Quantitative Economics 13 (2022), 1297–1318 1759-7331/20221297

Choice, deferral, and consistency

Miguel A. Costa-Gomes
School of Economics & Finance, University of St. Andrews

Carlos Cueva
Fundamentos del Análisis Económico (FAE), Universidad de Alicante

Georgios Gerasimou
School of Economics & Finance, University of St. Andrews

Matúš Tejiščák
Chordify

We report on two novel choice experiments with real goods where subjects in one treatment are forced to choose, as is the norm in economic experiments, while in the other they are not but can instead incur a small cost to defer choice. Using a variety of measures, we find that the active choices (i.e., those that exclude the deferral outside option) of subjects in the nonforced-choice treatment are generally more consistent. We also find that the combined deferral and active-choice behavior of subjects in that treatment is explained better by a model of dominant choice with incomplete preferences than it is by rational choice. Our results suggest that nonforced-choice experiments and models can be helpful in separating people's rational, hesitant/not-yet-rational and genuinely irrational behavior, and can potentially offer important new insights in revealed preference analysis.

Keywords. Choice deferral, active choices, choice consistency, revealed preferences, decision difficulty, experiments.

JEL classification. C91, D01, D03, D11, D12.

Miguel A. Costa-Gomes: mcg5@st-andrews.ac.uk
Carlos Cueva: carlos.cueva@ua.es
Georgios Gerasimou: gg26@st-andrews.ac.uk
Matúš Tejiščák: ziman@functor.sk

This is a significantly revised version of a working paper with the same title that was originally circulated in December 2014. We thank many anonymous referees, Syngjoo Choi, Eric Danan, Itzhak Gilboa, Jan Heufer, Larry Samuelson, Adam Sanjurjo, and audiences at the Stanford Institute for Theoretical Economics (2019), Risk, Uncertainty, and Decision (2017), Asian Meetings of the Econometric Society (2017), Annual Congress of the European Economic Association (2017), Workshop on Behavioural Game Theory (2016), World Congress of the Econometric Society (2015), Bounded Rationality in Choice (2014), Foundations of Utility and Risk (2014), Behaviour, Incentives and Contracts (2014), UEA, Georgetown, Johns Hopkins, William & Mary, Durham, Bath, WZB Berlin, Cyprus, Technion, Alicante, Aberdeen and St. Andrews for helpful feedback. Gerasimou (the corresponding author) and Costa-Gomes gratefully acknowledge financial support from the British Academy (2012–13 SRG). Cueva acknowledges support from the Spanish Ministry of Science and Innovation (Grant PID2019-108193GB-I00) and the Generalitat Valenciana (Grant SEJI/2019/005). Any errors are our own. Authors are ordered alphabetically.

© 2022 The Authors. Licensed under the Creative Commons Attribution-NonCommercial License 4.0. Available at http://qeconomics.org. https://doi.org/10.3982/QE1806

1. Introduction

Experiments and surveys on individual decision making typically require people to choose a market alternative from those that the researcher makes available to them. A possibility that arises when choice is forced in this way is that people may be asked to "actively" choose an alternative in situations where they would have otherwise opted for the "choice deferral" outside option instead (Tversky and Shafir (1992), Anderson (2003), Bhatia and Mullett (2016)). For example, Shafir, Simonson, and Tversky (1993) illustrated such experimental findings in psychology with the following story:

> *"At the bookstore, [Thomas Schelling] was presented with two attractive encyclopedias and, finding it difficult to choose between the two, ended up buying neither—this, despite the fact that had only one encyclopedia been available he would have happily bought it."*

When an individual is forced to choose in a situation like this she may do so in a way that is incompatible with utility maximization. Indeed, as pointed out by Luce and Raiffa (1957), *"intransitivities often occur when a subject forces choices between inherently incomparable alternatives."* This argument naturally leads to the following question:

> Do people make more consistent choices when they are not forced to choose?

If a decision maker is rational in the sense that her behavior is consistent with the maximization of a stable, complete, and transitive preference relation, then she will always choose a most preferred feasible alternative from every menu. Considering, however, the large body of work which shows that people often do not behave in such a utility-maximizing fashion, the above question becomes pertinent. Yet despite its methodological, theoretical, and empirical significance, no experimental/empirical study that we are aware of has investigated it before.

In this paper, we raise and attempt to answer this question for the first time. In summary:

1. We report on two lab experiments that implemented a new between-subjects binary-treatment design that elicited forced active choices (control treatment) and deferral-permissive/nonforced active choices (target treatment) from menus of up to 5 real goods.

2. Using several consistency metrics, we find that nonforced choice subjects were generally more likely to be consistent and their active choices were closer to being rational compared to forced-choice subjects. In particular, we find that (not) allowing subjects to defer may affect their choice reversals and general active-choice consistency in environments of riskless choice over real goods. Moreover, some evidence in favor of introspective preference learning (i.e., learning without new exogenous information) for forced-choice subjects further suggests that the active-choice consistency gap between treatments is decreasing over the course of the experiment.

3. The primary source of deviation from utility maximization for most deferring nonforced-choice subjects is their occasionally hesitant behavior that manifests

itself in deferrals, and not the potential inconsistency of their active choices. In addition, most of these subjects' "not-yet-rational" behavior is better explained by a model of dominant choice with transitive but incomplete preferences (Gerasimou (2018), Section 2) than it is by rational choice. This model portrays a decision maker who is occasionally unable to compare some alternatives as making an active choice at a menu if there is a most preferred alternative at that menu and as deferring otherwise. It therefore predicts fully consistent active choices.

4. Analyzing additional data from an indecisiveness personality questionnaire (Germeijs and De Boeck (2002)) we also find that:

   (a) Within the nonforced-choice treatment, subjects who occasionally deferred had higher indecisive-personality scores than those who did not.

   (b) Indecisive forced-choice subjects were more likely to make inconsistent choices than decisive ones.

Our primary behavioral findings suggest that nonforced-choice economic experiments and models can complement the standard forced-choice ones in a fruitful way. In particular, our analysis shows that theory-guided revealed-preference analyses that are appropriate for nonforced-choice data can be helpful in distinguishing between rational, hesitant, and genuinely inconsistent behaviors. Our findings also have implications for experimental as well as survey design. More specifically, by not offering respondents the possibility to opt for an *"I don't know"* option, many experiments and surveys do not currently allow respondents to express their potential reservations in some questions they are presented with. Our results suggest that a more accurate understanding of respondents' preferences, attitudes, or views could be achieved when such a possibility is available to them. In this case, additional reliable information about the respondents' preferences can also be obtained when they are allowed to return to some or all of these questions once they have thought more about the possible active choices, a dimension which we also explore in our study.

The remaining part of the paper is organized as follows. Section 2 describes the experimental design and its implementations. Section 3 provides some theoretical background on rational and active-choice consistent decision making under forced- and nonforced-choice environments. Section 4 presents our main results on differences between the forced-choice and nonforced-choice treatments. Section 5 discusses two potential drivers of these differences in behavior: indecisiveness (incomplete preferences) and learning through introspection. The Appendix in the Online Supplementary Material (Costa-Gomes, Cueva, Gerasimou, and Tejiščák (2022)) contains additional material that also includes details and findings from a third experiment that compared nonforced with forced decision making when the choice alternatives are money lotteries.

2. Experimental design

2.1 *An approach to eliciting incomplete preferences*

We conducted two between-subjects choice experiments with real goods at the University of St. Andrews Experimental Economics Lab in September 2013–February 2014

(henceforth Exp1) and again in September–October 2018 (Exp2). We recruited subjects using ORSEE (Greiner (2015)). In both experiments, we used a between-subjects design with two treatments: Forced-Choice (FC) and Nonforced-Choice (NFC).

In the main phase of the experiment, all subjects were presented with a sequence of menus that were generated from a set of five headsets. Their brand names and models were chosen so that the products' prices were similar but their attributes differed in ways that made comparisons between them nontrivial. For instance, some headsets were basic but with well-known brand names, whereas others were more sophisticated or had some superior or distinctive features but were associated with less recognizable brand names (e.g., the headset with the less commonly known brand name was wireless whereas all others were not). In order to make the decision problems as realistic as possible, the short description of each headset's main features reproduced exactly the same information (in bullet-point form) that the large online retailer from which the headsets were purchased had chosen to provide on the relevant product's web page.[1]

The order in which menus appeared was random and varied across sessions but was fixed within sessions. Each item appeared top-left, middle, top-right, etc. in an even manner across menus. Subjects in the FC treatment were asked to choose an item from all menus, without being able to defer choice. Subjects in the NFC treatment had the opportunity to choose one item or to select *"I'm not choosing now"* in each menu. No subject in either treatment could go back to review and change their decision after they were past a menu during this "main phase" of the experiment.

Once all subjects in both treatments completed this task they proceeded to the "final phase" of the experiment, where one menu was randomly selected separately for each of them. Subjects were then reminded of the decision they had made at that menu in the "main phase" and were asked to make an active choice at that menu. Subjects knew from the beginning that 1 out of every 4 of them (this feature of the design was entirely driven by the experimenters' budget constraints) would be randomly selected to win the item of their *final* choice from their randomly selected menu at the end of the experiment. We refer to such subjects as *"winners."* Participants also knew from the beginning that winners might face some costs which would be deducted from their initial monetary endowment, I. These costs depended on the subjects' decisions at the randomly selected menu in the "main" and "final" phase of the experiment.

In particular, if a subject that was later drawn as winner had decided in the "final phase" to choose an option other than the one she selected from this menu in the main phase, an amount $c_r < I$ was taken away from her initially allocated I. In contrast, there was no deduction if the subject opted for the same headset when she chose from that menu in the final phase. These features were shared by both the FC and NFC treatments. Subjects in the NFC treatment who deferred at their randomly selected menu in the main phase and were later drawn as winners incurred the cost of a $c_d < c_r$ deduction from the initial allocation of I. Finally, participants were told from the beginning that

[1]Subjects could not access the internet during the experiment and, therefore, could not read product reviews either through the PCs they were using or through their smartphones, as using the latter was not allowed.

if they were *not* selected to win a headset they would receive their endowment, I, irrespective of their main- and final-phase decisions at the randomly selected menu.

The FC treatment follows procedures used in standard forced-choice experiments. On the other hand, the NFC treatment is structured in a way that parallels real-world situations where delaying an active choice is costly but changing such a choice is even costlier, for example, the decision to delay or buy immediately from a store with a restrictive returns policy.

All 26 menus with 2 to 5 headsets were shown to subjects in both experiments, using the same z-Tree experimental interface (Fischbacher (2007)). Singleton menus were also included in Exp1, even though they involved no meaningful choice (note that deferral is a dominated decision in such menus). Those menus were not shown to subjects in Exp2.

Exp1 subjects in both treatments were told that they would be allowed to try out the headphones in the randomly selected menu before making their second and final choice from that menu. It is possible, therefore, that the decision to opt for costly deferral in Exp1 could be interpreted as the rational decision to buy more information before making an active choice. We stress, however, that we did not design the treatment with that in mind, and no additional information was given to subjects in writing or in some other objective way.[2] By contrast, our motivation in allowing subjects to try out the headphones was based on intuition and experimental evidence suggesting that inadequate information is a potential cause of limited comparability and deferral-seeking behavior. Sen (1997), for example, argued that preference incompleteness *"can arise from limited information, or from 'unresolved' value conflicts,"* and referred to the former and latter kinds as *"tentative"* and *"assertive"* incompleteness, respectively. In the personality psychology literature, moreover, indecisiveness has been approached as a phenomenon that results from decision impediments such as lack of information, valuation difficulty, and outcome uncertainty. Among other things, this literature finds that individuals with high scores in indecisiveness personality questionnaires are significantly more likely to delay making a decision and to seek additional information.[3] We therefore hypothesized that decision makers with (possibly tentatively) incomplete preferences might be willing to acquire costly additional information about the available options before making an active choice, although we acknowledge that identifying such preferences from deferring behavior alone is not possible.

Subjects in Exp2 on the other hand could not inspect the headphones in their randomly selected menu between their first and final choice, and neither did they receive any additional information. Deferring in this case can in principle be driven by an urge to avoid the task of making a preference-guided active choice when the menu in question generates decision difficulty. We set $c_r = £4$, $c_d = £1$, $I = £7$ in Exp1 and $c_r = £6$, $c_d = £0.5$, $I = £8$ in Exp2 to compensate for this fact and to account for inflation.

In Exp1, we also tried to elicit subjects' potential indifference between alternatives.[4] Following their choices at each binary menu, we asked whether their choice reflected

[2]We also note that, upon entering the lab, subjects could see the different headsets displayed on a table at the front of the room.

[3]See, for example, Germeijs and De Boeck (2002), Rassin (2007), and references therein.

[4]In an earlier version of this paper, we used these responses to transform the single- or empty-valued choice functions into multi- or empty-valued choice correspondences, aiming to give both FC and NFC

TABLE 1. Summary description of the two experiments' similarities and differences.

Design	*Experiment 1*	*Experiment 2*
Choice alternatives	5 headsets	5 headsets (as in Exp1)
Decision problems	31 (5 singletons + 26 nonsingletons)	26 (nonsingletons)
Initial endowment	£7	£8
Choice-reversal cost	£4	£6
Choice-deferral cost	£1	£0.50
Info. about alternatives after main phase	Yes	No
Rewards for correct quiz responses	No	£2
Trial round with different alternatives	Yes	Yes
Survey-based indifference elicitation	Yes	No
Location	St. Andrews	St. Andrews
Period	Sept 2013–Feb 2014	Sept–Oct 2018

a preference for their chosen alternative over the other, whether they found both to be equally good and so chose randomly, or whether they chose for another reason. Subjects were told that if they stated that they found both to be equally good, we would pick one of the alternatives at random for them at that menu, and they would not have the opportunity to revise their choice should this menu be selected in the final phase of the experiment. Subjects, therefore, did not have a strict incentive to state "indifference" in their responses, although it would be a dominated action for one to ever say they were indifferent if they were not so.[5] To simplify and facilitate the subjects' understanding of the experimental instructions we did not elicit indifferences in Exp2. Table 1 summarizes the similarities and differences in the implementation of our design in Exp1 and Exp2.

2.2 *Alternative approaches in the experimental literature*

We outline two broad approaches in the existing experimental economics literature through which other scholars have attempted to understand whether subjects' behavior may have been influenced by preference incompleteness (sometimes referred to as *imprecision* instead). Although these approaches differ, one feature they have in common—and which also distinguishes them from our approach—is that they were all

subjects the benefit of the doubt and maximize their behavioral consistency. We acknowledge, however, that any choice-augmentation method of this kind is bound to rely on assumptions, which in some cases may involve giving an equal weight to the nonstrictly incentivized indifference data and the payoff-relevant choice data. For this reason, we are focusing here on the primitive choice data only. More information about the above analysis is available upon request.

[5]It is possible, however, that some subjects who occasionally declared themselves to be indifferent did so because of a "preference for randomization" that was analogous to the one reported in Agranov and Ortoleva (2017). Subjects in these authors' experiments—which were conducted independently of our own and with a different focus—were faced with the same decision problem several times and, in addition to making different choices in the same menu, they were often willing to incur a small cost to have their choice made randomly. Unlike their experiments, ours was not specifically designed to test for such a preference. This is reflected in our nonrepeated decision problems and nonstrictly incentivized indifference statements.

implemented in environments of binary choice under risk and/or ambiguity. In the case of the first approach, this feature of their designs is key for their findings to have a relevant imprecise/incomplete-preference interpretation.

1. *Preference imprecision and preference for randomization in binary menus of money lotteries or uncertain acts.*

Cubitt, Navarro-Martinez, and Starmer (2015) provide a detailed overview of the various applications of the imprecise-preference elicitation approach and offers a new design and additional results.[6] The key feature in this design is that subjects are given several lists of binary menus where each row in a given list comprises a certain amount and a fixed non-degenerate lottery. The certain amounts between consecutive rows/menus in the list increase by a small fixed increment. Subjects are asked to state in each row whether they are: (i) sure they prefer the lottery; (ii) sure they prefer the certain amount; or (iii) unsure about their preference. Finally, subjects with a nondegenerate imprecision interval in some list (i.e., where preference uncertainty was stated in more than one row) are asked to identify a row/binary menu within that interval that corresponds to their own "best estimate" of their certainty equivalent for the nondegenerate lottery in that list. This best estimate acts as a "switching point" that guides the subjects' reward in case their payoff-relevant randomly selected menu fell within that list and within that interval. This design does not make it incentive-compatible for subjects to only report their true preference-imprecision interval within a list, because submitting any other interval that contains the same switching point will result in the same reward.

In related but distinct recent work, Agranov and Ortoleva (2021) proposed a different design to study choices from a list of binary menus of money lotteries where one of them is kept fixed while a parameter in the other lottery varies throughout the list. In particular, subjects were asked to either choose one of the two lotteries from every menu or to *randomize* their choice between them by selecting the relevant choice probability in increments of 0.1. The authors found evidence for significant randomization in subjects' behavior. When the fixed lottery was nondegenerate and the variable lottery was a certain amount of money, this randomization translated into large ranges for the certainty equivalent of the former lottery that was elicited in this incentivized way. The authors also pointed out that this behavior might be thought of as being suggestive of incomplete preferences, but also clarified that this interpretation cannot be formally separated from those of nonexpected utility models of complete preferences that violate the Independence axiom instead.

In choice under uncertainty, moreover, Cettolin and Riedl (2019) reported findings from an experiment where subjects were asked to either (i) choose one item from every binary menu in a sequence of such menus that feature a lottery and an ambiguous monetary act over the same two certain amounts, or (ii) choose the "mix" option of delegating choice between these prospects to a randomization device. They found that half

[6]As the authors noted, this literature has produced mixed results on whether preference imprecision—as measured in those studies—contributes importantly to violations of expected-utility theory under risk such as preference reversals or disparities between willingness-to-pay and willingness-to-accept (Butler and Loomes (2007, 2011)).

of the subjects opted for the mix option more than once, and explained that such behavior in this particular sequence of menus is incompatible with models of complete preferences under ambiguity. The authors also reported findings from a subsequent experiment suggesting that, from those subjects who chose the mix option more than once, half of them were always unwilling to pay a small amount to do so, and about one-third were always willing to do so. The authors concluded that the first class of these repeatedly-mixing subjects exhibit behavior that is consistent with Bewley's (2002) incomplete-preference model of indecisiveness in beliefs, and that the latter class are consistent with related experimental findings and models of preference for randomization in Agranov and Ortoleva (2017) and Cerreia-Vioglio, Dillenberger, Ortoleva, and Riella (2019), respectively.

2. *Commitment and flexibility/deferral in binary menus of money lotteries.*

Danan and Ziegelmeyer (2006) proposed a design that links incomplete preferences under risk with a *preference for flexibility* over menus (Kreps (1979); Dekel, Lipman and Rustichini (2001)) whereby the decision maker prefers menu $\{p, q\}$ to both menus $\{p\}$ and $\{q\}$. With p being a risky lottery and q a certain cash amount, and with p^* and q^* being p and q when their respective prizes are augmented by a fixed small amount, the authors interpreted subjects as revealing incompleteness between p and q if they chose the "flexibility" menu $\{p, q\}$ over the two "commitment" menus $\{p^*\}$ and $\{q^*\}$. A week later some decision problem was selected at random for each subject and, depending on their first-stage decision there, they were rewarded with the lottery/cash in their commitment menu or the lottery/cash that they were then asked to choose from their flexibility menu. The main differences between our NFC treatment design and these authors' can be summarized as follows: (i) we framed decision problems as ones where subjects had to choose *from* menus, not *between* them; (ii) we presented subjects with decision problems of many sizes, not only binary ones; (iii) we allowed costly switching to a different active choice in the final phase of the experiment for subjects who had previously made such a choice at their randomly selected menu [switching was not possible in Danan and Ziegelmeyer (2006)]; (iv) we introduced explicit deferral costs that were taken off some initial endowment instead of making the rewards marginally more appealing with nondeferral.

3. Rationality and active-choice consistency

In this section, we introduce the formal criteria we use to compare the choice consistency between FC and NFC subjects. Given a finite set X of general choice alternatives and a collection of nonempty subsets of X (to be called *menus*), the decision maker's observable behavior is described by a choice correspondence C that satisfies $C(A) \subseteq A$ for every menu A in that collection. By letting $C(A) = \emptyset$, here we model the situation where the decision maker, by opting for the no-choice/choice-deferral outside option, has chosen none of the feasible *market* alternatives. The agent's weak, strict, and indirect weak revealed preference relations $\succsim^R$, $\succ^R$ and $\succsim^{\widehat{R}}$ are defined, respectively, by $x \succsim^R y$ if there is a menu A such that $x \in C(A)$ and $y \in A$; $x \succ^R y$ if there is a menu A such that $x \in C(A)$ and $y \in A \setminus C(A)$; and $x \succsim^{\widehat{R}} y$ if there are alternatives $x_1, \ldots, x_n$ such

that $x = x_1$, $y = x_n$, and $x_i \succsim^R x_{i+1}$ for all $i = 1, \ldots, n-1$. Given these concepts and notation, we can now introduce compactly the following fundamental principles of choice consistency (see also Chambers and Echenique (2016) and references therein):

Weak Axiom of Revealed Preference (WARP).

$$x \succ^R y \Longrightarrow y \not\succsim^R x.$$

Congruence/Strong Axiom of Revealed Preference (SARP).

$$x \succ^R y \Longrightarrow y \not\succsim^{\widehat{R}} x.$$

In words, Congruence/SARP amounts to active-choice acyclicity, whereas WARP is its special case that rules out choice reversals/choice cycles of length two. WARP violations such as $C(\{w, x, y\}) = \{x\}$ and $C(\{x, y, z\}) = \{y\}$ amount to *direct choice reversals* that involve two market alternatives x and y. On the other hand, violations of Congruence/SARP that are not themselves WARP violations capture more general revealed-preference cycles such as $C(\{x, y, z\}) = \{x\}$, $C(\{y, z, v\}) = \{y\}$ and $C(\{x, v\}) = \{v\}$. Therefore, analysing the subjects' conformity with respect to both these principles offers complementary ways in which one can understand how consistent their active choices are.

A choice correspondence C is *rational* if there exists a complete and transitive preference relation $\succsim$ on X such that, for every menu A,

$$C(A) = \{x \in A : x \succsim y \text{ for all } y \in A\} \tag{1}$$

Rationality implies conformity with WARP and Congruence/SARP. In our environment of nonforced choice, however, the converse is not necessarily true: even when active choices satisfy these axioms, the overall behavior will be incompatible with that model if the decision maker defers at some menu(s). Two equally important questions naturally arise in this case for such an agent's behavior:

1. Are *all* her decisions compatible with (1)?
2. If not, are her *active-choice* decisions compatible with (1)?

Clearly, when the answer to the first question is positive, the agent behaves as if she were a utility maximizer. But even when that's not true and the answer to the second question is positive instead, rationality is not contradicted in a strong sense because an incomplete but transitive preference relation is recoverable from her active choices and, by standard extension results, this can in theory be extended at a later stage to a complete relation. Therefore, rather than suggesting "irrationality", this case signifies a "not-yet-rational" behavior. Our primary aim is to investigate the existence and frequency of precisely this unexplored kind of behavior.

Now, when an agent's active-choice decisions are incompatible with rational choice, one is naturally interested in measuring the severity of this incompatibility. Two such measures are suggested by the classic Houtman–Maks (1985) (henceforth "HM") index

and the more recently proposed Apesteguia–Ballester (2015) (henceforth "Swaps") index. The former corresponds to the smallest number of choices that need to be removed from this data set in order for the remaining ones to be perfectly consistent with some instance of utility maximization, that is, some complete and transitive relation $\succsim$ that explains the data as in (1). The Swaps index instead is computed by minimizing over all such instances the sum of the total number of alternatives in each menu that are better than the agent's chosen alternative.

Let us illustrate the differences between the two measures with the following example (we write $C(\{x, y\}) \equiv C(x, y)$, etc., to ease notation): $C(x, y) = x$, $C(x, z) = z$, $C(y, z) = z$, $C(x, y, z) = y$ on $X := \{x, y, z\}$. The corresponding HM and Swaps scores here are 1 and 2, respectively: changing $C(x, y, z)$ from y to z makes that dataset compatible with the preference order $z \succ x \succ y$, that is, suggested by the binary choices; however, although $z \succ x \succ y$ is indeed the best-matching ordering for that data set, Swaps also takes into account that the "erratic" choice of y in $C(x, y, z)$ is two ranks away from the optimal choice of z in that menu, and is therefore associated with a score of 2. Thus, Swaps incorporates menu-specific cardinal information on the agent's deviations from rational choice that in some cases might be interpretable as quantifying the welfare loss associated with the agents' "wrong" choices.

4. Main behavioral findings

The data analysis comprises 75 FC, 86 NFC subjects in Exp1 and 53 FC, 68 NFC such subjects in Exp2. Some additional participants were excluded from the analysis due to: (i) failing to demonstrate that they understood the instructions (12 and 17 subjects, resp.); (ii) deferring at one or more singleton menus in Exp1 (58 NFC subjects); (iii) behaving randomly (1 FC and 3 NFC subjects in Exp1; 1 FC subject in Exp2). A participant was thought to behave randomly if the HM score on their active choices was weakly greater than the 2.5% cut-off value of the subject-specific HM distribution that resulted from simulated uniform-random active choices at only those menus where the participant made their active choices. Subjects deferring at singleton menus were excluded because such decisions were dominated and rendered the interpretation of deferrals at larger menus unclear. Our main results are robust—and in fact with higher levels of statistical significance—without the latter two kinds of exclusions (details are available in the Appendix in the Online Supplementary Material). Table 2 summarizes this information.

Figure 1 shows the distribution of deferrals of NFC subjects in Exp1 and Exp2. We find no statistical difference between them ($p = 0.2$, two-sided Mann–Whitney U-test). In particular, Exp1 and Exp2 subjects deferred on average at 3.8 and 4.8 menus, that is, in about one out of every six menus. Moreover, 35% of these subjects deferred at least once, and did so at an average of 12.5 (Exp1) and 11.5 (Exp2) menus, with most subjects deferring at between 1 and 16 menus, and 8 subjects deferring at either all or all but one or two menus. We found no differences in deferring behavior between Exp1 and Exp2, both in terms of the proportions of deferring NFC subjects ($p = 0.2$, two-sided Fisher's exact test) and in their overall distribution of deferrals ($p = 0.9$, Mann–Whitney U-test).

TABLE 2. Summary description of the two experiments' sample sizes and subject exclusions.

	Experiment 1	*Experiment 2*
Subjects in the original sample	235	139
Subject exclusions due to quiz failures	12	17
Subject exclusions due to indistinguishability from random behavior	4	1
Subject exclusions due to singleton deferrals	58	N/A
Subject exclusions—total	74	18
Subjects after all exclusions (% of original sample)	161 (68.5%)	121 (92%)
Subjects after all exclusions—FC	75	53
Subjects after all exclusions—NFC	86	68

Taking both experiments together, we found that around one-third of the subjects deferred at least once, and that the rate of deferrals at nonsingleton menus among these subjects was close to one-half.

We now compare FC and NFC subjects' rationality/conformity with (1) and their active-choice consistency/conformity with WARP and Congruence/SARP within each experiment separately, as well as when the data from the two experiments are pooled. Although the two experiments differed in some important ways, their primary features (namely their choice menus, experimental interface, and general reward structure) were identical. Together with the results from the previous tests, which show that the distribution of deferrals in the two experiments were not significantly different, and further tests showing that the same is true for the various measures of choice consistency used throughout the paper, this fact suggests that the pooled data—and the resulting larger sample size—allow for an additional informative comparison.[7]

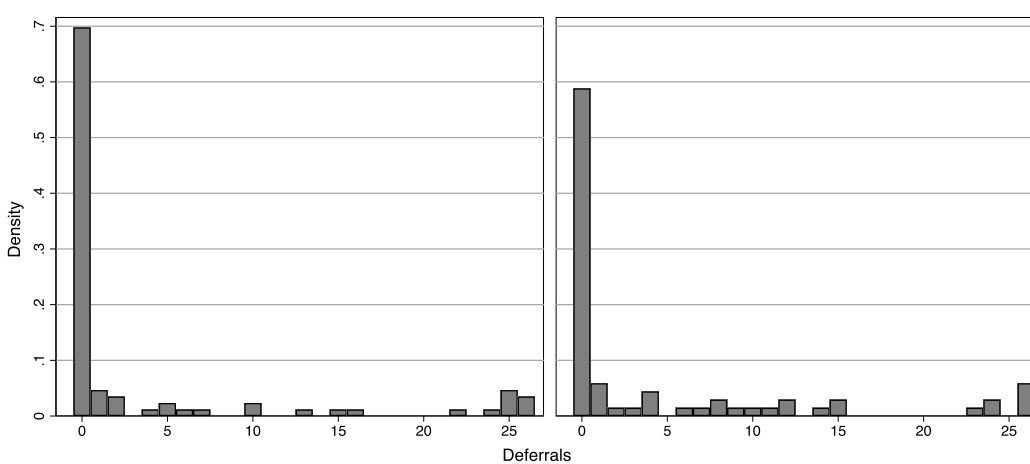

FIGURE 1. Deferral relative frequencies in Experiment 1 (left panel) and Experiment 2 (right panel).

[7]In particular, we find no statistically significant differences between Exp1 and Exp2 in the proportions of WARP/Congruence-violating subjects in either treatment (FC: $p = 0.588$; NFC: $p = 0.593$; Fisher's exact

Table 3. Proportions of subjects with zero active-choice cycles.

	Proportions of Subjects Conforming With the Weak and Strong Axiom of Revealed Preference/Congruence		
	Exp1	Exp2	Pooled
Forced choice	55% (41/75)	60% (32/53)	57% (73/128)
Nonforced choice	69% (59/86)	74% (50/68)	71% (109/154)
p-value	0.075	0.170	0.018
N	161	121	282

Note: (i) p-values from two-sided Fisher exact tests; (ii) the average proportion of simulated random-behaving subjects satisfying Congruence/SARP conditional on the menus where Exp1-Exp2 subjects made their active choices is 0% for FC and 9.8% in NFC.

Table 3 presents the results from comparing the proportions of subjects who violated WARP and Congruence/SARP in the two treatments. Every Congruence/SARP violator turned out to also violate WARP, in both experiments and treatments. Moreover, the proportion of subjects who violated these axioms is 14 percentage points higher in the FC treatment in each of Exp1 and Exp2. The difference between the proportions is statistically significant at the 5% level when the data from the two experiments are pooled, even if not within each experiment.[8] Furthermore, Table 4 presents the results

Table 4. Average WARP and Congruence/SARP violations.

	Weak Axiom of Revealed Preference			*Strong Axiom of Revealed Preference/ Congruence*		
	Exp1	Exp2	Pooled	Exp1	Exp2	Pooled
Forced choice	3.20 (4)	4.25 (7)	3.63	12.61 (7)	18.43 (12)	15.02
Nonforced choice	2.24 (3)	3.16 (1.5)	2.65	5.42 (3)	20.76 (1.5)	12.19
p-value	0.098	0.124	0.020	0.076	0.107	0.014
N	161	121	282	161	121	282

Note: (i) values in parentheses under the Exp1 and Exp2 columns show the 3rd quartiles (all medians are zero); (ii) p-values from two-sided Mann–Whitney U tests; (iii) the average number of violations of the simulated uniform-random (over alternatives) subjects when simulations consider only the menus where Exp1-Exp2 subjects made their active choices are 52, 42 for WARP violations and 2698, 2014 for SARP/Congruence violations in FC and NFC environments, respectively.

tests), nor in their Houtman–Maks or Swaps scores on active choices (HM-FC: $p = 0.926$; HM-NFC: $p = 0.860$; Swaps-FC: $p = 0.956$; Swaps-NFC: $p = 0.830$; Mann–Whitney U-tests).

[8]Prior to conducting Exp2, the weak evidence for the existence of binary choice cycles (formally, Congruence/SARP violations restricted to menus with two alternatives) in the 10 binary menus over 5 goods in Exp1 prompted us to investigate further the comparative incidence of such cycles in a more focused experiment that only featured choices from the 15 possible binary menus over a set of 6 lotteries. These lotteries had three strictly positive monetary outcomes and were pairwise unrelated by second-order stochastic dominance. That experiment was conducted at the University of Alicante Experimental Economics Lab in January–March 2018. Binary cycles were more frequent in the two treatments of that experiment and deferring NFC subjects were significantly more consistent in this sense than both nondeferring NFC subjects and FC subjects. However, there was no statistically significant overall difference in consistency between NFC and FC subjects. Because the structure of that experiment was very different from those of Exp1 and Exp2, we report on it in more detail in the Appendix of the Online Supplementary Material.

TABLE 5. Subjects' average Houtman–Maks and Swaps indices on active choices.

	Houtman–Maks index			*Swaps index*		
	Exp1	Exp2	Pooled	Exp1	Exp2	Pooled
Forced choice	0.80 (75)	0.98 (53)	0.88 (128)	0.88 (75)	1.09 (53)	0.97
Nonforced choice	0.58 (86)	0.75 (64)	0.65 (150)	0.62 (86)	0.86 (64)	0.72
p-value	0.088	0.205	0.032	0.108	0.199	0.035
N	161	117	278	161	117	278

Note: (i) number of subjects in parentheses; (ii) p-values from two-sided Mann–Whitney U-tests; (iii) the average number of violations of the simulated uniform-random (over alternatives) subjects when simulations consider only the menus where Exp1-Exp2 subjects made their active choices are 10.89, 8.98 for HM and 16.33, 13.45 for Swaps in FC and NFC environments, respectively.

from comparing the distributions of subjects' WARP and Congruence/SARP violations in the two treatments, which point in the same direction.

Complementing the analysis based on the above metrics, Table 5 shows the results from comparing the distributions of HM and Swaps indices across treatments. These results reinforce the ones reported in Table 3. In particular, NFC subjects' active choices are closer to the rational choice model relative to FC subjects according to both the HM and Swaps indices, and the difference between the respective distributions in each case is significant at the 5% level in the pooled data. Notably, the average scores are generally lower than or very close to 1 in both experiments and treatments, suggesting that subjects were on average less than one decision away from perfect conformity with utility maximization, conditional on the menus where they made active choices.

The findings in Tables 3–5 provide indications that subjects who are not forced to choose make more consistent active choices than subjects who are forced to do so. Importantly, however, although rational decision makers would have zero axiom violations and HM/Swaps indices regardless of whether they operated in an FC or an NFC environment, for agents whose behavior is incompatible with this model one naturally expects that the incidence and magnitude of deviations from it will depend on the number of their active choices and on the specific menus where these choices were made. To address this comparability issue, we also perform a "like-for-like" cross-treatment comparison of Selten's (1991) measure of *predictive success*, the relevance of which in revealed preference analysis is increasingly appreciated since Beatty and Crawford (2011).

To this end, let us first recall that, when applied to the rational choice model, Selten's measure is defined by subtracting from the proportion of actual subjects whose behavior is perfectly explained by it (called the "pass rate") the proportion of uniform-random simulated subjects that are also explained by that model perfectly (called the "area"). The latter proportion is informative of the *power* of the revealed-preference test, a concept that originated in Bronars (1987). The difference between these two proportions lies in the $[-1, 1]$ interval, and higher values point to a higher predictive power for the model relative to the uniform-random model, conditional on the given decision environment.

Let us denote by m^i_{FC}, p^i_{FC}, and a^i_{FC} the predictive success rate, pass rate, and area that correspond to the rational choice model in the FC treatment of experiment $i \in \{1, 2\}$.

By definition,

$$m^i_{\mathrm{FC}} = p^i_{\mathrm{FC}} - a^i_{\mathrm{FC}}.$$

To apply this method to the NFC treatment as well, we proceed as follows. For a given subject n in that treatment, we first find the proportion of random-behaving simulated subjects whose active choices are rational once they are confined to the same menus as those where subject n made their own active choices. We then define the pass rate p^i_{NFC} in that treatment and experiment to be the proportion of actual NFC subjects whose corresponding active choices were rational, and the area a^i_{NFC} to be the average proportion of the random-behaving subjects—conditioned as above—that were rational. The predictive success rate of rational choice in the NFC treatment is then defined by

$$m^i_{\mathrm{NFC}} = p^i_{\mathrm{NFC}} - a^i_{\mathrm{NFC}}.$$

A comparison between m^i_{FC} and m^i_{NFC} therefore accounts for the fact that subjects in the latter treatment make fewer active choices in general, and corrects the possibly higher pass rates that may emerge as a result of this fact by conditioning the area-determining simulations accordingly.

Because our subjects made decisions from the full set of 26 possible nonsingleton menus that can be derived from a set of 5 alternatives, the power of the revealed-preference test for this model is very high in the FC environment, with $a^i_{\mathrm{FC}} \approx 0$. That is, such randomly-made forced-choice decisions can be consistent with rational choice very rarely (at a roughly 1 out of 2.83 billion rate). It follows therefore that $m^i_{\mathrm{FC}} \approx p^i_{\mathrm{FC}}$ in our case or, equivalently, that the rational choice model's predictive success rates in the FC treatment can be taken to coincide with the proportion of such subjects that this model explains *perfectly*. Because approximately 54.6% and 60.4% of FC subjects in Exp1 and Exp2, respectively, behaved as if they were perfect utility maximizers, this in turn translates into the Selten rates $m^1_{\mathrm{FC}} \approx 0.546$ and $m^2_{\mathrm{FC}} \approx 0.604$ for utility maximization in the two experiments, with $m_{\mathrm{FC}} \approx 0.570$ in the pooled data.

In the NFC treatment on the other hand, some of the collections of menus where subjects made active choices were associated with nonnegligible areas in the corresponding subject-specific simulations. Taking both these areas and the pass rates of utility maximization into account, the Selten measure in the NFC treatment was $m^1_{\mathrm{NFC}} \approx 0.686 - 0.103 = 0.583$ in Exp1 and $m^2_{\mathrm{NFC}} \approx 0.735 - 0.092 = 0.643$ in Exp2, with $m_{\mathrm{NFC}} \approx 0.609$ in the pooled data. Therefore, rational choice predicts much better than random behavior in both treatments, and more so in the NFC than in the FC treatment.

5. Possible interpretations of the main findings

5.1 *Indecisiveness and incomplete preferences*

The results presented so far suggest that a significant fraction of subjects are willing to opt for costly choice deferral when given the opportunity to do so, and that active choices are more likely to be consistent with rationality in such an environment. A possible interpretation of these findings is that some subjects in our experiment may have

TABLE 6. Active-choice consistency for decisive and indecisive subjects.

	Weak and Strong Axiom of Revealed Preference/Congruence			*Houtman–Maks Index*			*Swaps Index*		
	Decisive	Indecisive	p-Value	Decisive	Indecisive	p-Value	Decisive	Indecisive	p-Value
Forced choice	70% (28/40)	47% (23/49)	0.033	0.43 (40)	1.10 (49)	0.012	0.48	1.24	0.013
Nonforced choice	72% (41/57)	78% (38/49)	0.655	0.52 (56)	0.53 (47)	0.499	0.59	0.60	0.447
p-value	1.00	0.003		0.928	0.002		0.871	0.002	
N	97	98		96	96		96	96	

Note: (i) pooled data; (ii) number of subjects in parentheses; (iii) decisive (indecisive) subjects correspond to bottom (top) tertile scorers in the indecisiveness personality questionnaire; (iv) the first column block lists the percentage of subjects with zero WARP (and also Congruence/SARP) violations (p-values from two-sided Fisher's exact tests); the second and third column blocks list, respectively, the Houtman–Maks and Swaps mean scores on active choices (p-values from two-sided Mann–Whitney U-tests).

been *indecisive* at some menus and so opted to defer when this option was available to them. The increase in inconsistencies observed in the FC treatment could then be driven by the fact that such potentially indecisive subjects who are forced to choose may be more likely to violate rationality.

To test this prediction, we first examine the relationship between subjects' behavior in the experiment and their responses to the *indecisive personality* questionnaire of Germeijs and De Boeck (2002) that we administered at the end of each session. The questionnaire elicited 0–7 Likert-scale responses in 11 original statements and also in 11 statements that made the exact opposite claim. Examples include *"I find it easy to make decisions," "It is hard for me to come to a decision," "I delay deciding," "I don't postpone making decisions to a later date."* Following Germeijs, Verschueren, and Soenens (2006), the indecisiveness score was calculated as the average response to the 22 items and was normalized to lie between 0 and 1, with a higher score indicating higher indecisiveness. In what follows, we refer to subjects with relatively higher scores as *indecisive*, those with relatively lower scores as *decisive*, and those who deferred at least once as *deferring*. We ask:

1. Are deferring subjects more indecisive?

2. When forced to choose, are indecisive subjects more inconsistent than decisive ones?

3. Is the observed forced-choice treatment effect driven by indecisive subjects?

We find evidence suggesting a positive answer to each of these questions. Specifically, the average indecisive-personality score of the 54 deferring and 100 nondeferring NFC subjects in the two experiments were 0.435 and 0.360, respectively ($p = 0.026$; two-sided Mann–Whitney U-test). In addition, Table 6 shows that: (i) decisive FC subjects—defined as those in the bottom third of the indecisive-personality distribution—were significantly more consistent than indecisive ones—defined as those in the top third; (ii)

TABLE 7. Subjects' average distance scores for the models of rational choice and dominant choice with incomplete preferences over the combined active-choice and deferral decisions.

	Rational Choice	*Dominant Choice With Incomplete Preferences*	*Average Best Score per Subject Over the Two Models*
Experiment 1: Forced choice	0.80 (0–5; 0; 55%)	1.80 (1–6; 1; 0%)	0.80 (0–5; 0; 55%)
Experiment 1: Nonforced choice	4.36 (0–26; 1; 44%)	1.86 (0–7; 1; 10%)	1.12 (0–7, 0; 54%)
Experiment 2: Forced choice	0.98 (0–5; 0; 60%)	1.98 (1–6; 1; 0%)	0.98 (0–5; 0; 60%)
Experiment 2: Nonforced choice	5.46 (0–26; 1.5; 43%)	2.18 (0–10; 1; 10%)	1.53 (0–10; 0; 53%)
Pooled: Forced choice	0.88 (0–5; 0; 57%)	1.88 (1–6; 1; 0%)	0.88 (0–5; 0; 57%)
Pooled: Nonforced choice	4.84 (0–26; 1; 44%)	2.00 (0–10; 1; 10%)	1.30 (0–10; 0; 54%)

Note: Ranges, medians, and proportions of zero scores/perfect fits in parentheses.

indecisive subjects were significantly more consistent in NFC than in FC while no treatment differences were observed for decisive subjects. We obtain similar results when we categorize subjects as decisive or indecisive using a median-split instead.

Taken together, the preceding results naturally invite an investigation of the possibility that NFC subjects' deviation from rational choice is due to their occasionally hesitant behavior that manifests itself in deferrals, as opposed to the potential inconsistency of their active choices. As we pointed out in Section 3, although such behavior is incompatible with rationality, it is perhaps better viewed as "not-yet-rational" rather than as "irrational." Observing such "not-yet-rational" behavior raises the question of whether any additional information may also be recovered about deferring subjects' decision processes and preferences.

A natural starting point in this regard is to investigate whether the deferring subjects' behavior is potentially compatible with the preference-maximization principle laid out in (1) but when preferences are strictly incomplete. This model of *(maximally) dominant choice with incomplete preferences* (Gerasimou (2018), Section 2) portrays the individual as making an active choice at a menu if and only if there is a most preferred alternative

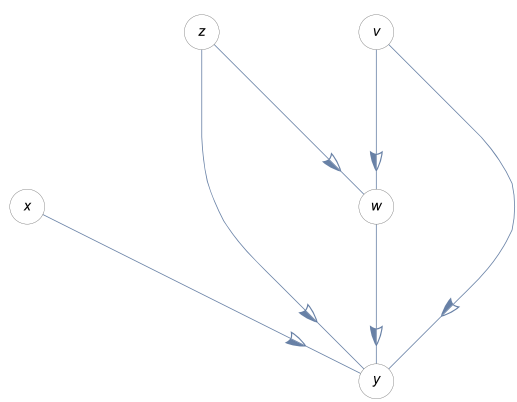

FIGURE 2. The incomplete preference relation recovered from a nonforced-choice subject via the dominant-choice model.

TABLE 8. Proportions of subjects whose combined active-choice and deferral decisions are best explained by rational choice and dominant choice with incomplete preferences.

	Rational Choice	*Dominant Choice With Incomplete Preferences*
Experiment 1: Forced choice	100% (55%)	0%
Experiment 1: Nonforced choice	76% (44%)	24% (11%)
Experiment 1: Nonforced choice—deferring	19% (0%)	81% (35%)
Experiment 2: Forced choice	100% (60%)	0%
Experiment 2: Nonforced choice	66% (43%)	34% (10%)
Experiment 2: Nonforced choice—deferring	18% (0%)	82% (25%)
Pooled: Forced choice	100% (57%)	0%
Pooled: Nonforced choice	71% (44%)	29% (10%)
Pooled: Nonforced choice—deferring	19% (0%)	81% (30%)

Note: (i) proportion of perfect fits in parentheses; (ii) all ties (one subject in each of Exp1 and Exp2) are broken in favor of rational choice.

in that menu and as deferring otherwise. As such, it predicts behavioral patterns like the one quoted in the opening paragraph. When we analyze the deferring NFC subjects' *combined* deferral and active-choice data we find that this model provides a better fit than rational choice for more than 80% of those subjects in the two experiments. Goodness of fit in this comparison is captured by the extension of the HM method that detects how close the combined deferral and active-choice data of a given subject are with (1) when preferences are potentially incomplete (Tables 7 and 8). This fit (referred to as the model's *distance score*) was perfect for some subjects and allowed for the recovery of a unique partial preference ordering (see Figure 2 for an example[9]). Although these facts do not prove that those subjects deferred because their preferences over the five headsets were indeed incomplete, they do suggest that their behavior can be thought of *as if* it had been generated by such a decision process and preference relation.

To relate the results from this analysis to those that are based on the indecisive personality questionnaire, we also compared the indecisiveness scores of all 109 NFC subjects who were best explained by rational choice with those of the 45 NFC subjects who were best explained by dominant choice with incomplete preferences. Although the average scores in these two groups are 0.37 and 0.42, respectively, and hence in the expected direction, the difference is not significant ($p = 0.126$; Mann–Whitney U-test).

We finally note that because our experimental data feature single- or empty-valued choices, we cannot use them to conduct a proper test of the complementary class of undominated-choice models of incomplete preference maximization such as those analyzed in Schwartz (1976) and Eliaz and Ok (2006). We leave the task of testing these no-deferral models for future work.

[9]In this subject's case, for example, $C(v, w, y) = v$ and $C(z, w, y) = z$ reveal a preference for v and z over both w and y, respectively, while $C(v, z) = \emptyset$ reveals indecisiveness/incomparability between v and z, and this in turn is consistent under this model with the observed $C(v, w, y, z) = \emptyset$. Similarly, x is revealed preferred to y and incomparable to everything else.

TABLE 9. Logit regression of the determinants of deferral.

Variable	Coef.	Std Err.	p-Value	Odds Ratio
Indecisiveness	1.721	0.944	0.068	5.590
Menu position	−0.004	0.005	0.345	0.996
Menu size 3	0.008	0.080	0.924	1.008
Menu size 4	−0.085	0.113	0.454	0.919
Menu size 5	−0.031	0.150	0.838	0.970
Experiment 2	0.181	0.380	0.635	1.198
Observations	4004			
Number of clusters	154			

Note: NFC data only. Dependent variable = 1 if the subject defers at a menu, = 0 otherwise. Standard errors clustered at the subject level.

5.2 *Introspective preference learning*

In addition to the indecisiveness and incomplete-preferences hypotheses, we also examine the possibility of introspective (i.e., information-free) preference learning taking place during the course of the main phase of the experiment. Intuitively, the introspective preference learning hypothesis would be supported in our data if the rates of deferrals and/or active-choice inconsistencies decrease as subjects progress through the sequence of menus.

To test for a learning effect on deferrals, we estimated a logit model of the probability of deferral as a function of the position of the menu in the sequence, the size/cardinality of the menu, the experiment (Exp1/2), and the indecisive personality score. This allows us to evaluate simultaneously the relative contributions to deferral of learning (via menu position), choice overload (via menu size), external information acquisition (via Exp1/2), and indecisive personality.

As shown in Table 9, the estimated coefficients of all variables in the model except indecisiveness are very close to 0 and far from significant. This suggests that learning, choice overload, or information acquisition had little impact on NFC subjects' deferral decisions. Moreover, although only significant at the 10% level, the estimated effect of indecisiveness is in line with our previously reported positive correlation between deferrals and indecisive personality scores.

We next evaluate the effect of introspective learning on active choice consistency. We note, however, that if such learning exists, its impact on choice consistency should intuitively be stronger in the FC treatment than in the NFC treatment. Since FC subjects are forced to choose at every menu in the sequence, introspective learning would in principle render them more vulnerable to mistakes early on compared to NFC subjects, who have the option to defer whenever they are unsure about which option to choose.

To test this hypothesis, we used some additional information from the computation process of the HM indices that we reported in Section 4. Let us first recall that a subject's HM index is the smallest number of active choices that are incompatible with the maximization of some strict linear order over the alternatives. In what follows, we refer to any such linear order as an *inferred order* for that subject. In this test, we look not only at the

TABLE 10. Logit regression of subjects' position-adjusted HM index (standard errors clustered at the subject level).

Variable	Coef.	Std Err.	p-Value	Odds Ratio
FC	0.580	0.287	0.043	1.787
Menu position	−0.018	0.012	0.141	0.983
FC x Menu position	−0.034	0.018	0.059	0.967
Menu size 3	0.468	0.108	0.000	1.597
Menu size 4	0.556	0.144	0.000	1.743
Menu size 5	0.400	0.302	0.185	1.491
Experiment 2	−0.025	0.244	0.920	0.976
Observations	7332			
Number of clusters	282			

count of incompatible choices but also at the menus in which these choices were made. We define the *position-adjusted HM* index as a binary variable that takes a value of 1 if the choice made by a subject at a given menu is incompatible with any of her inferred orders, and a value of 0 otherwise. In other words, although the original HM metric is a nonbinary measure defined at the subject level, the position-adjusted HM measure is a subject as well as menu-specific binary measure that reflects whether removing an observation from a subject's choice dataset reduces her HM score or not.

We used this position-adjusted HM metric as the dependent variable in a logit regression that includes as regressors treatment, experiment, the menu's size/cardinality, and its position in the sequence. In order to test whether FC subjects improve their active-choice consistency faster than NFC subjects through introspective learning, the regression also includes an interaction term between treatment and menu position. The results from this regression, which are displayed in Table 10, offer suggestive evidence in favor of the differential learning conjecture. While the estimated learning effect for NFC subjects (given by the coefficient on menu position) is not significantly different from zero, the estimated learning effect for FC subjects (given by the *sum* of the coefficients on menu position and the interaction term FC $\times$ menu position, or by the product of their respective odds ratios) is negative and highly significant (Wald test, $p = 0.0001$). Thus, while NFC subjects do not appear to become more consistent over time, the odds that FC subjects make an inconsistent active choice decrease by 5% for each new menu in the sequence.

The second finding from this regression is that menus with 3 or 4 alternatives have a highly significant effect on the position-adjusted HM measure. According to the estimated odds ratios, such menus are 60% and 74% more likely to be implicated in a choice error than binary menus, respectively. This result, however, should not necessarily be interpreted as evidence of context-dependent active-choice effects. First, note that, given the nature of the 5 headsets in our experiment, these menus did not feature objectively identifiable decoy or compromise effects. Second, only one mistake is possible at a binary menu (i.e., a choice reversal between its two alternatives), while more such mistakes are possible in larger menus. Hence, for any particular choice error in our

experiment, more opportunities to make it will generally occur at menus with sizes 3 and 4 than at the binary menus.

Finally, to further explore the notion that "learning by doing" in the FC condition makes subjects comparatively more vulnerable to mistakes early on, we evaluated the impact that ignoring the subjects' first $n \leq 5$ decisions (which amount to up to 20% of all) has on the proportion of active-choice consistent subjects across treatments. As expected, we find that progressively removing each of these 5 decisions monotonically increases the proportion of both FC and NFC subjects whose active choices conform with rationality. Moreover, consistent with the previous conjecture, we find that this increase tends to be more pronounced in the FC treatment. Thus, although a higher proportion of NFC subjects are rational compared to their FC counterparts in this sense for all $n \leq 5$, the difference between these two groups decreases from 14 percentage points when $n = 0$ to 8 percentage points when $n = 5$.

6. Concluding remarks

In this paper, we raised the question of whether not forcing experimental subjects to make active choices from the menus they are presented with, and allowing them instead to delay making such choices, could enable researchers to recover more consistent behaviors and stable preferences from such subjects, possibly over only a subset of the original set of alternatives. To answer this question, we implemented a new experimental design with a forced- and a nonforced-choice treatment. Analyzing the data from two experiments that implemented this design on 26 menus over 5 real goods, and using several nonparametric methods that have been suggested in the literature, we found evidence that higher proportions of subjects make consistent active choices when not forced to choose, and that the active-choice behavior of subjects in that treatment is generally closer to being considered rational.

In support of our intuition about the mechanism underlying the treatment effect, we found that (i) subjects categorized as indecisive according to a personality questionnaire were more likely to defer and to violate rationality in the forced-choice treatment than decisive subjects, and (ii) the combined deferral and active-choice decisions of nonforced-choice subjects was better explained by a model of dominant choice with incomplete preferences than by rational choice. These facts point to the potential usefulness of nonforced-choice experiments and models for revealed-preference analysis, and highlight the relevance of aiming to separate people's rational, hesitant/not-yet-rational and genuinely irrational behavior in such analysis.

Finally, we found that subjects in the forced-choice treatment were more likely to make inconsistent choices in earlier menus than in later ones, suggesting that introspective preference learning may play a role in driving down choice inconsistencies in forced-choice environments. We note, however, that whether introspective preference learning generally happens in choice experiments is an important question that goes beyond the boundaries of our study and the motivation underlying our experimental design. Both the learning hypothesis and the robustness of our main finding (the negative effect of forced choice on consistency) are in our view worthy of additional investigation in future studies.

References

Agranov, Marina and Pietro Ortoleva (2017), "Stochastic choice and preferences for randomization." *Journal of Political Economy*, 125, 40–68. [1302, 1304]

Agranov, Marina and Pietro Ortoleva (2021), "Ranges of randomization." Working Paper. [1303]

Anderson, Christopher J. (2003), "The psychology of doing nothing: Forms of decision avoidance result from reason and emotion." *Psychological Bulletin*, 129, 139–167. [1298]

Apesteguia, Jose and Miguel A. Ballester (2015), "A measure of rationality and welfare." *Journal of Political Economy*, 123, 1278–1310. [1306]

Beatty, Timothy K. M. and Ian A. Crawford (2011), "How demanding is the revealed preference approach to demand." *American Economic Review*, 101, 2782–2795. [1309]

Bewley, Truman F. (2002), "Knightian decision theory. Part I." *Decisions in Economics and Finance*, 25, 79–110. [1304]

Bhatia, Sudeep and Timothy L. Mullett (2016), "The dynamics of deferred decision." *Cognitive Psychology*, 86, 112–151. [1298]

Bronars, Stephen G. (1987), "The power of non-parametric tests of preference maximization." *Econometrica*, 55, 693–698. [1309]

Butler, David and Graham Loomes (2007), "Imprecision as an account of the preference reversal phenomenon." *American Economic Review*, 97, 277–297. [1303]

Butler, David and Graham Loomes (2011), "Imprecision as an account of violations of independence and betweenness." *Journal of Economic Behavior & Organization*, 80, 511–522. [1303]

Cerreia-Vioglio, Simone, David Dillenberger, Pietro Ortoleva, and Gil Riella (2019), "Deliberately stochastic." *American Economic Review*, 109, 2425–2445. [1304]

Cettolin, Elena and Arno Riedl (2019), "Revealed preferences under uncertainty: Incomplete preferences and preferences for randomization." *Journal of Economic Theory*, 181, 547–585. [1303]

Chambers, Christopher P. and Federico Echenique (2016), *Revealed Preference Theory*. Econometric Society Monograph, Vol. 56. Cambridge University Press, Cambridge. [1305]

Costa-Gomes, Miguel A., Carlos Cueva, Georgios Gerasimou, and Matúš Tejiščák (2022), "Supplement to 'Choice, deferral, and consistency'." *Quantitative Economics Supplemental Material*, 13, https://doi.org/10.3982/QE1806. [1299]

Cubitt, Robin, Daniel Navarro-Martinez, and Chris Starmer (2015), "On preference imprecision." *Journal of Risk and Uncertainty*, 50, 1–34. [1303]

Danan, Eric and Anthony Ziegelmeyer (2006), "Are preferences complete? An experimental measurement of indecisiveness under risk." Working Paper. [1304]

Dekel, Eddie, Bart L. Lipman, and Aldo Rustichini (2001), "Representing preferences with a unique subjective state space." *Econometrica*, 69, 891–934. [1304]

Eliaz, Kfir and Efe A. Ok (2006), "Indifference or indecisiveness? Choice-theoretic foundations of incomplete preferences." *Games and Economic Behavior*, 56, 61–86. [1313]

Fischbacher, Urs (2007), "z-tree: Zurich toolbox for ready-made economic experiments." *Experimental Economics*, 10, 171–178. [1301]

Gerasimou, Georgios (2018), "Indecisiveness, undesirability and overload revealed through rational choice deferral." *Economic Journal*, 128, 2450–2479. [1299, 1312]

Germeijs, Veerle and Paul De Boeck (2002), "A measurement scale for indecisiveness and its relationship to career indecision and other types of indecision." *European Journal of Personality Assessment*, 18, 113–122. [1299, 1301, 1311]

Germeijs, Veerle, Karine Verschueren, and Bart Soenens (2006), "Indecisiveness and high school students' career decision-making process: Longitudinal associations and the mediational role of anxiety." *Journal of Counseling Psychology*, 53, 397–410. [1311]

Greiner, Ben (2015), "Subject pool recruitment procedures: Organizing experiments with ORSEE." *Journal of the Economic Science Association*, 1, 114–125. [1300]

Houtman, Martijn and Johannes A. H. Maks (1985), "Determining all maximal data subsets consistent with revealed preference." *Kwantitatieve Methoden*, 19, 89–104. [1305]

Kreps, David M. (1979), "A representation theorem for 'preference for flexibility'." *Econometrica*, 47, 565–577. [1304]

Luce, R. Duncan and Howard Raiffa (1957), *Games and Decisions*. Dover, New York. [1298]

Rassin, Eric (2007), "A psychological theory of indecisiveness." *The Netherlands Journal of Psychology*, 63, 2–13. [1301]

Schwartz, Thomas (1976), "Choice functions, "rationality" conditions, and variations of the weak axiom of revealed preference." *Journal of Economic Theory*, 13, 414–427. [1313]

Selten, Reinhart (1991), "Properties of a measure of predictive success." *Mathematical Social Sciences*, 21, 153–167. [1309]

Sen, Amartya (1997), "Maximization and the act of choice." *Econometrica*, 65, 745–779. [1301]

Shafir, Eldar, Itamar Simonson, and Amos Tversky (1993), "Reason-based choice." *Cognition*, 11, 11–36. [1298]

Tversky, Amos and Eldar Shafir (1992), "Choice under conflict: The dynamics of deferred decision." *Psychological Science*, 3, 358–361. [1298]

Co-editor Christopher Taber handled this manuscript.

Manuscript received 9 January, 2021; final version accepted 13 October, 2021; available online 1 November, 2021.