mid C_O - C_{O\$} \mid \leq 0.05$.			○			
5*	$0 \leq P_N = P_{N\$} \leq 1$ and $0 \leq P_O \leq P_{O\$} \leq 1$.				○		
5	$\mid P_N - P_{N\$} \mid \leq 0.05$ and $0 \leq P_O \leq P_{O\$} \leq 1$.					○	
6	$0 \leq C_N \leq C_{N\$} \leq 1$ and $0 \leq C_O \leq C_{O\$} \leq 1$.				○	○	○
7	$(P_{N\$} - P_N) \leq (P_{O\$} - P_O)$.						○

Note. P denotes the probability of performing highly and *C* the probability of cheating. The subscript denotes the test version (*O* for original, *N* for new) and the incentive treatment (*$* indicates the presence of piece-rate incentives). The asterisk (*) indicates more restrictive model versions.

idea that prior exposure to the CRT affects participants' responses (Haigh, 2016; Stieger & Reips, 2016; Woike, 2019; but see also Bialek & Pennycook, 2018; Meyer et al., 2018). First, there are reasons to expect more people to perform highly on the original CRT than on the new version. Some participants who previously worked on the CRT may simply remember the correct answers. Similarly, some participants could remember receiving negative feedback on their incorrect responses in a previous study, which could encourage deeper reflection. In contrast, the new version was developed for this study, hence could not be subject to such memory effects. Therefore, we formulate constraint set 1 to capture the hypothesis that a randomly sampled person is more likely to perform highly on the original than the new CRT, within each incentive structure:

$$0 \leq P_N \leq P_O \leq 1 \text{ and } 0 \leq P_{N\$} \leq P_{O\$} \leq 1. \quad (1)$$

Note that, in order to have a well-stated constraint set on probabilities, every probability must be bounded from below by either 0 or another probability, and every probability must be bounded from above by either 1 or another probability.

Prior exposure could similarly affect cheating rates. If some participants remember having solved the test before, this could raise suspicion that the answers can be found online. Moreover, participants could be more easily tempted to look up the correct answers on the internet when they remember previous negative feedback. Therefore, we formulate constraint set 2 to represent the hypothesis that cheating rates are higher for the original CRT than the new version, within each incentive structure:

$$0 \leq C_N \leq C_O \leq 1 \text{ and } 0 \leq C_{N\$} \leq \mathrm{C}_{O\$} \leq 1. \quad (2)$$

We summarize the first two constraint sets in *Model 0*. Hence, this model predicts a conjunction of two scientific hypotheses. The first is that performance is higher in the original version relative to the new one, within incentive structure. This, in turn, yields a conjunction of constraints (constraint set 1). The second hypothesis is that cheating is more common in the original version than in the new one, within incentive structure, which yields another conjunction of constraints (constraint set 2).

By combining the probabilities of high performance and cheating in four experimental treatments, Model 0 considers altogether eight binary probabilities. By forming the combination of constraint sets 1 and 2, Model 0 is a conjunction of twelve inequalities on those eight probabilities. We use QTest to evaluate Model 0 as a single hypothesis. To test this model, QTest requires the order constraints on the eight probabilities (here: all constraints described in Model 0, that is, constraint sets 1–2) to be converted into a single set of non-redundant constraints. The QTest input takes the form of a matrix as shown in the Appendix. For further detail on matrix calculation and more comprehensive examples, see the Appendix and Regenwetter and Cavagnaro (2019).

Model 0 is also implied by each of three following models. It represents two basic predictions that we expect to hold regardless of whether reflection (Model 1), cheating (Model 2), or both (Model 3) best explain performance improvements under incentives. Beyond these common predictions, Models 1–3 encompass different additional constraints on how the probabilities of performing highly and cheating within CRT versions are affected by incentives. Models 1–3 can therefore be considered as nested submodels of Model 0. As we see next, each model adds a unique set of further constraints on the probabilities of high performance and cheating.

Model 1 explains performance improvement under incentives by more effort. It comprises a conjunction of four scientific hypotheses that specify constraints on the probabilities of high performance and cheating across the four treatments. Four sets of order constraints together capture the hypotheses (see also Table 1). The first two hypotheses are represented in constraint sets 1 and 2. Third, under Model 1, incentives should improve performance on both CRT versions through increased reflection. This should be the case if effort and reflection are the main drivers of higher performance under incentives. Constraint set 3 predicts that the probability of scoring high on the CRT increases with incentives in both CRT versions:

$$0 \leq P_N \leq P_{N\$} \leq 1 \text{ and } 0 \leq P_O \leq P_{O\$} \leq 1. \quad (3)$$

Fourth, under Model 1, cheating should be equally likely to occur in both incentive treatments. The probability that a randomly sampled participant cheated should remain unaffected by the incentive treatment, regardless of the test version. This motivates constraint set 4*:

$$0 \leq C_N = C_{N\$} \leq 1 \text{ and } 0 \leq C_O = C_{O\$} \leq 1. \quad (4^*)$$

We flag constraint sets and models with an asterisk (*) if they include an equality constraint. Because the equality constraints in these hypotheses may be overly restrictive, we also consider versions with a bit of wiggle space, where we require only approximate equality among the pertinent probabilities. In the case of constraint set 4, we add constraints on the two probabilities C within test versions to express that, while C_N, $C_{N\$}$ and C_O, $C_{O\$}$ may lie anywhere within the range of 0 and 1 (subject to constraint set 2), they should not differ by more than some threshold, here 5 percentage points, within test version. This gives constraint set 4:

$$\mid C_N - C_{N\$} \mid \leq 0.05 \text{ and } \mid C_O - C_{O\$} \mid \leq 0.05. \quad (4)$$

To sum up, Model 1 is a conjunction of four hypotheses, captured by constraint sets 1-4, respectively. First, irrespective of incentive structure, performance improves on the original compared to the new CRT. Second, more people engage in one or more tab changes under original CRT than under new CRT. Third, for the same version of the CRT, performance improves under piece-rate incentive relative to fixed rate. Finally, for a given version of the CRT, the probability of one or more tab changes does not differ by more than 0.05 between incentive structures.

Model 2, on the other hand, posits that higher performance under piece-rate incentivization is best explained by more cheating. It does this

Table 2
Binomial counts of pertinent behavioral outcomes to be input into QTest.

Experimental treatment (CRT version – incentive)	*n*	Performance		Cheating	
		(high, low)	QTest	(cheater, honest)	QTest
Original CRT – fixed rate	76	29, 47	A, B	15, 61	I, J
Original CRT – piece-rate	83	45, 38	C, D	28, 55	K, L
New CRT – fixed rate	73	31, 42	E, F	13, 60	M, N
New CRT – piece-rate	67	20, 47	G, H	14, 53	O, P

Note. Colum QTest refers to the labels of these binomials used in QTest in the next section, see Fig. 1, Screenshot of QTest 2.1 GUI.

with two key changes from Model 1. First, whereas Model 1 predicts improved performance with piece-rate incentives in both the new and original CRT (constraint set 3), Model 2 instead predicts improved performance only in the original CRT, where web searches are likely to be successful. It does not predict improved performance in the new CRT, where cheating in search of answers is likely to be futile. Secondly, whereas Model 1 predicts comparable cheating rates within given CRT types (constraint set 4 or 4*), Model 2 predicts more frequent cheating under piece-rate incentives within each CRT type.

We define constraint set 5* (and 5) to capture that (a) new CRT performance does not differ (much) across incentive treatments, and (b) original CRT performance increases under incentives.

$$0 \le P_N = P_{N\$} \le 1 \text{ and } 0 \le P_O \le P_{O\$} \le 1. \tag{5*}$$

$$\mid P_N - P_{N\$} \mid \le 0.05 \text{ and } 0 \le P_O \le P_{O\$} \le 1. \tag{5}$$

We define constraint set 6 to capture the prediction of more frequent cheating under piece-rate incentives within each CRT type.

$$0 \le C_N \le C_{N\$} \le 1 \text{ and } 0 \le C_O \le C_{O\$} \le 1. \tag{6}$$

In all, Model 2 (or 2*) forms the conjunction of four hypotheses captured by the constraint sets 1 and 2 together with 5 (or 5*) and 6 (Table 1). It differs from Model 1 by replacing constraint sets 3 and 4 with constraint sets 5 (or 5*) and 6. The QTest-compatible input matrices for Models 2 and 2* are given in the Appendix.

Next, *Model 3* captures the hypothesis that performance improves under incentives through both cheating and cognitive reflection. Model 3 differs from Models 1 and 2 in that it predicts incentives to improve performance on both CRT versions (constraint set 3; like Model 1, but unlike Model 2). Model 3 further predicts that incentives increase cheating rates in both CRT versions (constraint set 6; like Model 2, but unlike Model 1).

Model 3 also predicts a difference in the performance increment under incentives between the new and original versions of the CRT. Model 3 predicts that both cognitive reflection and cheating may drive performance improvements. However, whereas both factors may affect performance in the original CRT, only cognitive reflection would be expected to contribute to performance improvements in the new version. This is because cheating is unlikely to uncover answers to the new CRT. Therefore, if performance improves under incentives through both cheating and cognitive reflection, then the increment should be higher in the original than in the new CRT. We capture this with constraint set 7:

$$(P_{N\$} - P_N) \le (P_{O\$} - P_O) \tag{7}$$

Together, Model 3 forms the conjunction of the constraint sets 1–3 and 6–7, see also Table 1.

3.2.2. Data preparation: recoding and pre-processing

Making the dataset amenable to order-constraint analysis with QTest (Regenwetter et al., 2014; Zwilling et al., 2019) only requires a few simple steps of recoding and data pre-processing. Order-constrained inference with QTest rests on estimating the probabilities of one or more behavioral outcomes based on the number of 'successes' in each of the binomials of interest.

The original study recorded performance as an ordinal variable with range [0,4]. To facilitate order-constrained analytics, we have dichotomized this measure. To that end, we categorized participants who solved at least two out of the four decision problems correctly as "high performers" (successes, coded 1) and participants with one or no correct response as "low performers" (coded 0). The resulting proportions per treatment are given in Table 2. In the results section, we report a robustness check using a different cutoff for high performance (at least three correct). That analysis generates similar results.

Regarding cheating, this outcome already had the correct form in the original dataset, as every participant was categorized as either "honest" (coded 0) or as a "potential cheater" (coded 1). We can simply count the number of binomial 'successes' (in this case: the number of potential cheaters) per treatment and feed this information into QTest, see Table 2.

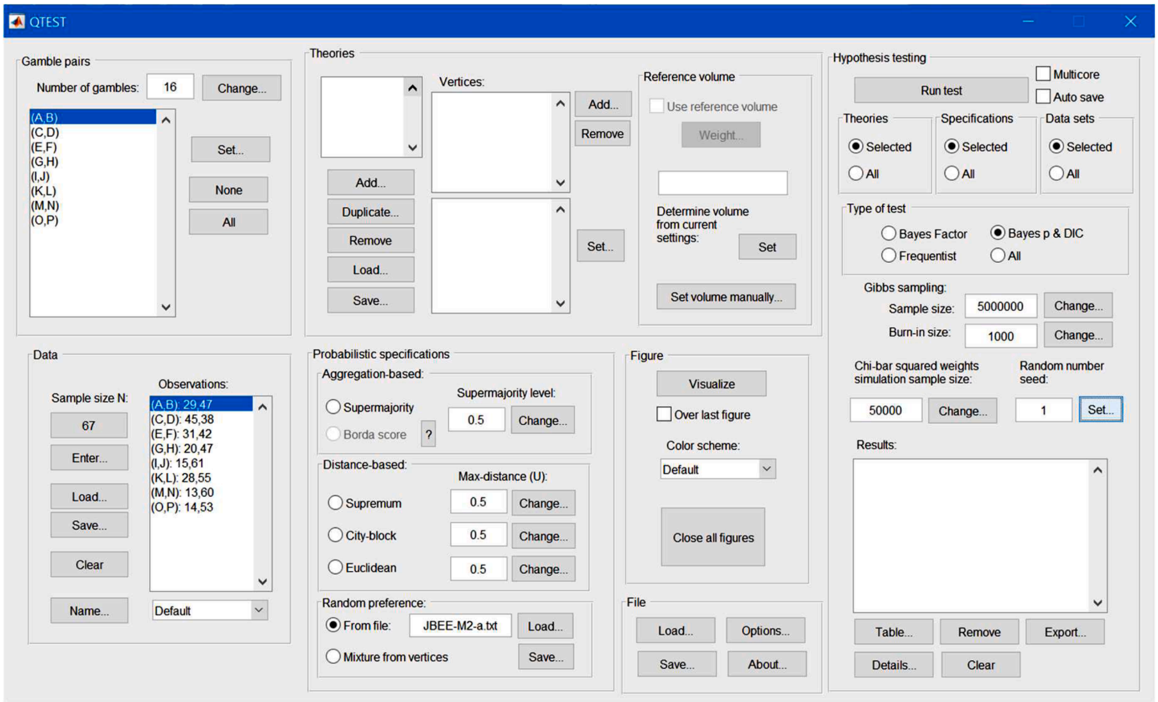
3.2.3. Feeding and operating QTest: data input and configurations

Fig. 1 shows a screenshot of the QTest 2.1 GUI (Zwilling et al., 2019). We now review some of its panels more closely. Because QTest was originally designed to model and analyze risky choice behavior, the current interface uses somewhat idiosyncratic terminology. Under *Gamble pairs* in the upper left, the user specifies the behavioral outcomes to be modeled. Performance and cheating measures in each treatment add up to eight binary variables, or eight pairs of 'gambles.' For instance, (A, B) refers to high (A) and low performance (B) in the Original CRT – fixed rate treatment; (O, P) refers to cheaters (O) and honest participants (P) in the New CRT – piece-rate treatment, and so forth. The labels (letters A-P) can be adjusted under "Set…".

Immediately below the definition of *Gamble pairs* is the *Data* input field. The data (see Table 2) can be entered manually or loaded from a text file. QTest can handle multiple datasets at once. Datasets can be saved under a unique "Name…" for future reference. Next, in the bottom row, the panel *Random preference*[2] specifies the predicted set of order-constraints, specified via a matrix (see Appendix) and loaded from a text file. We do not use any of the other "Probabilistic specifications" offered by the QTest GUI.

Below (see Table 3) we look at the frequentist *p*-value and the Bayesian *p*-value to evaluate model fit. We also consider the Bayes factor for the comparison between models. But let us first consider the informational value of these statistics. The frequentist and Bayesian *p*-values for order-constrained inference are derived based on pioneering work in statistics and mathematical psychology (Davis-Stober, 2009; see also Gelman et al., 1996; Meng, 1994; Silvapulle & Sen, 2005). While requiring advanced methods to compute, the frequentist *p*-value follows the conventional interpretation of *p*-values in null hypothesis significance testing. The Bayesian *p*-value works similarly. Small frequentist or Bayesian *p*-values < 0.05 result in the rejection of the model. In the terminology of model fitting, a *p*-value of at least 0.05 indicates an "adequate fit." Very well-fitting models will generate a frequentist

[2] Just like we are not using "gambles," so are our models technically not what is commonly referred to as "random preference" models. Conveniently, our models can 'mimic' random preference models in that one can also specify the latter via a matrix of order-constraints as input (see Zwilling et al., 2019).



Number of gambles: 16

(A,B)
(C,D)
(E,F)
(G,H)
(I,J)
(K,L)
(M,N)
(O,P)

☐ Use reference volume

☐ Multicore
☐ Auto save

Theories: ● Selected ○ All

Specifications: ● Selected ○ All

Data sets: ● Selected ○ All

Type of test: ○ Bayes Factor ● Bayes p & DIC ○ Frequentist ○ All

Gibbs sampling:
Sample size: 5000000
Burn-in size: 1000

Chi-bar squared weights simulation sample size: 50000

Random number seed: 1

Sample size N: 67

Observations:
(A,B): 29,47
(C,D): 45,38
(E,F): 31,42
(G,H): 20,47
(I,J): 15,61
(K,L): 28,55
(M,N): 13,60
(O,P): 14,53

Name: Default

Supermajority level:
○ Supermajority 0.5
○ Borda score

Max-distance (U):
○ Supremum 0.5
○ City-block 0.5
○ Euclidean 0.5

Random preference:
● From file: JBEE-M2-a.txt
○ Mixture from vertices

☐ Over last figure

Color scheme: Default

Fig. 1. Screenshot of the QTest 2.1 GUI (Zwilling et al., 2019).
The reader is directed to four panels in particular: *Gamble pairs, Data* (see also Table 2), *Random preference*, and *Hypothesis testing*.

Table 3
Results of order-constrained inference for the competing models in three analyses.

	Model	Frequentist *p*-value	Bayesian *p*-value	Bayes Factor Comparison to Encompassing Model	Bayes Factor Comparison to Model 0
Analysis 1	0	0.093	0.494	2.8	1
	1*	–	0.078	4.1	1.4
	1	0.037	0.221	4.8	1.7
	2*	–	0.441	29.0	10.2
	2	0.224	0.489	30.3	10.7
	3	0.020	0.361	4.8	1.7
Analysis 2 (High performance cutoff $\geq$ 3)	0	0.236	0.511	3.5	1
	1*	–	0.123	9.4	2.7
	1	0.100	0.308	11.0	3.2
	2*	–	0.542	65.3	18.9
	2	0.440	0.572	63.8	18.4
	3	0.098	0.459	10.8	3.1
Analysis 3 ($N = 255$)	0	0.090	0.480	2.9	1
	1*	–	0.063	3.9	1.4
	1	0.151	0.215	4.7	1.7
	2*	–	0.469	42.2	14.8
	2	0.248	0.508	42.6	15.0
	3	0.097	0.384	8.3	2.9

Note. The Models 1*, 2* are nested in their respective parent model and apply an equality constraint instead of allowing for a 0.05 divergence (see Hypotheses). Computing a frequentist *p*-value is not possible for these models.

p-value near 1 and a Bayesian *p*-value around 0.50.

When more than one model adequately fits the data based on the Bayesian (or frequentist) *p*-value, we can use the Bayes factor to determine which model provides the best explanation for the data. More precisely, the Bayes factor is an evidence ratio. It quantifies the strength of evidence for or against one model relative to another, given the data. The QTest software provides the Bayes factor for each model under consideration (the "constrained" model) relative to a common "encompassing" model. This encompassing model includes the same binomial probabilities as the model in question, but only constrains these probabilities to be between 0 and 1. The larger the Bayes factor, the stronger the evidence in favor of the constrained model. Following a common convention in Bayesian statistics (Andraszewicz et al., 2015; Jeffreys, 1998), we interpret Bayes factors larger than three as evidence for the model. If the Bayes factor is three, there is three times more evidence for the constrained model than for the encompassing model. Similarly, a Bayes factor smaller than one third suggests evidence against the model in question. If the Bayes factor is one third, the encompassing model is three times more likely than the constrained model. For more details on fit statistics and model selection tools, as well as their interpretation in order-constrained inference, see Regenwetter and Cavagnaro (2019) or another tutorial by Regenwetter (2020).

Returning to QTest, for *Hypothesis testing* in the upper right side, the user must input one or more "Theories," "Specifications," and "Datasets" for analysis. Because we specify one model at a time via a matrix specification of order-constraints, this analysis employs one "theory" and one "specification." While QTEST allows to run multiple tests and datasets at the same time, some computations can be rather slow. Some (primarily the Bayes factor computation) will converge only when setting the underlying algorithms to high "Sample size" in the "Gibbs sampler." This sometimes renders larger-size models (i.e., with more binomial parameters) intractable on a personal computer in reasonable time. To avoid unnecessary strain on computational demand, we recommend conducting each *Type of test* separately and, for the Bayesian analyses, adjusting *Sample size* in *Gibbs sampling* as needed, based on convergence of the individual test. For instance, smaller *Gibbs sample* sizes will usually suffice for the Bayesian *p*-value than for the Bayes factor. Use "Change…" in the *Gibbs sampling* panel to adjust the sampling size (see also Zwilling et al., 2019, for much more detailed information). In our analysis, we used five million draws for the Bayesian *p*-value computation, and one billion draws for the Bayes factors (see Table 3). In the frequentist analysis we used the default setting of 50,000 draws for *Chi-bar squared weights simulation sample size.*

3.2.4. *Results of order-constrained inference with QTEST*

We consider three statistics for each competing model: the frequentist *p*-value, Bayesian *p*-value, and Bayes factor. We assess these statistics on each of four analyses. For each statistic, we first review results for the main analysis of the full sample, based on the performance coding described above (high performance if at least two CRT items were solved correctly, "Analysis 1″). We then move on to two robustness checks. One of these repeats the analysis using a different cutoff for high performance (at least three CRT items solved correctly, "Analysis 2″). The second one is based on a subsample ($N = 255$, "Analysis 3″). Here, we have excluded all participants who self-reported prior knowledge with the CRT. Finally, to discuss further modeling options, the next section considers gender differences in CRT performance and cheating behavior ("Analysis 4″).

Table 3 summarizes the results of order-constrained likelihood-based inference for Analyses 1–3. Remember that Models 1–3 differ in how they specify the role of incentives. Model 1 predicts performance improvement under incentives through more effort. Model 2 predicts the same performance improvement, but due to cheating. Model 3 posits that both effort and cheating jointly contribute to better performance under incentives. In addition, all three models have in common that they predict more high performers and more cheaters in the original CRT, as captured by Model 0. Note that because the nested submodels Models 1* and 2* combine equality and inequality constraints, frequentist tests are not possible.

Frequentist and Bayesian p-values. Judging by the frequentist *p*-value, Models 1 and 3 are rejected, while Model 0 and Model 2 fit the data ("Analysis 1″ panel in Table 3). Hence, the frequentist analysis suggests that Models 0 and 2 account for the observed behavior adequately. However, the result does not fully repeat in the robustness checks. In both Analyses 2 and 3, all models fit the data according to frequentist analytics. Moreover, the Bayesian *p*-value does not fully align with the frequentist analysis. The Bayesian *p*-value is greater than 0.05 for all models in all analyses, suggesting reasonable fit throughout. Since multiple models fit and since the fit varies across robustness checks, it is unclear, from these analyses alone, which model best describes the data. In the next steps, we rely on the Bayes factor for model selection.

Bayes factor. The Bayes factor quantifies the amount of evidence in support of, or against, a given model compared to another model (here the encompassing model). Consistent with the frequentist and Bayesian *p*-value analysis, the Bayes factors indicate that there is evidence in the data to support each constrained model against the encompassing model. Model 2 (or Model 2*) has the highest Bayes factor in each of Analyses 1–3. Its Bayes factor in Analysis 1 is 30.3. This means that there is around 30 times more evidence for the constrained model than for the encompassing model. Hence, the hypothesis that just cheating predicted performance improvements under incentives receives substantial support.

Since there is also support for other models, we need to carry out model selection. We can compare two constrained models directly by taking the ratio of their Bayes factors. This provides a quantitative evidence ratio for one constrained model over another. For instance, comparing Model 2 to either Model 1 or 3, in Analysis 1, there is $30.3/4.8 = 6.3$ times more evidence in favor of Model 2.

To evaluate whether the overall good performance of Models 1–3 is due to the constraints they all share, which form Model 0, we can compare each model against Model 0. We do this as before by taking ratios of Bayes factors, see the last column in Table 3. From this perspective, Model 2 best describes the data. In Analysis 1, there is around ten times more evidence in support of Model 2 against Model 0 (18 times more in Analysis 2). Evidence for Models 1, 1*, and 3 (relative to Model 0) is rather weak. This pattern of results is robust across the three analyses. In sum, compared to Model 0, there is about six to nine times more evidence supporting Model 2 than supporting Model 1, and still five to six times more evidence for Model 2 than for Model 3. We thus conclude that Model 2 most adequately describes the data. Its more restrictive version, Model 2*, has very similar Bayes factors. The latter finding suggests that adding equality constraints instead of permitting five percentage points of wiggle space does not substantially alter model performance.

3.2.5. *Gender differences: exploring further modeling opportunities*

The present dataset offers opportunities to expand the order-constrained analysis in interesting ways. For instance, Ludwig and Achtziger (2021) reported that gender was an important predictor of both performance and cheating. Order-constrained methods can test precise theories about gender differences in relation to incentives, performance, and cheating. To do so, we split each of the eight probabilities described above into two, by gender (e.g., the probability that a randomly selected female/male cheated in the original CRT under piece-rate incentives). This results in 16 binomials (or QTEST "Gamble pairs"), which, in turn, substantially increases computational cost of order-constrained inference.

While additional factors can be accounted for in this way, it is generally advisable to limit the number of binomials in an order-constrained model. The main reason is that larger models (25 binomials or more) often become intractable with desktop computing resources.[3] In some situations, one can partition the collection of binomials into smaller subsets such that each model only states constraints within each subset, but not between binomials of different subsets. In such a case, one can perform separate analyses on the subsets of binomials. The product of these Bayes factors is the Bayes factor for the full model on all binomials. We use this strategy in the following Analysis 4, see the online supplement for more detail.

Studies related to Ludwig and Achtziger's experiment have commonly reported two findings of gender differences. First, males typically outperform females on the CRT (e.g., Alós-Ferrer et al., 2016; Brañas-Garza et al., 2019; Ring et al., 2016). It is well-known that this pattern need not translate into higher cognitive abilities of males. Rather, it points to several problems with the construction of the test (see also Juanchich et al., 2020; Sirota et al., 2020). By its nature, the test appears to favor individuals with high confidence in their numerical ability. Such positive self-appraisal, in turn, is more common among males. Second, in related studies, males typically cheat more than females (Abeler et al., 2019; Gerlach et al., 2019; Leib et al., 2021).

[3] Earlier work relied on supercomputers to test models of large size that involved computing very many expensive Bayes factors, e.g., Guo and Regenwetter (2014) or Regenwetter et al. (2017); see also Sarafoglou et al., (2023b) for an alternative approach for efficient order constraint evaluation based on a bridge sampling procedure.

Table 4
Additional constraint sets for the analysis of gender differences. F = females, M = males.

Constraint set	Order constraints	Description
8	$0 \leq P_{\text{O-F}} \leq P_{\text{O-M}} \leq 1$ and $0 \leq P_{\text{O\$-F}} \leq P_{\text{O\$-M}} \leq 1$ and $0 \leq P_{\text{N-F}} \leq P_{\text{N-M}} \leq 1$ and $0 \leq P_{\text{N\$-F}} \leq P_{\text{N\$-M}} \leq 1$.	Directed hypothesis
8b	$0 \leq P_{\text{O-F}} = P_{\text{O-M}} \leq 1$ and $0 \leq P_{\text{O\$-F}} = P_{\text{O\$-M}} \leq 1$ and $0 \leq P_{\text{N-F}} = P_{\text{N-M}} \leq 1$ and $0 \leq P_{\text{N\$-F}} = P_{\text{N\$-M}} \leq 1$.	Typical null hypothesis
8c	$\mid P_{\text{O-F}} - P_{\text{O-M}} \mid \leq 0.05$ and $\mid P_{\text{O\$-F}} - P_{\text{O\$-M}} \mid \leq 0.05$ and $\mid P_{\text{N-F}} - P_{\text{N-M}} \mid \leq 0.05$ and $\mid P_{\text{N\$-F}} - P_{\text{N\$-M}} \mid \leq 0.05$.	Approximate equality
9	$0 \leq C_{\text{O-F}} \leq C_{\text{O-M}} \leq 1$ and $0 \leq C_{\text{O\$-F}} \leq C_{\text{O\$-M}} \leq 1$ and $0 \leq C_{\text{N-F}} \leq C_{\text{N-M}} \leq 1$ and $0 \leq C_{\text{N\$-F}} \leq C_{\text{N\$-M}} \leq 1$.	Directed hypothesis
9b	$0 \leq C_{\text{O-F}} = C_{\text{O-M}} \leq 1$ and $0 \leq C_{\text{O\$-F}} = C_{\text{O\$-M}} \leq 1$ and $0 \leq C_{\text{N-F}} = C_{\text{N-M}} \leq 1$ and $0 \leq C_{\text{N\$-F}} = C_{\text{N\$-M}} \leq 1$.	Typical null hypothesis
9c	$\mid C_{\text{O-F}} - C_{\text{O-M}} \mid \leq 0.05$ and $\mid C_{\text{O\$-F}} - C_{\text{O\$-M}} \mid \leq 0.05$ and $\mid C_{\text{N-F}} - C_{\text{N-M}} \mid \leq 0.05$ and $\mid C_{\text{N\$-F}} - C_{\text{N\$-M}} \mid \leq 0.05$.	Approximate equality

We now discuss how to test these two predictions with order-constrained methods. Table A2 in the online supplement shows how we recoded the data for this analysis. The dataset and QTest files to reproduce the following analysis are available on OSF (https://osf.io/3va2w). We captured the predictions for this analysis with additional constraint sets, see Table 4. Constraint set 8 represents the idea that males outperform females on the CRT, in each treatment (e.g., $P_{\text{O \$-females}} \leq P_{\text{O\$-males}}$). Constraint set 9 formalizes the hypothesis that males cheat more than females, in each treatment. We test the hypothesis that constraint sets 8 and 9 hold jointly. We also test the competing hypothesis of no gender differences (constraint sets 8b and 9b hold jointly). This hypothesis predicts equal probabilities for females/males within treatments. It forms the typical null hypothesis in standard approaches. Finally, we replace the equality constraints in the former step by five percent wiggle space to model approximate gender equality (constraint sets 8c and 9c).

We are interested in two sets of questions. First, how well do the earlier Models 0–3 perform on separate samples, namely the female subsample, the male subsample, or both jointly? Second, which of the three hypotheses on gender differences (directed hypothesis, typical null hypothesis, approximate equality, see Table 4) best describes the data? The results of Analysis 4 are summarized in Table 5. Its left panel ("Evidence for Models 0–3″) addresses the first set, the right panel ("Gender differences") considers the second.

To tackle the first question, we repeated the main analysis (see above "Analysis 1″) separately for the female/male subsamples. When considering females and males jointly (see "Analysis 1″ and Table 5 column "combined"), the Bayes factor indicates substantial support for all Models. But the evidence in favor of Model 2 (and 2*) is much stronger.

From some perspectives, the results shown in Table 5 indicate important gender differences. Model 1 (and 1*: Incentives increase CRT performance through reflection rather than cheating) fits the male subsample reasonably well. But there is circumstantial evidence away from Model 1 (Bayes factor of about 2 in favor of the encompassing model) in the female subsample. While Model 2 (and 2*: Incentives increase CRT performance through cheating rather than reflection) is supported similarly well in both female and male subsamples, the processes captured in Model 1 seem to be more descriptive of males than females. Strikingly, Model 1 describes males' performance improvements under incentives better than Model 2, while the reverse is true for females.

On the second question, we evaluated models that added constraint sets 8–8c and 9–9c to our Models 0–3, see the right panel of Table 5. It stands out that the Bayes factors are generally much higher in this analysis. This is mainly due to the additional constraints making these models more parsimonious in the right side of the table, compared to the left. We can look at the right-most three columns of Table 5 in two different ways. One is to compare values within each column. This perspective reveals that regardless of our hypothesis about gender differences, we find strong support for Model 2 over any of the other models. The other is to compare values within each row. This comparison shows that, regardless of our hypotheses about CRT version and incentives, we reach one and the same conclusion about gender differences: the typical null hypothesis is supported over the directed hypothesis by about 4:1.

Comparing these results to the Bayes factor for Model 2 without any constraints regarding gender (from Table 3, it is 30.3), we can conclude that adding constraint sets 8 and 9 does improve the model substantially (the corresponding Bayes factor for Model 2 in Analysis 4 is 730.8, see Table 5). In other words, the hypothesis that there are gender differences in all treatments outperforms the encompassing model that is fully unconstrained. At the same time, however, comparing the directed hypothesis against the hypothesis of gender equalities (constraint sets 8b and 9b), we find even stronger support for gender equality. The typical null hypothesis is favored by around 4:1 over the directed hypothesis, and around 96:1 (2700/30.3) over Model 2 without any constraints on gender. Notably, applying Occam's razor to these models (see e.g., Myung & Pitt, 1997), the gender equality hypothesis is rewarded for its extreme parsimony. This is one of the notable advantages of an inference framework that comparatively weighs evidence and parsimony against each other, via Bayes factors among pairs of models, rather than merely looking for enough evidence to claim the presence of an effect.

This analysis only scratches the surface of possibilities to address gender differences. Much more nuanced hypotheses are testable with QTest. For instance, our models spell out conjunctions of constraints (summarized in Table 4) that capture the hypotheses that (a) there are gender differences in performance and cheating in all treatments, (b) there are no gender differences in any of the treatments, and (c) there are only negligible differences between females and males in all treatments. Models 2* and 2 fit the data particularly well under the hypothesis that all response probabilities are invariant across gender. We

Table 5
Results of order-constrained inference: bayes factors in analysis 4.

		Evidence for Models 0–3 in			Gender differences		
	Model	Female subsample	Male subsample	combined	Directed hypothesis	Typical null hypothesis	Approximate equality
Analysis 4 (Gender differences)	0	4.7	1.2	5.4	59.2	277.4	274.6
	1*	20.0	0.5	9.5	82.1	403.5	314.8
	1	20.0	0.6	12.4	101.7	472.7	388.3
	2*	9.7	11.1	107.4	758.0	2868.4	2825.5
	2	9.4	11.0	104.4	730.8	2900.1	2837.2
	3	3.6	2.2	7.9	70.0	490.6	375.7

Note. The left panel shows Bayes factors for Models 0–3 in female and male subsamples and combined. The right panel contains Bayes factors for Models 0–3 when constraints are added that capture different hypotheses on gender differences (cf. Table 4).

find substantial evidence in favor of this hypothesis. However, this raises the question whether gender differences may occur in some treatments, but not in others. If there were a theory about why performance or cheating should increase more strongly for females or males in one particular treatment, QTest could evaluate the corresponding collection of order-constraints.

For instance, performance-based incentives might trigger different behavioral responses among females and males. Related research reported that females and males reacted differently to competition (Niederle & Vesterlund, 2011). Males were more eager to compete, and their performance tended to benefit more strongly from competitive environments than females' performance. If performance-based pay produces a similar gender difference, then CRT performance increments under incentives should be larger for males than for females. Order-constrained modeling can capture this pattern by adding further order-constraints.

For example, $0 \leq (P_{\text{O\$-female}}$ - $P_{\text{O-female}}) \leq (P_{\text{O\$-male}}$ - $P_{\text{O-male}}) \leq 1$ captures the idea that incentives produce stronger performance improvements for males than for females (on the original CRT). One may or may not predict this hypothesis also on the new CRT. Females and males may also differ in how strongly incentives tempt them to cheat. To capture the idea that incentivization increases cheating more strongly for males than females, one can spell out similar constraints for cheating probabilities. It is also possible to require that the difference between certain probabilities should exceed some threshold.

In our dataset, order-constrained inference offers great flexibility to model a variety of hypotheses on the behavioral patterns of high performance and cheating on the CRT. Given a plausible theory of gender differences, it is possible to test highly nuanced and precise predictions. This emphasizes the great potential of order-constrained inference for more nuance and precision in theory testing. At the same time, this flexibility may tempt excessive exploration. We would emphasize that order-constrained inference is intended to serve as a theory testing tool, and not for exploratory analysis. Theory testing is where it best plays out its advantages over conventional approaches. Order-constrained inference metrics like Bayes factors are uninterpretable in exploratory settings.

4. Discussion

Together, these results make a convincing argument that cheating drove the performance increment under incentivization in Ludwig and Achtziger's (2021) experiment, rather than increased effort, or both effort and cheating combined. While Ludwig and Achtziger suggested that such an interpretation of their data was plausible, the original publication was lacking the analytical means to provide strong evidence in support of that claim. Here, relying on order-constrained inference, we were able to bridge this gap and ground that argument in more convincing data analytics.

The re-analysis addressed four limitations of the original publication. First, cheating was a measured variable, not an experimental one. Therefore, the authors could not rule out alternative explanations for correlations between performance and cheating behavior. Relying on Bayesian statistics, order-constrained inference offers advanced tools for quantitative competition among explanatory models. We leveraged this unique quality and obtained more convincing evidence in support of the claim that incentives increased performance merely through cheating.

Second, the re-analysis eschewed standard auxiliary assumptions of regression analysis (distributional assumptions, functional relation between variables, see, e.g., Regenwetter & Cavagnaro, 2019). Many statistical models force auxiliary assumptions that are arbitrary because they do not follow from the scientific theory under investigation. Adding or relaxing constraints on the theoretical level just to satisfy the statistical model can negatively affect the interpretation and reproducibility of empirical analysis (Regenwetter & Cavagnaro, 2019). Order-constrained inference is built on the far weaker assumption that the data generating process is a product of binomials. Since this assumption translates into drawing respondents independently from the population and counting how many show a given behavior, this requirement has allowed us to avoid a-theoretical distributional assumptions completely and stay true to Ludwig and Achtziger's theory.

Third, we have considered two behavioral outcomes jointly within one analytical step. Order-constrained inference has allowed us to formulate nuanced hypotheses for these outcomes based on the predictions in the original publication. By placing constraints on both outcomes and predicting that the hypotheses will hold jointly, we improved the precision of the theory test.

Finally, unbalanced data (e.g., large differences in experimental group sizes) may cause problems in conventional analysis (e.g., related to heteroscedasticity, see Klein et al., 2016). In order-constrained inference, different sample sizes merely imply that the samples differ in how much evidence they can generate for or against one model versus another, at a maximum. The higher the sample size, the higher the power to reject a hypothesis in a frequentist test and the higher the potential Bayes factors. Relying on Bayesian statistics, our analysis of gender differences showcases how order-constrained inference can be leveraged to improve the level of nuance and precision in experimental data analytics with unbalanced samples, especially Occam's razor, according to which statistical fit should be balanced against theoretical parsimony.

This analysis also highlights the potential of order-constrained methods to generate new hypotheses. For instance, our analysis seems to suggest that females' performance improvement under incentives is best explained by more reflection (Model 1). On the other hand, more cheating (Model 2) captured males' performance improvement better than any of the alternative models. We can combine this into a new hypothesis, essentially predicting that Model 1 applies to females (e.g., because their inclination to cheat is generally lower), while Model 2 applies to males (who, in addition to cheating more than females, might also be more strongly tempted to do so by the financial incentive). Running an order-constrained analysis posthoc on the same data would be hard to interpret because the data would be used twice: Once in generating the hypothesis, and again in calculating analytical metrics. Such an analysis would not be interpretable. Instead, it would be best to subject this new hypothesis to a new experiment.

Like any other methodology, our approach also has limitations. First, we dichotomized the ordinal performance measure to facilitate order-constrained modeling with QTest. In the process, some valuable information may have gotten lost. We sought to address this limitation by providing a robustness check based on a different cut-off for high performance. While we acknowledge that dichotomization presents a limitation, we also emphasize that, on the other hand, our approach completely avoids making any scale assumptions, in contrast to routine conventional analysis. A second limitation concerns assumptions about independence between the probabilities of performing highly and of cheating. Because participants who cheated on the CRT can be expected to also have a higher probability of performing highly, it is reasonable to assume some level of dependence among these behavioral outcomes. While assumptions of independence between binomials may be viewed as a limitation of the methodology more generally, it is important to note that inter-dependent empirical observations present a challenge for almost any analytical approach. For a more comprehensive discussion of different forms of independence, and their implications for analysis and theory development, see e.g., Regenwetter and Cavagnaro (2019, pp. 138–140), and Regenwetter and Davis-Stober (2018).

Beyond the methodological contribution, our results inform an ongoing debate, in experimental research on dishonesty, about the link between financial incentives and cheating (Abeler et al., 2019; Gerlach et al., 2019; Kajackaite & Gneezy, 2017). In some experimental paradigms (e.g., sender-receiver game with cheap-talk element, Gneezy, 2005) lying will typically increase with incentives. But studies with many other procedures, like the coin-flip or matrix tasks (Abeler et al., 2014; Mazar et al., 2008), rarely report a similar increase (see also Fischbacher & Föllmi-Heusi, 2013). Our re-analysis supports an important role of financial incentives in predicting dishonest behavior.

We emphasize that CRT performance improvements under incentives, in Ludwig and Achtziger’s (2021) dataset, were best explained by more cheating. This adds an interesting perspective on inconsistencies recently reported regarding the relation between financial incentives and CRT performance (see Brañas-Garza et al., 2019; Yechiam & Zeif, 2022). To the extent that cheating occurred not only in Ludwig and Achtziger’s experiment, but also in other CRT web studies, participants’ dishonest behavior could have distorted the findings. Future online research with the CRT should consider these influences.

Declarations of Competing Interest

None.

Data availability

Data, code, and materials are available on the OSF, see https://osf.io/3va2w/

Appendix

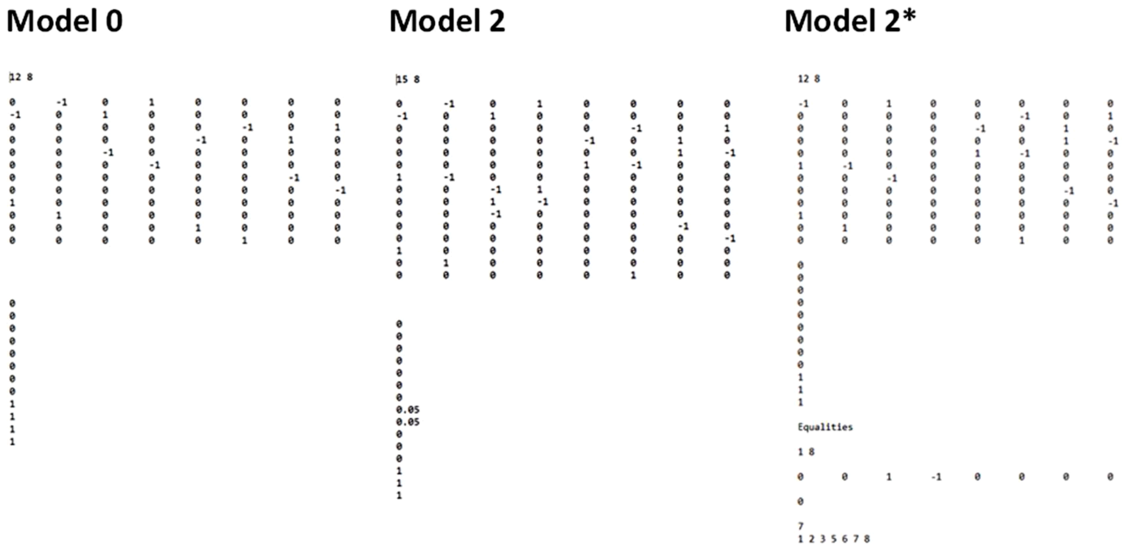
QTest input: text files

The order of eight probabilities is the same in all files: 4 x Performance (denoted by *P*), 4 x Cheating (denoted by *C*), and ordered within these groups according to the experimental treatments: Original-fixed, Original-piece-rate, New-fixed, New-piece-rate.

P_O $P_{O\$}$ P_N $P_{N\$}$ C_O $C_{O\$}$ C_N $C_{N\$}$

The text files (see Fig. A1) first state the number of rows (constraints) and columns (probabilities) in the document. While we list 15 constraints in Model 2, there are additional constraints implicit in our model, such as $P_{N\$} \geq 0$. Indeed, QTest requires the order-constraints to be non-redundant. To better understand the composition of these files, let us consider the example of Model 2. The individual constraints contained in Model 2 (constraint sets 1–2 and 5–6, see Table 1) can be spelled out as seen in Table A1.

Note that rows 13 and 14 are redundant. Constraint set 1 requires $P_{N\$}$ to be smaller than $P_{O\$}$ (row 1 in Table A1). Because row 11 specifies that $P_{O\$} \leq 1$, row 13 contains a redundant constraint. The same is the case for row 14. Because constraint set 6 requires that $C_O \leq C_{O\$}$ (row 9), and row 15 states that $C_{O\$} \leq 1$, the constraint in row 14 is redundant. Removing the two rows results in a matrix with 15 rows, as shown in Fig. A1. Note that the order of the rows is irrelevant, but the order of columns is fixed.



Model 0

```
12 8

0   -1  0   1   0   0   0   0
-1  0   1   0   0   0   0   0
0   0   0   0   0   -1  0   1
0   0   0   0   -1  0   1   0
0   0   -1  0   0   0   0   0
0   0   0   -1  0   0   0   0
0   0   0   0   0   0   -1  0
0   0   0   0   0   0   0   -1
1   0   0   0   0   0   0   0
0   1   0   0   0   0   0   0
0   0   0   0   1   0   0   0
0   0   0   0   0   1   0   0


0
0
0
0
0
0
0
0
1
1
1
1
```

Model 2

```
15 8

0   -1  0   1   0   0   0   0
-1  0   1   0   0   0   0   0
0   0   0   0   0   -1  0   1
0   0   0   0   -1  0   1   0
0   0   0   0   0   0   1   -1
0   0   0   0   1   -1  0   0
1   -1  0   0   0   0   0   0
0   0   -1  1   0   0   0   0
0   0   1   -1  0   0   0   0
0   0   -1  0   0   0   0   0
0   0   0   0   0   0   -1  0
0   0   0   0   0   0   0   -1
1   0   0   0   0   0   0   0
0   1   0   0   0   0   0   0
0   0   0   0   0   1   0   0


0
0
0
0
0
0
0
0.05
0.05
0
0
0
1
1
1
```

Model 2*

```
12 8

-1  0   1   0   0   0   0   0
0   0   0   0   0   -1  0   1
0   0   0   0   -1  0   1   0
0   0   0   0   0   0   1   -1
0   0   0   0   1   -1  0   0
1   -1  0   0   0   0   0   0
0   0   -1  0   0   0   0   0
0   0   0   0   0   0   -1  0
0   0   0   0   0   0   0   -1
1   0   0   0   0   0   0   0
0   1   0   0   0   0   0   0
0   0   0   0   0   1   0   0

0
0
0
0
0
0
0
0
0
1
1
1

Equalities

1 8

0   0   1   -1  0   0   0   0

0

7
1 2 3 5 6 7 8
```

Fig. A1. Screenshots of matrix-formated data input for QTest. The figure shows Models 0, 2 and 2* as examples (see also Table 1). All text files are available on the OSF.

Table A1
Model 2 in matrix format.

Row	Constraint set	Constraint	P_O	$P_{O\$}$	P_N	$P_{N\$}$	C_O	$C_{O\$}$	C_N	$C_{N\$}$	Upper bound
1	1	$(-1 * P_{O\$}) + (1 * P_{N\$}) \leq 0$	0	−1	0	1	0	0	0	0	0
2	1	$(-1 * P_O) + (1 * P_N) \leq 0$	−1	0	1	0	0	0	0	0	0
3	2	$(-1 * C_{O\$}) + (1 * C_{N\$}) \leq 0$	0	0	0	0	0	−1	0	1	0
4	2	$(-1 * C_O) + (1 * C_N) \leq 0$	0	0	0	0	−1	0	1	0	0
5	5	$(1 * P_O) + (-1 * P_{O\$}) \leq 0$	1	−1	0	0	0	0	0	0	0
6	5	$(-1 * P_N) + (1 * P_{N\$}) \leq 0.05$	0	0	−1	1	0	0	0	0	0.05

(*continued on next page*)

Table A1 (*continued*)

Row	Constraint set	Constraint	P_O	$P_{O\$}$	P_N	$P_{N\$}$	C_O	$C_{O\$}$	C_N	$C_{N\$}$	Upper bound
7	5	$(1 * P_N) + (-1 * P_{N\$}) \leq 0.05$	0	0	1	−1	0	0	0	0	0.05
8	6	$(1 * C_N) + (-1 * C_{N\$}) \leq 0$	0	0	0	0	0	0	1	−1	0
9	6	$(1 * C_O) + (-1 * C_{O\$}) \leq 0$	0	0	0	0	1	−1	0	0	0
10		$(1 * P_O) \leq 1$	1	0	0	0	0	0	0	0	1
11		$(1 * P_{O\$}) \leq 1$	0	1	0	0	0	0	0	0	1
12		$(-1 * P_N) \leq 0$	0	0	−1	0	0	0	0	0	0
13		$(-1 * P_{N\$}) \leq 0$	0	0	0	−1	0	0	0	0	0
14		$(1 * C_O) \leq 1$	0	0	0	0	1	0	0	0	1
15		$(1 * C_{O\$}) \leq 1$	0	0	0	0	0	1	0	0	1
16		$(-1 * C_N) \leq 0$	0	0	0	0	0	0	−1	0	0
17		$(-1 * C_{N\$}) \leq 0$	0	0	0	0	0	0	0	−1	0

References

Abeler, J., Becker, A., & Falk, A. (2014). Representative evidence on lying costs. *Journal of Public Economics, 113*, 96–104. https://doi.org/10.1016/j.jpubeco.2014.01.005

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica : journal of the Econometric Society, 87*(4), 1115–1153. https://doi.org/10.3982/ECTA14673

Alós-Ferrer, C., Fehr, E., & Netzer, N. (2021). Time will tell: Recovering preferences when choices are noisy. *Journal of Political Economy, 129*(6), 1828–1877. https://doi.org/10.1086/713732

Alós-Ferrer, C., Garagnani, M., & Hügelschäfer, S. (2016). Cognitive reflection, decision biases, and response times. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01402, 1402–1402.

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., & Wagenmakers, E. J. (2015). An introduction to Bayesian hypothesis testing for management research. *Journal of Management, 41*(2), 521–543. https://doi.org/10.1177/0149206314560412

Bialek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods, 50*(5), 1953–1959. https://doi.org/10.3758/s13428-017-0963-x

Brañas-Garza, P., Kujal, P., & Lenkei, B. (2019). Cognitive reflection test: Whom, how, when. *Journal of Behavioral and Experimental Economics, 82*, Article 101455. https://doi.org/10.1016/j.socec.2019.101455

Davis-Stober, C. P. (2009). Analysis of multinomial models under inequality constraints: Applications to measurement theory. *Journal of Mathematical Psychology, 53*(1), 1–13. https://doi.org/10.1016/j.jmp.2008.08.003

Davis-Stober, C. P., & Regenwetter, M. (2019). The 'paradox' of converging evidence. *Psychological Review, 126*(6), 865–879. https://doi.org/10.1037/rev0000156

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise—An experimental study on cheating. *Journal of the European Economic Association, 11*(3). https://doi.org/10.1111/jeea.12014

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4). https://doi.org/10.1257/089533005775196732

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*(4), 733–760.

Gerlach, P., Teoderescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin, 145*(1), 1–44. https://doi.org/10.1037/bul0000174

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review, 95*(1), 384–394.

Gu, X., Hoijtink, H., Mulder, J., & Rosseel, Y. (2019). Bain: A program for Bayesian testing of order constrained hypotheses in structural equation models. *Journal of Statistical Computation and Simulation, 89*(8), 1526–1553. https://doi.org/10.1080/00949655.2019.1590574

Guo, Y., & Regenwetter, M. (2014). Quantitative tests of the perceived relative argument model: Comment on Loomes (2010). *Psychological Review, 121*, 696–705. https://doi.org/10.1037/a0036095

Haigh, M. (2016). Has the standard cognitive reflection test become a victim of its own success? *Advances in Cognitive Psychology, 12*(3). https://doi.org/10.5709/acp-0193-5

Heck, D. W., & Davis-Stober, C. P. (2019). Multinomial models with linear inequality constraints: Overview and improvements of computational methods for Bayesian inference. *Journal of Mathematical Psychology, 91*, 70–87. https://doi.org/10.1016/j.jmp.2019.03.004

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods, 24*(5), 539–556. https://doi.org/10.1037/met0000201

Jeffreys, H. (1998). *The theory of probability*. Oxford University Press.

Juanchich, M., Sirota, M., & Bonnefon, J. F. (2020). Anxiety-induced miscalculations, more than differential inhibition of intuition, explain the gender gap in cognitive reflection. *Journal of Behavioral Decision Making, 33*(4), 427–443. https://doi.org/10.1002/bdm.2165

Kahneman, D. (2012). *Thinking, fast and slow*. Penguin.

Kajackaite, A., & Gneezy, U. (2017). Incentives and cheating. *Games and Economic Behavior, 102*, 433–444. https://doi.org/10.1016/j.geb.2017.01.015

Klein, A., Gerhard-Lehn, C., Büchner, R., Diestel, S., & Schermelleh-Engel, K. (2016). The detection of heteroscedasticity in regression models for psychological data. *Psychological Test and Assessment Modeling, 58*(4), 567–592.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods, 10*(4), 477–493. https://doi.org/10.1037/1082-989X.10.4.477

Leib, M., Köbis, N., Soraperra, I., Weisel, O., & Shalvi, S. (2021). Collaborative dishonesty: A meta-analytic review. *Psychological Bulletin, 147*(12), 1241–1268. https://doi.org/10.1037/bul0000349

Ludwig, J., & Achtziger, A. (2021). Cognitive misers on the web: An online-experiment of incentives, cheating, and cognitive reflection. *Journal of Behavioral and Experimental Economics, 94*, Article 101731. https://doi.org/10.1016/j.socec.2021.101731

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*. https://doi.org/10.1509/jmkr.45.6.633

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological), 42*(2), 109–142.

Meng, X. L. (1994). Posterior predictive p-values. *The Annals of Statistics, 22*(3), 1142–1160. https://doi.org/10.1214/aos/1176325622

Meyer, A., Zhou, E., & Frederick, S. (2018). The non-effects of repeated exposure to the cognitive reflection test. *Judgment and Decision Making, 13*(3), 246–259.

Moffatt, P. G. (2016). *Experimetrics: Econometrics for experimental economics*. Palgrave Macmillan.

Mulder, J., Williams, D. R., Gu, X., Tomarken, A., Böing-Messing, F., Olsson-Collentine, A., et al. (2021). BFpack: Flexible bayes factor testing of scientific theories in R. *Journal of Statistical Software, 100*, 1–63. https://doi.org/10.18637/jss.v100.i18

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4*(1), 79–95. https://doi.org/10.3758/BF03210778

Niederle, M., & Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics, 3*(1), 601–630. https://doi.org/10.1146/annurev-economics-111809-125122

Regenwetter, M. (2020). Tutorial: "With sufficient increases in X, more people will engage in the target behavior. *Journal of Mathematical Psychology, 99*, Article 102457. https://doi.org/10.1016/j.jmp.2020.102457

Regenwetter, M., & Cavagnaro, D. R. (2019). Tutorial on removing the shackles of regression analysis: How to stay true to your theory of binary response probabilities. *Psychological Methods, 24*(2), 135–152. https://doi.org/10.1037/met0000196

Regenwetter, M., Cavagnaro, D. R., Popova, A., Guo, Y., Zwilling, C., Lim, S. H., et al. (2017). Heterogeneity and parsimony in intertemporal choice. *Decision, 5*(2), 63–94. https://doi.org/10.1037/dec0000069

Regenwetter, M., & Davis-Stober, C. P. (2018). The role of independence and stationarity in probabilistic models of binary choice. *Journal of Behavioral Decision Making, 31*(1), 100–114. https://doi.org/10.1002/bdm.2037

Regenwetter, M., Davis-Stober, C. P., Lim, S. H., Guo, Y., Popova, A., Zwilling, C., et al. (2014). QTest: Quantitative testing of theories of binary choice. *Decision, 1*(1), 2–34. https://doi.org/10.1037/dec0000007

Ring, P., Neyse, L., David-Barett, T., & Schmidt, U. (2016). Gender differences in performance predictions: Evidence from the cognitive reflection test. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01680

Sarafoglou, A., Aust, F., Marsman, M., Bartoš, F., Wagenmakers, E. J., & Haaf, J. M. (2023a). *Multibridge: An R package to evaluate informed hypotheses in binomial and multinomial models* (pp. 1–26). Behavior Research Methods. https://doi.org/10.3758/s13428-022-02020-1

Sarafoglou, A., Haaf, J. M., Ly, A., Gronau, Q. F., Wagenmakers, E. J., & Marsman, M. (2023b). Evaluating multinomial order restrictions with bridge sampling. *Psychological Methods, 28*(2), 322–338. https://doi.org/10.1037/met0000411

Silvapulle, M. J., & Sen, P. K. (2005). *Constrained statistical inference: Inequality, order and shape restrictions*. Wiley. https://doi.org/10.1002/9781118165614

Sirota, M., Dewberry, C., Juanchich, M., Valuš, L., & Marshall, A. C. (2020). Measuring cognitive reflection without maths: Development and validation of the verbal cognitive reflection test. *Journal of Behavioral Decision Making, 34*(3), 322–343. https://doi.org/10.1002/bdm.2213

Stieger, S., & Reips, U. D. (2016). A limitation of the cognitive reflection test: Familiarity. *PeerJ, 4*, e2395. https://doi.org/10.7717/peerj.2395

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning, 20* (2). https://doi.org/10.1080/13546783.2013.844729

Woike, J. K. (2019). Upon repeated reflection: Consequences of frequent exposure to the cognitive reflection test for mechanical Turk participants. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.02646

Yechiam, E., & Zeif, D. (2022). Revisiting the effect of incentivization on cognitive reflection: A meta-analysis. *Journal of Behavioral Decision Making*. https://doi.org/10.1002/bdm.2286 *(Online early view)*.

Zwilling, C. E., Cavagnaro, D. R., Regenwetter, M., Lim, S. H., Fields, B., & Zhang, Y. (2019). QTest 2.1: Quantitative testing of theories of binary choice using Bayesian inference. *Journal of Mathematical Psychology, 91*, 176–194. https://doi.org/10.1016/j.jmp.2019.05.002