Reconsidering mistakes: reproduction and replication of Nielsen and Rehbeck (2022)

Gabriel Bayle, Dimitri Dubois, Simon Varaine

CEE-M
CENTER FOR ENVIRONMENTAL ECONOMICS – MONTPELLIER

Reconsidering mistakes: reproduction and replication of Nielsen and Rehbeck (2022)

Gabriel Bayle,
Dimitri Dubois,
&
Simon Varaine

CEE-M Working Paper 2026-02

CEE-M

INRAE

cnrs

L'INSTITUT agro Montpellier

Reconsidering mistakes: reproduction and replication of Nielsen and Rehbeck (2022)*

Gabriel Bayle[1,2], Dimitri Dubois[1] & Simon Varaine[1,3]

[1]CEE-M, University of Montpellier, CNRS, Institut Agro, INRAE

[2]GATE, CNRS, Université Lumière Lyon 2,

Université Jean-Monnet Saint-Etienne, emlyon business school

[3]Univ. Grenoble Alpes, IEPG, CNRS, INRAE, Grenoble INP, GAEL

January 14, 2026

Abstract

There is a long-standing debate as to whether violations of rational choice axioms reflect normative deviations from the theory or simply mistakes. We contribute to this debate by reproducing and replicating Nielsen and Rehbeck's (2022) experimental study, with a new focus on heterogeneity across axioms. We conduct a three-part analysis comprising a direct computational reproduction, a robustness reproduction, and a high-powered preregistered replication (N = 451) focusing on the Independence of Irrelevant Alternatives (IIA). We find robust evidence that individuals express a desire to follow canonical axioms, but perceive violations as mistakes only for a subset of them—specifically, IIA and Transitivity. In contrast, we find no evidence that violations of Independence, First-Order Stochastic Dominance, Branching, or Consistency are perceived as mistakes. We discuss these findings through the lens of cognitive complexity, suggesting that individuals may fail to recognize violations of more demanding axioms even when prompted.

Keywords: decision theory, choice axioms, complexity, reproducibility, replication

JEL Codes: D01, C12, C91, D81, D91

*Corresponding author: Gabriel Bayle, GATE, 35 rue Raulin, 69007 Lyon, France, gabriel.bayle.econ@gmail.com. This replication was part of the "Grenoble Replication Games", organized by the Institute for Replication (I4R; Replication Games). We thank Abel Brodeur, Anna Dreber, Paolo Crosetto, Margaux Sinceux and the Institute for Replication for the organization and the funding of the Online experiment. We also thank Kirby Nielsen and John Rehbeck for their comments on a previous version of this paper, which led to several corrections. An early version of this paper has been presented in the 2025 ASFEE conference in Nancy (France) and the 2025 ESSCA workshop on replication and reproducibility. Pre-registration, replication data, analysis and oTree codes are available in the OSF repository: https://osf.io/csgv4/. Following the recommendation of Altonji et al. (2025), we want to specify the individual contributions of the authors. G.B. took the lead in managing the project, G.B., D.D. and S.V. worked on the original paper to plan the project and determine the focus and the necessary steps for the replication, G.B. and S.V. wrote the preregistration and conducted the power analysis, S.V. conceived the data analysis codes, G.B. and S.V. conceived and conducted the two reproductions, D.D. coded the experiment and conducted the pilots, D.D. managed the data collection, G.B. performed the citation-context review, G.B. and S.V. took the lead in writing the manuscript. All authors provided critical feedback and helped shape the research, analysis and manuscript. Authors may have also made minor contributions to categories beyond those listed. Authors declare no conflicts of interest. Authors used AI solely for language editing and improvements to writing clarity and readability, all content has been reviewed by the authors.

1 Introduction

Violations of canonical axioms of rational choice, such as transitivity, independence, and stochastic dominance, have been a central finding of experimental economics for several decades (*e.g.* May, 1954; Tversky, 1969; Slovic and Tversky, 1974; Huber et al., 1982; Loomes et al., 1991; Camerer, 1995; Li, 1996).

In response, an early strand of the literature in the 1960s and 1970s sought to defend the normative appeal of these axioms. A common view was that observed violations reflected misunderstanding, and that once the axioms were properly explained, reasonable individuals would wish to comply with them (Raiffa, 1961; Borch, 1968). This view motivated experimental studies examining whether individuals would endorse the axioms when invited to reflect on their own choices or to reconsider them in light of explicit discussion of the axioms (MacCrimmon, 1968; MacCrimmon and Larsson, 1979).

These early attempts, however, suffered from important methodological limitations, including small samples, weak or absent incentives, and largely unstructured or experimenter-guided reconciliation procedures. Moreover, subsequent experimental evidence documented persistent violations even when subjects were presented with explicit normative arguments in favor of the axioms (Slovic and Tversky, 1974). As a result, the prevailing interpretation in the literature has been that such violations primarily reflect stable and meaningful preferences rather than mistakes, thereby challenging both the descriptive accuracy and the normative relevance of rational choice theory.

Nielsen and Rehbeck (2022) fundamentally revisit this conclusion. They directly assess whether individuals themselves regard violations of rational choice axioms as *mistakes*. To do so, they introduce an incentivized experimental framework that separates preferences over axioms from realized lottery choices and examines how individuals reconcile conflicts between the two.

In a first stage of the experiment, participants are asked whether they want their choices to satisfy a given decision rule, corresponding to a canonical axiom, namely Transitivity (TRANS), First-Order Stochastic Dominance (FOSD), Independence of Irrelevant Alternatives (IIA), Independence (IND), Branching (BRANCH), or Consistency (CONS). These axioms are presented in an intuitive, graphical form,

and the elicitation is fully incentivized: if an axiom is selected, it may later be implemented to determine payoffs, replacing the participant's own choice in relevant decision problems. Participants can also choose not to follow the axiom and instead retain control over their choices. Importantly, participants are also presented with "control" axioms that are constructed as the "opposite" of each axiom under study and therefore run counter to rational choice theory, in order to control for demand effects.

In a second stage, participants make a series of standard lottery choices designed to generate potential violations of the axioms. Finally, in a reconciliation stage, participants are confronted with any inconsistencies between the axioms they had endorsed ex ante and their actual lottery choices. At this point, they are given a neutral opportunity to revise their decisions: they may change their lottery choices, renounce the axiom, do both, or leave the inconsistency unresolved. This design allows the researcher to observe not only whether inconsistencies arise, but also how individuals themselves interpret and resolve them.

Using this framework, Nielsen and Rehbeck (2022) report that a large majority of participants express a desire to follow canonical axioms ex ante and that, when confronted with inconsistencies, many revise their lottery choices in order to restore compliance. They interpret this pattern as evidence that a substantial share of observed violations should be understood as mistakes rather than as expressions of alternative, stable preferences.

These findings are potentially far-reaching, and they have quickly begun to shape how researchers interpret and operationalize axiom violations (further details in Section 2). Given the theoretical and empirical importance of this claim, it is essential to assess the reliability of Nielsen and Rehbeck (2022)'s findings. Are their results robust to replication by independent researchers? Can the main patterns be reproduced using their data, and replicated with new subjects and experimental sessions? The present study aims to address these questions.

Our study computationally and experimentally replicates and extends the main findings of Nielsen and Rehbeck (2022). Using the authors' original data and code, we first confirm the reproducibility of their main results. We then show that these results are robust to a range of reasonable inferential corrections.

We also investigate whether the results hold for the different axioms studied by Nielsen and Rehbeck (2022). This focus is motivated by the fact that the experimental and theoretical literature on rational choice has historically devoted uneven attention to different axioms, and that these axioms may differ substantially in their intuitive appeal and normative force. Some axioms, such as TRANS, have been extensively studied and debated both theoretically and experimentally (e.g. May, 1954; Tversky, 1969). The IND axiom has also played a central role in the literature on the Allais paradox (e.g. Slovic and Tversky, 1974; Humphrey and Kruse, 2024). Importantly, Humphrey and Kruse (2024) found that violations of the IND axiom persist even after Savage's normative arguments in favor of the axiom are provided to the subjects, replicating the original findings of Slovic and Tversky (1974). This contrasts with the more general pattern documented by Nielsen and Rehbeck (2022). A possible reason for these different results is that the normative appeal of rational choice axioms is heterogeneous, and that conclusions drawn from pooling across axioms may depend critically on which axioms are considered.

Our results align with this view and reveal substantial heterogeneity across axioms. In particular, we replicate the preference for canonical over control axioms and find no systematic differences in violation rates conditional on axiom choice. However, the central result of the original study, that individuals are more likely to revise choices violating canonical axioms than control axioms, does not hold uniformly across axioms. Instead, this reconciliation pattern is driven by a subset of axioms, namely IIA and TRANS. By contrast, no robust reconciliation effect emerges for FOSD, BRANCH, CONS, as well as IND, which is consistent with the results of Slovic and Tversky (1974) and more recent studies (Humphrey and Kruse, 2024, 2025).

Motivated by this axiom-level heterogeneity, we conduct a preregistered experimental replication using newly collected online data ($N = 451$), focusing on IIA. This replication confirms all three main results of the original study, albeit with smaller effect sizes, thereby providing complementary experimental support for the normative interpretation of IIA violations.

Taken together, our findings suggest that individuals' willingness to treat violations as mistakes is selective rather than universal. While some canonical axioms,

such as IIA and TRANS, appear to retain strong normative appeal, others do not consistently trigger reconciliation. We interpret these patterns as reflecting differences in cognitive accessibility and complexity across axioms, with violations involving compound lotteries (such as FOSD, IND, and BRANCH) being less likely to be detected and corrected.

The rest of the paper is organized as follows. In section 2 we use a citation-context review to motivate this replication. Section 3 details our replication methodology. Section 4 reports the computational reproduction: both the direct reproduction based on the authors' code and robustness analyses. Section 5 details the experimental replication. We conclude in Section 6 with a discussion of how axiom complexity shapes these findings.

2 Citation-context review

Deciding what to replicate is inherently a choice under resource constraints, and recent work formalizes replication study selection as a decision problem aimed at maximizing the expected utility ("replication value") of increasing certainty about a claim (Isager et al., 2023). In this spirit, we use a citation-context review as a transparent, descriptive proxy for the perceived value component of replication, capturing how strongly Nielsen and Rehbeck (2022) is already shaping subsequent research, and we present the review protocol as part of our methods.

We assess how Nielsen and Rehbeck (2022) is used in all Google Scholar citing documents. For each citing paper, we verify that Nielsen and Rehbeck (2022) is genuinely referenced and code the paper's topic, the section in which the citation appears, and the role the citation plays in the argument. Due to the recent publication of the original paper, we limit our study to the first generation citing papers. See the online appendix A for a detailed description of the method used for the review.

Since its original publication as a working paper, Nielsen and Rehbeck (2022) has been widely cited across economics and adjacent fields. In our citation-context review of all citing papers identified (N = 107), we exclude 21 documents under the criteria described in the online appendix A, leaving 86 papers in the analytic sample. Most citing papers use the study as evidence or motivation for interpreting

axiom violations as mistakes, and as evidence that individuals are willing to revise choices toward axiom-consistent behavior in a reconciliation stage (55 papers; 64%), and/or as a methodological template for eliciting preferences over decision rules and reconciling inconsistencies (15 papers; 17%). Only a small minority of citing papers directly examine robustness across axioms, samples, or implementation details. This pattern underscores the importance of assessing the reliability and scope of Nielsen and Rehbeck (2022)'s central claims.

In terms of subject matter, citations come from multiple clusters. A first set of papers in behavioral and experimental economics cites Nielsen and Rehbeck (2022) in connection with classic questions about rationality and consistency, often using it to motivate renewed attention to whether observed violations reflect mistakes rather than stable preferences. A second cluster links the result to mechanisms of bounded rationality, especially complexity, cognitive noise, and the use of rules or procedures, treating Nielsen and Rehbeck (2022) as evidence that deviations from axioms may arise from implementational frictions. A third, smaller set draws on the paper as a methodological reference point, emphasizing the separation between preferences over decision rules and realized choices, and the role of reconciliation designs in eliciting "second thoughts." Finally, citations also appear in adjacent domains (e.g., normative welfare discussions, policy design, and even applications outside economics), where the study is invoked as a general argument for interpreting inconsistencies as correctable errors.

In terms of citation placement, a large majority of papers (70%) cite Nielsen and Rehbeck (2022) in the introduction, typically when framing the contribution or motivating the research question.

Taken together, these patterns indicate that Nielsen and Rehbeck (2022) is frequently used to support broad claims about axiom violations reflecting mistakes. This widespread and often general use makes it especially important to subject the findings to independent replication and to clarify the boundaries of the effect across axioms.

3 Replication method

Our paper is situated within the growing literature on best practices in reproducibility and replication in economics. Following recent methodological contributions, a fundamental distinction is commonly drawn between *reproduction* and *replication* (e.g. Dreber and Johannesson, 2019; Bayle et al., 2026). Reproduction refers to the reuse of the original data and code to verify that reported results can be obtained as stated, whereas replication tests the robustness and external validity of these results by applying the same analytical framework to new data.

Each of these approaches can further be subdivided depending on whether the original methodology is strictly followed or deliberately extended. On the reproduction side, a distinction is made between *direct computational reproduction*, which aims to exactly reproduce the original results using the original data and code, and *robustness reproduction*, which evaluates the sensitivity of these results to reasonable methodological modifications or corrections. On the replication side, experimental studies can be replicated either through *direct replications*, which closely mirror the original protocol, or through *conceptual replications*, which test the same underlying hypothesis using a modified design, population, or context.

Our contribution follows this structured approach and combines three complementary steps. First, we conduct a *direct computational reproduction*, reproducing the original results of Nielsen and Rehbeck (2022) using their data and code without modification. This step serves as a validation exercise, ensuring that the original analysis is computationally reproducible and that the replication materials function as intended.

Second, we perform a *robustness computational reproduction*, which constitutes the core methodological contribution of our paper. Using the same data as the original study, we introduce a set of robustness corrections that we consider necessary for valid inference, including adjustments related to statistical testing, dependence across observations, and subsample analyses. Beyond these standard robustness checks, a central feature of our reproduction is to explicitly examine heterogeneity in the results across axioms.

Accordingly, our robustness analysis is explicitly conducted on an axiom-by-

axiom basis. Rather than treating canonical axioms as a homogeneous class, we assess whether the patterns documented by Nielsen and Rehbeck (2022) emerge uniformly across axioms or are driven by a subset of them. From a normative perspective, this distinction is important: if some axioms elicit systematic reconciliation while others do not, this provides information about which axioms genuinely retain normative appeal in the eyes of decision makers.

The third step of our analysis is a preregistered *conceptual experimental replication* based on newly collected online data. This experimental replication does not aim to reproduce the entire original study, but instead focuses on a single axiom. This design choice allows us to maximize statistical power and to directly test, in a new sample, the implications suggested by the axiom-level patterns identified in the robustness analysis. The criteria guiding the selection of this axiom are discussed in detail in the corresponding section. To streamline terminology, we refer throughout to reproduction (encompassing both direct and robustness, with the direct reproduction viewed as the preliminary stage of the reproduction exercise) and replication (corresponding to our conceptual experimental replication).

4 Reproduction

We focus on the first three main results reported by Nielsen and Rehbeck (2022). The first result shows that individuals select canonical axioms more frequently than control axioms. The second result indicates that individuals are equally likely to violate a canonical axiom regardless of whether they initially selected it or not. The third result demonstrates that individuals are more likely to revise their choices to reconcile them with canonical axioms when they violate them than when they violate control axioms. This last result is particularly important, as it sheds light on whether axiom violations can be interpreted as mistakes from the subjects' own perspective. These results are based on analyses pooling the six axioms studied in the laboratory experiment and on the IND axiom in the online experiment.

For each of these results, we first conduct our direct reproduction using the package attached to the original publication (https://www.openicpsr.org/openicpsr/project/164661/version/V1/view). The authors provide all materials necessary

to reproduce the original results (Results 1 to 3) based on the raw data from both the laboratory and online experiments. To the best of our knowledge, the authors did not register a pre-analysis plan. Without any modification, we successfully reproduce all main results computationally from the raw data. We do not encounter any coding errors. Our direct reproduction yields the same proportions and p-values as those reported in the paper for all main results. Table A3 in the online appendix summarizes the reproduced findings. We also reimplemented the authors' Stata code in R by strictly following their data analysis procedures and obtained consistent results. The R code is provided in our OSF depository.

We now turn to our robustness reproduction. For each experiment, we first present the authors' original procedures and results, and then introduce our methodological modifications along with the results they produce. We proceed separately for the laboratory and the online experiments.

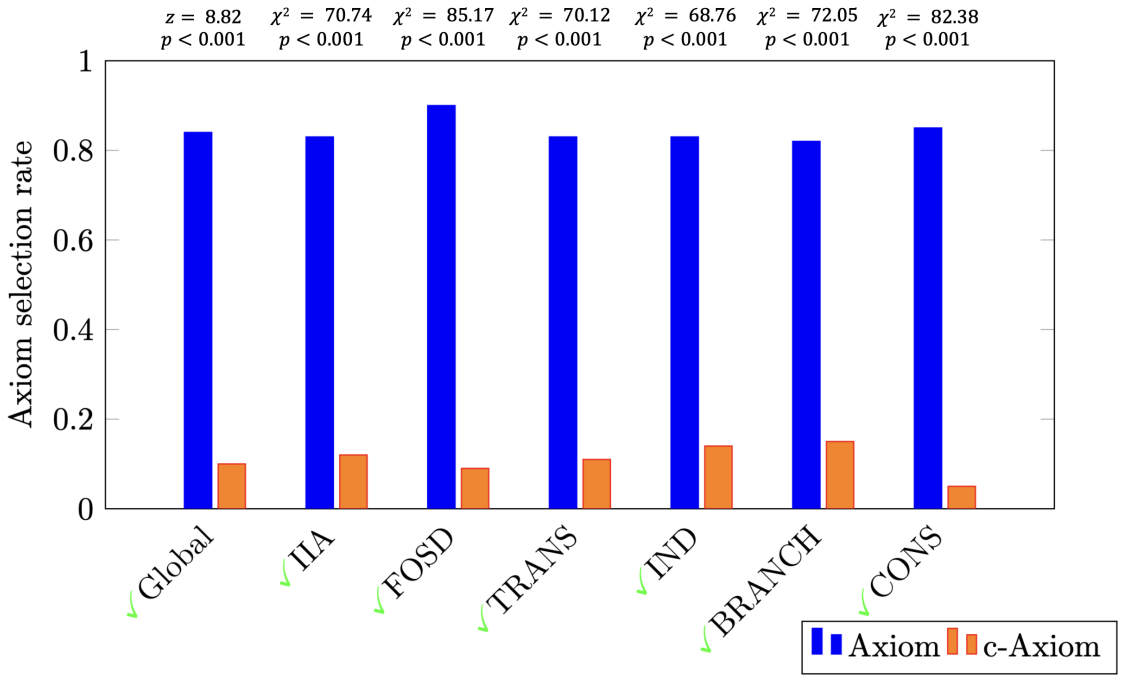
4.1 Lab experiment

4.1.1 Result 1: Do people select axioms more than their control axioms (c-axioms)?

In the original study, subjects are asked, in a first stage, to select decision rules corresponding to canonical axioms from rational choice theory and control axioms (referred to as c-axioms in the paper). This stage consists of a randomized set of eighteen rules: the six canonical axioms under study, the six corresponding c-axioms, and six distractor rules. The authors report the frequency with which individuals select canonical axioms from rational choice theory (85%) and control axioms (10%). They use these descriptive proportions to conclude that individuals select canonical axioms more frequently than control axioms. No statistical test is reported, possibly due to the substantial difference between the two frequencies.

For this result, each observation in the data corresponds to one pair of axiom and c-axiom for a single subject, resulting in six observations per subject. To statistically test the difference while accounting for the dependency of individual observations, we propose the following procedure: For each subject, we calculate the proportion of selected axioms and c-axioms, compute the overall means, and then perform a Wilcoxon signed-rank test for the difference between these two means. This method yields a highly significant difference ($z = 8.80, p < 0.001$), aligning

Figure 1: Summary of Reproduction of Result 1



Note: Proportion of selection of (1) axioms and (2) their respective c-axioms. At the global level, we run the Wilcoxon signed-rank test and report the z and p-value. At the axiom level, we run McNemar's χ^2 and report the χ^2 and p-values.

with their first result.

Additionally, we test the difference in selection for each pair of axiom and c-axiom using McNemar's χ^2 test for paired proportions, finding that the difference is significant for each pair of axioms and c-axioms. Figure 1 displays the summary of the reproduction of Result 1.

Conclusion 1 (on Result 1)**.** *People select canonical axioms more than control axioms. This holds true for all individual pairs of axioms. This result is fully reproduced.*

4.1.2 Result 2: Do violations occur more or less frequently depending on whether people chose or not an axiom?

In the original study, in stage 2, participants make a series of randomized lottery decisions. This stage allows the authors to identify violations of canonical axioms. Among the 17 lottery decisions used to identify axiom violations, four correspond to IIA, four to FOSD, three to TRANS, three to IND, two to CONS, and one to BRANCH. In the dataset analyzed, each observation corresponds to one lottery decision for one axiom made by one subject

(17 observations per subject, yielding a total of 1,870 observations). The authors report the proportion of lottery decisions violating the corresponding axiom among individuals who selected the axiom (30%) and those who did not (24%). They then perform Fisher's exact test to compare these proportions.

We identify two issues with this approach. First, statistical independence is violated because observations are nested at both the axiom and the individual level, a feature that is not accounted for by the authors' statistical test. Second, this method assigns unequal weight to the different axioms. Axioms associated with a larger number of lotteries in the experimental design mechanically contribute more to the reported proportions. For instance, the IIA and FOSD axioms each receive four times more weight than the BRANCH axiom. While one could argue that the importance of each axiom is not necessarily equal, this issue is not discussed in the paper. Moreover, the overall motivation of the study suggests that each axiom is of equal conceptual importance.

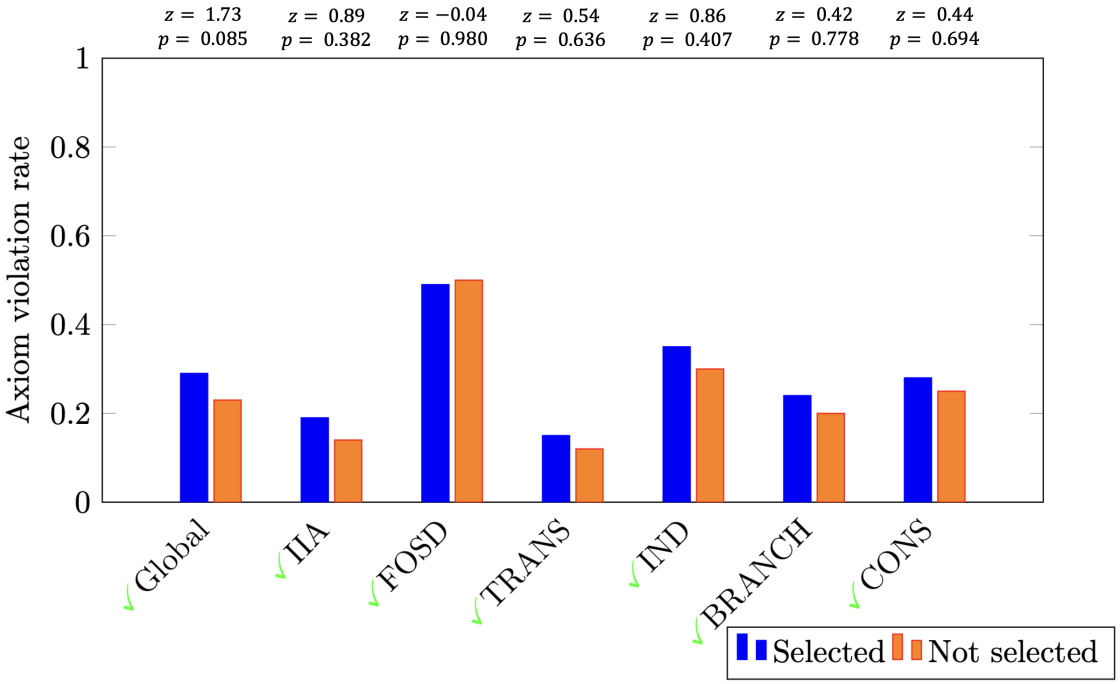
To address these two issues, we propose the following procedure. For each subject, we compute the proportion of violations per axiom. At this stage, the data consists of 6 observations per subject (6 axioms per subject), thus giving equal weight to each axiom. We then run an OLS regression with random effects at the individual level to account for the dependency between observations to explain the proportion of violations.

$$Y_{ik} = \beta_0 + \beta_k X_{ik} + u_i + \varepsilon_{ik} \tag{1}$$

where Y_{ik} is the dependent variable (Proportion of violations) for individual i for axiom k. β_0 is the global intercept, applicable to all observations. X_{ik} represents the dummy variable (Axiom selection) coded 1 if individual i selected axiom k and 0 otherwise. β_1 is the coefficient for the dummy variable (Axiom selection). u_i is the random effect associated with each individual, capturing the unobserved heterogeneity that is constant over axioms for each individual but varies across individuals. ε_{ik} is the error term.

Doing so, we observe no significant difference between selected and non-selected axioms in the proportion of violations ($z = 1.73, p = 0.085$). Table A4 in the online appendix displays the OLS results, which are in line with Result 2 of the original

Figure 2: Summary of Reproduction of Result 2



Note: Proportion of violation of canonical axioms depending on whether they were selected or not selected. At the global level, we run the Random effects Linear regression and report the z and p-value. At the axiom level, we run Wilcoxon rank-sum tests and report the z and exact p-values.

paper.

We also test whether this result is observed at the axiom level. We consistently find no significant difference across the axioms. Figure 2 displays the summary of the reproduction of Result 2.

Conclusion 2 (on Result 2)**.** *Whether people choose or not a canonical axiom, the violation rate is not significantly different. This is true for each individual axiom. We therefore reproduce the result.*

4.1.3 Result 3: When people violate a canonical axioms, do they change more their lottery decision to reconcile with the axioms than with the control axioms?

In the original study, in stage 3, subjects are presented with all inconsistencies between their lottery choices and the rules they previously selected. At this stage, they can either change their lottery choice and/or unselect the rule, or leave both unchanged. The authors use these decisions to test whether subjects are more likely to revise their choices to reconcile them with canonical axioms than with c-axioms.

As for Result 2, the original dataset contains 17 lottery decisions per individual,

for a total of 1,870 observations. For each lottery decision, it is known whether a subject selected an axiom and/or its corresponding c-axiom. For this result, the authors first restrict attention to lottery decisions for which an individual selected at least one of the corresponding axiom–c-axiom pair (1,584 observations for axioms and 201 for c-axioms, for a total of 1,785 observations). From this subsample, they then retain only the lottery decisions that violate the selected rule (468 for axioms and 124 for c-axioms, totaling 592 observations). Finally, they classify a violation as a mistake when the individual either changes the lottery choice to comply with the axiom or unselects the axiom to avoid the violation (277 for axioms and 69 for c-axioms, totaling 346 observations). The main analysis for this result is conducted on this final subsample, which represents 18.5% of the original sample.

From this subsample, they then compare the relative proportions of these two decisions, considering that deciding to change the lottery choice is a reconciliation with the canonical axiom, as opposed to the unselection of the axiom. For the canonical axiom, 218 lottery decisions (79%) out of this sample are corrected to fit the axiom, and the axiom is unselected for 59 lottery decisions. For the control axiom, 25 lottery decisions (36%) out of this sample are corrected to fit the axiom, and the axiom is unselected for 44 lottery decisions. To test the difference between these two proportions, they compute a Wilcoxon rank-sum test and report a statistically significant difference, showing that people, when violating axioms, favor the reconciliation with canonical axioms.

We identify several issues here. Firstly, the subsample selection leads to an overweighting of some individuals – those who selected multiple axioms and committed multiple violations. Secondly, similar to Result 2, the analysis gives extra weight to axioms for which more lotteries were played. Thirdly, and again similar to previous results, the analysis does not account for the dependency of observations across axioms and individuals. Finally, the statistical tests used are inappropriate: the non-parametric test they employed is suitable for continuous dependent variables, whereas they are testing for differences in proportions (binary dependent variable). In this case, a Chi-square or Fisher exact test is more appropriate.

We were not able to fully address the first issue, as by design their analysis focuses on individuals who selected axioms and violated them – thus excluding

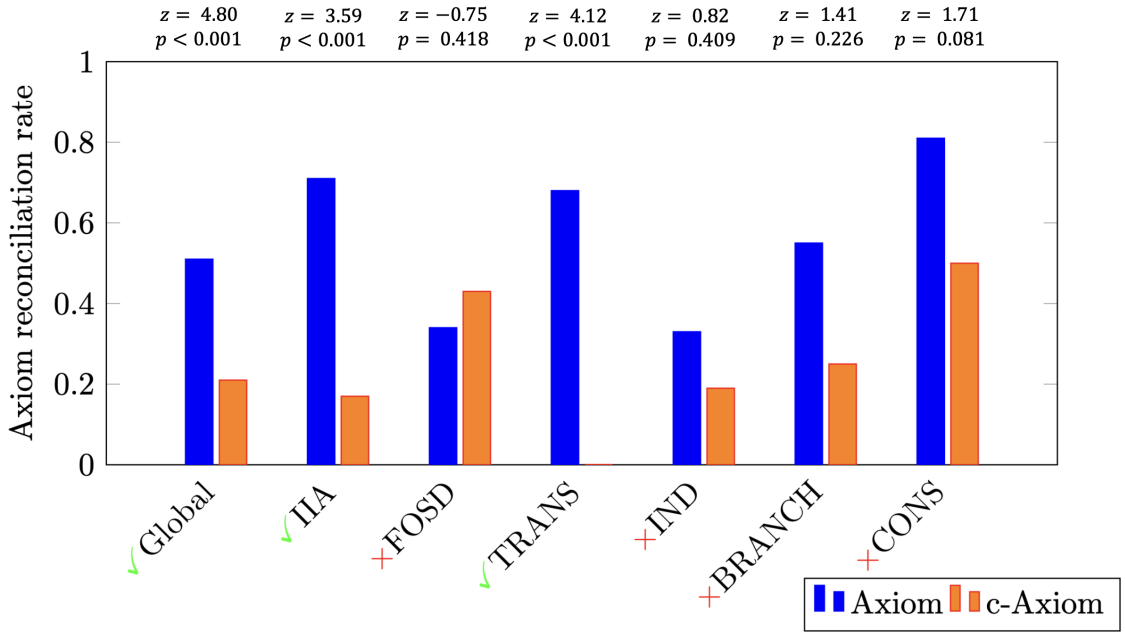
subjects who do not enter this category. However, we try in our method to give equal weight to each individual having selected and initially violated at least one given axiom – independently of the number of those violations. Regarding the weight of different axioms, we are not able to fix the issue in the general test – contrary to the previous test – due to the subsample selection inherent to this result. The issue is, however, addressed when looking at the separate tests of the result on each axiom we present below. Finally, we try to take into account the dependency of individual observations and computed an appropriate statistical test.

For each individual, we proceed as follows. First, we consider that an observation is an individual and axiom/c-axiom pair (12 observations per subject, 1320 observations in total). Second, we keep only the axioms or c-axioms that have been chosen by the subject (628 observations). For each of these, we compute the proportion of initial violations. This corrects the issue of overweighting the axioms with more lotteries. We keep only the axioms or c-axioms that are violated at least once (349 observations). This proportion of violations for each axiom/c-axiom per individual is our first variable (θ^{pre}). We subtract from this variable the proportion of cases in which individuals decided to change their lottery decision to reconcile with the corresponding chosen axiom (or c-axiom). This proportion of violations post-reconciliation is our second variable (θ^{post}). We then compute the proportion of violations that were corrected by changing the lotteries – hereafter denoted as the reconciliation rate – as follows: $Y = (\theta^{pre} - \theta^{post})/\theta^{pre}$. Finally, we compare this reconciliation rate with the same proportion for the c-axioms using an OLS regression with random effects per individual.

$$Y_{ij} = \beta_0 + \beta_1 X_j + u_i + \varepsilon_{ij} \tag{2}$$

where Y_{ij} is the dependent variable (reconciliation rate) for individual i for the canonical axiom or control axiom j. β_0 is the global intercept, applicable to all observations. X_j represents the dummy variable (Axiom type) coded 1 if j is a canonical axiom and 0 if it is a control axiom. β_1 is the coefficient for the dummy variable (Axiom type). u_i is the random effect associated with each individual, capturing the unobserved heterogeneity that is constant over axioms for each individual but varies across individuals. ε_{ij} is the error term.

Figure 3: Summary of Reproduction of Result 3



Note: Proportions of reconciliation with axioms and c-axioms when they have been both selected and violated. At the global level, we run the Random effects Linear regression and report the z and p-value. At the axiom level, we run Wilcoxon rank-sum tests and report the z and exact p-values.

We also test whether this result is observed at the axiom level. At this stage, the data is partly paired. For some individuals, we have observations for axioms and c-axioms (if they selected both and violated both at least once). For the remaining, the data is unpaired. Accordingly, we simply use a Wilcoxon rank-sum test for the difference in the reconciliation rate between canonical and control axioms for each axiom. Results are summarized in Figure 3.

Conclusion 3 (on Result 3)**.** *At a global level, people violating axioms they selected, tend to reconciliate more with the axiom when it is canonical rather than control. However, this result is driven by two out of the six axioms: the Independence of the Irrelevant Alternative (IIA) and the Transitivity (TRANS). We therefore partially reproduce the result.*

While we believe our method is more appropriate and should be used for the analysis of the new data in our replication study, we want to report that we also obtain similar results when using the original method from the paper for each individual axiom. The results are presented in Table A5 in the online appendix. The results are comparable to those of our reproduction for all axioms, except for the

BRANCH axiom. Their method indicates that reconciliation is also significantly more frequent with the BRANCH axiom than with the c-axiom. Thus, the global result relies on three out of the six axioms following their method, and on two following our method.

4.2 Online experiment: Independence axiom

The additional online experiment served several purposes in the study of Nielsen and Rehbeck (2022): testing the relationship between rule preferences and personality and cognitive traits. More relevant to our case, it also serves to replicate the results on a larger sample for a specific axiom. Here, we computationally reproduce the three main results of their online experiment using the same data analysis procedure as described for the lab experiment.

4.2.1 Result 1: Do people select the axiom more than its control axiom? In the original study, the authors report the proportion of subjects selecting the axiom (75%) and the proportion selecting the c-axiom (25%). To compare these two proportions, they use a Wilcoxon signed-rank test and conclude that the difference is highly significant ($p < 0.0001$).

We apply the same procedure as we did for individual axioms in their lab experiment. We follow their procedure, except that we rely on a McNemar's χ^2 test, which is appropriate for the comparison of paired proportions. The result is reported in Figure 4 and is consistent with the original finding.

Conclusion 4 (on online Result 1)**.** *People select the IND canonical axiom more than its control axiom. This result is fully reproduced.*

4.2.2 Result 2: Do violations occur more or less frequently depending on whether people chose or not the IND-axiom? In the original dataset, one observation corresponds to one lottery decision for the IND-axiom for one subject (3 observations per subject). The authors report the proportion of lottery decisions violating the IND-axiom among people who choose the axiom (42%) or not (29%). They then perform a Fisher's exact test on these proportions.

As there is only one axiom under study, there is no weighting issue of different

axioms, in contrast with analyses of the lab experiment. However, as with the lab experiment, there is a lack of statistical independence because observations are nested at both the axiom and individual levels, which is not taken into account by the authors' statistical test.

We again apply the same procedure as we did for individual axioms in their lab experiment, addressing this issue. We compute the frequency of violations of the IND axiom per individual and run a Wilcoxon rank-sum test depending on whether the subjects selected the axiom or not. The result is displayed in Figure 4 and confirms the original result.

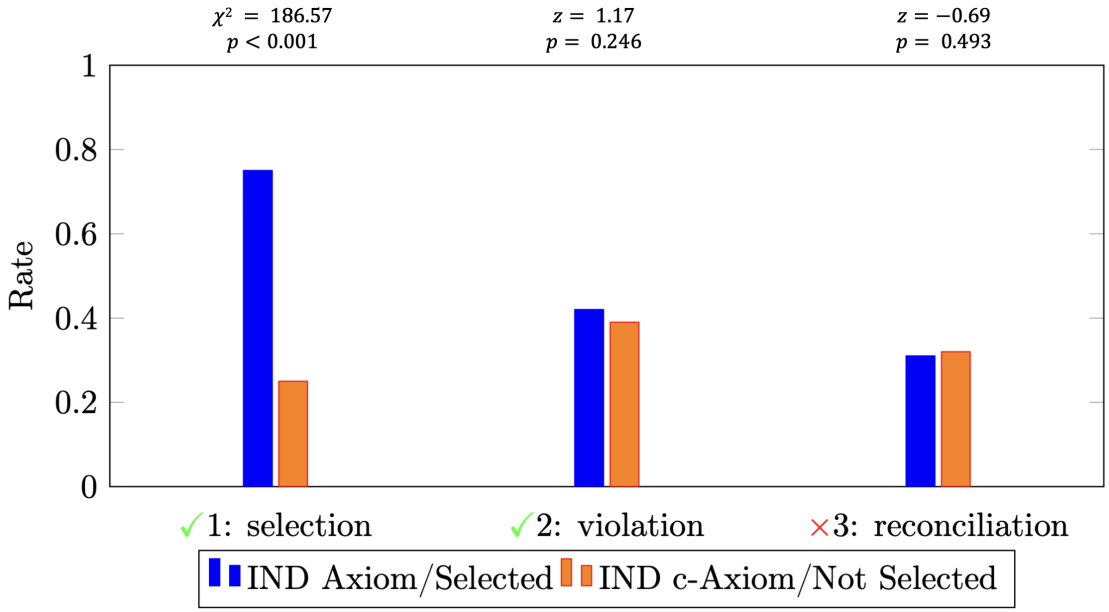
Conclusion 5 (on online Result 2)**.** *Whether people choose or not the IND canonical axiom, the violation rate is not significantly different. We therefore reproduce their result.*

4.2.3 Result 3: When people violate the IND canonical axiom, do they change their lottery decision more to reconcile with the axiom than with the control axiom?

As for the laboratory experiment, the authors consider a dataset of lottery decisions per individual from which they extract individuals who selected the IND axiom or c-axiom and violated it. They compare the relative proportions of four outcomes: change in lottery choice and still inconsistent, consistent change in lotteries, unselection of the axiom, and absence of change. This is slightly different from their procedure in the lab experiment, in which they only consider the mistakes (violations for which an individual decided either to consistently change the lottery choice or unselect the axiom). They find no significant difference in the distribution of these outcomes according to whether the IND axiom is canonical or not, using a Fisher's exact test. For instance, their table indicates that 31% of lottery choices are changed to be consistent for both the IND axiom and c-axiom. Thus, the third result from the lab experiment is not replicated in their online study.

We again apply the same procedure as we did for individual axioms in their lab experiment. We keep only the IND axiom or c-axiom that were chosen and the IND axiom or c-axiom that were violated at least once. We first compute the proportion of violations for the IND axiom/c-axiom per individual as our first variable (θ^{pre}). We subtract from it the proportion of cases in which individuals

Figure 4: Summary of Reproduction of the Online experiment



Note: For the result 1-online, we run McNemar's χ^2 test and report the χ^2 and p-value. For the results 2-online and 3-online, we run Wilcoxon rank-sum tests and report the z and p-values. The blue bar corresponds to the proportion for the IND axiom and the orange bar to the IND c-axiom for the three results.

decided to change their lottery decision to reconcile with the selected axiom (or c-axiom), to get the proportion of violations post-reconciliation (θ^{post}). We then compute the reconciliation rate $Y = (\theta^{\text{pre}} - \theta^{\text{post}})/\theta^{\text{pre}}$. Finally, we compare this reconciliation rate for the IND canonical axiom versus the IND control axiom by running a Wilcoxon rank-sum test. The result is displayed in Figure 4 and confirms their null result in the online setting.

Conclusion 6 (on online Result 3)**.** *For the IND axiom, people violating axioms they selected do not reconcile differently with the canonical and the control axioms. We therefore reproduce the original online study's null result, which contradicts the authors' hypothesis.*

4.3 Reproduction discussion

The results from our reproduction exercises are mixed, with some findings fully replicated and others requiring more nuanced interpretation. The direct reproduction, which strictly follows the original methods and utilizes the same dataset, confirms

that the package provided by the authors is well-constructed and fully reproducible. This initial step ensures that the original results can be accurately reproduced using the provided data and analysis code.

The more insightful and revealing part of our study lies in the robustness reproduction, where we employ methods we deem more appropriate to assess the robustness of the original findings. This approach addresses several key methodological concerns.

First, we observe that the statistical tests used in the original study are not always appropriate, in particular for the comparison of proportions. Second, the original analysis treats observations as independent, whereas our method accounts for the statistical dependency between observations. While our method reveals similar overall results, the statistical power is slightly reduced, indicating that the original method might overestimate the precision of the results. Third, the original study's focus on increasingly specific subsets of data could limit the generalizability of its findings. For instance, the key result concerning errors in choosing the wrong lottery when trying to follow an axiom involved only 15.3% of the sample. This highly specific focus suggests that the original sample size may not be sufficient to generalize the findings broadly.

Our reproduction reaffirms the result that subjects significantly preferred canonical axioms over control axioms, with improved statistical testing reinforcing this finding. This is true both in the lab and online for all axioms. Additionally, the original results suggests no significant difference in violation rates between selected and unselected axioms. Again, this is true both in the lab and online for all axioms.

However, our conclusions are mixed regarding the third result, i.e., that individuals are more likely to change their choices to be consistent with canonical axioms than with control axioms. This is noteworthy given that this result is arguably the main claims of the paper. For instance, the authors argue that their "main interest is in studying how individuals reconcile inconsistent choices" (p.2246) and highlight Result 3 in the first paragraph of their conclusion (p.2263):

> "In directly eliciting preferences over axioms, we find that individuals view them as rules that they want their choices to follow. When lottery choices conflict with stated axiom preferences, individuals often change

their choices to be consistent with the axiom, rather than inferring from their choices that the axiom is not desirable."

Our reproduction partially confirms the finding that participants were more likely to reconcile their choices with canonical axioms than with control axioms. However, this effect is only driven by the IIA and TRANS axioms in the lab. For the other axioms, the tendency to reconcile is not statistically significant, suggesting that the original finding is overstated when applied uniformly to all axioms. Their online study also contradicts this result for the IND axiom.

5 Replication

5.1 Experimental design

While most of the results are reproduced, our reproduction reveals important discrepancies depending on the axioms. Our replication is aimed at testing whether the results that are most successfully reproduced can also be experimentally replicated. More specifically, we choose to focus on one axiom in the same vein as their online experiment. Our computational reproduction indicates that targeting the FOSD, IND, BRANCH, and CONS axioms is not promising, as the main results does not hold consistently for these axioms. However, the results is replicated for the IIA and TRANS axioms.

In deciding for which axiom we want to replicate the study, we aim to minimize issues related to subsample analysis. Therefore, we select the axiom where both the canonical axiom and its control axiom (c-axiom) are frequently chosen and violated, ensuring a sufficient number of individuals had the opportunity to reconcile their axiom choices. The TRANS axiom is unsuitable because its c-axiom is rarely chosen and violated, making observable reconciliation difficult. In contrast, the IIA axiom and its c-axiom are neither trivially chosen nor consistently followed, making IIA a strong candidate for experimental replication. By focusing on the IIA axiom, we aim to confirm with greater power than in their lab study that the original paper's results are replicated online.

5.1.1 Data Collection To conduct this replication, we ran the experiment using the Prolific platform with a US sample of 451 participants with at least high school education (mean age = 31 years, 57% female). The online experiment on Prolific lasted 15 minutes on average. The average payoff was £2.14 per participant plus a participation fee of £1.5. Unlike the original study, which paid 1 out of every 10 subjects (in addition to the participation fee paid to everyone), we paid each participant because the payments for the IIA lotteries have relatively low variance compared to other axioms.

We perform a power analysis specifically for the IIA axiom and estimate that we need approximately 431 subjects (compared to 110 in the original lab study) to achieve 90% power and a 5% significance rate, while accounting for issues of independence.

5.1.2 Procedure: Online experiment The design is similar to the original design of the online experiment of Nielsen and Rehbeck (2022). The three main differences are the axiom studied (IIA instead of IND), some changes in lotteries (detailed and justified in the next section) and the experimental software. We used oTree (**?**) to code the online experiment.

5.1.3 Change in IIA lotteries In the original design of Nielsen and Rehbeck (2022), the four decision problems intended to test the IIA axiom each involve a choice between a triple of lotteries $\{A, B, C\}$, combined with a single corresponding binary choice $\{A, B\}$. A violation of IIA is identified when a participant selects A over B in one context but B over A in the other. Importantly, while each $\{A, B\}$ choice is presented only once, the same lotteries A and B appear repeatedly across the four different $\{A, B, C\}$ choice sets.

We modify the lottery values in three out of the four IIA choice pairs, affecting six out of eight individual lottery options (see the lotteries used in the online appendix H). This modification is motivated by two design considerations specific to our experimental context.

First, our preregistered online replication focuses exclusively on the IIA axiom and involves a substantially smaller total number of lottery choices than the original experiment, which jointly tested multiple axioms. In this setting, repeatedly

embedding the same lotteries A and B within different $\{A, B, C\}$ choice sets risks making the structure of the IIA tests overly salient to participants. This could encourage pattern recognition or strategic responding unrelated to genuine normative reflection. In contrast, in the original experiment, IIA-related choices are interspersed among a much larger and more diverse set of lotteries testing multiple axioms, making such repetition less conspicuous.

Second, and more importantly, the repeated use of the same lotteries across multiple IIA tests complicates the interpretation of inconsistencies during the reconciliation stage. When a participant's single $\{A, B\}$ choice is compared to several distinct $\{A, B, C\}$ choices involving the same lotteries, multiple inconsistencies can arise simultaneously. In such cases, it becomes ambiguous which specific decision should be revised to restore consistency with the axiom. For example, an inconsistency could be resolved either by revising the binary choice or by revising one or several of the ternary choices, leading to multiple plausible reconciliation paths. This ambiguity makes it more difficult to interpret choice revisions as clear evidence of perceived mistakes. By introducing slight variations in the lotteries across IIA choice pairs, we ensure that each potential violation maps onto a unique reconciliation problem.

5.1.4 Power Analysis To conduct our power analysis, we follow the recommendations from the I4R. Our primary goal is to achieve at least 90% power for a 5% significance level by considering two-thirds of the original effect size. This approach is intended to mitigate any issues related to effect size inflation. We calibrate our power analysis specifically for the IIA axiom based on effect size and proportions derived from the lab experiment.

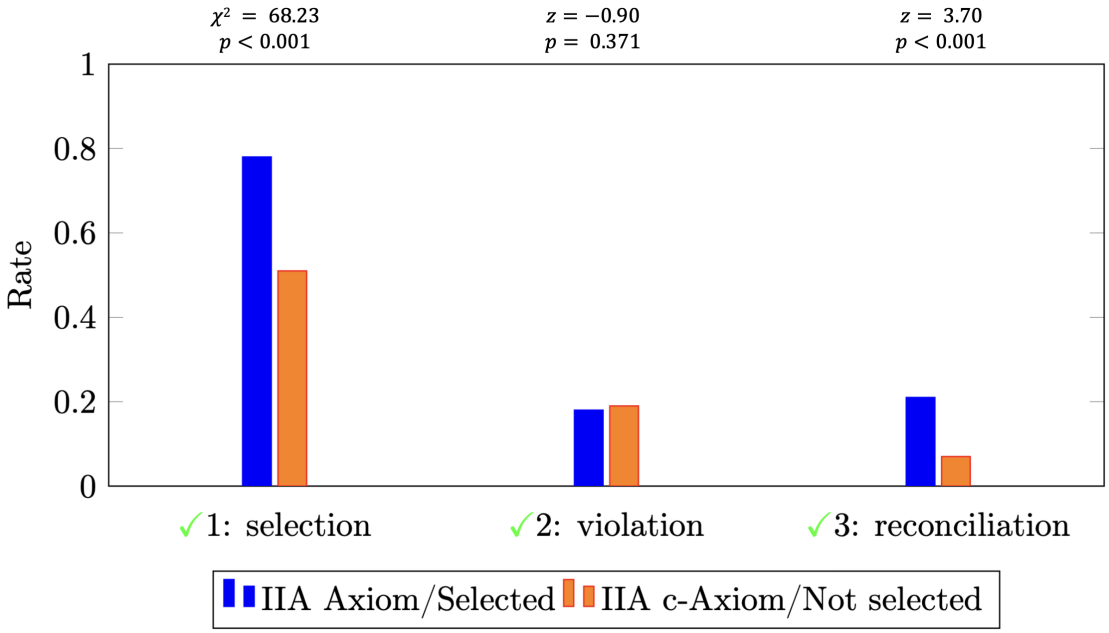
Additionally, we recognize that the sample size needed for Results 1 and 2 is lower than for Result 3. Therefore, we base our power calculation on the test that we plan to use to test Result 3. As a reminder, for Result 3, we rely on a subset of the subjects for the IIA canonical axiom and for its control axiom. For each type of axiom (canonical and control), our test is based only on subjects who selected the axiom and violated it at least once. In the original lab data, this lead to a subset of $n^a = 35$ individuals for the IIA canonical axiom and $n^c = 13$ individuals for the IIA

control axiom. For our power analysis, we assume that we will obtain a similar ratio between the two groups $\kappa = n^a/n^c = 35/13 = 2.692$. We plan on a similar attrition rate due to subsample selection: we denote δ the share of the initial sample finally present in the first subset n^a: $\delta = n^a/N = 35/110 = 0.318$.

We estimate the sample size needed for the subset n^a. We rely on the means and standard deviations of the reconciliation rate we derive from the lab data for the IIA canonical axiom $\mu^a = 0.7095$, $\sigma^a = 0.4453$ and the IIA control axiom $\mu^c = 0.1667$, $\sigma^c = 0.3333$. We reduce the difference $\mu^a - \mu^c$ by a third in order to follow the recommendations from the I4R. We then reduce this difference again to account for the increase in noise resulting from online data collection. To estimate noise, we rely on the discrepancy in the differences reported between their first result online and in the lab for the IND axiom. Specifically, they find that 83% chose the IND axiom and 14% the c-IND axiom in the lab, while 75% chose the IND axiom online and 25% the c-IND axiom online. The difference is thus 27.54% smaller online compared to the lab. We use this as an estimate of the reduction in effect size we should observe from replicating the lab result online. In sum, we reduce by 1/3 and then 27.54% the expected difference in means between IIA and c-IIA. More specifically, we assume $\bar{\mu^a} = \mu^a - (1/3) \times (\mu^a - \mu^c) - 0.2754 \times [\mu^a - (1/3) \times (\mu^a - \mu^c) - \mu^c]$. We assume $\bar{\mu^c} = \mu^c$, $\bar{\sigma^a} = \sigma^a$, $\bar{\sigma^c} = \sigma^c$.

We compute simulations with these parameters to assess the n^a needed to achieve 90% power when testing a difference between the reconciliation rate for IIA and c-IIA using a Wilcoxon rank-sum test. We then simply calculate the total sample size N required, given that $N = (n^a + n^c)/\delta$ and $n^c = n^a/\kappa$. Figure A1 in the online appendix presents the power achieved depending on the total sample size. We obtain $n^a = 90$, resulting in a total sample size of $N = 388$. In order to reach this sample size on Prolific while taking into account the fact that some participants leave the experiment before its end (we used 10% as a benchmark), we planned on recruiting at least 431 subjects.

Figure 5: Summary of Replication on the IIA axiom



Note: For the result 1-online, we run McNemar's χ^2 test and report the χ^2 and p-value. For the results 2-online and 3-online, we run Wilcoxon rank-sum tests and report the z and p-values. The blue bar corresponds to the proportion for the IIA axiom and the orange bar to the IIA c-axiom for the three results.

5.2 Experimental Results

Our preregistered online replication took place in March 2025, with a total sample of size N = 451. The statistical testing procedure follows the same methodology as our robustness reproduction of the Online experiment, addressing the concerns described in the reproduction section.

5.2.1 Result 1 We first test whether participants prefer the canonical version of the IIA axiom over its control. We ask participants to make binary choices for both the IIA axiom and its control counterpart, alongside several filler axioms used as decoys. Using McNemar's χ^2 exact test to compare paired proportions, we find that 78% of participants selected the canonical IIA axiom, while 51% selected the control ($\chi^2 = 68.23$, $p < 0.001$; Figure 5). This replicates the original finding with high statistical significance.

5.2.2 Result 2 We then test whether an individual selected or not the IIA axiom changes its violation rate. As preregistered, we compute the frequency of violations

of the IND axiom per individual and run a Wilcoxon rank-sum test depending on whether the subjects selected the axiom or not. The result is displayed in Figure 5 and confirms the original result. The violations rate of the IIA axiom does not significantly differ depending on whether the subject selected the axiom (18%) or not (19%). This supports the original paper's conclusion for Result 2.

5.2.3 Result 3 Finally, we test whether participants who initially selected and violated the IIA axiom are more likely to revise their lottery choices to align with the axiom compared to those who selected and violated its control. As preregistered, we compute the reconciliation rate for people who selected and violated at least one the IIA (and c-IIA) axiom. Among participants who selected and violated the IIA axiom at least once, the individual reconciliation rate is 21%, while is only 7% for individuals who selected and violated the control axiom at least once. A Wilcoxon rank-sum test confirms that this difference in individual reconciliation rate is statistically significant ($z = 3.70$, $p < 0.001$; Figure 5). This result replicates the key finding that individuals are more likely to correct violations of the canonical axiom than of the control axiom.

Comparison of Effect Sizes

Are the effect sizes of our replication comparable to those of the original study? For Result 1, we find a difference in the proportion of subjects selecting the IIA and c-IIA axioms of $0.78 - 0.51 = 0.27$. For a standardized comparison, we compute Cohen's $h \approx 0.56$, as presented in the online appendix I, suggesting a medium-sized effect. In their lab experiment, this difference was much larger: $0.83 - 0.12 = 0.71$ ($h \approx 1.62$).

For Result 2, the magnitude of the effect is not relevant, as the result indicates an absence of effect.

For Result 3, we compute Cohen's d using the reconciliation rates and pooled standard deviations for the canonical and control axiom groups. For transparency, we also compute Cohen's h using the authors' original statistical method. This measure, also presented in the online appendix I, incorporates the robustness corrections applied in our reproduction. In the original study, the effect size for the

IIA axiom was large ($d \approx 1.30$), indicating a strong difference in reconciliation rates between the canonical and control conditions. In our replication, the effect remains statistically significant but is notably smaller ($d \approx 0.50$), suggesting a medium-sized effect.

The smaller effect sizes in our replication may be partly due to the fact that our experiment is conducted online. As noted by Nielsen and Rehbeck (2022), online samples are more prone to inattention and noise; they also observe an attenuation of effects in their own online module (see Section V of their paper). However, the attenuation of Result 1 from their lab to online experiment for the IND axiom was smaller, from $h \approx 1.52$ to $h \approx 1.05$. Thus, the effect size for Result 1 is more reduced in our online replication of IIA than it is for their online replication of IND. Taken together, the evidence suggests that while the IIA effect is robust, its magnitude may be context-sensitive or somewhat inflated in the original lab setting.

5.2.4 Authors' statistical method. For the sake of comprehensiveness, we also conduct the analysis using the authors' original statistical approach. The results are consistent with those obtained using our preregistered robustness-corrected method and confirm the main findings (see Table A6).

6 General Discussion

Our study provides a multi-faceted assessment of Nielsen and Rehbeck (2022)'s influential claim that individuals treat violations of rational choice axioms as mistakes. We conducted three complementary exercises: a direct computation reproduction, a robustness computational reproduction, and a preregistered conceptual experimental replication.

The direct reproduction confirms the full reproducibility of the original results with the authors' data and code. The robustness reproduction, using more appropriate statistical methods, confirms the first two results: participants tend to prefer canonical axioms over control axioms, and violations occur at similar rates regardless of initial axiom selection. However, the third and most central result, participants correcting their violations to align with chosen axioms, receives only

partial support: the reconciliation effect holds robustly for only two out of six axioms, TRANS and IIA. Finally, our high-powered online replication focusing on the IIA axiom confirms all three of the original study's results, albeit with a smaller effect size.

These findings offer a more nuanced interpretation of the original conclusions. The results suggest that people perceive their violations as mistakes only for two axioms: Transitivity and Independence of Irrelevant Alternatives. The former has received substantial attention in the literature (May, 1954; Tversky, 1969; Loomes et al., 1991; Birnbaum and Schmidt, 2008; Regenwetter et al., 2011). Prior to Nielsen and Rehbeck (2022)'s study, early experimental work had already suggested that individuals find the Transitivity axiom normatively compelling and want to correct violations when prompted (MacCrimmon and Larsson, 1979; MacCrimmon, 1968). Our replication adds novel evidence for the IIA axiom, indicating that individuals also regard it as normatively relevant and are more likely to revise their choices when violating it.

In contrast, our analyses do not support the idea that violations of the four other axioms (FOSD, CONS, IND, and BRANCH) are perceived as mistakes. This conclusion is particularly informative for the IND axiom, which has been extensively studied through the Allais paradox. In a classic study, Slovic and Tversky (1974) found that violations of IND persist even when decision makers are presented with explicit normative arguments in its favor, suggesting that such violations do not stem from mistakes. This result has recently been replicated and extended in two high-powered studies by Humphrey and Kruse (2024) and Humphrey and Kruse (2025). These authors further discuss how differences in experimental design may account for the apparent discrepancy between their findings and those of Nielsen and Rehbeck (2022). Our reproduction results help reconcile these seemingly conflicting conclusions. We find that violations of IND are in fact not robustly perceived as mistakes in either laboratory or online samples of Nielsen and Rehbeck (2022). Rather than contradicting the earlier literature, this pattern aligns closely with the evidence from the Allais paradox tradition: for this axiom, violations tend to persist and are rarely corrected, even when individuals are made aware of the inconsistency. Taken together, these findings suggest that violations of Independence are better

understood as reflecting a normative rejection of the axiom, whereas violations of other axioms—such as Transitivity and IIA—are more likely to be experienced by decision makers as genuine mistakes.

This raises a key question for decision theory and behavioral economics: why are some axioms (such as IIA and TRANS) treated by individuals as more normatively binding than others?

A possible explanation for these findings relates to the cognitive complexity of the axioms and the corresponding lottery decisions (Oprea, 2020, 2024; Enke and Graeber, 2023; Oberholzer et al., 2024). Nielsen and Rehbeck (2022) mention this aspect: "some features of axioms might be more compelling to individuals, or alternatively certain aspects of a rule might be particularly complex." Intuitively, the more cognitively demanding an axiom is, the less likely people should violate it. In addition, the more cognitively demanding an axiom is the less likely individuals may be able to recognize violations as mistakes – even when explicitly reminded of the rule they previously endorsed. Among the six axioms, FOSD, IND, and BRANCH arguably involve the most complex reasoning, as they all require evaluating compound lotteries in contrast with IIA, TRANS and CONS. Evidence indicates that people are less able to compute expected values for compound compared to simple lotteries (Enke and Shubatt, 2023) and that people with higher cognitive abilities are more able to reduce compound lotteries (Prokosheva, 2016). Even if these axioms are appealing in principle, people may struggle to grasp their implications. As a result, they may be less inclined to revise violations, even when confronted with the normative rule they themselves selected.

The results regarding reconciliation patterns are consistent with this explanation for five out of the six axioms. Our replication results are positive for IIA and TRANS – axioms involving simple lotteries – and null for FOSD, BRANCH, and IND – axioms involving compound lotteries. The results do not align with this interpretation for CONS, which is arguably the simplest axiom, and yet the reconciliation rate is not significantly higher for participants who selected the axiom compared to those who selected the c-axiom. However, very few participants (5%) selected the CONS c-axiom, making it difficult to reject the null. This suggests that this result warrants further investigation with a larger sample size.

As mentioned, if this interpretation is correct, the violation rate should also be lower for simpler axioms compared to more complex ones. This is indeed the case for IIA and TRANS – which exhibit the lowest violation rates – compared to FOSD and IND, which show the highest. However, CONS has a higher violation rate than BRANCH, which does not align with this interpretation. Nonetheless, caution is warranted when interpreting the violation rate for the BRANCH axiom, as participants were exposed to only one lottery testing this axiom, compared to three or four lotteries for most others (and two for CONS). This limited exposure yields an estimate based on a sample likely too small for meaningful inference.

In short, our findings are compatible with the idea that for relatively simple axioms (such as TRANS, IIA and possibly CONS), people commit fewer errors and are able to recognize and correct them; and for more complex axioms (like FOSD, IND and BRANCH), individuals commit more errors and generally fail to identify them as mistakes, even when explicitly confronted with their own prior endorsement of the rule.

Normatively, this would imply that interventions designed to help individuals align their choices with their normative commitments – for instance, through deliberation or feedback – may only be effective for rules that are cognitively accessible. For more complex axioms, such interventions may prove insufficient, as individuals may fail to recognize the violation even when prompted. This would point to a limitation in efforts to "de-bias" choice behavior: some errors may not be not merely inattentive, but structurally resistant to introspective correction.

This hypothesis deserves further experimental investigation. It is debatable whether violation and reconciliation rates are truly comparable across axioms, as the structure of the lotteries — including their specific probabilities and payoffs — likely plays a significant role in shaping observed behavior. For example, violations of transitivity are rare in the TRANS1 and TRANS2 lotteries used by Nielsen and Rehbeck (2022), probably because one option clearly dominates the others in terms of expected value and downside risk. This clarity limits the scope for preference-driven choices, making any observed violations more likely due to inattention — and thus more easily recognized and corrected upon reflection. By contrast, this clarity is lacking in TRANS3 and in the CONS1 and CONS2 lotteries, where options tend

to be closer in expected value and more divergent in risk profile. In such cases, choices are more susceptible to individual preferences, such as risk attitudes or a desire for diversification, making violations more likely and less easily perceived as errors. These design features may play a significant role in driving observed violation and reconciliation rates. Future research should therefore take care in designing axiom tests that allow for meaningful comparison across axioms, ensuring that variation in results reflects cognitive demands linked to each axiom rather than incidental features of the lottery structure.

References

Altonji, J., G. Imbens, K. Lang, E. Luttmer, I. Rasul, S. Stantcheva, and R. Wacziarg (2025). Report on improving the publication process in economics. Technical report, Ad-hoc Joint AEA-EEA-ES-RES Committee.

Bayle, G., D. Dubois, and M. Willinger (2026). L'économie à l'ère de la science ouverte: un nouvel élan pour la reproductibilité. *Forthcoming in Revue Economique*.

Birnbaum, M. H. and U. Schmidt (2008). An experimental investigation of violations of transitivity in choice under uncertainty. *Journal of Risk and Uncertainty 37*, 77–91.

Borch, K. (1968). The allais paradox: A comment. *Behavioral Science 13*(6), 488–489.

Camerer, C. (1995). Individual decision making. *The handbook of experimental economics 1*, 587–704.

Dreber, A. and M. Johannesson (2019). Statistical significance and the replication crisis in the social sciences. In *Oxford research encyclopedia of economics and finance*.

Enke, B. and T. Graeber (2023). Cognitive uncertainty. *The Quarterly Journal of Economics 138*(4), 2021–2067.

Enke, B. and C. Shubatt (2023). Quantifying lottery choice complexity. Technical report, NBER Working Paper 31677.

Huber, J., J. W. Payne, and C. Puto (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of consumer research 9*(1), 90–98.

Humphrey, S. J. and N.-Y. Kruse (2024). Who accepts savage's axiom now? *Theory and Decision 96*(1), 1–17.

Humphrey, S. J. and N.-Y. Kruse (2025). The normative force of savage's axiom and the allais paradoxes: more evidence. *Theory and Decision*, 1–25.

Isager, P. M., R. C. M. Van Aert, Bahník, M. J. Brandt, K. A. DeSoto, R. Giner-Sorolla, J. I. Krueger, M. Perugini, I. Ropovik, A. E. Van 'T Veer, M. Vranka, and D. Lakens (2023, April). Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints. *Psychological Methods 28*(2), 438–451.

Li, S. (1996). An additional violation of transitivity and independence between alternatives. *Journal of Economic Psychology 17*(5), 645–650.

Loomes, G., C. Starmer, and R. Sugden (1991). Observing violations of transitivity by experimental methods. *Econometrica: Journal of the Econometric Society*, 425–439.

MacCrimmon, K. R. (1968). Descriptive and normative implications of the decision-theory postulates. In *Risk and uncertainty: Proceedings of a conference held by the International Economic Association*, pp. 3–32. Springer.

MacCrimmon, K. R. and S. Larsson (1979). Utility theory: Axioms versus 'paradoxes'. In *Expected utility hypotheses and the allais paradox: Contemporary discussions of the decisions under uncertainty with allais' rejoinder*, pp. 333–409. Springer.

May, K. O. (1954). Intransitivity, utility, and the aggregation of preference patterns. *Econometrica: Journal of the Econometric Society*, 1–13.

Nielsen, K. and J. Rehbeck (2022, July). When choices are mistakes. *American Economic Review 112*(7), 2237–68.

Oberholzer, Y., S. Olschewski, and B. Scheibehenne (2024). Complexity aversion in risky choices and valuations: Moderators and possible causes. *Journal of Economic Psychology 100*, 102681.

Oprea, R. (2020). What makes a rule complex? *American economic review 110*(12), 3913–3951.

Oprea, R. (2024). Decisions under risk are decisions under complexity. *American Economic Review 114*(12), 3789–3811.

Prokosheva, S. (2016). Comparing decisions under compound risk and ambiguity: The importance of cognitive skills. *Journal of Behavioral and Experimental Economics 64*, 94–105.

Raiffa, H. (1961). Risk, ambiguity, and the savage axioms: comment. *The Quarterly Journal of Economics 75*(4), 690–694.

Regenwetter, M., J. Dana, and C. P. Davis-Stober (2011). Transitivity of preferences. *Psychological review 118*(1), 42.

Slovic, P. and A. Tversky (1974). Who accepts savage's axiom? *Behavioral science 19*(6), 368–373.

Tversky, A. (1969). Intransitivity of preferences. *Psychological review 76*(1), 31.

CEE-M Working Papers[1] - 2026

WP 2026-01	**Gabriel Bayle, Nicolas Quérou, Mickael Beaud, Dimitri Dubois, Alexis Lefebvre1 & Marc Willinger1** « Spatial externalities in renewable resource management: Experimental evidence on fragmented property rights »
WP 2026-02	**Gabriel Bayle, Violette Pinçon**, Gladys Barragan-Jason, **C´ecile Bazart**, **Lisette Ibanez**, **S´ebastien Roussel**, Arielle Syssau-Vaccarella, **Dimitri Dubois** & **Marc Willinger** « Intragenerational conflict undermines cooperation with the future »
WP 2026-03	Gabriel Bayle, **Dimitri Dubois** & Simon Varaine « Reconsidering mistakes: reproduction and replication of Nielsen and Rehbeck (2022) »

[1] **CEE-M Working Papers / Contact** : laurent.garnier@inrae.fr

- RePEc https://ideas.repec.org/s/hal/wpceem.html
- HAL https://halshs.archives-ouvertes.fr/CEE-M-WP/