Can commitments cause counterpreferential choices?

Messerli, Michael ; Reuter, Kevin

DOI: https://doi.org/10.1080/1350178X.2022.2077407

JOURNAL OF ECONOMIC METHODOLOGY
https://doi.org/10.1080/1350178X.2022.2077407

Routledge
Taylor & Francis Group

Can commitments cause counterpreferential choices?

Michael Messerli and Kevin Reuter

Department of Philosophy, University of Zurich, Zurich, Switzerland

ABSTRACT
Commitments are crucial for our lives but there is no consensus on how commitments and preferences relate to each other. In this paper, we present three empirical studies that provide evidence that people sometimes choose a less preferred option when they have made a commitment.

ARTICLE HISTORY
Received 4 May 2021
Accepted 10 May 2022

KEYWORDS
Commitments; counterpreferential choice; empirical studies

1. Introduction

Consider the following case: Yesterday, a colleague of John has asked John whether he would help him move some furniture the next day. John told his colleague that he would be at his place to help. This morning, friends of John ask John whether he would like to join them for a beautiful day at the lake. When John considers all the positive and negative aspects of both options (including telling his colleague that he cannot come), he prefers going to the lake more than helping his colleague move furniture. Nonetheless, he chooses to help his colleague move furniture.

Does John's choice make any sense? You might share the intuition that his decision does not seem that far-fetched and unreasonable, even though John makes a decision that potentially goes *against* his own preferences. Thus, this scenario seems to open the intuitive possibility that agents make reasonable counterpreferential decisions under specific circumstances.

Psychologists have discussed various phenomena that bear on the question of whether counterpreferential decisions exist. Widely studied cases of this sort are weakness of will and addiction. In typical cases of weakness of will, an agent has a preference for doing one thing but is moved by her desires to do another, e.g. people might believe that staying faithful to their partner is better than having an affair, but then find themselves being successfully seduced by a third person (for experimental-philosophical approaches to weakness of will, see, e.g. May & Holton, 2012; Mele, 2010; Messerli et al., n.d.). People who have an addiction usually have considerably less control over their behavior than weak-willed people. Berridge and Robinson (2016) capture the putatively counterpreferential nature of addiction and food reward by distinguishing 'wanting' from 'liking'. The processes of 'wanting' (i.e. the incentives that motivate us) and 'liking' (i.e. the pleasantness of a stimulus) often work together, but they can come apart, and have been shown to be encoded in different brain areas.

These phenomena are different from our case under investigation. In the cases discussed above, people may behave counterpreferentially because they are driven by desires like craving for food, sex, or drugs, which often move a person unconsciously (Berridge, 1996). Thus, the actual trigger of the putative counterpreferential choice is very different from our case, in which having made a commitment is the decisive factor. Arguably, investigating the possibility of counterpreferential

CONTACT Michael Messerli michael.messerli4@uzh.ch, kevin.reuter@uzh.ch

choices that are based on commitments can show that counterpreferential decisions are *rational* choices.[1]

So, can commitments lead to counterpreferential choices? The economist and philosopher Amartya Sen (1977) has most famously argued for a positive answer to this question. He lists a number of reasons that might lead people to make counterpreferential decisions: People are committed to being truthful when talking to others, committed to being polite, and committed to uphold their moral obligations, among others. When those commitments are at stake, people might well decide in favor of a less-valued option in order to satisfy the commitments they have made. Hausman (2007) believes economists *should not* model preferences in a way that commitments can lead to counterpreferential choices.[2] Instead, he argues for conceiving of preferences as *total comparative evaluations* which, by definition, include commitments that have been made in the past. At the same time, Hausman is explicit in stating that the ordinary concept of preference may allow for commitments to cause counterpreferential choices (2007, pp. 3–4, 35, 62–64). His claim rests on a simple intuitive observation (Hausman, 2007, p. 3). According to Hausman, it sounds more natural to say 'Jill drank water rather than wine with dinner, despite preferring to drink wine, because she promised her husband she would stay sober.', than to say 'Jill drank water with dinner because she preferred to do so. But for the promise she made her husband to stay sober, she would have preferred to drink wine rather than water with dinner.' We agree, but this hardly provides any substantial evidence that commitments are conceived to potentially go against all-things-considered preferences, as ordinarily understood.

As far as we know, nobody so far has *directly* tried to settle the question of the possible existence of commitment-based counterpreferential decisions empirically. While there is substantial empirical research on commitments (see, e.g. Charness & Dufwenberg, 2006; Kiesler & Corbin, 1965; Kiesler & Sakumura, 1966; Michael et al., 2016; Székely & Michael, 2018; Vanberg, 2008), this branch of research does not focus on whether commitments can lead to counterpreferential choices. Instead, appeals to intuition do most of the work in the literature that examines the relation between preferences and commitments (Cudd, 2014; Hausman, 2007; Heath, 2008; Herfeld, 2009; Peacock, 2020; Peter & Schmid, 2007; Pettit, 2007; Reiss, 2013; Sen, 1977).

This paper takes on the task of providing empirical results in support of the claim that commitments can lead to counterpreferential choices. Here is how we will proceed: In Section 2, we expound on how the notions of preference and choice are understood for the purposes of this paper, and state the hypothesis we empirically investigated. In Section 3, we present Study 1, the results of which suggest that commitment-based deliberations can lead to counterpreferential choices. Study 2 and Study 3 (in Section 4) respond to two objections regarding the significance of our results from Study 1. We challenge those objections empirically. In Section 5, we summarize our results and briefly discuss the reasonableness of commitment-based counterpreferential choices.

2. Preference and choice

Our investigation presupposes some technical qualifications. Specifically, we need to be clear on how the notions of preference and counterpreferential choice are interpreted. Obviously, we do not use a behaviorist conception of preference, that is, preferences are not explicated as choices from choice sets. We rather use a mentalist conception of preferences. Accordingly, preferences are understood as mental states expressed by psychological judgments. These judgments express value rankings after all things relevant for the ranking of alternatives have been considered. Furthermore, the rankings are interpreted on an ordinal scale and the concept of value or utility is used as in modern rational choice theory. We can put this more specifically in the following way:

Preference

A person prefers x *if she judges* x *best, all-things-considered.*

This characterization is equivalent to how Hausman illustrates the notion of preference:' to say that Jill prefers *x* to *y* is to say that when Jill has thought about everything she takes to bear on how much she values *x* and *y*, Jill ranks *x* above *y*' (Hausman, 2012, p. 34).[3] Given such a characterization, we can now state what it means to make a counterpreferential choice:

Counterpreferential Choice

A person makes a counterpreferential decision if that decision goes against that person's preference.

Here, preferences are interpreted in the way defined above, i.e. an all-things-considered judgment about what is best. We believe that Hausman's conception is useful as an operationalization of *preference* for two reasons: First, Hausman's conception takes a clear stance in favor of *stated preference* compared to *revealed preference*, thereby opening a gap between *preference* and *choice*, such that counterpreferential decisions become at least a theoretical possibility.[4] Second, Hausman's concept of preference is specifically effective because it illustrates why choosing a worse option apparently does not make sense within a rational choice perspective: Only by following her preference ordering will an agent do what she believes is best for her.[5]

Importantly, our characterization of preference does not allow us to distinguish between what Hausman calls 'total comparative evaluations' and 'overall comparative evaluations'. Only in the latter case, 'people regard some of the factors that affect their evaluation of alternatives as *competing with* preferences rather than as influencing preferences' (Hausman, 2007, p. 3). It is part of our investigation to find out whether participants consider commitments to be competing with preferences, or whether they take commitments to be influencing decisions only via preferences. Such an investigation is only feasible, if we leave it open whether the ordinary concept of preferences aligns with 'total comparative evaluations' or 'overall comparative evaluations'.

In this paper, we aim to provide evidence for the claim that commitments can lead to counterpreferential choices. Given our methodological approach, we do, however, empirically investigate a weaker null hypothesis, which can be stated as follows:

Hypothesis: *Imagined commitments do not lead people to favor counterpreferential choices.*

The results we present below provide evidence that this hypothesis is wrong. In other words, the data we collected suggest that imagined commitments can trigger people to favor counterpreferential choices, and thus, that they can influence their decisions independently of preferences – as understood by laypeople.[6] We will argue that this result speaks in favor of the claim that commitments can lead to counterpreferential choices. However, this conclusion will be somewhat restrained, given that we only measure people's ratings and judgments but not actual decisions (see also Section 5). In principle, we can test this hypothesis by a wide variety of cases but we do expect that in *standard* cases, people will not make counterpreferential choices. Instead, there will need to be very specific reasons for why an agent should even contemplate a counterpreferential decision. We believe that the example of John, as mentioned in the introduction, might be a case that lends itself to the investigation of the possibility of commitment-based counterpreferential choices.[7]

3. Empirical study 1

One hundred and twenty participants were recruited on Amazon Mechanical Turk and paid a small fee for their participation; 5 participants were excluded for not having completed the survey. The remaining 115 participants (51 women, $M_{age} = 39.09$, $SD = 15.69$) all indicated that they were native English speakers. All participants were randomly assigned to one of three conditions, two test conditions (*Acquaintance*, *Colleague*) and one control condition. The vignettes of the two test conditions read as follows:

Test condition: Imagine that an acquaintance (a colleague) of yours asks you whether you would help him move some furniture and household appliances into his new apartment. You agree to be at his place at 10am the next day. The next morning, it is a beautiful warm summer day. At 8 am you get a call from friends who ask you whether you would like to join them for a nice day at a lake. All things considered and independent of how you decide in the end, how do you value each of the two options:

- Spending the day at the lake and tell my acquaintance (colleague) that I cannot come.
- Moving furniture and household appliances and tell my friends that I cannot join them.[8]

The control condition was designed as a contrast case in which only preferences (should) determine the decision. The vignette for the control condition read:

Control condition: Imagine that you plan your yearly holidays. On the one hand, you could go to the seaside and spend a week relaxing at the beach. On the other hand, you could book a trip to a city you have not seen before and experience some cultural highlights. All things considered and independent of how you decide in the end, how do you value each of the two options [9]:

- Spending the holidays on the beach and not going to a city.
- Spending the holidays in a city and not going to the beach.

After the participants rated both options, they were then directed to the second question reading:

Decision Question: You have just valued each of the two options. But how do you decide in the end? Please tell us what you will do:

For the two test conditions, the participants were presented with two options: (1) I choose to go to the lake and tell my acquaintance (colleague) I cannot come. (2) I choose to move furniture and household appliances and tell my friends I cannot join them. In the control condition, the options were: (1) I choose to go to the seaside and spend a week relaxing at the beach. (2) I choose to go to a city and experience some cultural highlights. Participants had to choose which decision they would take.

After the two main questions, people were also asked two further questions. First, they were prompted to give an explanation for the decision they had made. Second, we asked participants the following question: 'Lastly, do you think it is reasonable to sometimes choose an option that you consider to be worse than another?' People answered on a 7-point Likert scale anchored at −3 meaning 'Not at all reasonable' and 3 meaning 'Totally reasonable'.

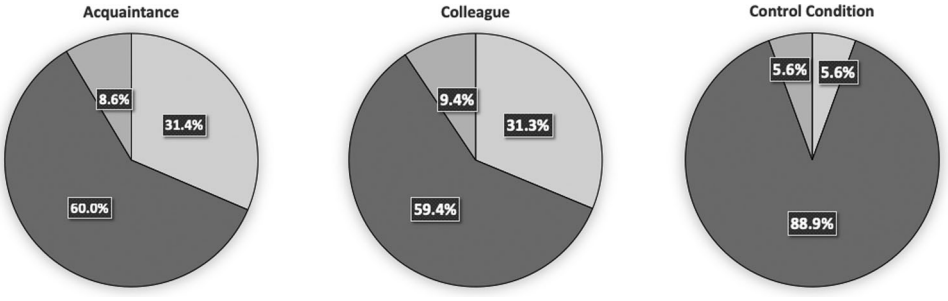

Figure 1. Responses in % to the two test conditions and the control condition. Dark gray depicts the percentage of participants who would decide in line with their preferences. Light gray represents the percentage of participants who would decide against their preferences. A few participants (medium gray) indicated an equal preference for both options.

3.1. Results

The results of people's responses are summarized in Figure 1. Almost all participants (67 out of 77) in the two test conditions chose to help their acquaintance or colleague to move furniture. In both the *Acquaintance* as well as the *Colleague* condition, around 31% of the participants (21 out of 67) who stated that they would decide in favor of helping to move furniture, considered going to the lake more valuable. The response profiles were significantly different between the test condition *Acquaintance* and *Control*, $\chi^2 = 8.70$, $p = .013$, as well as *Colleague* and *Control*, $\chi^2 = 8.64$, $p = .013$. Of all the 21 participants who indicated counterpreferential decisions, only a single person considered it unreasonable to sometimes choose an option that is worse than another. The average result of the participants in regards to the 'unreasonable'-question was 1.62 ($SD = 1.24$).

3.2. Discussion

The results indicate that a substantial portion of the participants in the test conditions would decide in favor of a less valuable option when they consider the scenario we presented them with. Simply put, the results suggest that we tested a situation in which many people will decide against their own preferences. The study was not designed to investigate which percentage of the population are likely to make a decision against their preferences. Obviously, the scenarios were quite specific and for many the situation did not even present them with a 'difficult' choice. Thus, it is likely that for many more people than just the recorded 31%, there exist choices in which option *x* is preferred but they still decide in favor of option *y*. Of course, for most decisions, people's choices will nicely align with their preferences. In fact, the control condition was specifically designed as a base rate for decisions in which preferences alone determine the decision in question. The significant differences between the test conditions and the control condition demonstrate, however, that not all decisions are like that. Other factors may determine which option we are going to choose.

While the data seems to put pressure on those who deny the existence of counterpreferential decisions, one might question how we can be certain that our experiment delivered reliable responses? As stated above, we asked the participants of our experiment to explain their responses. Importantly, when we look at individual responses from those participants who suggested making counterpreferential decisions, many explanations reflect awareness of the very fact that they would make a counterpreferential choice. For example, participants stated that 'I would rather go to the lake, but I already made a commitment.', 'I may rather go to the lake but I wouldn't. Once I make a commitment I keep it.', 'I made a promise to help the friend move first, and I'm not going to break that promise even though I would rather go to the lake.', 'I don't want to break the promise I already made to my friend to help him move -- even if it's not fun, it's the right thing to help them.' These answers clearly suggest that many participants did not randomly indicate a counterpreferential choice but instead were aware of the tension between their preferences and the commitment they had made.

Before we discuss our hypotheses in light of the recorded data, let us first address potential worries against our study. In order to counter these objections, we present the results of two further studies.

4. Objections

4.1. Objection 1: the concept of value

The experiment reveals a potential decision against one's preferences, only if people gave all-things-considered value ratings when considering their options. It is indeed possible that some people merely considered the positive value of spending a day on the lake without considering the negative

value of telling one's acquaintance or colleague that one is not available for moving after all. If that were the case, then it would not surprise to see decisions made against one's *rated* preferences.

We do not believe, however, that this is a likely possibility. When we asked the participants to rate the value of the options, we specifically named the main positive as well as the main negative aspect of the choice, e.g. one of the options read: 'Spending the day at the lake and tell my colleague that I cannot come.' However, one might insist on the ambiguity of the term 'value'. Within the context of our investigation, the concept of value is understood in terms of an agent's ends and desires. Accordingly, if I judge that *a* is more valuable than *b*, then I believe that *a* is more valuable than *b* in terms of my ends and desires.[10] Admittedly, we have not ensured that participants entertain this understanding of value. Instead, they might have used a watered-down notion that is not as demanding, e.g. perhaps many participants have only considered short-term pleasures. If this were the case, we cannot conclude that participants do in fact decide counterpreferentially.

This objection should be taken seriously. We have, therefore, conducted a second experiment where we first explained to participants which concept of value is involved. We will see that our results are robust, even if we change the experimental setting in this way.

In order to address the objection stated above, we decided to rerun both test conditions (*Acquaintance*, *Colleague*) to see whether the results would change or remain robust. If the objection is correct, then we should see a substantial drop in the percentages of people who indicate decisions that go against their preferences.

4.1.1. Methods

One hundred participants were recruited on Amazon Mechanical Turk and paid a small fee for their participation; two participants were excluded for not having completed the survey. The remaining 98 participants (48 women, M_{age} = 36.92, SD = 12.38) all indicated that they were native English speakers. All participants were randomly assigned to one of two conditions (*Acquaintance*, *Colleague*). The vignettes of the two conditions were exactly the same as the vignettes of the test conditions in Experiment 1 with one exception: after participants had given their consent to this study, they were informed about the task ahead in the following manner.

Instructions: On the next screen we will ask you to value certain events. Before you do so, please consider the following example: Imagine you have to value a one-week trip to Europe. On the positive side, there might be aspects like relaxing, eating new and exciting food, being able to tell your friends of an amazing trip when you are back, etc. On the negative side, there might be aspects like being jetlagged, longing for your loved ones at home, missing an important meeting at work, etc. Thus, if you value an option or an event, you take into account all its positive and negative aspects and then make an overall judgment.

After these instructions, participants rated both options (see Experiment 1), and then answered the decision question (see also Experiment 1). Subsequently, the participants were prompted to provide an explanation for why they had made their respective decision.

4.1.2. Results

In the *Acquaintance* condition, 36.6% of the participants who decided in favor of helping to move furniture considered going to the lake more valuable. In the *Colleague* condition, 26.3% of the participants who decided in favor of helping to move furniture considered going to the lake more valuable. The results of people's responses are summarized in Figure 2.

4.1.3. Discussion

The data we received in Experiment 2 are highly similar to those we collected in Experiment 1. While the percentage of people who decided against their preference in the *Acquaintance* condition rose from 31.4% to 36.6%, the percentages in the *Colleague* condition decreased from 31.3% to 26.3%. Thus, overall the results of Experiment 1 were very robust. It seems therefore very likely, that the

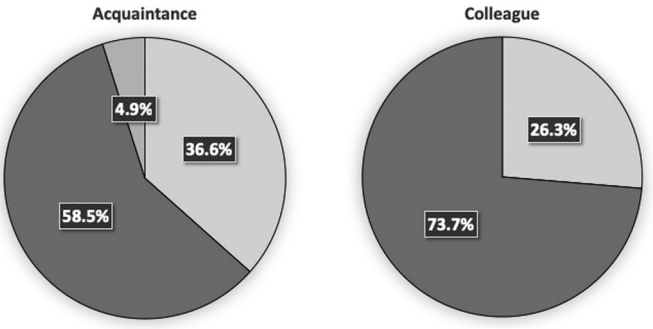

Figure 2. Responses in % to the two conditions. Dark gray depicts the percentage of participants who would decide in line with their preferences. Light gray represents the percentage of participants who would decide against their preferences. A few participants (medium gray) indicated an equal preference for both options.

participants in Experiment 1 entertained a notion of *value* not only similar to the one in Experiment 2, but also of the right kind.

4.2. Objection 2: wording

In the previous section, we discussed and countered the objection that in the design of our study ordinary people may have understood the notion of value differently. A further objection can be raised against the wording of the answer options that we presented participants with. More specifically, one might worry that an option like 'Spending the day at the lake and tell my colleague that I cannot come ', sounds far too benign for what is in reality the breaking of a promise. Thus, the objector might suggest using alternative wordings like 'Spending the day at the lake and break my promise to my colleague'. The reason why we opted for a different way to tackle the objection is that 'breaking a promise' or 'breaking a commitment' is a very negative trigger. The wording 'tell my colleague that I cannot come' is relatively neutral in this regard. However, we agree that the empirical evidence for decisions against preferences would be greater if the negative aspects of a decision would be highlighted even further. Thus, in the third experiment, we investigate the impact of the exact wording on our results. We will see that the empirical effect does not depend on the exact wording of the vignettes.

4.2.1. Methods

One hundred and seventy participants were recruited on Amazon Mechanical Turk and paid a small fee for their participation; 10 participants were excluded for not having completed the survey or for indicating that English was not their native language. The remaining 160 participants included 67 women and were on average 37.50 ($SD = 11.28$) years old. All participants were randomly assigned to one of three conditions (*telling*, *cancelling*, *breaking*). The vignettes of the three conditions were almost the same as the vignettes of the acquaintance condition in Experiment 2, i.e. we included the instructions page but focussed on *Acquaintance*. We only manipulated the response options to the value question as well as the decision question. In the previous two experiments, we gave the participants the following options:

- Spending the day at the lake and tell my acquaintance that I cannot come.
- Moving furniture and household appliances and tell my friends that I cannot join them.

To test the influence of the wording of the response options on people's value ratings and decisions, we opted for three different versions of the first option.

- (telling) Spending the day at the lake and tell the person that I cannot come.
- (cancelling) Spending the day at the lake and cancel the arrangement to move furniture.
- (breaking) Spending the day at the lake and break my commitment to help moving furniture.

Arguably, whereas the first option is fairly neutral and raises the worry that people do not really consider the impact of their decision, the latter two make it very clear that a decision to go to the lake with friends, means to *cancel the arrangement* and *break a commitment*.

4.2.2. Results

In the *telling* condition, 23.1% of the participants who decided in favor of helping to move furniture considered going to the lake more valuable. In the *cancelling* condition, the numbers were even higher, with 38.1% of the participants indicating a counter-preferential decision. In the *breaking* condition, the numbers were slightly lower, with 25.5% of the participants who decided in favor of helping to move furniture considered going to the lake more valuable. The results of people's responses are summarized in Figure 3. The response profiles were not significantly different between the three conditions: *telling* and *cancelling*, $\chi^2 = 2.51$, $p = 0.286$, as well as *telling* and *breaking*, $\chi^2 = 3.17$, $p = .205$.

4.2.3. Discussion

The data we received in Experiment 3 are highly similar to those we collected in Experiment 2. The percentage of people who decided against their preference in the *telling* condition was less than in Experiment 2, then 36.8%, now 23.1%. However, in the *cancelling* condition, the percentages were fairly high: 38.1%. In the, arguably, most extreme wording condition, i.e. *breaking*, the results were in the same region as in the *telling* condition. Thus, again, overall the results of Experiment 1 were very robust. Wording did not have a significant effect on the actual results of our experiment.

4.3. Objection 3: the double-counting-problem

As briefly described in Section 2, our operationalization of preference is ambivalent between 'total comparative evaluations' – which include commitments – and 'overall comparative evaluations' – which allow commitments to influence decisions independently of preferences. The results suggest that the ordinary concept of preference aligns more strongly with 'overall comparative evaluations'. However, some people explained their responses by stating that they valued going to the lake less than helping to move furniture, because they had made a commitment to their neighbor or colleague. Thus, for these people, it seems commitments influenced their decisions via their preferences.

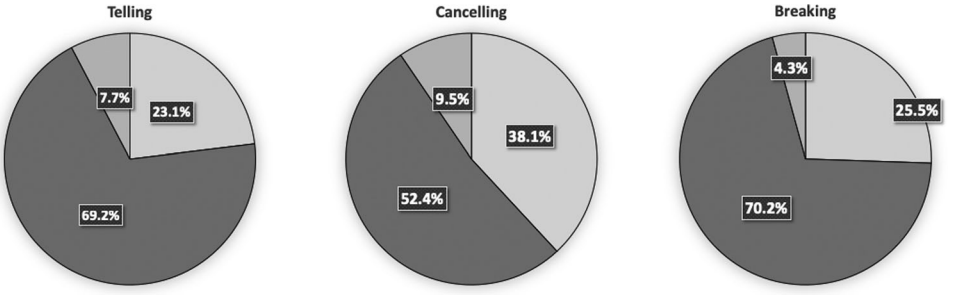

Figure 3. Responses in % to the three conditions. Dark gray depicts the percentage of participants who would decide in line with their preferences. Light gray represents the percentage of participants who would decide against their preferences. Some participants (medium gray) indicated an equal preference for both options.

If commitments can act on preferences as well as directly on decisions, then we face a problem of double-counting: Some participants might have considered commitments both do be competing with preferences as well as influencing preferences (i.e. the commitment is part of the preference ranking).[11] If that were the case, then it looks as if we cannot be sure that we have correctly identified the amount of people who are willing to make counterpreferential choices. Furthermore, it seems our studies are not well designed to single out and measure the importance of a commitment.

Such an objection, however, misses that commitments may enter the choice-deliberating process in two different ways. First, our empirical results suggest that commitments function as *constraints* for some participants. The qualitative responses we received indicate that these people think that commitments should not be weighed against other evaluative considerations, such as preference rankings, in a kind of weighing deliberation. Second, breaking a commitment is likely to feel bad for most people. Thus, commitments, or rather the breaking of commitments, also function as *negative unpleasant states* that influence people's preference ranking (some participants who stated both a preference and a decision in favor of moving furniture, often claimed that their prior commitment was an important factor in their deliberation. It seems quite likely – even if we cannot directly infer it from the data – that these participants referred to this negative unpleasant state). Consequently, even if people have considered the importance of commitments both for their preference as well as independently of their preferences, this does not mean that commitments have been counted twice. Admittedly, more work needs to be done to investigate this dual aspect of commitments in decision-making processes.

5. General discussion

5.1. Summary of the results and philosophical implications

The results of the three experiments we conducted and presented in Sections 3 and 4 suggest that a substantial number of people are willing to favor counterpreferential decisions when faced with the lake-or-furniture case. Compared to a control condition, in which people had to indicate a decision between going on a beach holiday or going on a city trip, the amount of stated counterpreferential decisions was significantly higher and robustly around the 30% mark. The specific wording of the vignette as well as clarifying in greater detail the notion of value had no significant effect on people's responses. In Section 2, we stated the empirical null hypothesis of our investigation ('Imagined commitments do not lead people to favor counterpreferential choices.') and stated the aim of providing evidence for the claim that commitments can lead to counterpreferential choices.

The experimental data from Studies 1–3 seem to support the claim that commitments can lead to counterpreferential choices. People's willingness to accept our conclusion will, however, largely depend on whether they agree with the operationalizations we used in those experiments. Some scholars will simply not accept the way we defined a counterpreferential choice, i.e. as a choice that goes against one's *stated* all-things considered preference. Hence, advocates of revealed preference theory will claim that we have not measured any preferences independent of decisions.

While we think that many readers will accept that preferences can indeed be measured by peoples' value ratings, a related objection can be mounted against the way we have operationalized people's choices. Instead of having measured real decisions, we merely asked people to *state* their decisions. Thus, the objection goes, participants are not actually making a decision but provide inconsequential responses after reading abstract descriptions of options. Accordingly, the gap between participants' rated decisions and their real behavior is too large to claim with any confidence that the results have any clear bearing on the claim under investigation. It is certainly correct that some studies have shown an inconsistency between people's rated decisions and real behavior (e.g. in moral contexts). However, a variety of empirical studies have also shown high consistency between people's ratings and their behavior (e.g. in family planning). Thus, we need to inquire whether there are any specific reasons for why people should give responses that deviate

from their actual behavior. One reason why some people might have given wrong indications of their actual behavior is that they do not want to be perceived as people who would not keep their promises. However, if that were the case, we would expect participants to also value moving furniture more strongly than going to the lake.

The outcome of our studies supports Sen's positions on the status of commitments for decision-making processes. According to Sen, narrow self-interest and sympathy directly affect a person's own welfare and should be reflected in people's value judgments. But self-interest and sympathy are not the only factors that influence people's decisions. Sen (1977, p. 326) characterizes *commitments* as altruistic attitudes towards others: A person who acts out of commitment chooses an option that she considers the right thing to do, even if that option is less preferable than an alternative.[12] As Peter and Schmid nicely sum up Sen's idea, '(…) one feels compelled to intervene in a certain matter, even if doing so leaves one worse off. What matters, however, is that increasing one's welfare is not the central motive ' (Peter & Schmid, 2007, p. 4).[13]

While our results indicate that preferences, as ordinarily understood, can run counter to people's decisions, what are the consequences for both revealed preference theory and stated preference theory? For the purpose of this paper, it is our understanding of revealed preference theory (hereafter RVT) that a preference can be defined in terms of a choice (if the choice behavior satisfies WARP).[14] Simply put, preferences derive from choices (see, e.g. Bernheim & Rangel, 2009). Our results on imagined commitment-based decisions indicate that preferring, as understood by laypeople, is not revealed by choosing. So, our results suggest that the folk concept of preference has a different meaning compared to the technical definition of preference within RVT. Advocates of RVT will not be worried about our results given the stipulative definition of preference they favor. That said, our results can be interpreted to re-iterate some of the critical points that have been voiced against RVP.[15]

In our paper, we have been explicit that we operate within the *stated preferences* framework. Our results are relevant for this framework in two important ways. First, even if theorists will not be surprised by the results of the experiments, it is an important desideratum of every framework that the concepts in use are empirically well-founded, i.e. that intuitions about the meaning of the central concepts used are representative of those people who use them on an everyday basis. Thus, our research provides support to those who allow for and factor in counterpreferential decisions. Second, researchers who work on social preferences need to be aware that they can improve their framework by including both preferences and commitments as predictors of choices. For example, to make better predictions of agents, we not only need to know which preferences agents have, but also which commitments they have made that potentially go against their preferences.

5.2. The rationality of commitment-based counterpreferential choices

Our empirical results indicate that many people make counterpreferential decisions when they have made a commitment that goes against their preferences. But can we consider these decisions to be rational? Rational choice theorists are likely to classify counterpreferential decisions as unreasonable or irrational; after all, such decisions go against what people actually prefer. Most of the choices that are usually considered in this regard are cases such as weak-willed decisions and decisions based on one's addictive behavior. And we do not deny that such counterpreferential choices are irrational: If a counterpreferential choice is based on weakness of will, for example, the decision is likely to be irrational (see also Reiss, 2013, p. 35).

The lake-or-furniture scenario is very different, however. It does not seem to be the case that the agent values *x* more than *y*, but 'lower' desires move the agent to do *y*. The person's judgment seems rather *well-considered* when acting against their most valued option. In fact, we have highlighted that participants justify their counterpreferential choice by stating the commitment they had previously made. Crucially, in the first experiment, we have additionally asked participants to tell us

how reasonable they believe it is to choose an option that they consider to be worse than another. The average ratings were way above the midline, indicating that people do not consider counterpreferential decisions that are based on commitments to be unreasonable.

Of course, this claim has to be taken with a grain of salt. Not only should we be careful to equate *unreasonable* with *irrational*, it is also invalid to move from people's subjective assessment to a claim about the irrationality of their decisions. Nonetheless, if people overwhelmingly consider their choices to be reasonable, then certainly more work needs to be done to uphold the claim that *all* counterpreferential decisions are irrational.

6. Conclusion

Our experiments were designed to test the possible existence of counterpreferential choices by pitching an option that is very valuable to most of us – like spending a nice day at a lake with friends – against another option that includes rather unpleasant experiences but also a *commitment* that the agent has made towards an acquaintance or colleague. We have conducted three experiments, the results of which strongly suggest that many people make decisions against their preferences. Crucially, such choices may be understood as decisions based on counterpreferential commitments. These results are important because they demonstrate that stated preferences are not a good predictor of choices when commitments are in play.

Notes

1. Researchers have also started to investigate various aspects of promise-keeping. This literature has shown that the exchange of promise has an important influence on cooperation in games (see, e.g. Bicchieri & Lev-On, 2007; Charness & Dufwenberg, 2006; Sally, 1995).
2. Lehtinen summarizes:

   > Hausman argues that considerations such as commitments should be included in the preferences rather than taken to compete with them. Sen argues for counter-preferential choice in his account of commitment. […] Such a choice is counter-preferential because the link between the person's choice and her preferences is broken. (2013, p. 207)

3. For a substantive criticism, see, e.g. Angner (2018).
4. Luce and Raiffa (1957, p. 50), for instance, claim that a statement according to which a person chooses the option that leads to the most preferred outcome is tautological. This is not quite right (see, e.g. Choi et al., 2014; Varian, 1982). Revealed preferences can be contradictory: If a subject chooses *x* over *y* and *y* over *x*, then choices cannot be in line with preferences.
5. Note that 'best for her' does not imply egoism, maximizing pleasure or the like. Take John's case. John may choose to honor his promise despite the fact that spending the day on the lake is more pleasurable. All-things-considered preferences address this issue head-on. John's all-things considered preference, for example, involves his own enjoyment but also the value of keeping his promise. We will see that our experiments are designed in a way that 'value' is not reduced to personal enjoyment.
6. Favoring a counterpreferential choice can be understood roughly as being disposed to make a counterpreferential choice.
7. Such decisions are also known as deontically constrained. A deontic constraint can be understood as a form of duty or rule following behavior to refrain from the pursuit of individual advantage (see, e.g. Heath, 2008).
8. Both options were presented in randomized order and participants were asked to rate the value of each option on an 11-point Likert scale anchored at 0 meaning 'Not at all valuable' and 10 meaning 'Extremely valuable'. As mentioned in the theoretical section, the strength of a value judgment can be understood in purely ordinal terms, i.e. the experiment is perfectly consistent with an ordinal interpretation.
9. The phrasing 'independent of how you decide in the end' is potentially problematic. We decided to include this phrase, because we did not want participants to pragmatically infer that the question aims at stated decisions.
10. In accordance with rational choice theory, we do not make any proposal concerning the content of these ends. This means that we recognize no distinction between goals such as making a million dollars, helping other people and being a sadist. Also note that there are no implications regarding risk-taking. The value judgment that *a* is more valuable than *b* might be risk-neutral such as in standard approaches or risk-averse such as in *prospect theory*.

11. We are grateful to an anonymous reviewer of this journal for pointing out this problem to us.
12. Sen has refined his view on commitments over time. Sen (1985) argues that commitments play a crucial role when violating *self-welfare goal*; see, e.g. Peacock (2020, chapter 7) for a discussion on Sen's view on commitments, self-welfare goal, self-centered welfare and self-goal choice.
13. Some researchers have started to question Sen's depiction of commitments as factors that may have a motivating force beyond expected advantage. Pettit (2007), for instance, believes that Sen's notion of commitment involves putting aside one's own goals. He considers this to be implausible, since pursuing one's goals is central to the concept of action (for a critical reply, see, e.g. Schmid, 2007; for further discussion on this issue, see, e.g. Cudd, 2014). Contra Pettit, Sen (2007) claims that acting out of commitment does not involve pursuing another person's goals. Instead, a person acts in a way that the other person can fulfill her desires and goals. Our data seem to show that moving furniture is indeed a reasonable restraint on one's behavior in this sense.
14. The so-called weak axiom of revealed preference theory (WARP), which can be traced back to Samuelson's (1938) work, specifies consistency restrictions on consumer choices based on observable quantities such as prices. Although Samuelson aimed to eliminate the concept of preference, many scholars argue that preferences are part of RVT. Sen (1973), for instance, argues that RVT should be interpreted in a way that an agent's choice reveals her preferences. The idea is that if two options, x and y, fall within an agent's budget set and the agent chooses x, then this choice reveals that she prefers x to y.
15. To name just two of those critical points. First, while RVT does not distinguish preferences from choices, preferences are considered to be causes of an agent's behavior inside the standard folk psychological theory. Second, communicating the predictions and assumptions is made difficult when concepts are used differently from how they are understood by the folk.

Acknowledgments

We would like to thank Dan Hausman, Catherine Herfeld, Jay Jian, Don Ross, Andrija Soc, Pascale Willemsen, as well as three anonymous reviewers for the *Journal of Economic Methodology* for their very helpful comments on previous versions of the manuscript. Thanks also to the audiences at the 'Soul of Economics' Conference and the 'Preferences, Commitments and Choice' Conference at which this paper was presented. We also very much appreciate the support of the guest editors Catherine Herfeld, Chiara Lisciandra and Carlo Martini.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Kevin Reuter was funded by an SNF Eccellenza Grant [PCEFP1_181082]. Michael Messerli was funded by an SNF Ambizione Grant [PCEFP1_186151].

Notes on contributors

Michael Messerli is currently Ambizione Fellow at the Ethics Research Institute, University of Zurich. He studied Philosophy and Economics at the University of Bern and obtained a PhD in Philosophy and Economics in 2017 at the LMU Munich. Afterwards, he joined the Department of Philosophy at the University of Sheffield (UK). In 2018, he was Adam Smith's Guest Professor in Bayreuth.

Kevin Reuter is currently SNSF Eccellenza Professor at the Institute of Philosophy, University of Zurich. Before that he worked as a Lecturer and Post-Doctoral Researcher at the Institutes of Philosophy in Bern and Bochum. He holds a PhD in Philosophy from the University of London, Birkbeck College and a MA in Physics from the LMU Munich.

ORCID

Michael Messerli http://orcid.org/0000-0002-2549-4559
Kevin Reuter http://orcid.org/0000-0003-2404-1619

References

Angner, E. (2018). What preferences really are. *Philosophy of Science*, *85*(4), 660–681. https://doi.org/10.1086/699193

Bernheim, D., & Rangel, A. (2009). Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *The Quarterly Journal of Economics*, *124*(1), 51–104. https://doi.org/10.1162/qjec.2009.124.1.51

Berridge, K. (1996). Food reward: brain substrates of wanting and liking. *Neuroscience & Biobehavioral Reviews*, *20*(1), 1–25. https://doi.org/10.1016/0149-7634(95)00033-B

Berridge, K., & Robinson, T. (2016). Liking, wanting, and the incentive-sensitization theory of addiction. *American Psychologist*, *71*(8), 670–679. https://doi.org/10.1037/amp0000059

Bicchieri, C., & Lev-On, A. (2007). Computer-Mediated communication and cooperation in social dilemmas: An experimental analysis. *Politics, Philosophy and Economics*, *6*(2), 139–168. https://doi.org/10.1177/1470594X07077267

Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, *74*(6), 1579–1601. https://doi.org/10.1111/ecta.2006.74.issue-6

Choi, S., Kariv, S., Müller, W., & Silverman, D. (2014). Who is (more) rational? *American Economic Review*, *104*(6), 1518–1550. https://doi.org/10.1257/aer.104.6.1518

Cudd, A. (2014). Commitment as motivation: Amartya Sen's theory of agency and the explanation of behaviour. *Economics and Philosophy*, *30*(1), 35–56. https://doi.org/10.1017/S0266267114000030

Hausman, D. (2007). Sympathy, commitment, and preference. In F. Peter & B. Schmid (Eds.), *Rationality and commitment* (pp. 49–69). Oxford University Press.

Hausman, D. (2012). *Preference, value, choice, and welfare*. Cambridge University Press.

Heath, J. (2008). *Following the rules. Practical reasoning and deontic constraint*. Oxford University Press.

Herfeld, C. (2009). The motive of commitment and its implications for rational choice theory. *Analyse & Kritik*, *31*(2), 291–317. https://doi.org/10.1515/auk-2009-0206

Kiesler, C., & Corbin, L. (1965). Commitment, attraction, and conformity. *Journal of Personality and Social Psychology*, *2*(6), 890–895. https://doi.org/10.1037/h0022730

Kiesler, C., & Sakumura, J. (1966). A test of a model for commitment. *Journal of Personality and Social Psychology*, *3*(3), 349–353. https://doi.org/10.1037/h0022943

Lehtinen, A. (2013). Preferences as total subjective comparative evaluations, a review of Hausman's *Preference, value, choice, and welfare*. *Journal of Economic Methodology*, *20*(2), 206–210. https://doi.org/10.1080/1350178X.2013.804743

Luce, R, & Raiffa, H. (1957). *Games and decisions: Introduction and critical survey*. John Wiley.

May, J., & Holton, R. (2012). What in the world is weakness of will? *Philosophical Studies*, *157*(3), 341–360. https://doi.org/10.1007/s11098-010-9651-8

Mele, A. (2010). Weakness of will and akrasia. *Philosophical Studies*, *150*(3), 391–404. https://doi.org/10.1007/s11098-009-9418-2

Messerli, M., Fink, J, & Reuter, K. (n.d.). *The varying rationality of weakness of the will: An empirical investigation and its challenges for a unified theory of rationality* [Manuscript submitted for publication].

Michael, J., Sebanz, N., & Knoblich, K. (2016). The sense of commitment: A minimal approach. *Frontiers in Psychology*, *6* (1968), 1968. https://doi.org/10.3389/fpsyg.2015.01968

Peacock, M. (2020). *Amartya Sen and rational choice. The concept of commitment*. Routledge.

Peter, F., & Schmid, B. (Eds.) (2007). *Rationality and commitment*. Oxford University Press.

Pettit, P. (2007). Construing Sen on commitment. In F. Peter & B. Schmid (Eds.), *Rationality and commitment* (pp. 28–48). Oxford University Press.

Reiss, J. (2013). *Philosophy of economics*. Routledge.

Sally, D. (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, *7*(1), 58–92. https://doi.org/10.1177/1043463195007001004

Samuelson, P. A. (1938). A note on pure theory of consumer's behaviour. *Economica*, *5*(17), 61–71. https://doi.org/10.2307/2548836

Sen, A. (1973). Behavior and the concept of preference. *Economica*, *40*(159), 241–259. https://doi.org/10.2307/2552796

Sen, A. (1977). Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & Public Affairs*, *6* (4), 317–344.

Sen, A (1985). Goals, commitment, and identity. *Journal of Law, Economics and Organization*, *1*(2), 341–355.

Sen, A. (2007). Rational choice: Discipline, brand name, and substance. In F. Peter & B. Schmid (Eds.), *Rationality and commitment* (pp. 339–361). Oxford University Press.

Schmid, B. (2007). Beyond self-goal choice: Amartya Sen's analysis of the structure of commitment and the role of shared desires. In F. Peter & B. Schmid (Eds.), *Rationality and commitment* (pp. 211–226). Oxford University Press.

Székely, M., & Michael, J. (2018). Investing in commitment: Persistence in a joint action is enhanced by the perception of a partner's effort. *Cognition*, *174*(1), 37–42. https://doi.org/10.1016/j.cognition.2018.01.012

Vanberg, C. (2008). Why do people keep their promises? An experimental test of two explanations 1. *Econometrica*, *76* (6), 1467–1480. https://doi.org/10.3982/ECTA7673

Varian, H. R. (1982). The nonparametric approach to demand analysis. *Econometrica*, *50*(4), 945–973. https://doi.org/10.2307/1912771