Revisiting the decoy effect: Replication and extension of Ariely and Wallsten (1995) and Connolly, Reb, and Kausel (2013)

Qinyu Xiao
Department of Psychology, University of Hong Kong, Hong Kong SAR
xqy1020@hku.hk / xqyvincent@gmail.com

Shiyuan Zeng
Department of Psychology, University of Hong Kong, Hong Kong SAR
shiyuanv@connect.hku.hk / vanessazeng21@gmail.com

^Gilad Feldman
Department of Psychology, University of Hong Kong, Hong Kong SAR
gfeldman@hku.hk / giladfel@gmail.com

In press at Comprehensive Results in Social Psychology

Accepted for publication on January 8, 2021

^Corresponding author
Word: abstract – 226, manuscript (excluding references, footnotes, tables, and figures) – 13,270

Corresponding author

Gilad Feldman, Department of Psychology, University of Hong Kong, Hong Kong SAR; gfeldman@hku.hk.

Author bios

Qinyu Xiao is an M.Phil. student at the University of Hong Kong's psychology department. Shiyuan Zeng is an undergraduate student at the University of Hong Kong's psychology department. Gilad Feldman is an assistant professor with the University of Hong Kong's psychology department. His research focuses on judgment and decision-making.

Declaration of conflict of interest

The authors declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Financial disclosure/funding

The authors received no financial support for the research and/or authorship of this article.

Authorship declaration

Qinyu designed the experiment, produced the study materials, wrote the data analysis code, conducted the analysis, and drafted the manuscript and the supplementary materials. Shiyuan did preliminary work in analyzing the target articles and producing the surveys. Gilad guided the replication efforts, supervised each step, and ran data collection.

Additional information

The current replication is part of the larger "mass pre-registered replications in judgment and decision-making" project lead by Gilad Feldman. The project aims to revisit well known research findings in judgment and decision making (JDM) and investigate the replicability of these findings. Further details can be found on: http://mgto.org/pre-registered-replications.

Acknowledgement

We thank Siu Kit Yeung and Jonathan Pettibone for valuable comments on earlier drafts of this article.

Abstract

The decoy effect refers to the phenomenon whereby an inferior, unpreferable option reverses people's preferences and increases the choice share of a targeted option. In two pre-registered experiments with an Amazon Mechanical Turk (MTurk) sample (*N* after exclusion = 1,001), we attempted to replicate Experiment 1 from Ariely and Wallsten (1995) (Study 1) and Experiment 1 from Connolly, Reb, and Kausel (2013) (Study 2). We failed to replicate the original results in Study 1. The observed effects were not in the predicted direction, and their sizes were trivial. We replicated the decoy effect in Study 2, yet with a much smaller effect size than in the original. In addition, we concluded inconclusive evidence for the central hypothesis of the original study that regret salience weakens the decoy effect. We found some indication for a weak reduction, yet our sample size did not provide adequate power to detect this difference. Extending the replication in Study 2, we tested whether making salient the low reversibility of decisions can have a similar impact as regret salience. We again found indication for an effect in the predicted direction, yet the effect was too weak to be detected given our sample size. We discuss potential reasons for the discrepancies between the original and the replication results, as well as the implications. All materials, data, and analysis codes are available at https://osf.io/vsbzk .

Keywords: decoy effect, decision reversibility, regret, attraction effect, replication

Revisiting the decoy effect:

Replication and extension of Ariely and Wallsten (1995) and Connolly, Reb, and Kausel (2013)

Human choice behaviors are susceptible to manipulations of choice settings. A classic example of this is the decoy effect. Known also as the attraction effect (Simonson, 1989) or the asymmetric dominance effect (Ariely & Wallsten, 1995; Huber et al., 1982), the decoy effect refers to the phenomenon whereby an option becomes more attractive due to the presence of another inferior option that, rationally, should not have an influence.[1] Initially demonstrated by Huber and colleagues (1982), the effect is an exemplar of contextual influences on decision-making, and of cases where rationality principles are violated in predictable ways (Ariely, 2008).

In its simplest form, the decoy effect emerges when two competing options are joined by a third option, referred to as a *decoy*, that is constructed to be inferior to one but competes with the other. The option superior to the decoy is commonly referred to as the *target*, whereas the other option is referred to as the *competitor*. The *target* and the *competitor* form a *core choice set*. With a decoy added to the core choice set, the target becomes more attractive and chosen more often than without the decoy. For instance, consider two cars. Car A is expensive but offers great comfort, whereas Car B is relatively more affordable but also less comfortable. With respect to

[1] Judgment and decision-making researchers use "decoy effects" to refer to a class of phenomena where preferences for two competing options are reversed due to the presence of a third option. The third option can be an asymmetrically dominated decoy, as mainly discussed in this article, a compromise decoy (an extreme option that makes its target look like a compromise between it and the other option; Simonson, 1989), or a phantom decoy (an option that dominates its target but is unavailable for choosing; Highhouse, 1996). Sometimes, however, researchers – particularly those whose expertise lies elsewhere – equate the decoy effect with the attraction or asymmetric dominance effect or at least make no explicit distinction (Reb et al., 2018; Stoffel et al., 2019). Also, the decoy effect almost always refers to the attraction effect in popular media. In this article, we did not distinguish between the two since our studies involved only the attraction effect. It also helped with the consistency in terminologies as we pre-registered our studies without making the distinction. Nonetheless, we would like our readers to be aware of the other decoy effects and the differences among them.

affordability and comfort, neither one of the cars is always preferred to the other, because each car excels in only one attribute (i.e., there is no *dominance relationship* between the two cars). If the choice set consisting of A and B now includes a third car, one that is less comfortable and more expensive than A but is still more comfortable than B (i.e., the car is dominated by A but not by B; it is hence called an *asymmetrically dominated* decoy), then the decoy effect predicts a rise in the choice share of A, the option that is targeted by virtue of being superior to the decoy on both attributes.

This article reports two very close replication studies of the decoy effect, targeting Experiment 1 in Ariely and Wallsten (1995) and Experiment 1 in Connolly et al. (2013). We had two clear goals in conducting these studies. The first was to conduct independent close replications of influential findings. The second was to examine an extension concerning the impact of decision reversibility on the decoy effect. We begin with a brief review of the decoy effect literature and the chosen studies for replication. Following that, we highlight our motivation for replicating these two specific studies and outline our extension plan and hypotheses.

Decoy effect: Background, findings, and robustness

Decision-making research has documented many choice behaviors that deviate from predictions of the rational choice theory. In the case of the decoy effect, the *regularity* axiom of rational decision-making is violated (Huber et al., 1982, 2014). The axiom specifies that the choice rate of an option from a certain choice set cannot be higher than that from any subset that includes the option. In other words, we cannot increase the choice rate of an option by expanding (i.e., adding new options to) the choice set to which that option belongs. This axiom was once said to be the only one in the rational choice theory that had not been violated (Luce, 1977), yet the decoy effect is a clear counterexample. Because the decoy is inferior to the target, it is expected to be chosen rarely, if at all. Yet past evidence showed that such an option can result in a substantial increase in the choice share of its target.

Although Huber and colleagues (1982, 2014) intended the decoy effect as a demonstration of how the regularity axiom may be violated, marketing scholars and practitioners have since then researched this effect extensively in the hope of uncovering the underlying mechanisms and utilizing the effect in real-world settings. The experimental paradigm born out of the effect, where choices are characterized as targets, decoys, and competitors, has also been applied extensively in not only behavioral economics and psychology but also many other fields, such as politics (e.g., Herne, 1997), public health (e.g., Stoffel et al., 2019), and human resources (e.g., Keck & Tang, 2020; Reb et al., 2018). Despite the widespread influence of the effect, no consensus, however, has been reached on the underlying mechanisms (Simonson, 2015; Sivakumar, 2016). Moreover, scholars have recently even started a debate about the robustness of the effect.

For over three decades, researchers tested the decoy effect with a wide range of option types, attributes (e.g., consumer goods; Ariely & Wallsten, 1995; jobs; Connolly et al., 2013), and modes of presentation (e.g., descriptive texts, tables, or pictures; Connolly et al., 2013; Yang & Lynn, 2014), as well as in different contexts (e.g., individual or collaborative decision-making; Slaughter et al., 2006). Some studies observed substantial influences of decoys, whereas some others failed to observe appreciable effects, leading to questions on the effect's robustness (for recent reviews, see Gaudeul & Crosetto, 2019; Lichters et al., 2015). In 2014, two researcher groups (Frederick et al., 2014; Yang & Lynn, 2014) conducted large-scale replications of the decoy effect and argued that the effect was confined to highly abstract, numerically presented attribute values. When the attribute values were presented more realistically (e.g., as pictures or in verbal descriptions), the effect was attenuated, diminished, or even reversed in some scenarios. In addition, these researchers showed that some successful demonstrations of the effect in the past failed to replicate, including a famous study that used real subscription plans of *The Economist* (Ariely, 2008; Kivetz et al., 2004), and therefore cast doubt on its robustness and practical significance. Also, a more recent large-scale replication study showed “a precisely zero attraction effect” in realistic choice tasks (Trendl et al., 2018), despite complying with the guidelines that Huber et al. (2014) proposed for successful replications. These findings posed a major challenge to the decoy effect, and the debate is ongoing (for criticisms of the challenge, see Huber et al., 2014; Lichters et al., 2015; Milberg et al., 2014; Simonson, 2014; see also Kaptein et al., 2016; Król & Król, 2019 for attempts to explain the replication failures).

Choice of studies for replication

The mixed evidence for the decoy effect and the debate about its robustness suggest the important of conducting more high-powered direct replications of previous findings, especially considering recent evidence indicating that the literature of psychology and behavioral sciences has been skewed by positive findings and insufficiently powered studies (Open Science Collaboration, 2015; Scheel et al., 2020; Smaldino & McElreath, 2016; Szucs & Ioannidis, 2017). We therefore proceeded to conduct such a replication. We selected Experiment 1 from Ariely and Wallsten (1995) and Experiment 1 from Connolly et al. (2013) as our targets for replication. In what follows, we introduce the designs and results of the target studies and outline the reasons why these specific studies were chosen.

The first target study, i.e., Experiment 1 of Ariely and Wallsten (1995), had participants complete item evaluation tasks where they assigned a certain sum of points to three items of the same category. Participants were asked to assign more points to an item if they find it more appealing relative to the others. Two items out of the three, referred to as Item A and C, formed a core choice set, and neither of them was superior to the other. In other words, there was no dominance relationship between Item A and Item C. The remaining item (Item B) was a decoy that targeted either A or C in two different conditions. If the decoy effect takes place, we expect an interaction between condition and item under evaluation: Item A would receive more points in the condition where it was targeted by Item B than in the condition where Item C was targeted, and vice versa for Item C. Ariely and Wallsten (1995) experimented with five product categories and found significant interactions in three of them.

We chose this study as a replication target because it had three features that distinguish it from many other studies on the decoy effect. First, the items used in this study had three attributes, or were evaluated on three attribute dimensions. In contrast, most studies in the decoy effect literature used options that were evaluated on two attribute dimensions (Lichters et al., 2015), and most theoretical work on the effect has focused on a two-dimensional attribute space (e.g., Kaptein et al., 2016; Sivakumar, 2016). Second, given that the task required participants to rate their preference for all available items, participants were to some extent forced to evaluate the items thoroughly, at least more so than in studies that only required a choice. Also, compared with asking for only a choice, the task design is more capable of quantifying the increase in the preference for targets due to decoys. Third, the decoys used in this study were not strictly dominated by their targets, but only subjectively. Ariely and Wallsten (1995) defined subjective dominance as a perceived relationship between two options, in which some of their differences are noticeable but considered unimportant, while others make one option clearly superior. In other words, subjectively dominated decoys are not strictly inferior to their targets, but their minor advantages are ignored or downplayed due to major disadvantages. Crucially, past evidence has suggested that subjectively dominated decoys suffice for inducing the decoy effect (Ariely & Wallsten, 1995; Huber & Puto, 1983; see also Simonson & Tversky, 1992; Tversky & Simonson, 1993; Wedell, 1991), though most studies on the effect used strictly dominated decoys. These unique features of the target study add to the value of a replication, which would contribute to the discussion about the robustness and generalizability of the decoy effect, especially considering that real-life decisions often involve a broader set of attribute dimensions and drive people to evaluate each option thoroughly before a final choice. We conducted a sensitivity analysis based on the design of the original study (see the supplementary for details) and found that the study was powered at .8 to detect an interaction effect size of 0.374 in Cohen's f. However, the observed effect sizes for the five product categories ranged from 0.120 to 0.563. No significant effect was found for two of the product categories, and only one original effect was above Cohen's f = 0.374 (running shoes). These numbers suggest that the original design lacked power, and consequently, it was possible that (1) there was an appreciable decoy effect, and the non-significant results were false negatives due to insufficient power, (2) there was only a small decoy effect, and the significant results had inflated effect sizes, or (3) the size of the decoy effect just varied across product categories. To arbitrate among these possibilities, we aimed to examine whether the decoy effect could be consistently observed across product categories when the study is sufficiently powered. We planned to power the same design at .99 to detect an f = 0.1, Cohen's (1988) small effect.[2]

[2] We deviated from this pre-registered plan due to a technical error. Please see the Participants section of Study 1 below.

In the second target study, i.e., Experiment 1 of Connolly et al. (2013), participants were asked to decide among three job options, including a decoy, a target, and a competitor. After the decision, they proceeded to imagine that they had chosen each of the three jobs and to evaluate how regrettable and justifiable the imagined choices were if the chosen job turned out to be dissatisfactory. Participants were randomly assigned to two conditions for the above tasks: a control condition, and a condition where the researchers made salient that participants may regret their choices. Regret salience was found to reduce the size of the decoy effect. The rate of choosing the target in the regret-salient condition was not affected by the decoy as much as in the control condition. Additionally, participants in the control condition found the imagined choice of the target to be less regrettable and more justifiable than that of the competitor when the choices did not end up being satisfactory. Such differences, however, were not observed in the regret-salient condition.

This study was chosen as a target because it not only examined the decoy effect but also proposed a simple technique to reduce its magnitude (see also Reb et al., 2018). According to Connolly et al. (2013), regret salience nudged participants towards more vigilant thinking and hence downplayed the nice sounding but shallow reasoning that choosing the target means choosing a winner (Janis & Mann, 1977; see also Simonson, 1989, p. 170). Anticipated regret is ubiquitous in decision-making, and people anticipate regret for some decisions more than others. We thought it worthwhile to examine anticipated regret as a potential moderator for the decoy effect given that studies on the effect have produced mixed evidence thus far. Before doing so, however, we need to first establish that regret salience indeed has an appreciable effect. The target study has provided pilot evidence in this regard. Nonetheless, the study was limited by its relatively low power. The target choice rate was 87% in the control condition and 61.5% in the regret-salient condition. A sensitivity analysis based on the original sample size, however, revealed that the study would be adequately powered (i.e., power = .8) only if the target choice rate in the regret-salient condition was 51.8% or lower. Furthermore, the post hoc power for the study to detect the observed effect size of regret salience was only .64. We therefore would like to explore whether the regret salience manipulation of the original study has a reliable effect, and how large that effect is, when the study design is better powered. Beyond the decoy effect, we believe this replication is also important for assessing the effectiveness of a regret salience manipulation that is also used elsewhere in the judgment and decision-making literature (see e.g., Connolly & Reb, 2012).

Extension – low decision reversibility

In replicating Experiment 1 from Connolly et al. (2013), we also extended the study by including an additional condition where the low reversibility of decisions was made salient. The reversibility of a decision refers to the extent to which its outcome can be modified or undone after the decision is made (C. J. Anderson, 2003, p. 151). For instance, most purchase decisions for consumer products are nowadays highly reversible, because many sellers accept unconditional returns within a certain timeframe. In contrast, the decision to have eye surgery is highly irreversible. Once eye tissues are removed, the surgeon cannot undo the removal. Like anticipated regret, decision reversibility is also ubiquitous in decision-making scenarios. Yet to the best of our knowledge, it has not been explicitly examined in the context of the decoy effect.

How may decision reversibility influence the decoy effect? First, decision reversibility has been negatively associated with anticipated regret: people predict that one experiences more regret with an undesirable outcome when it is less reversible (Tsiros & Mittal, 2000; Zeelenberg et al., 1996; cf. Gilbert & Ebert, 2002). If regret salience reduces the decoy effect, as shown in

Connolly et al. (2013), the salience of low decision reversibility may have a similar impact through inducing anticipated regret. Second, according to Connolly et al. (2013), regret salience reduces the decoy effect by nudging people to think more vigilantly. Meanwhile, it has long been established that people invest more time and effort and engage in more analytic thinking in decisions that are significant and cannot be reversed (McAllister et al., 1979). The parallel effects of regret salience and low decision reversibility on thinking style suggest that they may also influence the decoy effect similarly. In short, the negative association between anticipated regret and decision reversibility, as well as their parallel effects on thinking style, motivated us to hypothesize that the decoy effect can be reduced by making salient the low reversibility of decisions. Given that our primary aim in this research was to conduct direct replications, we were in a sense confined to the original study designs. Hence, we set out to test only the hypothesis that low decision reversibility has an appreciable effect. We left the question of whether the effect, if any, is mediated by anticipated regret to future investigations.

Overview of replication and extension

We attempted to replicate Experiment 1 in Ariely and Wallsten (1995) and Experiment 1 in Connolly et al. (2013) and added an extension to the latter. We targeted only the first experiments in these articles since we wanted to first establish that (1) the relatively uncommon task design in Ariely and Wallsten (1995) is suitable for demonstrating the decoy effect, and (2) the effect of regret salience on the decoy effect can be replicated (Connolly et al., 2013). We aimed at direct replications so that the original materials and procedures were adopted with minimal adjustments (see the supplementary where we documented all deviations). For both experiments, we attempted to power the original designs to .99 to detect a Cohen’s small effect (for Study 1; Cohen, 1988) or the original effects (for Study 2) (see the supplementary for details of the power analyses).

We recruited participants from the Amazon Mechanical Turk (MTurk) for both experiments, which were combined into a single data collection and presented in random order. Our prior experience of replicating judgment and decision-making findings has shown that combining different studies in one data collection and randomizing their presentation order does not affect the results (e.g., replicability, effect sizes), and doing so helps increase insights gained from the combined replication studies. For example, if the combined studies attempt to replicate the same phenomenon, and one succeeds but the other fails, sample differences can easily be ruled out as an explanatory factor for the discrepancy (Anvari et al., 2020; Chen et al., 2020; Ziano et al., 2020). We restricted our participants to be those that were both from and currently residing in the U.S. to minimize any geographical influences. Using an MTurk sample is a deviation from the original experiments, both of which recruited participants offline. Our experience has shown that MTurk is a reliable platform for conducting judgment and decision-making replications (Collaborative Open-science REsearch (CORE), 2020; see also C. A. Anderson et al., 2019; Buhrmester et al., 2011; Paolacci & Chandler, 2014; Thomas & Clifford, 2017). This deviation not only allows a reliable large-scale data collection but also contributes to examining the decoy effect in an online setting that is more natural in an era when many people browse information and make purchases on the internet. Successful demonstrations of the decoy effect with an online sample would also be in the interest of marketing professionals, as online shopping is becoming popular or even mainstream in some regions. Many past studies in the decoy effect literature also used samples from online panels (e.g., Pittarello et al., 2019; Reb et al., 2018) and successfully demonstrated the effect. Moreover, to the best of our knowledge, there has not been a systematic review suggesting that the decoy effect is sensitive to the difference between online and in-lab samples, and recent meta-evidence suggests that such deviations have little-to-no impact on findings (Olsson-Collentine et al., 2020). Therefore, we deemed that conducting the replications studies with online rather than offline or in-lab participants would not have any major impact on the replicability of the original findings.

The following were our pre-registered hypotheses:

H_1: The presence of a decoy makes the target more appealing than in its absence (baseline decoy effect; Study 1 and 2 replications).

H_2: Regret salience reduces the size of the decoy effect (Study 2 replication).

H_3: Choosing the target is more justified and less regrettable than choosing the non-target, or the competitor, if the choice turned out to be unsatisfactory (Study 2 replication).

H_4: The differences described in H_3 diminish when regret is made salient (Study 2 replication).

H_5: The salience of low decision reversibility reduces the size of the decoy effect (Study 2 extension).

Open science statement

This article was submitted as a Registered Report (Nosek & Lakens, 2014; Scheel et al., 2020; Wiseman et al., 2019). We made our raw data (after removing personal identifiers) and analysis codes available online (Open Science Framework: https://osf.io/vsbzk). Full open-science details and disclosures were provided in the supplementary. We confirm that all measures, manipulations, exclusions, and power analyses conducted for this investigation have been reported.

Replication evaluation

We evaluated our replication findings using the criteria set by LeBel et al. (2019) (see Figure 5S). Table 11S and Table 12S in the supplementary provide the classifications of the current replications using the criteria by LeBel et al. (2018) (Figure 6S). We summarized these two replications as *very close replications*.

Study 1: Replication of Ariely and Wallsten (1995)

Participants

A total of 1,100 American MTurk participants took part in this study and 1,053 completed the survey in exchange for $1.00. Fifty-two participants were excluded based on pre-registered criteria (please refer to the supplementary for details). Therefore, our final sample consisted of 1,001 participants (M_{age} = 42.24, *SD* = 12.97, two participants refused to disclose their ages; 501 (50.0%) males, 488 (48.8%) females, and 12 (1.2%) indicated their gender as others or preferred not to disclose this information). We present the results for the sample after exclusion in the main text and the full sample results in the supplementary, which also documents a comparison between the two samples (Table 13S).

We aimed at .99 power to detect Cohen's f = 0.1 when we planned our study, and our pre-registered power analysis suggested that we would need 462 participants. However, after we completed data collection and analysis, we realized an issue with the power analysis.[3] After correction, the analysis suggested a required sample size of 1,842 rather than 462. Based on the size of the sample that we actually collected (N = 1,001), a sensitivity analysis revealed that we had .80 power to detect Cohen's f = 0.089 or .99 power to detect Cohen's f = 0.136. We were able to detect Cohen's f = 0.1 with .88 power. Hence, we still had a satisfactory level of power despite not collecting enough to reach the original goal.

Participants gave their consent at the beginning of a Qualtrics survey, which was divided into two parts that collected data for Study 1 and 2, respectively. They then answered two

[3] Specifically, we did not select "as in SPSS" but proceeded with the default setting when using G*Power (Faul et al., 2007) to conduct power analyses for tests that involved repeated measures (Lakens, 2013).

confirmation questions that asked about their willingness and ability to participate in this research. If they did not respond positively to both questions, their sessions would be terminated, and they would be asked to return the task. This was for excluding those who did not pay attention and only randomly clicked through questionnaires. After participants completed Study 1 and 2 in random orders, they filled in a funneling section that asked them to guess the study hypotheses and report how serious they were about the survey (Aust et al., 2013). They provided demographic information, such as gender and age, in the end.

Design and procedure

Study 1 followed a two (*Condition*: 1A or 1C; between-subjects) by two (*Item*: Item A and Item C; within-subjects) mixed design (see also Table 9S in the supplementary for a tabular description). Participants performed five item evaluation tasks about different product categories, including running shoes, microwaves, computers, TVs, and bicycles, in a uniquely randomized order. In each task, participants were given three items of the same product category, whose attributes were measured on three dimensions. These items were presented in two conditions (1A and 1C) that differed with respect to which item in the core choice set (Item A or Item C) was targeted by a decoy (Item B). Participants randomly saw one condition for each product category, and the display of conditions was counterbalanced across subjects.

The following task instructions were presented at the beginning of this study: “Like the TV game, *Family Feud*, you are to answer the following questions in the same way that the average American would answer those. You have one hundred points and you are given information about three products. Please assign these points in a way that the number of points reflects their relative preference for the average American. Do it in a way that a higher number indicates a higher preference and give two items the same amount of points only if you think they are equally as appealing to the average American. All points must be used."

In the original study, Ariely and Wallsten (1995) had a pilot sample perform the same tasks. Whereas participants in the pilot sample were asked to evaluate the items based on personal values, those in the main study were told that their responses should reflect the pilot sample's preferences as accurately as possible (please refer to the instructions above for how this was implemented). The most accurate main study participant was rewarded $20. According to Ariely and Wallsten (1995), asking participants to approximate the preference of the pilot sample prevented them from engaging in the type of reasoning that arbitrarily downplays any dimension, e.g., "I don't care about dimension X" or "It doesn't matter how much the item costs." We followed this practice, yet instead of referring to an arbitrary sample, we asked participants to make evaluations as would be made by an average American.

After participants read the instructions, they answered two multiple-choice comprehension check questions that were presented on the same page: "How many points altogether are you going to assign to the items presented to you?" and "How are you supposed to evaluate the items presented to you and assign those points?" Participants had to answer these two questions correctly before they could proceed. The original study did not employ any comprehension checks. In our experience, these checks could help ensure participants understand the instructions, i.e., not to allocate points based on personal values but the preference of the general American population. They, therefore, function similarly to the $20 extra compensation in the original study, which incentivized participants not to assign points based on personal preferences, in that they both reinforced this idea.

Materials and manipulations

Table 1 presents the items used in both the original study and our replication. As shown, each product category had three items evaluated on three dimensions. Item A and Item C, which did not dominate each other, formed a core choice set. Item B was a decoy that targeted A in Condition 1A and C in Condition 1C. The dimensional values were determined systematically in Ariely and Wallsten (1995) (please refer to the supplementary for details).

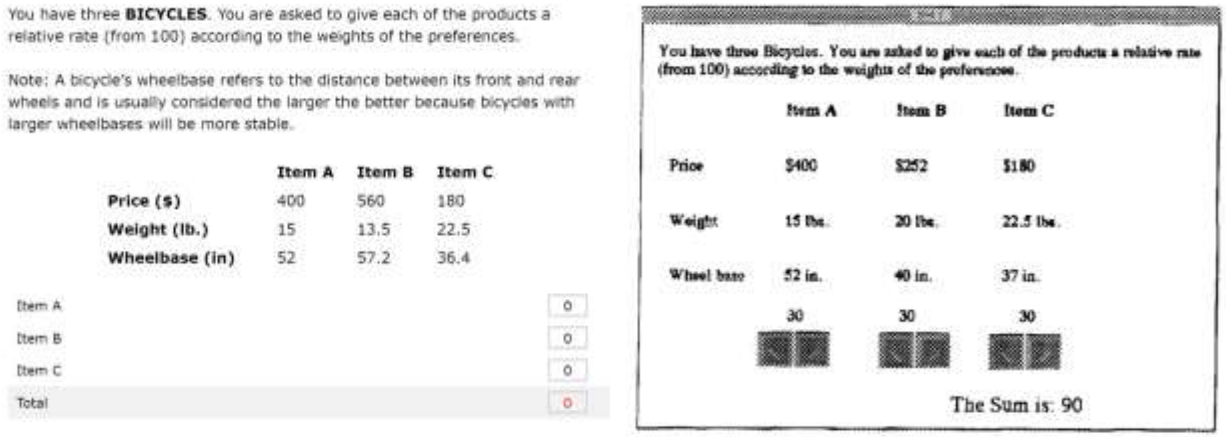
We mentioned two differences between the original study and our replication: the original was conducted in a lab with university students whereas we recruited participants from an online panel; the original had a pilot sample as a reference whereas we asked our participants to evaluate the items based on an average American's preference. Furthermore, our replication had four minor deviations from the original experiment. First, because the dimensional values of the TVs and the computers likely appear outdated, for these two product categories, we asked participants to imagine that they were doing the task in the 1990s, when technology was not as advanced. When conducting replications with stimuli that change over time, replicators often face a dilemma between following the original and risking the effect of changing times and changing the original and risking the effect of that change. We preferred following the original, as we believe that in its essence, the decoy effect is not about the items themselves, but the quantitative comparisons among them. This way, if we fail to replicate the effect with those time-sensitive items (computers and TVs) but not the less sensitive ones (running shoes, bicycles, and microwaves), we would be able to infer time as a possible moderating factor. Second, along with the items of each product category, we provided short explanations of those dimensions that might be difficult for participants to comprehend (including bicycles' wheelbase, microwaves' wattage and capacity, computers' speed and memory, and TVs' screen size and wattage). This deviated from the original experiment, which provided participants with a small brochure that explained all products and dimensions. Third, the original experiment presented assigned points horizontally and below the items under evaluation. In our replication, however, the assigned points were stacked vertically, and the item labels were indicated to their left (see Figure 1). We made this change as it was difficult to implement the original presentation format on Qualtrics, the survey platform. Implementing the original format would also be unnecessary, as replications of similar judgment and decision-making effects seem to indicate that changes in presentation format do not impact results (Collaborative Open-science REsearch (CORE), 2020). Lastly, the original experiment set a default of 30 points for each item, and participants adjusted these points (one point with each mouse click) so that they summed up to 100 in the end. We refrained from using defaults for concerns about potential anchoring effects. Instead, we asked for direct inputs in our replication. We figured that this change – if it had any impact at all – would most likely strengthen the effects, rather than weaken them, because participants would be able to input extreme values more easily in this way as compared with adding or deducting points with mouse clicks.

Table 1. *Descriptions of the items used in Study 1*

Products	Dimensions	Item A	Item C	Item B: B (1A)	Item B: B (1C)
Microwaves	Price ($) −	380	209	532	292.6
	Capacity (ft^3) +	1.8	1.2	2	1.3
	Wattage (W) +	1000	700	1100	770
Running shoes	Comfort +	8.5	5.5	5.1	3.3
	Durability +	6.8	4.4	7.5	4.9
	Price ($) −	90	58.5	81	52.7
Computers	Speed (Hz) +	33	21.5	16.5	10.8
	Memory (MB) +	8	4.4	8.8	4.8
	Price ($) −	1900	1235	1710	1111.5
TVs	Screen size (in) +	20	14	12	8.4
	Price ($) −	650	357.5	585	321.8
	Wattage (W) +	25	15	27.5	16.5
Bicycles	Price ($) −	400	180	560	252
	Weight (LB) −	15	22.5	13.5	20.3
	Wheelbase (in) +	52	36.4	57.2	40

Notes. (1) These items and their attributes were retrieved from Ariely and Wallsten (1995) without changes. (2) Unitless dimensions were measured on a 0-to-10 scale, where 10 indicated the highest desirability. (3) The +/− signs indicate whether the dimensions are positive, i.e., the higher the more desirable, or negative. (4) We followed Ariely and Wallsten (1995) to name the two conditions as 1A and 1C. The conditions were so named because the biggest difference between the target and the decoy lay on the first dimension presented in this table for each product category. (5) Please refer to the supplementary regarding how the dimensional values were determined.

Figure 1. *Sample task screens*



Sample task screens of our replication study (left) and the original study (right; Ariely & Wallsten, 1995, p. 227).

Results

Confirmatory analysis

Table 2 presents the descriptive statistics and the results of our confirmatory analyses. We conducted two-way mixed ANOVAs (*Condition*; between-subjects × *Item*; within-subjects) for each of the five product categories to examine the decoy effect (H_1), which, if took place, would manifest as an interaction between the two factors (i.e., Item A would receive more points in Condition 1A than in 1C, and vice versa for Item C). Contrary to our expectation, in none of these ANOVAs was the interaction term statistically significant, *p*s > .285. The interaction effect sizes were at most 0.001 in η^2_p, far below what Cohen (1988) considered to be a small effect (i.e., $\eta^2_p = 0.01$, roughly equivalent to Cohen's $f = 0.1$). Furthermore, none of the 90% upper limits for these effects reached this small effect benchmark. Figure 2 presents the interaction plots for these ANOVAs.

Additionally, we conducted Welch's independent-samples *t*-tests to compare the points that Item A or C received in the two conditions. As said above, the decoy effect predicts a cross-over type of interaction, where Item A/C receives more points in the condition in which they are targeted than in the other condition. As shown in Table 2, this pattern of results did not emerge for any of the product categories. We observed a general tendency for participants to have higher preferences for Item C in Condition 1C than in 1A (except for bicycles). Nonetheless, participants in general also preferred Item A more in Condition 1C than in 1A, which contradicted what the decoy effect would predict.

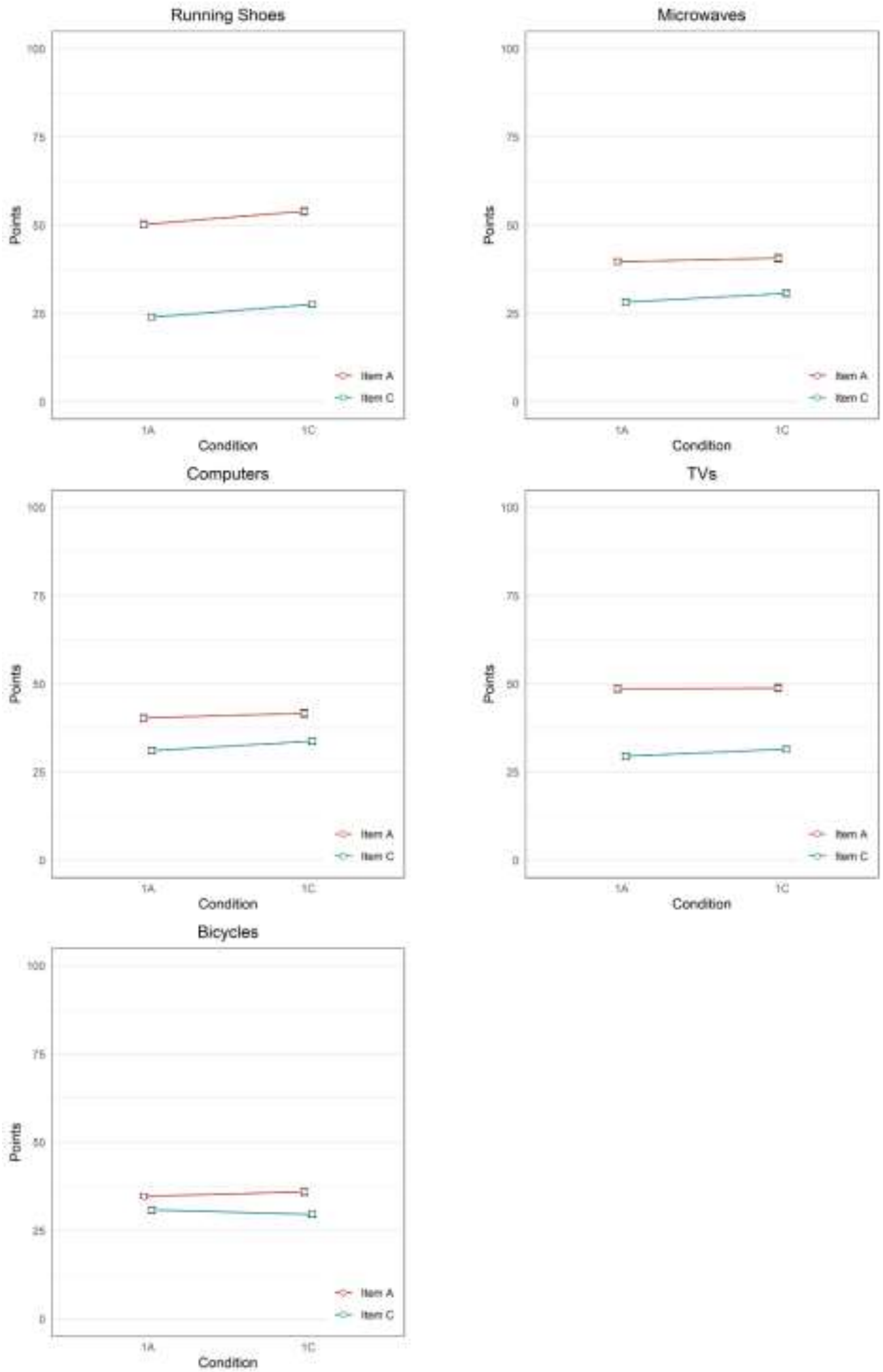
Following an anonymous reviewer's suggestion, we pre-registered to re-run the above analyses after removing extreme responses to examine the robustness of our results. In defining extreme responses, we considered the product categories separately. We calculated the variance of the assigned points by each participant (e.g., if a participant assigned 20, 30, and 50 to the items, the participant's variance would be 233.3). Because high variance indicated a strong preference for some items and low variance indicated indifference to the items, those responses with variances over or below three standard deviations from the mean variance were considered extreme and excluded. The results did not change substantially after the exclusion. Our analyses on the full sample also yielded similar results (please refer to the supplementary Table 15S and Table 16S for these results).

Table 2. *Study 1 descriptive statistics and confirmatory analysis results*

Product category	**Item**	**Points** M (SD) (n) Cond. 1A	Cond. 1C	t_{Welch} **(df)**	p	**Hedges' g (95% CI)**	F	p	η^2_p **(90% CI)**	η^2_G
Running shoes	A	50.26 (19.63) (502)	54.41 (21.03) (499)	-3.23 (993.46)	.001	-0.20 [-0.33, -0.08]	0.14	.707	< 0.001	< 0.001
	C	24.05 (15.51) (502)	27.41 (14.86) (499)	-3.50 (997.65)	< .001	-0.22 [-0.34, -0.10]			[0.000, 0.004]	
Microwaves	A	39.58 (17.16) (498)	40.92 (21.18) (503)	-1.10 (961.52)	.271	-0.07 [-0.19, 0.05]	0.05	.831	< 0.001	< 0.001
	C	28.65 (17.72) (498)	30.45 (19.17) (503)	-1.54 (994.37)	.124	-0.10 [-0.22, 0.03]			[0.000, 0.002]	
Computers	A	40.33 (19.39) (501)	41.62 (21.50) (500)	-1.00 (988.06)	.316	-0.06 [-0.19, 0.06]	0.36	.547	< 0.001	< 0.001
	C	31.06 (18.63) (500)	33.70 (17.19) (500)	-2.33 (992.87)	.020	-0.15 [-0.27, -0.02]			[0.000, 0.005]	
TVs	A	48.56 (19.65) (501)	48.79 (21.51) (500)	-0.18 (990.66)	.860	-0.01 [-0.13, 0.11]	0.68	.409	0.001	0.001
	C	29.48 (16.65) (501)	31.50 (16.59) (500)	-1.92 (999.00)	.055	-0.12 [-0.25, 0.00]			[0.000, 0.006]	
Bicycles	A	34.87 (14.37) (504)	35.72 (19.94) (497)	-0.77 (901.15)	.440	-0.05 [-0.17, 0.08]	1.14	.285	0.001	0.001
	C	30.92 (19.74) (504)	29.61 (17.46) (497)	1.11 (987.41)	.267	0.07 [-0.05, 0.19]			[0.000, 0.007]	

Notes. (1) The F-statistics, partial eta-squares, and generalized eta-squares pertained to the interaction effects. (2) Following Lakens' (2013) recommendation, we report both partial eta-squared and generalized eta-squared effect sizes. The former facilitates comparing our results with the target and the latter can easily be factored in future meta-analyses. (3) The Hedges' g effect sizes pertained to the between-condition comparisons for each item.

Figure 2. *Study 1 interaction plots*



Error bars represent ±1 SE. Graphs were produced with ggplot2 package in R (R Core Team, 2019; Wickham, 2016).

Exploratory analysis

To examine whether and to what extent participants perceived the decoys to be less preferable than their targets, we performed paired-samples *t*-tests to compare the points received by the targets and by the decoys. This analysis was not pre-registered. The results are presented in Table 3. In most cases, the targets were perceived to be significantly preferable to the decoys (exceptions were Condition 1A of bicycles and Condition 1C of microwaves, where the differences were not statistically significant, and Condition 1C of bicycles, where the difference was in the reverse direction). This suggests that the targets' superiority was perceived *in general*. At the very least, to some extent, the assigned points reflected the relative appeal of the targets and the decoys in a way that aligned with our expectations.

Table 3. *Comparing points received by targets and decoys*

Product category	**Cond.**	**Comparison**	***t***	**df**	***p***	***d***	**95% CI**
Running shoes	1A	A – B	18.30	501	< .001	0.82	[0.72, 0.92]
	1C	C – B	11.35	498	< .001	0.51	[0.41, 0.60]
Microwaves	1A	A – B	5.58	497	< .001	0.25	[0.16, 0.34]
	1C	C – B	1.55	502	.122	0.07	[-0.02, 0.16]
Computers	1A	A – B	8.72	500	< .001	0.39	[0.30, 0.48]
	1C	C – B	8.27	499	< .001	0.37	[0.28, 0.46]
TVs	1A	A – B	19.41	500	< .001	0.87	[0.76, 0.97]
	1C	C – B	11.18	499	< .001	0.50	[0.41, 0.59]
Bicycles	1A	A – B	0.57	503	.566	0.03	[-0.06, 0.11]
	1C	C – B	-4.29	496	< .001	-0.19	[-0.28, -0.10]

Note. All effects were expected to be positive, since the targets were intended to be perceived more attractive, and given more points, than their decoys.

Furthermore, we looked at the number of participants that rated the targets higher than the decoys. There were 720 such participants for running shoes (71.9%), 484 for microwaves (48.4%), 613 for computers (61.2%), 756 for TVs (75.5%), and 380 for bicycles (38.0%). Only 78 participants consistently rated the targets over the decoys across the five product categories. 248 did so for four product categories; 333 did so for three, 255 for two, 62 for one, and 25 for none of them. These results suggest that the extent to which participants' preference aligned with our expectation – that is, they should prefer the targets over the decoys – varied across product categories, and there were substantial individual differences in that extent of alignment.

We considered re-running the confirmatory ANOVAs after excluding those who did not rate the targets over the decoys. From these analyses, we obtained consistently significant and strong interaction effects in the predicted shape. These results, however, were of little interpretation value. To illustrate, if we exclude those in Condition 1A but gave Item A (the target) fewer points than Item B (the decoy), we would be left with those who gave Item A more points than the sample average. In other words, with the exclusion, we would raise the average point for Item A. Given that the total points for the three items were fixed to be 100, Item A points and Item C points were negatively correlated to a large extent. If we increase the average for A, we would decrease the average for C, thus creating the expected interaction. Therefore, the strong effects from these subgroup analyses could be an artifact of the experiment design (i.e., fixing the total points to 100) plus the exclusion, rather than reflect a true decoy effect. Still, it is hard to equate rating the target over the decoy with recognizing dominance, and we will revisit this point in the general discussion. Although we do not report these subgroup analysis results in detail here, we documented them in our analysis files for the reference of interested readers.

There is one final note. We pre-registered an exploratory direction, claiming that we would test whether the results would differ between the items that are time-sensitive and those that are less so. It is obvious that some of the original stimuli have outdated dimensional values (e.g., the specs of the computers), and we wondered whether such outdatedness would affect the replication results. We pre-registered that computers and TVs count as time-sensitive items and the others do not, and we planned to conduct a two (time sensitivity) by two (condition) by two (item) mixed ANOVA, collapsing the product categories on each level of time sensitivity, to answer that question. However, because we consistently observed small-to-no decoy effects across the product categories, we decided not to proceed in that exploratory direction as there was little ground for doing so.

Comparing replication and original findings

Comparing our results with the original (Table 4), we found no signal in the expected direction (i.e., a cross-over type of interaction) with the ANOVAs. Although no statistical tests were performed, the original data were mostly consistent with the prediction that Item A (C) would receive more points in Condition 1A (1C). However, our data did not conform to this prediction, which was the case for all product categories.

Table 4. *Comparison between the original and replication results of Study 1*

Product category	Original			Replication			Interpretation
	F	***p***	**η^2_p (90% CI)**	***F***	***p***	**η^2_p (90% CI)**	
Running shoes	18.36	< .01	0.240 [0.094, 0.378]	0.14	.707	< 0.001 [0.000, 0.004]	No signal – inconsistent
Microwaves	7.56	< .01	0.115 [0.018, 0.247]	0.05	.831	< 0.001 [0.000, 0.002]	No signal – inconsistent
Computers	0.84	.36	0.014 [0.000, 0.099]	0.36	.547	< 0.001 [0.000, 0.005]	No signal – inconsistent
TVs	4.922	.03	0.078 [0.004, 0.201]	0.68	.409	0.001 [0.000, 0.006]	No signal – inconsistent
Bicycles	1.02	.31	0.017 [0.000, 0.105]	1.14	.285	0.001 [0.000, 0.007]	No signal – inconsistent

Discussion

We conducted a very close direct replication of Experiment 1 in Ariely and Wallsten (1995). We failed to find support for the original results. We did not observe any statistically significant decoy effects, and the effect sizes were too trivial to be considered meaningful. Our results remained robust after excluding extreme responses and when the full sample was analyzed. Overall, we failed to find evidence for H_1 in this study. Our exploratory analyses suggested that in general, participants rated the targets to be preferable to their decoys. However, this does not constitute direct evidence that the dominance relationships were perceived and considered in participants' evaluation. There was a considerable proportion of participants that did not rate the targets over the decoys for each of the product categories, and only a minority of them consistently did this across the categories. Failing to detect the dominance relationship (a subjective one, to be precise) could be the main reason why the original results were not replicated, which we will revisit in the general discussion.

Study 2: Replication and extension of Connolly et al. (2013)

Participants

The same participants in Study 1 completed this study because the two studies were combined in a single data collection. Likewise, the results for the sample after exclusion are reported here, whereas the full sample results are documented in the supplementary. We aimed to achieve .99 power to detect the original effects, and our power analyses suggested that we would need 339 participants per condition (1,017 in total as we had three) for this purpose. Due to the same power-analysis issue mentioned in Study 1, following an adjusted reanalysis, we concluded that we achieved the intended power for all analyses but one (choice justifiability; see the supplementary for details).

Design and procedure

The experiment followed a two (*Choice Set*: Choice Set 1 or Choice Set 2) by three (*Condition*: Control, Regret-Salient, or Low-Reversibility) fully between-subjects factorial design (see also Table 10S in the supplementary for a tabular description). Participants were instructed to complete a decision task. The task instruction varied across conditions (see below), but the task was the same. After some comprehension and manipulation checks, participants were given three jobs that were described as similar in all aspects except two: promotion possibilities (rated on a 0-50 scale) and work interest (on a 0-100 scale), and they were to choose one from them. After the decision, participants were asked to imagine if they had chosen each of the three job options and found it dissatisfactory. For each imagined job choice, they rated their agreement with the following two statements: “I made a justifiable decision” and “I regret my decision.” The ratings were made on a seven-point Likert-type scale (1 = *Completely disagree*; 7 = *Completely agree*).

Materials and manipulations

Regret salience. Following Connolly et al. (2013), we presented participants in the *Regret-Salient* condition with the following paragraph after a general task introduction – which was the same across conditions – and before they saw the jobs: “As you make your decision, keep in mind that there is no guarantee that the job you pick will be right for you. You could find yourself in a job you don’t like, regretting the decision you made and wishing you had picked one of the other jobs.” Participants in the other two conditions did not see this paragraph.

Low decision reversibility. For participants in the *Low-Reversibility* condition, the above paragraph was replaced by “As you make your decision, keep in mind that in view of the current economic downturn, companies are restricting and even shrinking their headcounts, making job switching particularly difficult right now and in the coming years (but your current job options are firm and secure; you don’t need to worry that they will be retrieved or you will be fired soon).” Again, this paragraph was not shown to participants in the other two conditions.

Comprehension and manipulation checks. On the same page with the instruction, we asked participants three multiple-choice comprehension check questions: “how many job options do you have,” “in what aspect are the job options NOT different,” and “how should you evaluate the jobs presented to you.” Participants had to answer these questions correctly to proceed. On the next page, as manipulation checks, we asked participants in all three conditions to rate their agreement with the following two statements – “my job choice will surely be right for me” and “changing jobs will be easy in the future” – on a seven-point Likert scale (1 = *Completely disagree*, 7 = *Completely agree*). These questions were presented before participants saw the jobs.

Choice set. Table 5 presents the job options and their attributes used in the original research as well as in our replication. Two choice sets were prepared using these job options and presented randomly between subjects. The choice sets differed only with respect to which job within the core choice set (Job 1 or Job 2) was targeted by a decoy, which was presented as Job 3. Choice set was a method variable and bore no theoretical implications. As such, the two sets were not analyzed separately.

Table 5. *Job options*

Options	**Work interest (0 – 100)**	**Promotion opportunities (0 – 50)**
Job 1	83	34
Job 2	74	43
Job 3(a)	71	43
Job 3(b)	80	34

Note. Choice Set 1 comprised Job 1, Job 2, and Job 3(a) as a decoy targeting Job 2; Choice Set 2 comprised Job 1, Job 2, and Job 3(b) targeting Job 1.

Results

Manipulation checks. Participants who chose the decoys (11 out of 1001; 1.10%) were removed from the analysis (there were none in the original study). We conducted two Welch's one-way ANOVAs to compare participants' responses to the two manipulation check items across the conditions. We expected that *Regret-Salient* participants would agree less with the statement "my job choice will surely be right for me" than participants in the other two conditions, and *Low-Reversibility* participants would agree less with the statement "changing jobs will be easy in the future" than those in the other two conditions. The ANOVA on the regret salience manipulation check item revealed a main effect of condition, $F_{\text{Welch}}(2, 631.69) = 80.00$, $p < .001$, $\eta^2_p = 0.18$, 90% CI [0.15, 0.22]. Post hoc Games-Howell tests revealed that, as expected, participants in the *Regret-Salient* condition ($M = 4.29$, $SD = 1.60$) agreed with this item less than participants in the *Control* condition (M = 5.52, SD = 0.96), $p_{\text{Bonferroni}} < .001$, d = -0.93, 95% BCa CI [-1.08, -0.77], and those in the *Low-Reversibility* condition (M = 5.49, SD = 0.96), $p_{\text{Bonferroni}} < .001$, d = -0.91, 95% BCa CI [-1.08, -0.76]. We found no support for a difference between the *Control* and the *Low-Reversibility* conditions on this item, $p_{\text{Bonferroni}} = 1$, d = 0.03, 95% BCa CI [-0.13, 0.18].

The ANOVA on the low reversibility check item (i.e., "changing jobs will be easy in the future") also revealed a main effect of condition, $F_{\text{Welch}}(2, 648.78) = 51.32$, $p < .001$, $\eta^2_p = 0.11$, 90% CI [0.08, 0.14]. Post hoc Games-Howell tests revealed that, as expected, participants in the *Low-Reversibility* condition (M = 3.10, SD = 1.61) agreed with the check item less than those in the *Control* condition (M = 4.23, SD = 1.25), $p_{\text{Bonferroni}} < .001$, d = -0.78, 95% BCa CI [-0.96, -0.61], and those in the *Regret-Salient* condition (M = 3.94, SD = 1.25), $p_{\text{Bonferroni}} < .001$, d = -0.58, 95% BCa CI [-0.75, -0.42]. However, we also found support for a small difference between the *Control* and the *Regret-Salient* conditions, $p_{\text{Bonferroni}} = .03$, d = 0.23, 95% BCa CI [0.08, 0.38]. Overall, these results suggested that the manipulations worked as expected.

Confirmatory analysis

Decoy effect. Table 6 presents the descriptive statistics of Study 2, including the frequencies and rates of target choices. To test whether the decoy effect was present (H_1) and to estimate its size, we compared the overall choice rate of the target in each condition (Figure 3) against .5 with exact binomial tests. The decoy effect predicts that the overall target choice rate should be above .5, because a decoy increases the choice rate of its target regardless of choice set. We found support for the decoy effect in all three conditions (*Control*: 216 out of 337 participants chose the target (64.1%, 95% CI [58.7%, 69.2%]), exact binomial $p < .001$, Cohen's h = 0.29, 95% CI [0.18, 0.39]; *Regret-Salient*: 185 out of 326 participants chose the target

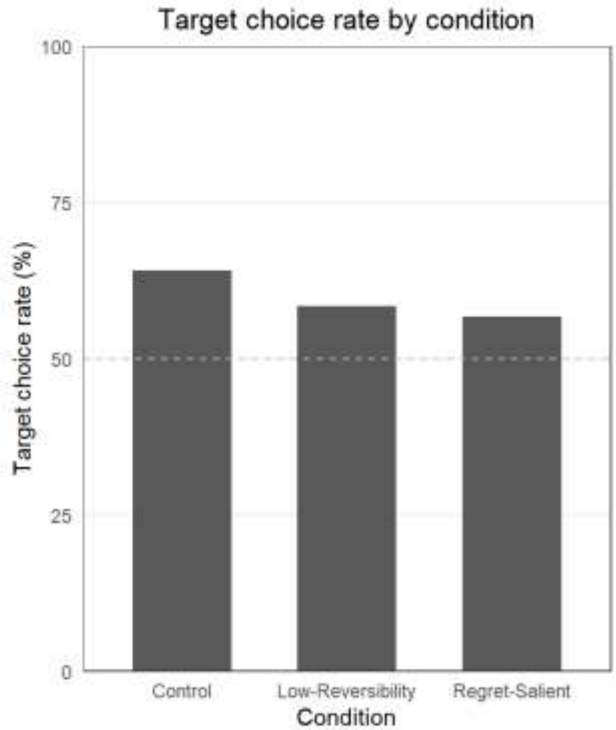
(56.7%, 95% CI [51.2%, 62.2%]), exact binomial *p* = .017, Cohen's *h* = 0.14, 95% CI [0.03, 0.24]; *Low-Reversibility*: 191 out of 327 participants chose the target (58.4%, 95% CI [52.9%, 63.8%]), exact binomial *p* = .003, Cohen's *h* = 0.17, 95% CI [0.06, 0.28]).

Table 6. *Study 2 descriptive statistics*

	Choice of the target option		Regret		Justifiability	
	N / Total No. of choices	(%)	Target *M* (*SD*)	Non-target *M* (*SD*)	Target *M* (*SD*)	Non-target *M* (*SD*)
Control (*n* = 339, 2 selected the decoy, 0.6%)						
Overall	216 / 337	64.1%	4.71 (1.68)	4.85 (1.76)	5.47 (1.32)	5.14 (1.41)
Choice set 1	104 / 165	63.0%	4.84 (1.64)	4.78 (1.71)	5.41 (1.33)	5.45 (1.37)
Choice set 2	112 / 172	65.1%	4.59 (1.72)	4.91 (1.81)	5.53 (1.31)	4.84 (1.39)
Regret-Salient (*n* = 331, 5 selected the decoy, 1.5%)						
Overall	185 / 326	56.7%	4.72 (1.71)	4.81 (1.74)	5.18 (1.48)	5.15 (1.46)
Choice set 1	86 / 170	50.6%	4.73 (1.71)	4.65 (1.73)	5.02 (1.37)	5.40 (1.31)
Choice set 2	99 / 156	63.5%	4.71 (1.72)	4.98 (1.74)	5.36 (1.57)	4.88 (1.56)
Low-Reversibility (*n* = 331, 4 selected the decoy, 1.2%)						
Overall	191 / 327	58.4%	4.86 (1.61)	4.89 (1.66)	5.42 (1.30)	5.20 (1.45)
Choice set 1	94 / 165	57.0%	4.91 (1.54)	4.70 (1.64)	5.25 (1.27)	5.51 (1.35)
Choice set 2	97 / 162	59.9%	4.81 (1.68)	5.09 (1.67)	5.59 (1.31)	4.88 (1.48)

Note. Descriptive means and standard deviations are reported above (i.e., they are not based on the ANOVA models in our analyses).

Figure 3. *Target choice rate by condition*



Condition effects. Connolly et al.'s (2013) results suggested that regret salience can reduce the decoy effect (H_2). We have also hypothesized that the salience of low decision reversibility has a similar influence (H_5). To test these two hypotheses, we compared the target choice rates among the conditions. A Fisher's exact test between the *Control* and the *Regret-Salient* conditions revealed weak support for a difference in the expected direction, $\chi^2(1) = 3.74$, $p = .057$, $\varphi = 0.075$, 95% CI [0.000, 0.151]. The difference between the *Control* and the *Low-Reversibility* conditions was smaller, $\chi^2(1) = 2.26$, $p = .151$, $\varphi = 0.058$, 95% CI [0.000, 0.134], and even smaller was the difference between the *Regret-Salient* and the *Low-Reversibility* conditions, $\chi^2(1) = 0.18$, $p = .693$, $\varphi = 0.017$, 95% CI [0.000, 0.092]. Overall, we observed effects in the predicted directions, but the effect sizes were smaller than what Cohen (1988) considered to be small effects ($\varphi = 0.1$), and far smaller than the original ($\varphi = 0.271$).

Choice regret and justifiability. Connolly et al. (2013) observed that *Control* participants indicated that an imagined choice of the target would be more justifiable and less regrettable than the competitor if the choice turned out to be dissatisfactory (H_3). These differences, however, were not observed in the *Regret-Salient* condition (H_4). We examined whether our data would reveal the same pattern of results, i.e., a “knock-out” type of interaction between condition (*Control* vs. *Regret-Salient*) and imagined choice (*target* vs. *competitor*) on choice regret and justifiability. We did not have any ad hoc hypothesis about the effect of low decision reversibility on regret and justifiability, but we included the *Low-Reversibility* condition in our analysis for exploratory purposes. To test H_3 and H_4, we conducted two three-way (*Condition* × *Choice Set* × *Imagined Choice*) mixed ANOVAs on choice regret and justifiability, respectively. Our focus was on the interaction between condition and imagined choice.

The three-way mixed ANOVA on choice regret revealed support for an interaction between choice set and imagined choice, $F(1, 984) = 15.31$, $p < .001$, $\eta^2_p = 0.015$, 90% CI [0.005, 0.030], and weaker support for a main effect of imagined choice, $F(1, 984) = 2.78$, $p = .096$, $\eta^2_p = 0.003$, 90% CI [0.000, 0.011]. We did not find support for an interaction between condition and imagined choice, $F(2, 984) = 0.32$, $p = .730$, $\eta^2_p = 0.001$, 90% CI [0.000, 0.004].

Breaking down the three-way ANOVA by condition, contrary to H_3, we did not find support for a main effect of imagined choice in the *Control* condition, $F(1, 984) = 2.23$, $p = .135$, $\eta^2_p = 0.002$, 90% CI [0.000, 0.010]. The evidence was insufficient for a difference in choice regret between the target ($M = 4.71$, $SE = 0.09$) and the competitor ($M = 4.84$, $SE = 0.09$), $d = -0.12$, 95% CI [-0.27, 0.04]. Also, we did not find support for a main effect of imagined choice in the *Regret-Salient* condition, $F(1, 984) = 1.07$, $p = .301$, $\eta^2_p = 0.001$, 90% CI [0.000, 0.007]. There was insufficient evidence supporting a difference in regret between choosing the target (*M*

= 4.72, *SE* = 0.09) and the competitor (*M* = 4.81, *SE* = 0.09), *d* = -0.08, 95% CI [-0.23, 0.07]. Lastly, we did not find support for a main effect of imagined choice in the *Low-Reversibility* condition, $F(1, 984) = 0.13$, $p = .716$, $\eta^2_p = 0.001$, 90% CI [0.000, 0.004]. Our data could not support a difference between choosing the target (*M* = 4.86, *SE* = 0.09) and the competitor (*M* = 4.89, *SE* = 0.09) in choice regret, *d* = -0.03, 95% CI [-0.18, 0.12]. Summarizing the above, we did not find sufficient evidence for H_3, the hypothesis that choosing the target is less regrettable than choosing the competitor. Since the baseline difference could not be established, we also concluded no evidence for H_4, that regret salience can reduce the difference in choice regret.

The three-way mixed ANOVA on choice justification revealed very weak support for an interaction between condition and imagined choice, $F(1, 984) = 2.51$, $p = .082$, $\eta^2_p = 0.005$, 90% CI [0.000, 0.014]. However, we obtained stronger support for an interaction between choice set and imagined choice, $F(1, 984) = 73.33$, $p < .001$, $\eta^2_p = 0.069$, 90% CI [0.046, 0.096], a main effect of imagined choice, $F(1, 984) = 16.08$, $p < .001$, $\eta^2_p = 0.016$, 90% CI [0.006, 0.032], as well as a main effect of choice set, $F(1, 984) = 4.99$, $p = .026$, $\eta^2_p = 0.005$, 90% CI [0.000, 0.015].

Breaking down the ANOVA by condition, we found support for a main effect of imagined choice in the *Control* condition, $F(1, 984) = 14.24$, $p < .001$, $\eta^2_p = 0.014$, 90% CI [0.005, 0.029]. Participants in the *Control* condition found choosing the target (*M* = 5.47, *SE* = 0.08) more justifiable than choosing the competitor (*M* = 5.15, *SE* = 0.08), *d* = 0.29, 95% CI [0.14, 0.44]. In contrast, we observed no main effect of imagined choice in the *Regret-Salient* condition, $F(1, 984) = 0.36$, $p = .548$, $\eta^2_p < 0.001$, 90% CI [0.000, 0.005]. Our data therefore failed to provide support for a difference between the justifiability of choosing the target (*M* = 5.19, *SE* = 0.08) and that of choosing the competitor (*M* = 5.14, *SE* = 0.08) in the *Regret-Salient* condition, *d* =

0.05, 95% CI [-0.11, 0.20]. We found support for a main effect of imagined choice in the *Low-Reversibility* condition, $F(1, 984) = 6.74$, $p = .010$, $\eta^2_p = 0.007$, 90% CI [0.001, 0.018]. Participants found choosing the target ($M = 5.42$, $SE = 0.08$) more justifiable than choosing the competitor ($M = 5.20$, $SE = 0.08$) in this condition, $d = 0.20$, 95% CI [0.05, 0.36]. Overall, our data supported the hypothesis that choosing the target is more justifiable than choosing the decoy in the *Control* condition (H_3). Although no significant main effect of imagined choice on justifiability was found in the *Regret-Salient* condition, which supported H_4, this should be interpreted with caution given the marginally significant interaction between condition and imagined choice in the grand ANOVA.

Exploratory analysis

An important facet of this experiment allowed us to test the decoy effect (H_1) by comparing the target choice rate in each condition against .5: there were two choice sets with decoys targeting different primary options. We could not test the effect in this way with only one choice set. Without aggregating such counterbalancing choice sets, the target choice rate that reflects zero decoy effect would not necessarily be 50%, because different choices have different baseline popularities. Also because of this design, we were able to test H_3 and H_4 by examining the main effects of imagined choice in each condition in the above ANOVAs. Aggregating the results of the two counterbalancing choice sets, we could account for the baseline differences in choice regret and justifiability of the primary options and focus on the effect of their roles in the choice set (i.e., whether an option was a target or a competitor) on the evaluations.

Recognizing this, we realized that the pre-registered H_3 probably did not spell out what we intended to test and what we (and the original authors) in fact tested. We hypothesized that "choosing the target is more justified and less regrettable than choosing the non-target, or the competitor, if the choice turned out to be unsatisfactory." This is apparently not always the case. Options vary in their baseline regret and justifiability, and hence we cannot expect to find evidence for the hypothesis with any choice set involving a target, a competitor, and a decoy. In fact, a closer look at our descriptive statistics (Table 6) suggested that when given Choice Set 1, participants across all conditions rated choosing the target to be more regrettable. Therefore, it makes little sense to test what H_3 literally claims, and a more accurate expression of this hypothesis should incorporate the element of controlling for baseline differences of options, like our experimental design did.

If we are to account for individual differences of options, it makes sense to examine the evaluations of *the same option* across choice sets, where it played different roles. Extending H_3, we found it reasonable to hypothesize that given two competing options, the same option would be evaluated differently depending on whether it was targeted by a decoy or not. To be more specific, our prediction was that an imagined choice of Job 1 would be less regrettable and more justifiable for those who were given Choice Set 2, where Job 1 was the target option, than for those who were given Choice Set 1, where Job 1 was the competitor; and vice versa for Job 2. Also, taking both H_3 and H_4 into consideration, we expected that the above differences would emerge in the *Control* condition but would be reduced in the *Regret-Salient* condition.

To test these two exploratory hypotheses, we analyzed the pertinent contrasts based on the confirmatory ANOVA models, and the results are presented in Table 7. We observed effects in the expected directions in the *Control* condition. However, we found support for only one of these effects. A quick sensitivity analysis suggested that with our sample sizes in the *Control* condition (n_1 = 165, n_2 = 172), an independent-samples *t*-test would have .80 power to detect a Cohen's d = 0.31. Therefore, even if there were indeed differences as suggested by these contrasts, we might not have enough power to detect them reliably. Given the magnitude of the observed point estimates, these differences are likely to be trivial in size.

Meanwhile, we conclude insufficient evidence to suggest that in the two experimental conditions, the same job differed in regret and justifiability depending on being the target or the competitor (one exception was about the justifiability of Job 2 in the *Low-Reversibility* condition). However, like what was found in the confirmatory analyses, because the baseline differences in the *Control* condition could not be reliably established, we concluded no sufficient evidence in support of the claim that these differences could be reduced by salience of regret or low decision reversibility.

Table 7. *Contrasts of evaluations on the same job in different choice sets*

Job	Evaluation	*MD*	*SE*	*t*	df	*p*	*d*	95% CI
Control condition								
Job 1	Regret	-0.19	0.19	-1.05	1526	.292	-0.17	[-0.48, 0.15]
	Justifiability	0.07	0.15	0.49	1739	.622	0.07	[-0.20, 0.34]
Job 2	Regret	-0.07	0.19	-0.38	1526	.702	-0.06	[-0.38, 0.25]
	Justifiability	0.57	0.15	3.77	1739	< .001	0.51	[0.25, 0.78]
Regret-Salient condition								
Job 1	Regret	0.06	0.19	0.34	1526	.732	0.06	[-0.26, 0.38]
	Justifiability	-0.04	0.15	-0.27	1739	.790	-0.04	[-0.31, 0.24]
Job 2	Regret	-0.25	0.19	-1.34	1526	.181	-0.22	[-0.54, 0.10]
	Justifiability	0.15	0.15	0.95	1739	.345	0.13	[-0.14, 0.40]
Low-Reversibility condition								
Job 1	Regret	0.11	0.19	0.60	1526	.551	0.10	[-0.22, 0.42]
	Justifiability	0.08	0.15	0.50	1739	.614	0.07	[-0.20, 0.34]
Job 2	Regret	-0.18	0.19	-0.95	1526	.344	-0.15	[-0.47, 0.17]
	Justifiability	0.37	0.15	2.43	1739	.015	0.34	[0.06, 0.61]

Note. We subtracted the evaluation of the job when it was a non-target/competitor from that when it was a target (i.e., evaluation as a target minus evaluation as a non-target, or competitor). We expected that under the *Control* condition we would obtain negative MDs in regret and positive MDs in justifiability.

MD = mean difference, SE = standard error of mean difference.

Comparing replication and original findings

Connolly et al. (2013) found a decoy effect of the size 0.832 in Cohen's h in their *Control* condition (95% CI [0.42, 1.24]), a large effect (Cohen, 1988). In comparison (see Table 6; interpretations are based on LeBel et al., 2019), the effect size in our replication was much smaller, Cohen's h = 0.29, 95% CI [0.18, 0.39], a small effect by Cohen's (1988) benchmark. The decoy effect in the *Regret-Salient* condition of our replication remained comparable with that in the same condition of the original study.

Table 6. *Comparison between original and replication results (Study 2)*

	Original				**Replication**				
Decoy effect	%	p	h	95% CI	%	p	h	95% CI	**Interpretation**
Control	87.0	< .001	0.83	[0.42, 1.24]	64.1	< .001	0.29	[0.18, 0.39]	Signal – inconsistent, smaller
Regret-Salient	61.5	.20	0.23	[-0.08, 0.55]	56.7	.017	0.14	[0.03, 0.24]	Signal – consistent
Low-Reversibility	-	-	-	-	58.4	.003	0.17	[0.06, 0.28]	-
Condition effect	χ^2	p	φ	95% CI	χ^2	p	φ	95% CI	
Regret salience	4.54	.04	0.27	[0.00, 0.52]	3.74	.057	0.075	[0.000, 0.151]	No signal – inconsistent
Low reversibility	-	-	-	-	2.26	.151	0.058	[0.000, 0.134]	-

Note. The decoy effect was tested with exact binomial tests, which compared the observed proportion to .5. The condition effects were tested by comparing the Control condition and the two experimental conditions with Fisher's exact tests. Interpretations are based on LeBel et al. (2019).

The original study found that regret salience reduced the decoy effect with an appreciable effect size (p = .04, φ = 0.27), whereas in our replication, the reduction was smaller, φ = 0.075, 95% CI [0.000, 0.151], p = .057. Despite being much more powerful than the original study, our replication was still not powered enough to reliably (power = .8) detect this effect size. A power analysis suggests that to detect this difference (64.1% in Control vs. 56.7% in Regret-Salient) with a one-tailed Fisher's exact test and at .8 power, we would need 569 participants in each condition. In comparison, we had 337 participants in the *Control* and 326 participants in the *Regret-Salient* condition (excluding those who chose the decoy). If we combine our data with the original, a Fisher's exact test would reveal a significant effect of regret salience (the target choice rates would be 65.6% and 57.3% in the two conditions), $\chi^2(1) = 5.26$, $p = .022$, $\varphi = 0.085$, 95% CI [0.008, 0.158]. To detect this effect at .8 power with a one-tailed Fisher's exact test, we would still need 445 participants in each condition.

Our results concerning evaluations of imagined choices were mixed. Contrary to the original study, we could not establish the baseline difference in regret between choosing the target and the competitor in the *Control* condition (the main effect of imagined choice was η^2_p = 0.002 in our replication vs. η^2_p = 0.278 in the original; see the supplementary for details). Our results pertaining to justifiability aligned with the original as well as the hypotheses (H_3 and H_4); however, these conforming results were qualified by an interaction between condition and imagined choice that was only marginally significant. Furthermore, like the case of choice regret, the baseline difference in the *Control* condition was much smaller in our replication (η^2_p = 0.014 here vs. η^2_p = 0.193 in the original).

Discussion

We conducted a very close direct replication and extension of Experiment 1 in Connolly et al. (2013). Our data provided support for a baseline decoy effect (H_1) in the *Control* condition (as well as in the two experimental conditions), though with a much smaller size than in the original study. The salience of both regret and low decision reversibility reduced the decoy effect as compared with the *Control*. The reductions, however, were very weak to the extent that our sample size, which was more than ten times larger than the original one (62 in the *Control* and *Regret-Salient* conditions in the original vs. 663 in our replication), did not provide sufficient power to detect them. Therefore, we found our evidence inconclusive for H_2 and H_5.

Our evidence was mixed for H_3 and H_4. In the absence of any experimental manipulation, our data failed to provide support for the hypothesis that choosing the target is less regrettable than choosing the competitor, but provided some support for the hypothesis that choosing the target is more justifiable than choosing the competitor (corroborating previous findings on choice justifiability; see Wedell & Pettibone, 1996; Pettibone & Wedell, 2000). Regret salience seemed to reduce the difference in choice justification, but the reduction was qualified by a very weak interaction between condition and imagined choice.

We examined whether low decision reversibility reduces the decoy effect. Our evidence was inconclusive because of the small effect size, which was nonetheless in the predicted direction, and because of its being statistically non-significant, despite what seemed to be an effective manipulation.

We explored whether the same job choice would be less regrettable and more justifiable when the job was a target than when it was a competitor. Again, we had inconclusive evidence because most of the effects in the *Control* condition had small sizes despite all being in the predicted directions. It would be unreasonable to claim any condition effects here (i.e., effects of regret salience or low reversibility) given that the baseline differences were not reliably established.

Based on the above, we summarize this replication as partly successful. We discuss potential reasons for the discrepancies between the original results and ours and their implications in the general discussion.

Joint Analysis

Having the same participants take part in multiple studies on the same effect has two advantages: first, if some studies show an effect but others do not, we can reasonably rule out sample difference as an explanatory factor and focus more on the differences in procedures, designs, and materials; second, if we observe consistent effects across studies, we can explore whether there are stable individual differences that make some participants more prone to display that effect than others regardless of study contexts. Even though the overall decoy effect was not consistent across our two studies, we would still like to leverage the fact that the same participants took part in both and examine whether they behaved consistently, i.e., whether it was the same subgroup of participants that showed the decoy effect in both Study 1 and Study 2.

For this purpose, we calculated a measure of the decoy effect in Study 1 for each participant. We subtracted the points for the competitors from the points for the targets, totaled them up across the five product categories, and assumed that the higher this sum was, the more the participant had preferred the targets relative to the competitors (i.e., the stronger the decoy effect was that the participant had displayed). This measure was then regressed on participants' choice in Study 2 (two levels: target or competitor; those who chose the decoys were excluded from this analysis). We failed to find an association between participants' choice in Study 2 and their decoy effect measure in Study 1, $F(1, 988) = 0.002$, $p = .963$. We further included Study 2 condition (three levels) as an explanatory variable and the interaction between condition and choice. Nonetheless, the overall model still had little-to-no explanatory power, $F(5, 984) = 0.45$, $p = .811$. Hence, we found no evidence suggesting that relatively higher preferences for the targets in Study 1 were associated with choosing the target in Study 2.

In a separate analysis, we calculated, for each participant, the number of product categories for which they rated the targets over the decoys (the descriptives are presented in Study 1's exploratory analysis section). This number reflects how frequently participants preferred the targets over the decoys, and it ranged from 0 to 5. We reasoned that the more frequently one showed such a preference in Study 1, the more likely he or she would also appreciate the dominance relationship in Study 2 and be driven to choose the target job option. However, when this calculated measure was regressed on participants' choice in Study 2 (again, excluding those who chose the decoy), there was no evidence suggesting an association, $F(1, 988) = 1.40$, $p = .237$. Therefore, choosing the target in Study 2 was not positively associated with higher frequencies of preferring the targets over the decoys in Study 1.

To summarize our exploratory analyses, we found no evidence for individual differences in showing the decoy effect. In other words, we failed to identify a subgroup of participants who were consistently driven to prefer the targets more by the decoys across our studies, or a subgroup that was consistently immune to the decoy effects. Furthermore, exhibiting higher preference for the targets over the decoys – an intended preference pattern – in Study 1 was not associated with choosing the target in Study 2. Because of the largely null results in Study 1, we caution our readers not to overgeneralize these findings.

General Discussion

We found little-to-no evidence for the decoy effect in Study 1, which was a very close replication of Experiment 1 in Ariely and Wallsten (1995). In contrast, we found evidence for the decoy effect in Study 2, which attempted to replicate and extend Experiment 1 in Connolly et al. (2013). The effect, however, was much smaller than the original one (Cohen's h = 0.29 here vs. 0.83 in Connolly et al., 2013). Replicating the original study, in Study 2, we also found some indication that regret salience reduced the decoy effect. Nonetheless, the reduction was much smaller in size.

As an extension, we investigated whether low decision reversibility can also reduce the decoy effect. Our data revealed an effect in the predicted direction. The effect, however, was also too trivial to be detected reliably here. Furthermore, we failed to obtain consistent evidence in support of the hypothesis that choosing a target is more justifiable and less regrettable than a competitor, and no sufficient evidence supports that regret salience can reduce or eliminate such differences.

Evaluating replication results

What might explain the discrepancies between the original and the replication results? First, we believe that differences in samples and settings are most likely unable to fully account for the discrepancies. In the past decade, online panel data, especially those from MTurk, have become an important source of evidence in behavioral research (C. A. Anderson et al., 2019), and there is no evidence suggesting that using online participants compromises the replicability of findings (Olsson-Collentine et al., 2020). Also, the decoy effect has been studied multiple times in an online setting (e.g., Pittarello et al., 2019; Reb et al., 2018), suggesting that the absence of effect in Study 1 and the reduced effect size in Study 2 could not be just attributed to sample and setting differences. Moreover, it would undermine the generalizability of the original findings by arguing that the decoy effect is this sensitive to sample and setting differences. Data quality is also an unlikely reason, as we employed multiple comprehension checks that either required correct answers for participants to proceed or excluded inattentive participants right away due to incorrect answers. Our results stood robustness checks, especially in Study 1.

We noted a procedural deviation from the original experiment in Study 1. In contrast to the original, we did not ask participants to adjust points from a default of 30 for each item; instead, we asked for direct inputs. It is unlikely, however, that this deviation could fully account for the vast differences between our results and the original ones. First, as we discussed in the introduction, with our input method, participants could input more extreme values with relative ease than in the original experiment. This would only enlarge but not reduce the size of the decoy effect. Consider a situation where, due to the presence of a decoy, participants thought that Item A should be preferred strongly by the average American and decided to give it more points. With the original input method, the easiest way to express a strong preference was to click the mouse ten times to give A 40 points and keep the default 30 points for B and C (so that the points add up to 100). Participants could go beyond this by intentionally deducting the points for B or C and further increase the points for A. However, these adjustments take time and effort and are by no means easier than with our input method, by which participants could give A 100 points and C 0 points in seconds if they think A is strongly preferred. Second, it should be noted that, despite this ease to allocate extreme points, few participants really allocated 100, or numbers close to this maximum, to any of the items, as can be seen in the visualized data (provided along with the analysis codes on OSF). This suggested that most of our participants did consider the relative appeal of the items and did not blindly assign all points to one item. Furthermore, if we were to argue that this deviation accounts for the result discrepancies, we would risk the generalizability of the original findings because we are seldom in a situation where we have explicit and equal priors for all available options to start with, like the equal defaults in the original experiment.

One probable explanation for the replication failure in Study 1 is that participants had difficulty noticing the dominance relationship between the targets and the decoys. As Huber et al. (2014) have reminded researchers of the decoy effect, the phenomenon can only be properly studied when the dominance relationship can be identified “quickly and unambiguously” (p. 522). If the dominance relationship is too difficult for decision-makers to perceive and register in their minds, we cannot expect that they will further act on it. Like the original experiment, we did not have a direct probe for the successful perception of dominance relationships. But the difference in points for a target and for a decoy could be a proxy. In our exploratory analyses, we tested whether more points were assigned to the targets than to the decoys. Largely conforming with our expectation, the results suggested that participants in general gave the targets more points. This evidence, however, needs to be interpreted cautiously, because a higher preference for one item over another does not necessarily mean that the former is perceived to dominate – that is, be superior in every aspect to – the latter. We believe that the decoy effect literature would benefit from clear and valid criteria that define a threshold for differences in quantitative measures of preferences to be taken as an indicator of a consciously perceived and registered dominance relationship. The fact that there was a considerable proportion of participants rating the decoy over its target was itself a sign of an inconspicuous dominance relationship, which might have resulted in the absence of evidence for the decoy effect.

Although we concluded a successful replication of the decoy effect in Study 2, its size was much smaller than that in the original. According to LeBel et al.’s (2019) criteria for evaluating replication outcomes, our result was an inconsistent weaker signal. Compared with Study 1, it was unlikely that failure to notice the dominance could explain the discrepancy between the results. Across the conditions, at most 1.5% of the participants chose the decoys, suggesting that most of them did notice the inferiority of the decoy options. Because we employed a much larger sample size and therefore had higher power, we may infer that the original effect size was an overestimate, and the true effect size is smaller.

We found limited support for the effect of regret salience. This, however, should not be taken to mean that anticipated regret does not have an effect; instead, we suggest that the manipulation might not be as effective as we (and the original authors) expected. The regret salience manipulation attempted to induce anticipated regret by reminding participants that their choice might not be the right one for them, and they might end up regretting the choice. Despite that the manipulation check results suggested a successful manipulation, the check item does not tap anticipated regret directly. Instead, it asked participants to what extent they would agree that their choice would surely be the right one. The extent of (dis)agreement with this item may not reflect participants’ anticipated regret (it may reflect, for instance, participants’ self-efficacy in making decisions), and hence, we see a need for more evidence to examine the effectiveness of this regret salience manipulation. We also caution against overclaiming the practical significance of the manipulation given the weak effects we obtained, and we recommend researchers seek to compare different regret salience manipulations and employ the one with the strongest effect in their regret studies.

Because the regret salience manipulation had such a small influence on the decoy effect, it was not surprising that we found inconclusive evidence concerning its effect on the evaluations of imagined choices, particularly on choice regret. It must be noted that we did not find sufficient evidence even for the baseline difference in choice regret or, in other words, for the claim that participants in the *Control* condition would find choosing the targets more regrettable than choosing the competitors. In our exploratory analyses, we also found insufficient evidence for the hypothesis that the same option within a certain core choice set would be considered less regrettable when targeted by a decoy than when not. Without a solid baseline difference to begin with, we cannot expect to find enough evidence in support of a reduction by the regret salience manipulation (part of H_4).

Implications

As an exemplar of contextual effects on decision making, the decoy effect has received widespread attention from researchers since it was initially demonstrated (Huber et al., 1982; Kaptein et al., 2016; Król & Król, 2019; Lichters et al., 2015; Sivakumar, 2016). In recent years, the robustness of this effect has been debated (Frederick et al., 2014; Huber et al., 2014; Milberg et al., 2014; Yang & Lynn, 2014). The results of our two replication studies imply that while the effect itself remains solid (successfully replicated in Study 2), whether and to what extent it emerges could be sensitive to an array of factors including task nature (e.g., indicating preference vs. making a choice), properties of and relationships between the options (e.g., numbers of dimensions on which they are evaluated; subjective vs. strict dominance), and many others (Milberg et al., 2014). Therefore, a careful examination of the robustness of the effect is warranted if this effect is to be applied in real-world settings, which vary greatly from each other. Still, researchers should carefully consider the possible range of effect sizes of the decoy effect in planning their studies. A recent meta-analysis summarizing 244 decoy effects revealed an overall effect size of Hedges' g = 0.14 (Milberg et al., 2014, p. 1420), lower than what is conventionally considered a small effect (Cohen, 1988). Our effect size estimate in Study 2 was close to this magnitude, and it was by no means an effect size that can be studied reliably with only tens of participants each making one decision. Considering that the choice sets used in Study 2 were of the kind that makes dominance relationships particularly noticeable (the decoy option was equivalent with the target on one dimension but clearly inferior on the other), it is reasonable to expect that the effect would likely be even smaller if the dominance is more obscure, like in Study 1, and when real options are involved (Frederick et al., 2014; Huber et al., 2014; Yang & Lynn, 2014). Researchers studying the application of the decoy effect thus need to properly calibrate their expectations about the magnitude of the effect and plan their studies with caution.

A single experiment with only hundreds of participants does not provide conclusive evidence for any phenomena, and given the important role regret plays in decision-making (Connolly & Reb, 2012; Connolly & Zeelenberg, 2002; Tsiros & Mittal, 2000), it is unreasonable to assert that regret salience does not influence the decoy effect. Nevertheless, it is reasonable to speculate that anticipated regret has a trivial impact on the type of decision that participants had made in Study 2, or, even more likely, that the manipulation in Study 2 was not as effective in inducing anticipated regret as we expected it to be. Despite that our manipulation check appeared to suggest a successful manipulation, the results should not be taken at face value, particularly considering that the check item did not tap regret directly. Unfortunately, our data do not allow us to arbitrate between the two possibilities (i.e., a small effect vs. an ineffective manipulation). But we may draw the bottom line that if we consider the manipulation effective, its effect is so small that we can hardly recommend researchers studying regret to induce the emotion in this way.

The hypothesis that regret salience reduces the decoy effect may be at odds with the findings based on computational modeling, which has recently been applied to the study of decoy effects (Trueblood et al., 2013, 2014). Backing the hypothesis is the rationale that the decoy effect emerges due to a heuristic processing of information (called the *dominance heuristic*; Chatterjee et al., 2011; Mourali et al., 2007; Simonson, 1989), which could be inhibited by more vigilant and effortful thinking and, possibly, by spending more time on decisions (Gigerenzer & Gaissmaier, 2011). Most recent computational modeling studies, however, revealed the decoy effect to be the result of an evidence accumulation process and predicted that spending more time on decisions leads to a larger effect (Roe et al., 2001; Trueblood et al., 2014; Trueblood & Pettibone, 2017; Usher & McClelland, 2004). This prediction has received some empirical support (Pettibone, 2012; c.f., Gaudeul & Crosetto, 2019). Computational approaches have become increasingly powerful and dominant in decision-making research. If anticipated regret influences the decoy effect as predicted by the hypothesis, there may be a need for a reconciliation between the different accounts on the mechanism of the decoy effect.

Not much can be drawn from our extension given the limited effect of the low reversibility manipulation on the decoy effect, despite that it was in the predicted direction. It was also possible that anticipated regret partially mediates the effect of low decision reversibility, as evidenced by the results of the manipulation checks (i.e., *Regret-Salient* participants also scored significantly lower on the low reversibility check item compared with *Control* participants). Nevertheless, we need more evidence for both claims, one concerning the effect of decision reversibility, and the other concerning the mediating effect of regret.

Limitations and future directions

Our studies were limited in several ways. Because we aimed at direct replications, we followed the original designs and procedures as closely as possible. The original studies were not without room for improvement, and so were ours given that we did not deviate much from the originals. An example was the absence of a measure for dominance identification in Study 1. Based on the original design, we could not get a sense of the extent to which participants perceived the relative inferiority of the decoys to their targets. As the result, we could not rule out the possibility that the absence of evidence for the decoy effect in Study 1 could be attributed to failure of perceiving dominance. Hence, we suggest future studies implement such measures/checks as a routine and evaluate perceptions of dominance relationships as a moderating factor.

Based on our results regarding the effect of regret salience, we call for a reexamination, ideally in the form of highly-powered direct replications, of other studies that used similar manipulations (e.g., Connolly & Reb, 2012; Reb, 2008), with the aim of (1) evaluating the effectiveness of these manipulations across different decision tasks and contexts and (2) using alternative manipulations to examine the robustness of previous findings and comparing the effectiveness of different manipulations. Such alternatives include telling participants that they will learn the outcomes of unchosen options (e.g., Zeelenberg et al., 1996) and directly asking participants how much they would regret given unsatisfactory decisions (e.g., Reb, 2008; Simonson, 1992). It is possible that because our manipulation here did not refer to alternative outcomes, it was not as effective as the other manipulations in the literature.

Conclusion

We conducted highly powered direct replications of two studies on the decoy effect. We failed to replicate the effect in one study but successfully replicated it in the other, though with a smaller effect size compared with the original. Our reexamination of the effect of regret salience suggests that anticipated regret has a weaker effect on the decoy effect than expected. We call for more direct replications of related studies to assess the factors that influence the size or even existence of the decoy effect, and to compare the effectiveness of different regret salience manipulations in terms of affecting reason-based decision-making.

References

Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin*, *45*(6), 842–850. https://doi.org/10.1177/0146167218798821

Anderson, C. J. (2003). The psychology of doing nothing: Forms of decision avoidance result from reason and emotion. *Psychological Bulletin*, *129*(1), 139–167. https://doi.org/10.1037/0033-2909.129.1.139

Anvari, F., Olsen, J., Hung, W. Y., & Feldman, G. (2020). *Misprediction of affective outcomes due to different evaluation modes: Replication and extension of two distinction bias experiments by Hsee and Zhang (2004)*. https://doi.org/10.13140/RG.2.2.29694.74564/1

Ariely, D. (2008). *Predictably irrational: The hidden forces that shape our decisions*. Harper.

Ariely, D., & Wallsten, T. S. (1995). Seeking subjective dominance in multidimensional space: An explanation of the asymmetric dominance effect. *Organizational Behavior and Human Decision Processes*, *63*(3), 223–232. https://doi.org/10.1006/obhd.1995.1075

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*(2), 527–535. https://doi.org/10.3758/s13428-012-0265-2

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon’s Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *9*(1), 3–5. https://doi.org/10.1177/1745691610393980

Chatterjee, S., Roy, R., & Malshe, A. V. (2011). The role of regulatory fit on the attraction effect. *Journal of Consumer Psychology*, *21*(4), 473–481. https://doi.org/10.1016/j.jcps.2010.05.001

Chen, J., Hui, L. S., Yu, T., Feldman, G., Zeng, S., Ching, T. L., Ng, C. H., Wu, K. W., Yuen, C. M., Lau, T. K., Cheng, B. L., & Ng, K. W. (2020). Foregone opportunities and choosing not to act: Replications of inaction inertia effect. *Social Psychological and Personality Science*. https://doi.org/10.1177/1948550619900570

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

Collaborative Open-science REsearch (CORE). (2020). *Replications and extensions of classic findings in social psychology and judgment and decision making*. https://doi.org/10.17605/OSF.IO/5Z4A8

Connolly, T., & Reb, J. (2012). Regret aversion in reason-based choice. *Theory and Decision*, *73*(1), 35–51. https://doi.org/10.1007/s11238-011-9269-0

Connolly, T., Reb, J., & Kausel, E. E. (2013). Regret salience and accountability in the decoy effect. *Judgment and Decision Making*, *8*(2), 136–149.

Connolly, T., & Zeelenberg, M. (2002). Regret in decision making. *Current Directions in Psychological Science*, *11*(6), 212–216. https://doi.org/10.1111/1467-8721.00203

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Frederick, S., Lee, L., & Baskin, E. (2014). The limits of attraction. *Journal of Marketing Research*, *51*(4), 487–507. https://doi.org/10.1509/jmr.12.0061

Gaudeul, A., & Crosetto, P. (2019). *Fast then slow: A choice process explanation for the attraction effect*. https://hal.archives-ouvertes.fr/hal-02408719/

Gilbert, D. T., & Ebert, J. E. J. (2002). Decisions and revisions: The affective forecasting of changeable outcomes. *Journal of Personality and Social Psychology*, *82*(4), 503–514. https://doi.org/10.1037/0022-3514.82.4.503

Herne, K. (1997). Decoy alternatives in policy choices: Asymmetric domination and compromise effects. *European Journal of Political Economy*, *13*(3), 575–589. https://doi.org/10.1016/S0176-2680(97)00020-7

Highhouse, S. (1996). Context-dependent selection: The effects of decoy and phantom job candidates. *Organizational Behavior and Human Decision Processes*, *65*(1), 68–76. https://doi.org/10.1006/obhd.1996.0006

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, *9*(1), 90–98. https://doi.org/10.1086/208899

Huber, J., Payne, J. W., & Puto, C. P. (2014). Let's be honest about the attraction effect. *Journal of Marketing Research*, *51*(4), 520–525. https://doi.org/10.1509/jmr.14.0208

Huber, J., & Puto, C. (1983). Market boundaries and product choice: Illustrating attraction and substitution effects. *Journal of Consumer Research*, *10*(1), 31–44. https://doi.org/10.1086/208943

Janis, I. L., & Mann, L. (1977). *Decision making: A psychological analysis of conflict, choice, and commitment*. Free Press. https://psycnet.apa.org/record/1978-00284-000

Kaptein, M. C., van Emden, R., & Iannuzzi, D. (2016). Tracking the decoy: Maximizing the decoy effect through sequential experimentation. *Palgrave Communications*, *2*, 16082. https://doi.org/10.1057/palcomms.2016.82

Keck, S., & Tang, W. (2020). When “decoy effect” meets gender bias: The role of choice set composition in hiring decisions. *Journal of Behavioral Decision Making*, *33*(2), 240–254. https://doi.org/10.1002/bdm.2157

Kivetz, R., Netzer, O., & Srinivasan, V. (2004). Extending compromise effect models to complex buying situations and other context effects. *Journal of Marketing Research*, *41*(3), 262–268. https://doi.org/10.1509/jmkr.41.3.262.35993

Król, M., & Król, M. (2019). Inferiority, not similarity of the decoy to target, is what drives the transfer of attention underlying the attraction effect: Evidence from an eye-tracking study with real choices. *Journal of Neuroscience, Psychology, and Economics*, *12*(2), 88–104. https://doi.org/10.1037/npe0000104

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402. https://doi.org/10.1177/2515245918787489

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2019). A brief guide to evaluate replications. *Meta-Psychology*, *3*. https://doi.org/10.15626/MP.2018.843

Lichters, M., Sarstedt, M., & Vogt, B. (2015). On the practical relevance of the attraction effect: A cautionary note and guidelines for context effect experiments. *AMS Review*, *5*(1–2), 1–19. https://doi.org/10.1007/s13162-015-0066-8

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, *15*(3), 215–233. https://doi.org/10.1016/0022-2496(77)90032-3

McAllister, D. W., Mitchell, T. R., & Beach, L. R. (1979). The contingency model for the selection of decision strategies: An empirical test of the effects of significance, accountability, and reversibility. *Organizational Behavior and Human Performance*, *24*(2), 228–244. https://doi.org/10.1016/0030-5073(79)90027-8

Milberg, S. J., Silva, M., Celedon, P., & Sinn, F. (2014). Synthesis of attraction effect research: Practical market implications? *European Journal of Marketing*, *48*(7/8), 1413–1430. https://doi.org/10.1108/EJM-07-2012-0391

Mourali, M., Böckenholt, U., & Laroche, M. (2007). Compromise and attraction effects under prevention and promotion motivations. *Journal of Consumer Research*, *34*(2), 234–247. https://doi.org/10.1086/519151

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Olsson-Collentine, A., Wicherts, J. M., & van Assen, M. A. L. M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*. https://doi.org/10.1037/bul0000294

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188. https://doi.org/10.1177/0963721414531598

Pettibone, J. C. (2012). Testing the effect of time pressure on asymmetric dominance and compromise decoys in choice. *Judgment and Decision Making*, *7*(4), 513–523.

Pettibone, J. C., & Wedell, D. H. (2000). Examining models of nondominated decoy effects across judgment and choice. *Organizational Behavior and Human Decision Processes*, *81*(2), 300–328. https://doi.org/10.1006/obhd.1999.2880

Pittarello, A., Caserotti, M., & Rubaltelli, E. (2019). 'Three is better than two': Increasing donations with the attraction effect. *British Journal of Psychology*. https://doi.org/10.1111/bjop.12428

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Reb, J. (2008). Regret aversion and decision process quality: Effects of regret salience on decision process carefulness. *Organizational Behavior and Human Decision Processes*, *105*(2), 169–182. https://doi.org/10.1016/j.obhdp.2007.08.006

Reb, J., Li, A., & Bagger, J. (2018). Decoy effect, anticipated regret, and preferences for work–family benefits. *Journal of Occupational and Organizational Psychology*, *91*(3), 441–464. https://doi.org/10.1111/joop.12207

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, *108*(2), 370–392. https://doi.org/10.1037/0033-295X.108.2.370

Scheel, A. M., Schijen, M., & Lakens, D. (2020). *An excess of positive results: Comparing the standard Psychology literature with Registered Reports* [Preprint]. https://doi.org/10.31234/osf.io/p6e9c

Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research*, *16*(2), 158–174. https://doi.org/10.1086/209205

Simonson, I. (1992). The influence of anticipating regret and responsibility on purchase decisions. *Journal of Consumer Research*, *19*(1), 105–118. https://doi.org/10.1086/209290

Simonson, I. (2014). Vices and virtues of misguided replications: The case of asymmetric dominance. *Journal of Marketing Research*, *51*(4), 514–519.

Simonson, I. (2015). Mission (largely) accomplished: What's next for consumer BDT-JDM researchers? *Journal of Marketing Behavior*, *1*(1), 9–35. https://doi.org/10.1561/107.00000001

Simonson, I., & Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research*, *29*(3), 281–295. https://doi.org/10.1177/002224379202900301

Sivakumar, K. (2016). A unified conceptualization of the attraction effect. *AMS Review*, *6*(1–2), 39–58. https://doi.org/10.1007/s13162-016-0074-3

Slaughter, J. E., Bagger, J., & Li, A. (2006). Context effects on group-based employee selection decisions. *Organizational Behavior and Human Decision Processes*, *100*(1), 47–59. https://doi.org/10.1016/j.obhdp.2006.01.003

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Stoffel, S. T., Yang, J., Vlaev, I., & von Wagner, C. (2019). Testing the decoy effect to increase interest in colorectal cancer screening. *PLOS One*, *14*(3), e0213668. https://doi.org/10.1371/journal.pone.0213668

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184–197. https://doi.org/10.1016/j.chb.2017.08.038

Trendl, A., Stewart, N., & Mullett, T. L. (2018). *A zero attraction effect in naturalistic choice* (SSRN Scholarly Paper ID 3299524). Social Science Research Network. https://doi.org/10.2139/ssrn.3299524

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological Review*, *121*(2), 179–205. https://doi.org/10.1037/a0036137

Trueblood, J. S., Brown, S. D., Heathcote, A., & Busemeyer, J. R. (2013). Not just for consumers: Context effects are fundamental to decision making. *Psychological Science*, *24*(6), 901–908. https://doi.org/10.1177/0956797612464241

Trueblood, J. S., & Pettibone, J. C. (2017). The phantom decoy effect in perceptual decision making. *Journal of Behavioral Decision Making*, *30*(2), 157–167. https://doi.org/10.1002/bdm.1930rr

Tsiros, M., & Mittal, V. (2000). Regret: A model of its antecedents and consequences in consumer decision making. *Journal of Consumer Research*, *26*(4), 401–417. https://doi.org/10.1086/209571

Tversky, A., & Simonson, I. (1993). Context-dependent preferences. *Management Science*, *39*(10), 1179–1189. https://doi.org/10.1287/mnsc.39.10.1179

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamic models of multialternative choice. *Psychological Review*, *111*(3), 757–769. https://doi.org/10.1037/0033-295X.111.3.757

Wedell, D. H. (1991). Distinguishing among models of contextually induced preference reversals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(4), 767–778. https://doi.org/10.1037/0278-7393.17.4.767

Wedell, D. H., & Pettibone, J. C. (1996). Using judgments to understand decoy effects in choice. *Organizational Behavior and Human Decision Processes*, *67*(3), 326–344. https://doi.org/10.1006/obhd.1996.0083

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wiseman, R., Watt, C., & Kornbrot, D. (2019). Registered reports: An early example and analysis. *PeerJ*, *7*, e6232. https://doi.org/10.7717/peerj.6232

Yang, S., & Lynn, M. (2014). More evidence challenging the robustness and usefulness of the attraction effect. *Journal of Marketing Research*, *51*(4), 508–513. https://doi.org/10.1509/jmr.14.0020

Zeelenberg, M., Beattie, J., van der Pligt, J., & de Vries, N. K. (1996). Consequences of regret aversion: Effects of expected feedback on risky decision making. *Organizational Behavior and Human Decision Processes*, *65*(2), 148–158. https://doi.org/10.1006/obhd.1996.0013

Ziano, I., Yao, J. D., Gao, Y., & Feldman, G. (2020). Impact of ownership on liking and value: Replications and extensions of three ownership effect experiments. *Journal of Experimental Social Psychology*, *89*, 103972. https://doi.org/10.1016/j.jesp.2020.103972

Revisiting the decoy effect: Replication and extension of Ariely and Wallsten (1995) and Connolly, Reb, and Kausel (2013)

Supplementary Materials

Contents

Contributor Roles Taxonomy

The authors' respective contributions are identified in Table 1S based on CRediT (Contributor Roles Taxonomy) (please refer to https://www.casrai.org/credit.html on details and definitions of each of the roles listed below).

Table 1S. *Contributor role taxonomy*

Role	Qinyu Xiao	Gilad Feldman	Shiyuan Zeng
Conceptualization	X	X	X
Pre-registration (writing)	X		X
Data curation	X		
Formal analysis	X		
Funding acquisition		X	
Investigation	X	X	
Pre-registration peer review			X
Pre-registration verification	X		
Data analysis peer review			X
Data analysis verification	X		
Methodology	X	X	X
Project administration	X	X	
Resources		X	
Supervision		X	
Validation	X		
Visualization	X		
Writing – original draft	X		
Writing – review and editing	X	X	

Open Science Disclosures

A *Transparency Report* has been provided on the Open Science Framework.

Data, code, and experiment materials

The data, analysis codes, and other experiment materials and documents mentioned in the manuscript and this supplementary have been shared on the Open Science Framework (link: https://osf.io/vsbzk).

Procedure and data disclosures

Item	***Disclosure***
Data collection	Data analysis method and code were reported using computer-generated random datasets before formal data collection. These results were later replaced with real data.
Conditions reporting	All collected conditions have been reported.
Data exclusions	Details have been reported in this document.
Variables reporting	We confirm that all variables collected for this study have been reported and included in the provided data.

Analysis of the Target Articles

The target articles for this replication research were Ariely and Wallsten (1995) (Experiment 1: Item Preference; Study 1 in our replication) and Connolly, Reb, and Kausel (2013) (Experiment 1; Study 2 in our replication).

Citations of and links to the target articles

Ariely, D., & Wallsten, T. S. (1995). Seeking subjective dominance in multidimensional space: An explanation of the asymmetric dominance effect. *Organizational Behavior and Human Decision Processes*, *63*(3), 223–232. https://doi.org/10.1006/obhd.1995.1075

Connolly, T., Reb, J., & Kausel, E. E. (2013). Regret salience and accountability in the decoy effect. *Judgment and Decision Making*, *8*(2), 136–149. Available at: https://sjdm.org/journal/12/12613a/jdm12613a.pdf

Analysis of target articles

Experiment 1 – Ariely & Wallsten (1995)

Methods

Type of study	Experimental
Design	The target article reported that the experiment followed a two (*Condition*: 1A and 1C) by two (*Item*: Item A and Item C) design, without specifying whether the factors were between- or within-subjects.[1]
Independent variables	***Condition*** (1A and 1C): which of the two items (Item A or Item C) in the choice set was targeted by the decoy Item B. In Condition 1A, the decoy targeted Item A, meaning that the decoy was constructed to be inferior to Item A in one attribute but comparable in others. In Condition 1C, the decoy

[1] Ariely and Wallsten (1995) did not specify whether the factors were between- or within-subjects. From our understanding, *Item* was clearly within-subjects because participants evaluated and assigned points to all three items presented to them for each product category. We were, however, confused if *Condition* was also within-subjects.

The only clue was the *F*-statistics, particularly, the degrees of freedom. The authors reported *F*s for the interactions between the two factors, of which the degrees of freedom were (1, 116). With a sample size of 60 (which was the sample size of the target study), a two-way repeated measures ANOVA would have (1, 59) as the degrees of freedom for the interaction, and a two-way mixed ANOVA would have (1, 58). Neither matched the reported statistics. We suspected that there was something wrong in either the reporting or the calculation in the target paper.

We considered the design to be a mixed one, i.e., *Condition* was between-subjects. We deemed that a mixed design would make more sense than a repeated-measures one. We also took it that the original study followed a mixed design as well and conducted power analysis on that basis.

targeted Item C in a similar way. The way these decoys were constructed was as follows (Ariely & Wallsten, 1995, p. 226):

1. The value on one dimension (*Dimension 1*) of the decoy was made 40% or 50% of the corresponding value of the targeted item for positive dimensions (dimensions that the higher the better for a decision-maker, such as the durability of running shoes), and 140% or 150% for negative dimensions (the lower the better; e.g., price).

2. The values on the other two dimensions (*Dimension 2 and 3;* there were altogether three dimensions on which the items were evaluated) of the decoy were set to be 110% of the values of the targeted item for positive dimensions, and 90% for the negative dimensions.

3. So constructed, a decoy should be substantially less appealing than the targeted item on one dimension (*Dimension 1*). Although the decoy was superior to the targeted item on the other two dimensions, such superiority was less noticeable and was of less subjective importance for decision-makers, as the authors reasoned (Ariely & Wallsten, 1995, p. 224). The items and their attributes used in the experiment was shown in Figure 1S below:

Figure 1S. *Items used in Experiment 1 of Ariely and Wallsten (1995)*

TABLE 1
Descriptions of All Items Used for the Item-Preference Experiment and for the Five Different Item Categories

				Item B	
Products	Dimensions	Item A	Item C	B/1A (A′)	B/1C (C′)
Microwaves	Price ($)–	380	209	532	292.6
	Capacity (ft^3)+	1.8	1.2	2	1.3
	Wattage (W)+	1000	700	1100	770
Running shoes	Comfort+	8.5	5.5	5.1	3.3
	Durability+	6.8	4.4	7.5	4.9
	Price ($)–	90	58.5	81	52.7
Computers	Speed (Hz)+	33	21.5	16.5	10.8
	Memory (MB)+	8	4.4	8.8	4.8
	Price ($)–	1900	1235	1710	1111.5
TVs	Screen size (in)+	20	14	12	8.4
	Price ($)–	650	357.5	585	321.8
	Wattage (W)+	25	15	27.5	16.5
Bicycles	Price ($)–	400	180	560	252
	Weight (LB)–	15	22.5	13.5	20.3
	Wheel base (in)+	52	36.4	57.2	40

Note. Dimensions without units were used on a scale from 0 to 10, where 10 was the most desirable level. The plus and minus signs next to the dimensions indicate the direction of the dimensions, as it was presented to the subjects in the information booklet.

Item (A and C): For each product category in Figure 1S, participants evaluated and assigned a total of **100 points** to three items presented to them – Item A, B (the decoy), and C – based on how appealing these items were. Because participants evaluated both Item A and Item C, this factor was within-subjects.

Participants were asked to evaluate the items based on the preference of a sample rather than their own values. This was to prevent them from disregarding one dimension straightforwardly for being personally unimportant. The participant whose assignment of points reflected most accurately the sample's preference was rewarded $20.

Dependent variables	***Item Preference***: Item preference was operationalized as points that participants assigned to the items according to their relative appeal. Higher points meant higher appeal and equal points meant equal appeal. Participants were required to use up all 100 points.

Results

Sample size before and after exclusion	A total of 60 subjects participated in the main study. No information was provided on the number of subjects in each condition. No exclusion was reported.
Sample description	No information was provided on their ages (which were most likely to be between 17 and 22 as they were students from the introductory psychology pool at the University of North Carolina at Chapel Hill, U.S.) or gender. They received experimental credit for their time, and as said above, the most accurate subject was rewarded $20 in addition.
Analyses and results	The authors did not provide the descriptive statistics of the points assigned to Item A and C in each condition and for each product category. Only figures were available (p. 227), which were presented as Figure 2S below.

Figure 2S. *Descriptive figures of Ariely and Wallsten (1995) Experiment 1*

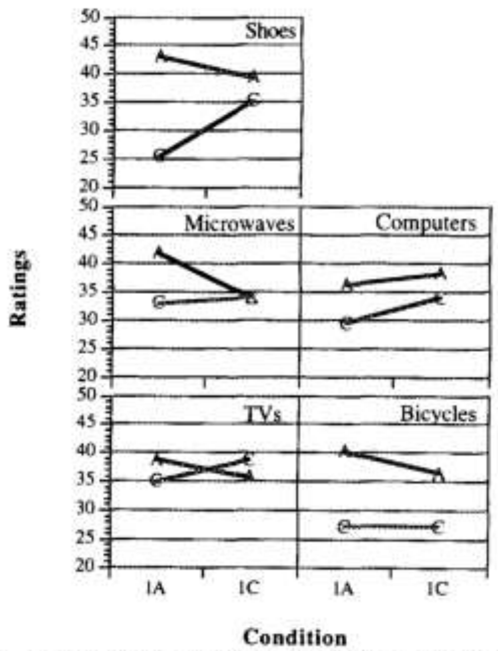


FIG. 3. Ratings of items A and C in condition 1A and 1C separately for each of the products in the Item-Preference Experiment.

We eyeballed these figures and converted them to the numbers in Table 2S.

Table 2S. *Estimated descriptives of Exp. 1 in Ariely and Wallsten (1995)*

		Mean points assigned	
Product category	**Item**	**Condition 1A**	**Condition 1C**
Running shoes	A	43	39
	C	25	35
Microwaves	A	42	33
	C	33	34
Computers	A	36	38
	C	30	34
TVs	A	39	36
	C	35	39
Bicycles	A	40	37
	C	27	27

Note. In Condition 1A/1C, Item A/C was targeted by a decoy.

Five two-way mixed ANOVAs were conducted for each of the product categories. The interactions, which implied reversals in relative preference, were taken to be a measure of the decoy effect size. Table 3S summarized the results (p. 227).

Table 3S. *Mixed ANOVA results of Exp. 1 in Ariely and Wallsten (1995)*

Product category	**Interaction *F*(1, 116)**	***p* value**
Running shoes	18.36	< .01
Microwaves	7.56	< .01
Computers	0.84	.36
TVs	4.922	.03
Bicycles	1.02	.31

Note. The dfs appeared to be problematic, as noted earlier.

The shape of the interactions was also examined. It was predicted that Item A would receive more points in Condition 1A than in Condition 1C, and vice versa for Item C (i.e., a cross-over interaction).

As shown in Table 2S, Item A received more points in Condition 1A than in 1C except for computers; Item C received more points in Condition 1C than in 1A except for bicycles. It should be noted that these conclusions were made by **comparing the means directly**, but not by conducting any statistical tests.

Experiment 1 – Connolly et al. (2013)

Methods

Type of study Experimental

Design A two (*Choice Set*: Set 1 or Set 2) by two (*Condition*: Regret-Salient or Control) fully between-subjects factorial design.

Independent variables ***Choice Set*** (Set 1 or Set 2): A method variable that bore no theoretical interest. It was introduced to counterbalance which option within the core choice set was targeted by a decoy. Figure 3S presents the options as well as their attributes used in the target experiment (Connolly et al., 2013, p. 138).

Figure 3S. *Job options used in Experiment 1 of Connolly et al. (2013)*

Each subject was presented with either Choice Set 1 (Job 1, Job 2, Job 3(a), a decoy targeting Job 2) or Choice Set 2 (Job 1, Job 2, Job 3(b), a decoy targeting Job 1).

Options	Work interest (0–100)	Opportunity for promotion (0–50)
Job 1	83	34
Job 2	74	43
Job 3(a)	71	43
Job 3(b)	80	34

In Choice Set 1, the decoy (Job 3(a)) targeted Job 2, the high-promotion prospect job, whereas in Choice Set 2, the decoy (Job 3(b)) targeted Job 1, the high-work interest job.

Condition (Regret Salient or Control): In the Regret-Salient condition, participants were presented with the text below, which was added after a general description of the decision-making scenario and before they saw the job offers. The text read, "As you make your decision, keep in mind that there is no guarantee that the job you pick will be right for you. You could find yourself in a job you don't like, regretting the decision you made and wishing you had picked one of the other jobs." The possibility of experiencing regret was thusly made salient. In the Control condition, the text was not presented.

Dependent variables

Job Choice: Participants made a choice from the job offers (Figure 3S). The percentages of participants choosing the targets in each condition were calculated and compared with each other and against .5, because a target choice rate of .5 was deemed to signal the absence of the decoy effect.

To illustrate, consider, for example, a hypothetical situation where 30% people in the population prefer Job 1 over Job 2. If there is no decoy effect whatsoever and nobody chooses the obviously inferior Job 3 regardless of which choice set they see, 70% of those presented with Choice Set 1 will choose the targeted option (i.e., Job 2) and 30% of those presented with Choice Set 2 will choose the targeted option (i.e., Job 1). Therefore, if we aggregate results from the two Choice Set conditions, that 50% people choose the targets implies the absence of the decoy effect. It should be noted that, however, drawing this implication presumes there are an equal number of people seeing each of the two choice sets. This assumption was not quite satisfied in the original research (see Table 4S below).

Regret and Justifiability: Subjects were asked to imagine that they had chosen each of the three jobs in the choice set and found their choice dissatisfactory. They then indicated their agreement with the following statements on a 7-point scale (1 = *Completely disagree*; 7 = *Completely agree*):

- "I made a justifiable decision."
- "I regret my decision."

Results

Sample size before and after exclusion

A total of 62 subjects participated in the original experiment. No exclusion was reported. Table 4S presents the sample sizes in each cell of the two-by-two design.

Table 4S. *Sample sizes of Experiment 1 in Connolly et al. (2013)*

	Choice Set 1	Choice Set 2	**Total**
Regret-Salient	22	17	39
Control	13	10	23
Total	35	27	62

As shown, the numbers of participants seeing each choice set were not the same, particularly considering the small overall sample size.

Sample description

Subjects were undergraduate business students, juniors, and seniors at a large Southwestern U.S. university, who participated for course credit (p. 139). Their ages were not specified but were likely to be between 17 and 22. The gender distribution was not reported.

Analyses and results

Descriptive statistics were summarized in Figure 4S below, retrieved from the target article (p. 140).

Figure 4S. *Descriptive statistics of Experiment 1 in Connolly et al. (2013)*

Table 2: Frequency of choosing targeted options by experimental conditions, Experiment 1.

	Choice of target option		Regret		Justifiability	
			Target	Non-target	Target	Non-target
	n	(%)	M (SD)	M (SD)	M (SD)	M (SD)
Control						
Overall	20 (out of 23)	(87.0%)	3.87 (1.96)	5.00 (1.62)	5.43 (1.53)	4.87 (1.29)
Choice set 1	12 (out of 13)	(92.3%)	4.00 (2.00)	5.62 (1.39)	5.00 (1.78)	4.92 (1.61)
Choice set 2	8 (out of 10)	(80.0%)	3.70 (2.00)	4.20 (1.62)	6.00 (.94)	4.80 (.79)
Regret-salient						
Overall	24 (out of 39)	(61.5%)	4.36 (1.86)	4.31 (1.72)	5.14 (1.58)	5.00 (1.43)
Choice set 1	15 (out of 22)	(68.2%)	4.73 (1.70)	4.32 (1.62)	5.27 (1.28)	5.32 (1.25)
Choice set 2	9 (out of 17)	(52.9%)	3.88 (2.00)	4.29 (1.90)	4.93 (1.98)	4.53 (1.60)

The numbers of subjects choosing the targets in each condition and for each choice set were indicated under the "Choice of target option" column. None of the subjects chose the decoys.

Two binomial tests were conducted comparing the target choice rates in the two conditions against .5:

1. For Control condition (20/23), exact binomial p = .000, there was a significant decoy effect.
2. For Regret-Salient condition (24/39), exact binomial p = .20, no significant decoy effect was found.

A Fisher's exact test was performed to compare the choice rates of the target options between the Regret Salient and the Control conditions. 20 out of 23 (87%) Control subjects chose the targeted option whereas 24 out of 39 (61.5%) Regret Salient subjects did so, $\chi^2(1, N = 62) = 4.54$, Fisher's exact test $p = .04$. Participants in the Regret-Salient condition were less likely to choose the targeted option than those in the Control condition.

Comparing the perceived regret and justifiability of target and non-target choices, two 2 (choice set) by 2 (imagined choice) mixed ANOVAs revealed that *Control* subjects judged the choice of the target to be less regrettable than the non-target or the competitor, $F(1, 21) = 8.10$, $p = .01$, $\eta^2_p = 0.28$, and more justifiable, $F(1, 21) = 5.02$, $p = .04$, $\eta^2_p = 0.19$. In contrast, no such differences were found in Regret-Salient participants' evaluations with respect to choice regret, $F(1, 37) = 0.00$, $p = 1.00$, $\eta^2_p = 0.00$, or justifiability, $F(1, 35) = 0.41$, $p = .53$, $\eta^2_p = 0.01$.

Nonetheless, despite that a three-way (condition × imagined choice × choice set) mixed ANOVA revealed a significant interaction between condition and imagined choice on choice regret, $F(1, 58) = 7.14$, $p = .01$, $\eta^2_p = 0.11$, the same analysis on justifiability did not find a significant interaction, $F(1, 56) = 1.22$, $p = .27$, $\eta^2_p = 0.02$. This non-significant interaction suggested that the conclusion with respect to the relationship between regret salience and decision justifiability should be made with caution.

Effect sizes, sensitivity analyses, and a priori power analyses

We calculated the observed effect sizes in the original articles with R and the results were documented in an R Notebook file. The file has been made available on the OSF (https://osf.io/vsbzk).

Experiment 1 – Ariely & Wallsten (1995)

Effect sizes

The effects of our interest were the interactions of the five two (*Condition*: 1A or 1C; between-subjects) by two (*Item*: A and C; within-subjects) mixed ANOVAs. The effect sizes for these interactions were calculated for power analyses. In view of possible errors in reporting the degrees of freedom for the *F*-statistics as mentioned in the previous section, we calculated and analyzed using both the degrees of freedom reported by the authors (1, 116) as well as those that should have applied if the authors in fact did a 2-by-2 mixed ANOVA, i.e., (1, 58).

Table 5S. *Interaction effect sizes in Experiment 1 in Ariely and Wallsten (1995)*

Product category	***F*(1, 116)**	***p***	η^2_p	**90% CI**	**Cohen's *f***
Running shoes	18.36	< .01	0.137	[0.053, 0.233]	0.398
Microwaves	7.56	< .01	0.061	[0.010, 0.142]	0.255
Computers	0.84	.36	0.007	[0.000, 0.053]	0.085
TVs	4.922	.03	0.041	[0.002, 0.113]	0.206
Bicycles	1.02	.31	0.009	[0.000, 0.056]	0.094
	***F*(1, 58)**				
Running shoes	18.36	< .01	0.240	[0.094, 0.378]	0.563
Microwaves	7.56	< .01	0.115	[0.018, 0.247]	0.361
Computers	0.84	.36	0.014	[0.000, 0.099]	0.120
TVs	4.922	.03	0.078	[0.004, 0.201]	0.291
Bicycles	1.02	.31	0.017	[0.000, 0.105]	0.133

Note. 90% CIs for partial eta-squared were calculated, as *F*-test is one-sided (Steiger, 2004).

As shown in Table 5S, some effects were not significant at $\alpha = .05$. Consequently, the 90% CIs for the corresponding partial eta-squares start from zero.

Sensitivity analysis

We conducted a sensitivity analysis on the original study with G*Power 3.1.9.4 (Faul et al., 2007):[2]

[2] "as in SPSS" was selected.

```
F tests - ANOVA: Repeated measures, within-between interaction
Analysis:    Sensitivity: Compute required effect size
Input:       α err prob                        =  0.05
             Power (1-β err prob)              =  0.8
             Total sample size                 =  60
             Number of groups                  =  2
             Number of measurements            =  2
             Nonsphericity correction ε        =  1
Output:      Noncentrality parameter λ         =  8.1170754
             Critical F                        =  4.0068729
             Numerator df                      =  1.0000000
             Denominator df                    =  58.0000000
             Effect size f(U)                  =  0.3740983
```

The analysis indicated that the original design (N = 60, balanced two-way mixed ANOVA) could detect a Cohen's f = 0.374 with .8 power at .05 significance level. All but one of the original effects were lower than this, suggesting a lack of power in the original study. Consequently, it was possible that (1) there was an appreciable decoy effect, and the non-significant results were false negatives due to insufficient power, (2) there was only a small decoy effect, and the significant results reflected overestimated effect sizes, or (3) the size of the decoy effect varied across product categories.

Power analysis

Note. After data collection, we realized that this analysis had issues. The correct analysis is in the supplementary power analysis section below. We keep these for our record and for transparency.

We decided that our Smallest Effect Size of Interest (SESOI) would be Cohen's f = 0.10 (η^2_p = 0.0099), which represents a small effect size (medium = 0.25, large = 0.4) by Cohen's (1988) benchmarks. If we fail to find a significant effect with a design that is highly powered (> .99) to detect f = 0.10, we can assert that the true effect size is so small that it is neither theoretically nor practically meaningful. Considering that our participants would perform the same evaluation task for all product categories, we did power analysis based on one SESOI only (calculation was done with G*Power 3.1.9.4; Faul et al., 2007):

```
F tests - ANOVA: Repeated measures, within-between interaction
Analysis:    A priori: Compute required sample size
Input:       Effect size f                     =  0.1
             α err prob                        =  0.05
             Power (1-β err prob)              =  0.99
             Number of groups                  =  2
             Number of measurements            =  2
             Corr among rep measures           =  0.5
             Nonsphericity correction ε        =  1
Output:      Noncentrality parameter λ         =  18.4800000
             Critical F                        =  3.8617534
             Numerator df                      =  1.0000000
             Denominator df                    =  460
             Total sample size                 =  462
             Actual power                      =  0.9900938
```

Our power analysis showed that, for a two-by-two mixed ANOVA to detect Cohen's $f = 0.1$ at a power of .99, we would need 462 participants (note that the correlation between repeated measures was set at .5).

Experiment 1 – Connolly et al. (2013)

Effect sizes

Table 6S presents the effect sizes and power analysis results for the comparisons between the target choice rates in each condition and the reference rate of .5.

Table 6S. *Effect sizes and power analyses results for the binomial tests in Exp. 1 of Connolly et al. (2013)*

Condition	**Choice of target (%)**	**Exact binomial *p* (compared against .5)**	**Cohen's *h***	**Required *n***
Control	20/23 (87.0%)	< .001	0.832	27
Regret-Salient	24/39 (61.5%)	.20	0.233	339

Note. Please refer to the R Markdown file for calculations. Required *n* were calculated at power = .99.

Table 7S presents the effect size and power analysis result for the Fisher's exact test that was conducted to compare the target choice rates in the Control and Regret-Salient conditions. The test result supported the hypothesis that regret salience reduces the decoy effect.

Table 7S. *Effect size and power analysis result for the Fisher's test in Exp. 1 of Connolly et al. (2013)*

Condition	**Choice of target**	$\chi^2(1, N = 62)$	***p***	φ^2	**Cohen's ω**	**Required *n***
Control	20/23 (87.0%)	4.54	.04	.073	0.271	252
Regret-Salient	24/39 (61.5%)					

Note. 252 is the total number of subjects we would need for the study, and hence for each condition, we would need 126. Required *n* were calculated at power = .99.

Table 8S presents the effect sizes and power analysis results for the mixed ANOVAs.

Note. After data collection, we realized that this analysis had issues. The correct analysis is in the supplementary power analysis section below. We keep these for our record and for transparency.

Table 8S. *Effect sizes and power analysis results for the ANOVAs in Exp. 1 of Connolly et al. (2013)*

	F	p	η^2_p (90% CI)	f	Required n
Regret					
Main effect of imagined choice in the 2 (*Choice Set*) by 2 (*Imagined Choice*) mixed ANOVA for the *Control* condition[3]	8.10	.01	0.278 [0.043, 0.480]	0.621	16 (control)
Interaction between condition and imagined choice in the 2 (*Condition*) by 2 (*Choice Set*) by 2 (*Imagined Choice*) mixed ANOVA[4]	7.14	.01	0.110 [0.015, 0.240]	0.351	52 (both conditions)
Justifiability					
Main effect of imagined choice in the 2 (*Choice Set*) by 2 (*Imagined Choice*) mixed ANOVA for the *Control* condition[5]	5.02	.04	0.193 [0.007, 0.405]	0.489	22 (control)
Interaction between condition and imagined choice in the 2 (*Condition*) by 2 (*Choice Set*) by 2 (*Imagined Choice*) mixed ANOVA[6]	1.22	.27	0.021 [0.000, 0.116]	0.148	276 (both conditions)

Note. Required *n*s were calculated using G*Power 3.1.9.4 at power = .99. Calculation protocols are copied-and-pasted below.

[3] It was observed that choosing the target was less regrettable than choosing the competitor in the *Control* condition.

[4] A knock-out interaction: No significant difference was observed in how regrettable it was to choose the target or the competitor when regret was made salient.

[5] It was observed that choosing the target was more justifiable than choosing the competitor in the *Control* condition.

[6] A knock-out interaction: No significant difference was observed in how justifiable it was to choose the target or the competitor when regret was made salient.

Sensitivity analysis

The original study concluded that regret salience could reduce the decoy effect, which was evidenced by the significant Fisher's exact test result. We conducted a sensitivity analysis on the design of the original study with G*Power 3.1.9.4 (Faul et al., 2007):

```
Exact - Proportions: Inequality, two independent groups (Fisher's exact test)
Options:     Exact distribution
Analysis:    Sensitivity: Compute required effect size
Input:       Tail(s)                          =  Two
             Effect direction                 =  p1≤p2
             Proportion p2                    =  0.87
             α err prob                       =  0.05
             Power (1-β err prob)             =  0.8
             Sample size group 1              =  39
             Sample size group 2              =  23
Output:      Proportion p1                    =  0.5183634
             Actual power                     =  0.8000000
             Actual α                         =  0.0298566
```

Note. Group 1 = Regret-Salience, Group 2 = Control

The sensitivity analysis indicated that the original design would be powered at .8 to detect a significant difference if the target choice rate were 51.8% in the *Regret-Salient* condition. It would have less power if the rate is actually higher (or conversely, the target choice rate in the *Control* condition is lower than 87%). The observed rate in the *Regret-Salient* condition, however, was 61.5% (post hoc power = .644). Despite that the original study found a significant difference, it was not powered enough to reliably detect this difference.

Power analysis for the ANOVAs

Note. After data collection, we realized that this analysis had issues. The correct analysis is in the supplementary power analysis section below. We keep these for our record and for transparency.

Main effect of imagined choice on choice regret:

```
F tests - ANOVA: Repeated measures, within factors
Analysis:    A priori: Compute required sample size
Input:       Effect size f                    =  0.621
             α err prob                       =  0.05
             Power (1-β err prob)             =  0.99
             Number of groups                 =  2
             Number of measurements           =  2
             Corr among rep measures          =  0.5
             Nonsphericity correction ε       =  1
Output:      Noncentrality parameter λ        =  24.6810240
             Critical F                       =  4.6001099
             Numerator df                     =  1.0000000
             Denominator df                   =  14.0000000
             Total sample size                =  16
             Actual power                     =  0.9959143
```

Interaction between condition and imagined choice on choice regret:

```
F tests - ANOVA: Repeated measures, within-between interaction
Analysis:    A priori: Compute required sample size
Input:       Effect size f                    =  0.351
             α err prob                       =  0.05
             Power (1-β err prob)             =  0.99
             Number of groups                 =  4
             Number of measurements           =  2
             Corr among rep measures          =  0.5
             Nonsphericity correction ε       =  1
Output:      Noncentrality parameter λ        =  25.6258080
             Critical F                       =  2.7980606
             Numerator df                     =  3.0000000
             Denominator df                   =  48.0000000
             Total sample size                =  52
             Actual power                     =  0.9901772
```

Main effect of imagined choice on choice justifiability:

```
F tests - ANOVA: Repeated measures, within factors
Analysis:    A priori: Compute required sample size
Input:       Effect size f                    =  0.489
             α err prob                       =  0.05
             Power (1-β err prob)             =  0.99
             Number of groups                 =  2
             Number of measurements           =  2
             Corr among rep measures          =  0.5
             Nonsphericity correction ε       =  1
Output:      Noncentrality parameter λ        =  21.0426480
             Critical F                       =  4.3512435
             Numerator df                     =  1.0000000
             Denominator df                   =  20.0000000
             Total sample size                =  22
             Actual power                     =  0.9917489
```

Interaction between condition and imagined choice on choice justifiability:

```
F tests - ANOVA: Repeated measures, within-between interaction
Analysis:    A priori: Compute required sample size
Input:       Effect size f                    =  0.148
             α err prob                       =  0.05
             Power (1-β err prob)             =  0.99
             Number of groups                 =  4
             Number of measurements           =  2
             Corr among rep measures          =  0.5
             Nonsphericity correction ε       =  1
Output:      Noncentrality parameter λ        =  24.1820160
             Critical F                       =  2.6377910
             Numerator df                     =  3.0000000
             Denominator df                   =  272
             Total sample size                =  276
             Actual power                     =  0.9908064
```

Sample size summary

Based on the analyses above:

- We would need 462 participants for a well-powered replication of Experiment 1 in Ariely and Wallsten (1995).
- We would need at least 339 participants for each condition in a well-powered replication of Experiment 1 in Connolly et al. (2013). Because we added an extension condition (the *Low-Reversibility* condition), we would therefore need 3 × 339 = 1,017 participants.

Because the two studies would be combined and conducted within a single data collection session, we therefore planned to recruit 1,050 participants (slightly more than 1,017 to allow room for exclusion). Our full sample in the end consisted of 1,053 participants, and our sample after exclusion had 1,001. We reported the results for the sample after exclusion in the main text of the published article; the full sample results are reported here in this document.

Supplementary power analysis

As mentioned in the manuscript and in the above, we realized that part of our power analyses had issues because we did not select "as in SPSS" when we used G*Power for repeated-measures designs (Lakens, 2013). The correct analyses based on our original plan should be as follows:

```
F tests - ANOVA: Repeated measures, within-between interaction
Analysis:    A priori: Compute required sample size
Input:       Effect size f(U)                =  0.10
             α err prob                      =  0.05
             Power (1-β err prob)            =  0.99
             Number of groups                =  2
             Number of measurements          =  2
             Nonsphericity correction ε      =  1
Output:      Noncentrality parameter λ       =  18.4000000
             Critical F                      =  3.8465176
             Numerator df                    =  1.0000000
             Denominator df                  =  1840
             Total sample size               =  1842
             Actual power                    =  0.9900258
```

We would need 1,842 participants to detect Cohen's *f* = 0.1 with .99 power. However, our sample after exclusion had only 1,001. We conducted a sensitivity analysis to see how small an effect size can be for our study to detect with .80 and .99 power:

```
F tests - ANOVA: Repeated measures, within-between interaction
Analysis:    Sensitivity: Compute required effect size
Input:       α err prob                      =  0.05
             Power (1-β err prob)            =  0.80
             Total sample size               =  1001
             Number of groups                =  2
             Number of measurements          =  2
             Nonsphericity correction ε      =  1
Output:      Noncentrality parameter λ       =  7.8639784
             Critical F                      =  3.8507840
             Numerator df                    =  1.0000000
             Denominator df                  =  999
             Effect size f(U)                =  0.0887234
```

```
F tests - ANOVA: Repeated measures, within-between interaction
Analysis:    Sensitivity: Compute required effect size
Input:       α err prob                      =  0.05
             Power (1-β err prob)            =  0.99
             Total sample size               =  1001
             Number of groups                =  2
             Number of measurements          =  2
             Nonsphericity correction ε      =  1
Output:      Noncentrality parameter λ       =  18.4078743
             Critical F                      =  3.8507840
             Numerator df                    =  1.0000000
             Denominator df                  =  999
             Effect size f(U)                =  0.1357435
```

Based on our analyses, with the sample we actually collected, we can detect Cohen's *f* = 0.089 with .80 power and Cohen's *f* = 0.136 with .99 power. Therefore, we still have a satisfactory level of power despite not collecting enough to reach the original goal.

For the ANOVAs in Study 2:

Main effect of imagined choice on choice regret:

```
F tests - ANOVA: Repeated measures, within factors
Analysis:    A priori: Compute required sample size
Input:       Effect size f(U)                 =  0.621
             α err prob                       =  0.05
             Power (1-β err prob)             =  0.99
             Number of groups                 =  2
             Number of measurements           =  2
             Nonsphericity correction ε       =  1
Output:      Noncentrality parameter λ        =  19.2820500
             Critical F                       =  4.0343097
             Numerator df                     =  1.0000000
             Denominator df                   =  50.0000000
             Total sample size                =  52
             Actual power                     =  0.9904966
```

Interaction between condition and imagined choice on choice regret:

```
F tests - ANOVA: Repeated measures, within-between interaction
Analysis:    A priori: Compute required sample size
Input:       Effect size f(U)                 =  0.351
             α err prob                       =  0.05
             Power (1-β err prob)             =  0.99
             Number of groups                 =  2
             Number of measurements           =  2
             Nonsphericity correction ε       =  1
Output:      Noncentrality parameter λ        =  18.7265520
             Critical F                       =  3.9033665
             Numerator df                     =  1.0000000
             Denominator df                   =  152
             Total sample size                =  154
             Actual power                     =  0.9903571
```

Main effect of imagined choice on choice justifiability:

```
F tests - ANOVA: Repeated measures, within factors
Analysis:    A priori: Compute required sample size
Input:       Effect size f(U)                 =  0.489
             α err prob                       =  0.05
             Power (1-β err prob)             =  0.99
             Number of groups                 =  2
             Number of measurements           =  2
             Nonsphericity correction ε       =  1
Output:      Noncentrality parameter λ        =  19.1296800
             Critical F                       =  3.9603524
             Numerator df                     =  1.0000000
             Denominator df                   =  80.0000000
             Total sample size                =  82
             Actual power                     =  0.9908802
```

Interaction between condition and imagined choice on choice justifiability:

```
F tests - ANOVA: Repeated measures, within-between interaction
Analysis:    A priori: Compute required sample size
Input:       Effect size f(U)                 =  0.148
             α err prob                       =  0.05
             Power (1-β err prob)             =  0.99
```

```
            Number of groups               =  2
            Number of measurements         =  2
            Nonsphericity correction ε     =  1
Output:     Noncentrality parameter λ      =  18.4431680
            Critical F                     =  3.8525265
            Numerator df                   =  1.0000000
            Denominator df                 =  842
            Total sample size              =  844
            Actual power                   =  0.9900886
```

Our final sample after exclusion had 326 to 337 in each condition (excluding those who chose the decoy). Therefore, we had the power as planned only except for the interaction effect between condition and imagined choice on choice justifiability, which requires 422 in each condition. We had the planned .99 power for the original decoy effects and condition effects.

Designs, Procedures, and Materials

Table of designs

Our study designs are presented below in a tabular format.

Table 9S. *Design of Study 1*

<table>
<tr><td>IV1: Condition (between)
IV2: Item (within)</td><td>IV1: Condition 1A
Item A within the core choice set is targeted by a decoy.</td><td>IV1: Condition 1C
Item C within the core choice set is targeted by a decoy.</td></tr>
<tr><td>IV2: Item A
Participants allocate preference points to Item A.</td><td colspan="2" rowspan="2">DV: Preference

Participants allocate a total of 100 points to three items, i.e., the target, the decoy, and the competitor. The points must be used up and should reflect the relative appeal of each item for an average American. The same number of points should be allocated if participants find two items equally appealing.</td></tr>
<tr><td>IV2: Item C
Participants allocate preference points to Item C.</td></tr>
</table>

Table 10S. *Design of Study 2*

<table>
<tr><td>IV1: Choice Set (between)
IV2: Condition (between)</td><td>IV1: Choice Set 1
Job 3 is a decoy targeting Job 2.</td><td>IV1: Choice Set 2
Job 3 is a decoy targeting Job 1.</td></tr>
<tr><td>IV2: Control condition
Only general introduction texts are presented.</td><td colspan="2" rowspan="3">DVs

DV1: Job Choice

The job chosen by participants is the primary dependent variable.

DV2: Justifiability and Regret of Imagined Choices

Subjects are asked to imagine if they have chosen each of the three options in the choice set and found their decision unsatisfactory. They then indicate their agreement with the following two statements on seven-point Likert scales (1 = Completely disagree; 7 = Completely agree):

1. “I made a justifiable decision.”
2. “I regret my decision.”</td></tr>
<tr><td>IV2: Regret-Salient condition
The possibility of feeling regret after an unsatisfactory decision is made salient.</td></tr>
<tr><td>IV2: Low-Reversibility condition
Low decision reversibility is made salient.</td></tr>
</table>

Procedures

1. Participants give informed consent.
2. They read a brief overview of the structure, length, and nature of the tasks and answer two questions to confirm their ability and willingness to participate. If they do not provide positive answers to both questions, their sessions are terminated, and they are asked to return the HIT.
3. They complete Study 1 and Study 2 in random order.
    a. Study 1:
        i. They read task background and instructions and answer two simple comprehension check questions (see *Materials* section below). They are not allowed to proceed if they fail to answer these two questions correctly. They can correct their answers, however.
        ii. They complete five item evaluation tasks, the order of which are randomized.
    b. Study 2:
        i. They read task background and instructions, the contents of which vary depending on the condition they are randomly assigned to (see *Materials* section below). They also answer two simple comprehension check questions. They are not allowed to proceed if they fail to answer these two questions correctly, but they can change their answers.
        ii. They perform a decision task.
        iii. They are asked to imagine if they have chosen each of the three options in the decision task and then to rate the justifiability of and regret for that hypothetical decision, given that the choice turns out to be unsatisfactory. The options were evaluated in random orders.
4. They complete a funneling section and provide demographic information in the end.

Materials

The Qualtrics survey files have been shared on the Open Science Framework (https://osf.io/vsbzk).

Overview

Overall introduction

<u>Study Outline</u>

This study comprises **two parts**. In one part, you will evaluate products from five product categories. Overall 5 pages. In another part, you will evaluate three job offers. Overall 3 pages.

Please follow the instructions and read the texts carefully.

There are **no right or wrong answers** to any of the questions. Try to choose the answer that **reflects your opinion best**.

Confirmation questions

Are you able to pay close attention to the details in scenarios and questions and carefully answer questions that follow?

- Yes (4)[7]
- I'm not sure, probably not (5)
- No (6)

WARNING: This survey includes attention and comprehension checks. If you do not like participating in surveys with checks, please return the HIT now.

Do you understand the study outline and are you willing to participate in a survey with attention/comprehension checks?

- Yes (1)
- I'm not sure, probably not (2)
- No (3)

---

[7] The numbers in brackets reflect value coding in our data sets.

Note. If participants did not answer "Yes" to both questions, their session would be terminated. These two questions were on the same page with the study outline.

Ending message for each study

After participants complete each study, they read the following before proceeding:

You've completed this part of the study. Thank you very much.

Study 1 – Replication of Experiment 1 in Ariely & Wallsten (1995)

Introduction

Product Evaluations

You are about to evaluate products from five product categories in this part of the study.

Please read the instructions below carefully and answer the questions that follow:

Like in the TV game, *Family Feud*, you are to answer the following questions **in the same way that the average American would answer those**.

You have **one hundred points** and you are given information about three products. Please assign these points in a way that **the number of points reflects their relative preference for the average American**.

Please do it in a way that **a higher number indicates a higher preference** and give two items the same amount of points only if you think they are equally as appealing to the **average American**. **All points must be used**.

Notes. Important information has been bolded and underlined to attract participants' attention.

Comprehension checks

How many points altogether are you going to assign to the items presented to you?

- 20 (1)
- 50 (2)
- 100 (3)

How should you evaluate the items presented to you and assign those points?

- I should evaluate and assign points based on **my own preferences**. (1)
- I should evaluate and assign points to reflect the **preferences of the average American**. (2)
- I am not sure. (3)

Notes. Participants must answer these questions correctly to proceed. They are presented on the same page as the background and instructions. The correct answers are (3) and (2). The order of the options is randomized.

Item evaluation tasks

Participants perform the same item evaluation task on five different product categories: bicycles, running shoes, microwaves, computers, and TVs. The materials are as follows:

<u>Bicycles</u>

You have three **BICYCLES**. You are asked to give each of the products a relative rate (from 100) according to the weights of the preferences.

Note: A bicycle's wheelbase refers to the distance between its front and rear wheels and is usually considered the larger the better because bicycles with larger wheelbases will be more stable.

Condition 1A

	Item A	Item B	Item C
Price ($)	400	560	180
Weight (lb.)	15	13.5	22.5
Wheelbase (in)	52	57.2	36.4

Condition 1C

	Item A	Item B	Item C
Price ($)	400	252	180
Weight (lb.)	15	20.3	22.5
Wheelbase (in)	52	40	36.4

<u>Microwaves</u>

You have three **MICROWAVES**. You are asked to give each of the products a relative rate (from 100) according to the weights of the preferences.

Note: Microwaves with higher wattages perform more efficiently. Those with higher capacities can cook more food at once. Please imagine that you are evaluating these products in the 1990s, when microwaves were more expensive.

Condition 1A

	Item A	Item B	Item C
Price ($)	380	532	209
Capacity (cubic feet)	1.8	2	1.2
Wattage (W)	1000	1100	700

Condition 1C

	Item A	Item B	Item C
Price ($)	380	292.6	209
Capacity (cubic feet)	1.8	1.3	1.2
Wattage (W)	1000	770	700

Running shoes

You have three **RUNNING SHOES**. You are asked to give each of the products a relative rate (from 100) according to the weights of the preferences.

Note: Comfort and durability were rated on a scale from 0 to 10 and a higher score indicates better comfort/durability.

Condition 1A

	Item A	Item B	Item C
Comfort	8.5	5.1	5.5
Durability	6.8	7.5	4.4
Price ($)	90	81	58.5

Condition 1C

	Item A	Item B	Item C
Comfort	8.5	3.3	5.5
Durability	6.8	4.9	4.4
Price ($)	90	52.7	58.5

Computers

You have three **COMPUTERS**. You are asked to give each of the products a relative rate (from 100) according to the weights of the preferences.

Note: Please imagine that you are evaluating these products in the 1990s, when computers were expensive and relatively unpowerful, as some of the figures below may appear outdated by today's standards. Also, computers are more powerful if they have higher speed and larger memory.

Condition 1A

	Item A	Item B	Item C
Speed (Hz)	33	16.5	21.5
Memory (MB)	8	8.8	4.4
Price ($)	1900	1710	1235

Condition 1C

	Item A	Item B	Item C
Speed (Hz)	33	10.8	21.5
Memory (MB)	8	4.8	4.4
Price ($)	1900	1111.5	1235

TVs

You have three **TVs**. You are asked to give each of the products a relative rate (from 100) according to the weights of the preferences.

Note: Please imagine that you are evaluating these products in the 1990s, when TVs were not so advanced, as some of the figures below may appear outdated by today's standards. Also, a TV is generally considered better if it has a larger screen size and a higher wattage.

Condition 1A

	Item A	Item B	Item C
Screen size (in)	20	12	14
Price ($)	650	585	357.5
Wattage (W)	25	27.5	15

Condition 1C

	Item A	Item B	Item C
Screen size (in)	20	8.4	14
Price ($)	650	321.8	357.5
Wattage (W)	25	16.5	15

Notes.

(1) Item B, the decoy, was always presented in the middle to facilitate comparison between the decoy and the target, regardless of whether the target was Item A or Item C.

(2) Explanations were provided on some dimensions, including wheelbase (bicycles), wattage and capacity (microwaves), durability and comfort (running shoes), speed and memory (computers), screen size and wattage (TVs).

(3) The original experiment asked participants to indicate the points that they assigned to each item directly below the corresponding column, like below (Fig. 2, p. 227):

You have three Bicycles. You are asked to give each of the products a relative rate (from 100) according to the weights of the preferences.

	Item A	Item B	Item C
Price	$400	$252	$180
Weight	15 lbs.	20 lbs.	22.5 lbs.
Wheel base	52 in.	40 in.	37 in.
	30	30	30

The Sum is: 90

We, however, found it difficult to implement this set-up in Qualtrics. Hence, in our study, participants indicated points to the side of vertically stacked item labels and below the table of attributes, as below:

You have three **BICYCLES**. You are asked to give each of the products a relative rate (from 100) according to the weights of the preferences.

Note: A bicycle's wheelbase refers to the distance between its front and rear wheels and is usually considered the larger the better because bicycles with larger wheelbases will be more stable.

	Item A	Item B	Item C
Price ($)	400	560	180
Weight (lb.)	15	13.5	22.5
Wheelbase (in)	52	57.2	36.4

Item A	0
Item B	0
Item C	0
Total	0

There was an additional "Total" field that was updated in real time, indicating how many points that participants have already used.

(4) The original experiment started with a default point of 30 for each item. These defaults did not sum up to 100, so participants were forced to adjust the points by mouse clicks. One click resulted in one-point increase/decrease. We refrained from using defaults due to potential anchoring effects. Also, the original experiment allowed only integer, whereas in our experiment, participants could input non-integers if they wish.

Study 2 – Replication and extension of Experiment 1 in Connolly et al. (2013)

Introduction

Job Evaluations

You will evaluate job options in this part of the study.

Please read the instructions below carefully and answer the questions that follow:

Control condition

Imagine that, after an extensive job search, you have narrowed your options to **three jobs**.

The three jobs are similar in every way except for two attributes:

1. Work interest (which you have rated on a **0-100 scale**)
2. Promotion possibilities (rated on a **0-50 scale**)

You will be asked to make a decision from these jobs.

In this part of the study, please evaluate the jobs according to <u>your own preferences</u>.

Regret-Salient condition

Imagine that, after an extensive job search, you have narrowed your options to **three jobs**.

The three jobs are similar in every way except for two attributes:

1. Work interest (which you have rated on a **0-100 scale**)
2. Promotion possibilities (rated on a **0-50 scale**)

You will be asked to make a decision from these jobs.

In this part of the study, please evaluate the jobs according to <u>your own preferences</u>.

<u>IMPORTANT</u>: As you make your decision, keep in mind that there is no guarantee that the job you pick will be right for you. You could find yourself in a job that you don't like, regretting the decision you made and wishing you had picked one of the other jobs.

Low-Reversibility condition

Imagine that, after an extensive job search, you have narrowed your options to **three jobs**.

The three jobs are similar in every way except for two attributes:

1. Work interest (which you have rated on a **0-100 scale**)
2. Promotion possibilities (rated on a **0-50 scale**)

You will be asked to make a decision from these jobs.

In this part of the study, please evaluate the jobs according to <u>your own preferences</u>.

<u>IMPORTANT</u>: As you make your decision, keep in mind that in view of the current economic downturn, companies are restricting and even shrinking their headcounts, making job switching particularly difficult right now and in the coming years (but your current job options are firm and secure; you don't need to worry that they will be retrieved or you will be fired soon).

Note. Condition-specific instruction texts are bolded for emphasis. The following comprehension questions were displayed on the same page as the instructions. Participants had to answer them correctly to proceed.

How many job options do you have?

- 2 (1)
- 3 (2)
- 4 (3)

In what aspect are the job options **NOT** different?

- Work interest (1)
- Promotion possibilities (2)
- Pay (3)

How should you evaluate the jobs presented to you?

- I should evaluate based on **<u>my own preferences</u>** (1)
- I should evaluate based on the **<u>preferences of the average American</u>** (2)
- I am not sure (3)

Note. The correct answers are (2), (3), and (1). The order of the options was randomized.

Manipulation checks

To check whether the manipulation was successful, we asked participants to rate their agreement (on 7-point Likert scales: 1 = *Completely disagree*, 2 = *Disagree*, 3 = *Somewhat disagree*, 4 = *Neither agree nor disagree*, 5 = *Somewhat agree*, 6 = *Agree*, 7 = *Completely agree*) with each of the following two statements:

- "My job choice will surely be right for me."
- "Changing jobs will be easy in the future."

Note. These items are presented on a separate page from the instruction texts.

Decision task

Choice Set 1

Below are the three jobs:

Option	Work interest (0 – 100)	Opportunity for promotion (0 – 50)
Job 1	83	34
Job 2	74	43
Job 3	71	43

Please make your decision.

- Job 1 (1)
- Job 2 (2)
- Job 3 (3)

Choice Set 2

Below are the three jobs:

Option	Work interest (0 – 100)	Opportunity for promotion (0 – 50)
Job 1	83	34
Job 2	74	43
Job 3	80	34

Please make your decision.

- Job 1 (1)
- Job 2 (2)
- Job 3 (3)

Regret and justification

<u>Choice Set 1</u>

Now, imagine how you would feel if you have chosen each of the three jobs and have found it to be unsatisfactory.

ATTENTION: The order of evaluation is NOT necessarily 1, 2, and 3.

Option	Work interest (0 – 100)	Opportunity for promotion (0 – 50)
Job 1	83	34
Job 2	74	43
Job 3	71	43

<u>Choice Set 2</u>

Now, imagine how you would feel if you have chosen each of the three jobs and have found it to be unsatisfactory.

ATTENTION: The order of evaluation is NOT necessarily 1, 2, and 3.

Option	Work interest (0 – 100)	Opportunity for promotion (0 – 50)
Job 1	83	34
Job 2	74	43
Job 3	80	34

Imagine that you have chosen **Job [1 / 2 / 3]** and have found it to be unsatisfactory. Please indicate your agreement with the following statements:

- "I made a justifiable decision."
- "I regret my decision."

Notes. The order in which participants evaluated imagined choices of Job 1, 2, and 3 was randomized uniquely. Only one question was presented on each page, and each presentation was accompanied with the above instruction texts and the table of job options (this was to ensure that participants can easily refer to the table when they make responses). They provided responses on 7-point Likert scales (1 = *Completely disagree*, 2 = *Disagree*, 3 = *Somewhat disagree*, 4 = *Neither agree nor disagree*, 5 = *Somewhat agree*, 6 = *Agree*, 7 = *Completely agree*).

Funneling and demographic information

Instruction

Now you have completed both parts of this study. Thank you much for your time and effort.

Before you leave, we would like you to answer a few questions about this survey (this page) and provide some demographic information (next page).

Funneling questions

How serious were you in filling out this questionnaire?

- 1 – Not at all (1)
- 2 (2)
- 3 (3)
- 4 (4)
- 5 – Very much (5)

What do you think was the purpose of the study? (short-answer question)

Have you ever seen the materials used in this study or similar before? If yes – please indicate where.

- No (1)
- Yes (if yes, please write in the box below regarding where) (2) (text entry box follows)

Help us improve for the next studies. Did you spot any errors? Is there anything missing or wrong, or something that we'd better pay attention to in our next run? Please indicate briefly. (short-answer question)

Please rate your satisfaction with the pay/compensation offered for this MTurk HIT. Your response will not impact your pay in any way.

- Extremely satisfied (1)
- Moderately satisfied (2)
- Slightly satisfied (3)
- Neither satisfied nor dissatisfied (4)

- Slightly dissatisfied (5)
- Moderately dissatisfied (6)
- Extremely dissatisfied (7)

Notes. The numbers in brackets indicate value coding. Response to the question that asks about errors in the questionnaire was optional; others were forced.

Demographic information

What is your age? Please answer using numbers, e.g., 30 (rather than thirty). Write down 99 if you do not wish to disclose this information. (short-answer question)

Please indicate your gender.

- Male (1)
- Female (0)
- Other/rather not disclose (2)

Which country are you originally from, i.e., your country of birth? (short-answer question)

Which country are you currently residing in? (short-answer question)

Please estimate your family's social class.

- Lower class (1)
- Working class (2)
- Lower middle class (3)
- Middle class (4)
- Upper middle class (5)
- Upper class (6)

How would you generally rate your understanding of the English used in this survey?

- Very bad (1)
- Bad (2)
- Poor (3)
- Neither good nor bad (4)
- Fair (5)

- Good (6)
- Very good (7)

Notes. The numbers in brackets indicate value coding. All responses were forced.

Exclusion criteria

We reported the results for the sample after exclusion in the main text. We also performed the same analyses on the full sample and reported the results here in the supplementary (in the *Full sample results* section).

Our exclusion criteria were:

1. Participants who indicated low proficiency of English (< 5 on a 1 – 7 scale)
2. Participants who self-reported not being serious about the survey (< 4 on a 1 – 5 scale)
3. Participants who correctly guessed the hypothesis of this study in the funneling section. To prevent arbitrariness, we decided to exclude those who mentioned the following keywords: ***decoy***, ***target***, or ***attraction***. Nonetheless, we checked all responses to see if there was any person who guessed the hypotheses right but did not mention the keywords. We also excluded these special cases and reported them in our analysis script (it turned out that no one mentioned anything close)
4. Participants who had already seen or done this survey or any similar surveys before (that is, to be excluded, they had to explicitly state that they did the same task before; we did not exclude those who only mentioned that they did some similar tasks)
5. Participants who failed to complete the survey (duration = 0, leave question blank)
6. Participants who were not from or not currently in the U.S.

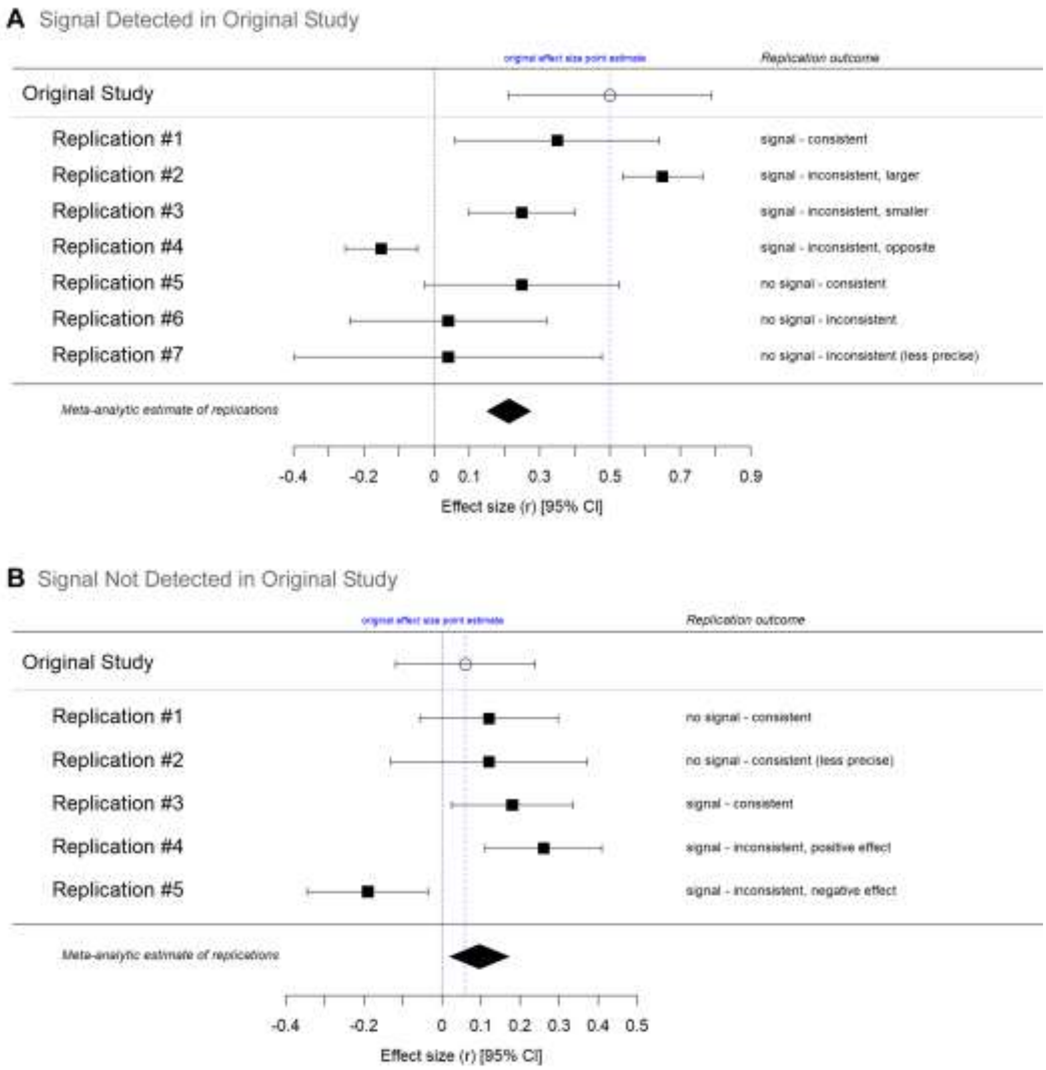
We also excluded participants based on Qualtrics' fraud detection metrics:

- Q_BallotBoxStuffing = 1
- Q_RecaptchaScore < 0.5
- Q_RelevantIDDuplicate = 1
- Q_RelevantIDDuplicateScore ≥ 75
- Q_RelevantIDFraudScore ≥ 30

Replication evaluation

We aimed to compare the replication effects with the original effects in the target article (Table 5S – 8S) using the criteria set by LeBel et al. (2019) (Figure 5S).

Figure 5S



Note. Criteria for evaluation of replication results by LeBel et al. (2019). A simplified replication taxonomy for comparing replication effects to target article original findings.

Figure 6S details the classification of replications using the criteria by LeBel et al. (2018). We summarized our two replications as *very close replications* (Table 11S and 12S).

Figure 6S. *Criteria for evaluating replications by LeBel et al. (2018)*

Target similarity	Highly similar				Highly dissimilar
Category	Direct replication				Conceptual replication
Design facet	**Exact replication**	**Very close replication**	**Close replication**	**Far replication**	**Very far replication**
Effect, Hypothesis	Same/similar	Same/similar	Same/similar	Same/similar	Same/similar
IV Construct	Same/similar	Same/similar	Same/similar	Same/similar	Different
DV Construct	Same/similar	Same/similar	Same/similar	Same/similar	Different
IV Operationalization	Same/similar	Same/similar	Same/similar	Different	
DV Operationalization	Same/similar	Same/similar	Same/similar	Different	
Population (e.g., age)	Same/similar	Same/similar	Same/similar	Different	
IV Stimuli	Same/similar	Same/similar	Different		
DV Stimuli	Same/similar	Same/similar	Different		
Procedural Details	Same/similar	Different			
Physical Setting	Same/similar	Different			
Contextual Variables	Different				

A classification of relative methodological similarity of a replication study to an original study. “Same” (“different”) indicates the design facet in question is the same (different) compared to an original study. IV = independent variable. DV = dependent variable. “Everything controllable” indicates design facets over which a researcher has control. Procedural details involve minor experimental particulars (e.g., task instruction wording, font, font size, etc.). “Similar” category was added to the Lebel et al. (2018) typology to refer to minor deviations aimed to adjust the study to the target sample that are not expected to have major implications on replication success.

Table 11S. *Classification of Study 1 replication, based on LeBel et al.'s (2018) criteria*

Design facet	**Replication (Ariely & Wallsten, 1995, Exp. 1)**
Effect, Hypothesis	Same
IV Construct	Same
DV Construct	Same
IV Operationalization	Same
DV Operationalization	Same
Population (e.g., age)	Similar
IV Stimuli	Same
DV Stimuli	Same
Procedural Details	Similar
Physical Setting	Different
Contextual Variables	Different
Replication classification	Very close replication

Table 12S. *Classification of Study 2 replication, based on LeBel et al.'s (2018) criteria*

Design facet	**Replication (Connolly et al., 2013, Exp. 1)**
Effect, Hypothesis	Same
IV Construct	Same
DV Construct	Same
IV Operationalization	Same
DV Operationalization	Same
Population (e.g., age)	Similar
IV Stimuli	Same
DV Stimuli	Same
Procedural Details	Similar
Physical Setting	Different
Contextual Variables	Different
Replication classification	Very close replication

Comparisons and Deviations

Target and replication samples

Note that the two replication studies used the same sample, because they were run in a single data collection.

Table 13S. *Comparing the sample of Exp. 1 in Ariely and Wallsten (1995) and the replication samples*

	Ariely and Wallsten (1995): Experiment 1	Replication (full sample)	Replication (after exclusion)
Sample size	60	1,053	1,001
Geographic origin	Not specified. Subjects were introductory psychology students at the University of North Carolina at Chapel Hill	US American	US American
Gender	Not specified	536 males (50.9%), 505 females (48.0%), 12 others/rather not disclose (1.1%)	501 males (50.0%), 488 females (48.8%), 12 others/rather not disclose (1.2%)
Median age (years)	Not specified	39	39
Average age (years)	Not specified	42.01	42.24
Age range (years)	Not specified	20 – 82 (two participants did not disclose age)	20 – 82 (two participants did not disclose age)
Medium (location)	Not specified	Computer (online)	Computer (online)
Compensation	Course credit	$1	$1
Year	Around 1994	2020	2020

Table 14S. *Comparing the sample of Exp. 1 in Connolly et al. (2013) and the replication samples*

	Connolly et al. (2013): Experiment 1	Replication (full sample)	Replication (after exclusion)
Sample size	62	1,053	1,001
Geographic origin	Not specified. Subjects were undergraduate business students, juniors, and seniors at a larger Southwestern U.S. university.	US American	US American
Gender	Not specified	536 males (50.9%), 505 females (48.0%), 12 others/rather not disclose (1.1%)	501 males (50.0%), 488 females (48.8%), 12 others/rather not disclose (1.2%)
Median age (years)	Not specified	39	39
Average age (years)	Not specified	42.01	42.24
Age range (years)	Not specified	20 – 82 (two participants did not disclose age)	20 – 82 (two participants did not disclose age)
Medium (location)	Not specified	Computer (online)	Computer (online)
Compensation	Course credit	$1	$1
Year	Around 2013	2020	2020

Original vs. replication

Study 1 – Replication of Experiment 1 in Ariely & Wallsten (1995)

	Original	Replication	Reason for change
Study design	-	-	-
Procedure	The participant who most accurately indicated the pilot sample's preference was awarded $20.	No extra compensation was offered. The participants were asked to make evaluations as an average American would do and refrain from answering based on their personal values. Comprehension checks were added to ensure that subjects understood this point.	It was not possible to follow the original procedure given that our questionnaire was administered online via MTurk. The participants were compensated as soon as they finish. To provide compensation as the authors did, we, however, must wait until data collection is completed (and then determine participants' accuracy). Only until then could we give out the extra award to the participant who had the best accuracy. We included comprehension checks to ensure participants' understanding of the nature of the task.
	The original study had default initial points/weights of 30 for each item. Participants	We did not use defaults. Participants had to make direct inputs.	Defaults may have anchoring effects.

	click buttons to increase or decrease the points. In addition, the original study allowed only integers.	Participants could also input non-integers if they wish.	
Conditions	-	-	-
Materials	Participants received a small brochure with descriptions of the choice items and a detailed description of the dimensions and their meaning. Participants were asked to indicate preference points just below the item columns.	Explanations on some dimensions were included in the scenario descriptions, just above the table where items and their attributes were presented. For three decision scenarios, we asked participants to imagine that they are in the 1990s to accommodate some attribute values that might seem outdated. The blank fields where participants filled in points were vertically stacked, instead of being placed under the item columns.	We adjusted these so that the study can be better administered online.

Study 2 – Replication and extension of Experiment 1 in Connolly et al. (2013)

	Original	Replication	Reason for change
Study design	-	We included a new condition as our extension, so there were more conditions. The original conditions remained the same. We added instruction for evaluation based on personal preferences.	 Since the two studies were combined, and Study 1 asked participants to evaluate based on an average American's preferences, we needed to address the possibility that participants will arrive at this part still using an average American as the reference point.
Procedure	-	We included comprehension checks to exclude nonattentive participants, and manipulation checks to check the effectiveness of manipulation.	This was to ensure that our manipulation worked. The original study did not include any manipulation check, which can undermine the validity of their conclusion.
Conditions	-	-	-

Full sample results

Study 1

Table 15S. *Study 1 descriptive statistics and confirmatory analysis results (full sample)*

Product category	**Item**	**Points** *M* (*SD*) (*n*) Cond. 1A	Cond. 1C	t_{Welch} **(df)**	***p***	**Hedges' *g* (95% CI)**	***F***	***p***	η^2_p **(90% CI)**	η^2_G
Running shoes	A	50.26 (19.60) (528)	53.98 (21.05) (525)	-2.97 (1044.78)	.003	-0.18 [-0.30, -0.06]	0.00	.955	< 0.001	< 0.001
	C	23.98 (15.41) (528)	27.59 (15.11) (525)	-3.84 (1050.80)	< .001	-0.24 [-0.36, -0.12]			[0.000, 0.000]	
Microwaves	A	39.67 (17.17) (524)	40.67 (21.22) (529)	-0.85 (1010.89)	.397	-0.05 [-0.17, 0.07]	0.48	.488	< 0.001	< 0.001
	C	28.25 (17.67) (524)	30.71 (19.37) (529)	-2.15 (1044.00)	.032	-0.13 [-0.25, -0.01]			[0.000, 0.005]	
Computers	A	40.50 (19.56) (523)	41.67 (21.56) (530)	-0.92 (1043.66)	.356	-0.06 [-0.18, 0.06]	0.34	.560	< 0.001	< 0.001
	C	30.85 (18.60) (523)	33.29 (17.16) (530)	-2.21 (1041.93)	.027	-0.14 [-0.26, -0.02]			[0.000, 0.005]	
TVs	A	48.45 (19.70) (530)	48.76 (21.58) (523)	-0.25 (1039.72)	.802	-0.02 [-0.14, 0.11]	0.90	.343	0.001	0.001
	C	29.18 (16.58) (530)	31.50 (16.61) (523)	-2.27 (1050.75)	.024	-0.14 [-0.26, -0.02]			[0.000, 0.006]	
Bicycles	A	34.72 (14.37) (528)	35.95 (20.19) (525)	-1.15 (946.03)	.252	-0.07 [-0.19, 0.05]	1.61	.205	0.002	0.001
	C	30.85 (19.74) (528)	29.58 (17.53) (525)	1.10 (1037.80)	.271	0.07 [-0.05, 0.19]			[0.000, 0.008]	

Note. The *F*-statistics pertained to the interaction between item and condition.

Table 16S. *Study 1 descriptive statistics and confirmatory analysis results (full sample; after removal of extreme responses)*

Product category	**Item**	**Points** M (SD) (n) Cond. 1A	Cond. 1C	t_{Welch} **(df)**	p	**Hedges' g (95% CI)**	F	p	η^2_p **(90% CI)**	η^2_G
Running shoes	A	49.69 (18.55) (518)	52.95 (19.46) (507)	-2.74 (1018.09)	.006	-0.17 [-0.29, -0.05]	0.19	.661	< 0.001	< 0.001
	C	24.24 (14.87) (518)	28.36 (14.15) (507)	-4.54 (1022.20)	< .001	-0.28 [-0.41, -0.16]			[0.000, 0.004]	
Microwaves	A	40.01 (16.26) (514)	40.03 (19.51) (511)	-0.02 (988.81)	.983	0.00 [-0.12, 0.12]	1.70	.192	0.002	0.001
	C	28.06 (16.56) (514)	30.65 (17.86) (511)	-2.40 (1016.26)	.016	-0.15 [-0.27, -0.03]			[0.000, 0.008]	
Computers	A	40.61 (18.71) (512)	40.97 (20.10) (515)	-0.29 (1020.57)	.770	-0.02 [-0.14, 0.10]	0.76	.383	0.001	0.001
	C	31.30 (18.16) (512)	33.48 (15.88) (515)	-2.05 (1005.47)	.041	-0.13 [-0.25, -0.01]			[0.000, 0.006]	
TVs	A	48.04 (18.82) (522)	48.07 (20.53) (512)	-0.03 (1020.61)	.979	0.00 [-0.12, 0.12]	1.74	.187	0.002	0.001
	C	29.44 (16.06) (522)	32.17 (16.15) (512)	-2.73 (1031.37)	.007	-0.17 [-0.29, -0.05]			[0.000, 0.008]	
Bicycles	A	35.14 (13.97) (521)	35.28 (18.31) (510)	-0.13 (951.72)	.893	-0.01 [-0.13, 0.11]	0.18	.672	< 0.001	< 0.001
	C	30.51 (18.86) (521)	29.85 (16.35) (510)	0.60 (1014.20)	.550	0.04 [-0.08, 0.16]			[0.000, 0.004]	

Note. The F-statistics pertained to the interaction between item and condition.

Table 17S. *Comparing points received by targets and by decoys (full sample)*

Product category	**Cond.**	**Comparison**	***t***	**df**	***p***	***d***	**95% CI**
Running shoes	1A	A – B	18.75	527	< .001	0.82	[0.72, 0.91]
	1C	C – B	11.20	524	< .001	0.49	[0.40, 0.58]
Microwaves	1A	A – B	5.51	523	< .001	0.24	[0.15, 0.33]
	1C	C – B	1.81	528	.071	0.08	[-0.01, 0.16]
Computers	1A	A – B	8.90	522	< .001	0.39	[0.30, 0.48]
	1C	C – B	7.64	529	< .001	0.33	[0.24, 0.42]
TVs	1A	A – B	19.29	529	< .001	0.84	[0.74, 0.94]
	1C	C – B	11.48	522	< .001	0.50	[0.41, 0.59]
Bicycles	1A	A – B	0.25	527	.804	0.01	[-0.07, 0.10]
	1C	C – B	-4.28	524	< .001	-0.19	[-0.27, -0.10]

Note. All effects were expected to be positive, since the targets should be perceived more attractive, and be given more points, than their decoys.

Table 18S. *Comparing points received by targets and by decoys (full sample; after removal of extreme responses)*

Product category	**Cond.**	**Comparison**	***t***	**df**	***p***	***d***	**95% CI**
Running shoes	1A	A – B	19.01	517	< .001	0.84	[0.74, 0.94]
	1C	C – B	12.62	506	< .001	0.56	[0.47, 0.65]
Microwaves	1A	A – B	6.12	513	< .001	0.27	[0.18, 0.36]
	1C	C – B	1.21	510	.227	0.05	[-0.03, 0.14]
Computers	1A	A – B	10.22	511	< .001	0.45	[0.36, 0.54]
	1C	C – B	7.64	514	< .001	0.34	[0.25, 0.43]
TVs	1A	A – B	19.56	521	< .001	0.86	[0.76, 0.96]
	1C	C – B	12.43	511	< .001	0.55	[0.46, 0.64]
Bicycles	1A	A – B	0.71	520	.475	0.03	[-0.05, 0.12]
	1C	C – B	-4.69	509	< .001	-0.21	[-0.30, -0.12]

Note. All effects were expected to be positive, since the targets should be perceived more attractive, and be given more points, than their decoys.

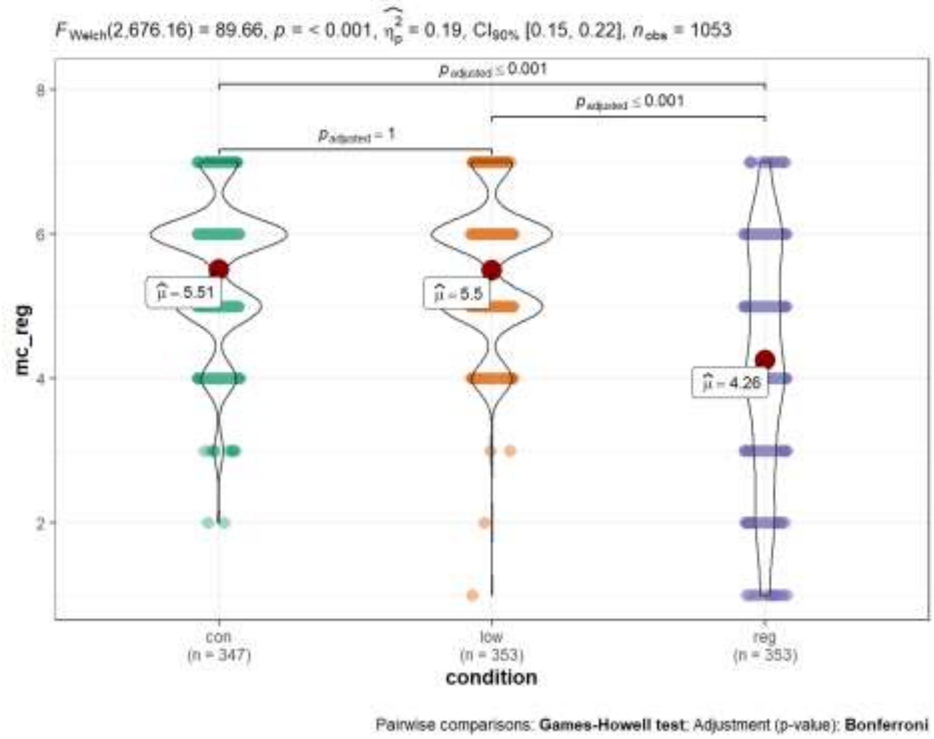
Study 2

Table 19S. *Descriptive statistics of Study 2 (full sample)*

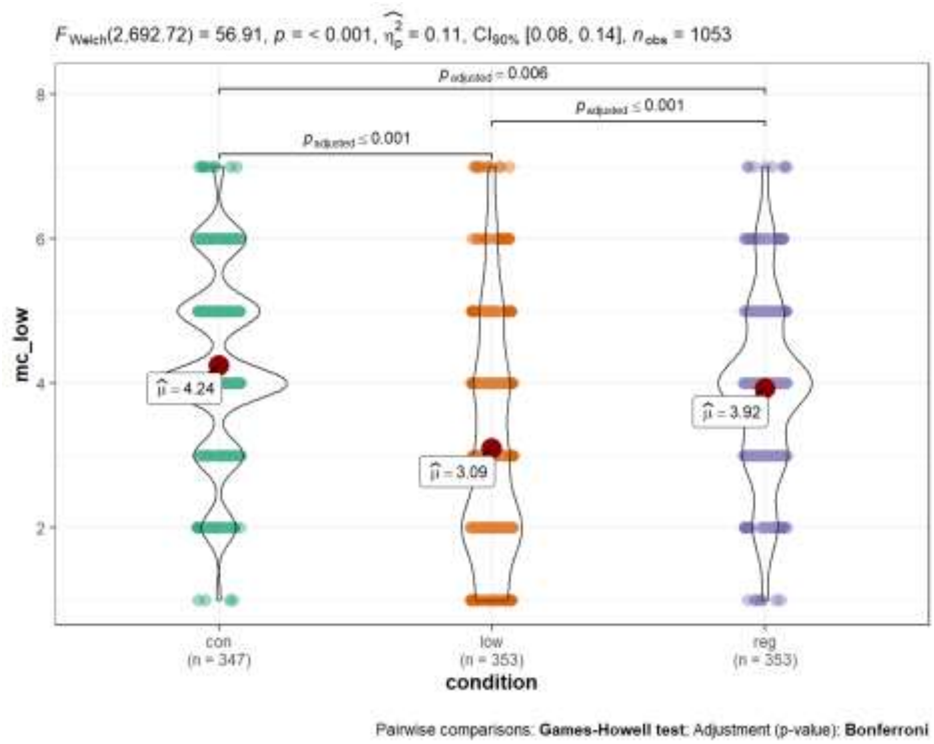
	Choice of the target option		Regret		Justifiability	
	N / Total No. of choices	(%)	Target *M* (*SD*)	Non-target *M* (*SD*)	Target *M* (*SD*)	Non-target *M* (*SD*)
Control (*n* = 347, 2 selected the decoy, 0.6%)						
Overall	221 / 345	64.1%	4.68 (1.69)	4.85 (1.75)	5.47 (1.32)	5.12 (1.42)
Choice set 1	107 / 170	62.9%	4.81 (1.65)	4.79 (1.70)	5.42 (1.32)	5.42 (1.38)
Choice set 2	114 / 175	65.1%	4.56 (1.72)	4.91 (1.80)	5.52 (1.32)	4.83 (1.40)
Regret-Salient (*n* = 353, 6 selected the decoy, 1.7%)						
Overall	196 / 347	56.5%	4.70 (1.71)	4.78 (1.73)	5.16 (1.49)	5.14 (1.44)
Choice set 1	91 / 179	50.8%	4.70 (1.71)	4.62 (1.74)	5.03 (1.36)	5.41 (1.28)
Choice set 2	105 / 168	62.5%	4.71 (1.72)	4.95 (1.72)	5.29 (1.61)	4.86 (1.55)
Low-Reversibility (*n* = 353, 5 selected the decoy, 1.4%)						
Overall	202 / 348	58.0%	4.83 (1.61)	4.87 (1.66)	5.39 (1.33)	5.18 (1.45)
Choice set 1	100 / 176	56.8%	4.85 (1.57)	4.66 (1.67)	5.24 (1.29)	5.48 (1.36)
Choice set 2	102 / 172	59.3%	4.81 (1.65)	5.08 (1.64)	5.53 (1.35)	4.87 (1.49)

Manipulation checks (full sample)

Regret salience



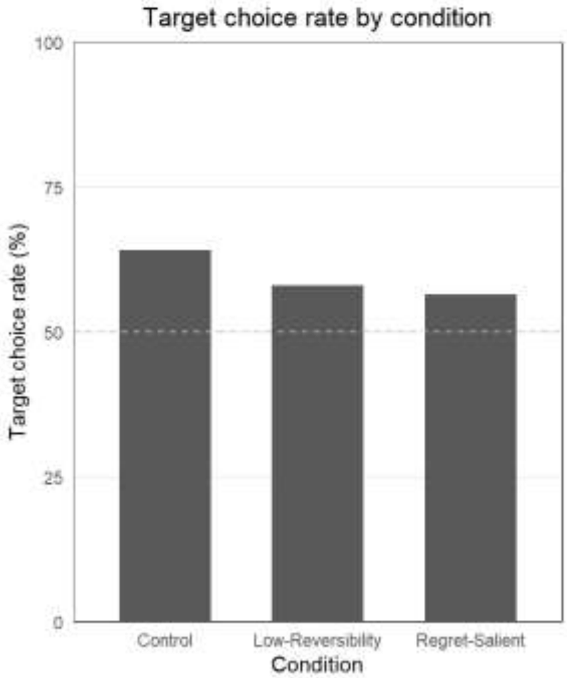
Low decision reversibility



Decoy effects

Table 20S. *Exact binomial tests in Study 2 (full sample)*

Condition	**Target choice**	**Total choice**	**%**	**95% CI**	***p***	**Cohen's *h***	**95% CI**
Con.	221	345	64.1	[58.7, 69.1]	< .001	0.29	[0.18, 0.39]
R.S.	196	347	56.5	[51.1, 61.8]	.018	0.13	[0.02, 0.24]
L.R.	202	348	58.0	[52.7, 63.3]	.003	0.16	[0.06, 0.27]



Condition effects

Table 21S. *Fisher's exact test in Study 2 (full sample)*

Condition 1	Condition 2	χ^2	*p*	φ	95% CI
R.S.	L.R.	0.17	.702	0.016	[0.000, 0.088]
Con.	L.R.	2.63	.119	0.062	[0.000, 0.136]
Con.	R.S.	4.14	.044	0.077	[0.000, 0.152]

Evaluation of imagined choices

Please refer to the R Notebook files shared on the OSF.

Additional tables and figures

Study 1

Table 22S. *Study 1 descriptive statistics and confirmatory analysis results (after exclusion and removal of extreme responses)*

Product category	**Item**	**Points *M* (*SD*) (*n*)** Cond. 1A	Cond. 1C	t_{Welch} **(df)**	***p***	**Hedges' *g* (95% CI)**	***F***	***p***	η^2_p **(90% CI)**	η^2_G
Running shoes	A	49.66 (18.52) (492)	53.32 (19.59) (483)	-2.99 (967.64)	.003	-0.19 [-0.32, -0.07]	0.03	.874	< 0.001	< 0.001
	C	24.33 (14.95) (492)	28.31 (14.24) (483)	-4.25 (972.14)	< .001	-0.27 [-0.40, -0.15]			[0.000, 0.002]	
Microwaves	A	39.94 (16.21) (488)	40.22 (19.69) (488)	-0.24 (939.43)	.813	-0.02 [-0.14, 0.11]	0.83	.364	0.001	0.001
	C	28.46 (16.55) (488)	30.58 (18.03) (488)	-1.91 (966.98)	.057	-0.12 [-0.25, 0.00]			[0.000, 0.007]	
Computers	A	40.52 (18.62) (491)	40.91 (20.16) (487)	-0.31 (968.81)	.757	-0.02 [-0.15, 0.11]	0.95	.330	0.001	0.001
	C	31.31 (18.05) (491)	33.78 (15.93) (487)	-2.27 (963.04)	.024	-0.14 [-0.27, -0.02]			[0.000, 0.007]	
TVs	A	48.14 (18.95) (495)	48.17 (20.51) (490)	-0.02 (975.33)	.981	0.00 [-0.13, 0.12]	1.38	.241	0.001	0.001
	C	29.63 (16.18) (495)	32.13 (16.15) (490)	-2.42 (982.93)	.016	-0.15 [-0.28, -0.03]			[0.000, 0.008]	
Bicycles	A	35.26 (14.00) (498)	35.27 (18.35) (484)	-0.01 (903.30)	.994	0.00 [-0.13, 0.12]	0.15	.698	< 0.001	< 0.001
	C	30.51 (18.83) (498)	29.77 (16.28) (484)	0.66 (966.91)	.511	0.04 [-0.08, 0.17]			[0.000, 0.004]	

Note. The *F*-statistics pertained to the interaction between item and condition.

Table 23S. *Comparing points received by targets and by decoys (after exclusion and removal of extreme responses)*

Product category	**Cond.**	**Comparison**	***t***	**df**	***p***	***d***	**95% CI**
Running shoes	1A	A – B	18.57	491	< .001	0.84	[0.74, 0.94]
	1C	C – B	12.86	482	< .001	0.59	[0.49, 0.68]
Microwaves	1A	A – B	6.23	487	< .001	0.28	[0.19, 0.37]
	1C	C – B	1.21	487	.227	0.05	[-0.03, 0.14]
Computers	1A	A – B	9.86	490	< .001	0.44	[0.35, 0.54]
	1C	C – B	7.97	486	< .001	0.36	[0.27, 0.45]
TVs	1A	A – B	19.26	494	< .001	0.87	[0.76, 0.97]
	1C	C – B	12.14	489	< .001	0.55	[0.45, 0.64]
Bicycles	1A	A – B	0.91	497	.365	0.04	[-0.05, 0.13]
	1C	C – B	-4.72	483	< .001	-0.21	[-0.30, -0.12]

Note. All effects were expected to be positive, since the targets should be perceived more attractive, and be given more points, than their decoys.

Pre-exclusion vs. post-exclusion

Table 24S. *Decoy effects in Study 1 (pre- vs. post-exclusion)*

Product category	Pre-exclusion			Post-exclusion		
	F	p	η_p^2 (90% CI)	F	p	η_p^2 (90% CI)
Running shoes	0.00	.955	< 0.001 [0.000, 0.000]	0.14	.707	< 0.001 [0.000, 0.004]
Microwaves	0.48	.488	< 0.001 [0.000, 0.005]	0.05	.831	< 0.001 [0.000, 0.002]
Computers	0.34	.560	< 0.001 [0.000, 0.005]	0.36	.547	< 0.001 [0.000, 0.005]
TVs	0.90	.343	0.001 [0.000, 0.006]	0.68	.409	0.001 [0.000, 0.006]
Bicycles	1.61	.205	0.002 [0.000, 0.008]	1.14	.285	0.001 [0.000, 0.007]

Table 25S. *Decoy effects in Study 2 (pre- vs. post-exclusion)*

Condition	Post-exclusion			Pre-exclusion		
	Target choice rate	Exact binomial p	Cohen's h (95% CI)	Target choice rate	Exact binomial p	Cohen's h (95% CI)
Control	216 / 337 (64.1%)	< .001	0.29 [0.18, 0.39]	221 / 345 (64.1%)	< .001	0.29 [0.18, 0.39]
R.-S.	185 / 326 (56.7%)	.017	0.14 [0.03, 0.24]	196 / 347 (56.5%)	.018	0.13 [0.02, 0.24]
L.-R.	191 / 327 (58.4%)	.003	0.17 [0.06, 0.28]	202 / 348 (58.0%)	.003	0.16 [0.06, 0.27]

Table 26S. *Condition effects in Study 2 (pre- vs. post-exclusion)*

Condition (vs. Control)	Post-exclusion			Pre-exclusion		
	χ^2	Fisher's exact p	Phi (95% CI)	χ^2	Fisher's exact p	Phi (95% CI)
R.-S.	3.74	.057	0.075 [0.000, 0.075]	4.14	.044	0.077 [0.000, 0.152]
L.-R.	2.26	.151	0.058 [0.000, 0.134]	2.63	.119	0.062 [0.000, 0.136]

Additional Information

Data collection procedures

The studies were conducted on Amazon Mechanical Turk via CloudResearch with U.S. participants. We imposed the following settings in recruiting our participants:

- Participants were paid $1.00 as a fixed participation compensation. This amount was determined by multiplying the expected completion time with the minimal federal wage in the U.S. (i.e., around $0.125 per minute).
- The expected completion time was set at 8 minutes in advance.
- The most time we allowed each worker to complete the study was 30 minutes.
- We limited all workers' HIT Approval Rate to be between 95% and 100%.
- We limited each worker's number of HITs approved to be between 5,000 and 100,000.
- We blocked Suspicious Geocode Locations and Universal Exclude List Workers.
- We blocked duplicate IP addresses and duplicate geolocation.
- We enabled HyperBatch so that all eligible workers were able to participate in our HIT immediately after the survey was launched.
- We restricted workers' location to be in the U.S.

References

Ariely, D., & Wallsten, T. S. (1995). Seeking subjective dominance in multidimensional space: An explanation of the asymmetric dominance effect. *Organizational Behavior and Human Decision Processes*, *63*(3), 223–232. https://doi.org/10.1006/obhd.1995.1075

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

Connolly, T., Reb, J., & Kausel, E. E. (2013). Regret salience and accountability in the decoy effect. *Judgment and Decision Making*, *8*(2), 136–149.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402. https://doi.org/10.1177/2515245918787489

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2019). A brief guide to evaluate replications. *Meta-Psychology*, *3*. https://doi.org/10.15626/MP.2018.843

Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*(2), 164–182. https://doi.org/10.1037/1082-989X.9.2.164