Reply to PCIRR decision letter reviews #657: Arkes (1996) replication

We would like to thank the editor and the reviewers for their useful suggestions and below we provide a detailed response to each item. We also provide a summary table of changes. Please note that the editor's and reviewers' comments are in bold with our reply underneath in normal script.

A track-changes comparison of the previous submission and the revised submission can be found on: https://draftable.com/compare/FpHOLspcOopU

A track-changes manuscript is provided with the file: "PCIRR-S1-RNR-Arkes-1996-replication-extension-registered-report-main-manuscript-track-changes.docx" (https://osf.io/fkjs9)

Summary of changes

Below we provide a table with a summary of the main changes to the manuscript and our response to the editor and reviewers:

Section	**Actions taken in the current manuscript**
General	Ed: We changed the description of our planned experiments from "studies" to "scenarios." Ed & R1: We changed the labels in Scenario 2 to "Waste (no rebate) v.s. No waste (rebate)" to keep consistency.
Introduction	Ed: We added a preview of the extensions and the overall motivation for the secondary replication goal. We edited content regarding Reasons to explain the limitations of the target study and content regarding Perceived Wastefulness in the "Extensions" section.
Methods	Ed: We added the description of the hypotheses in the "Data analysis strategy" section. We rewrote and consolidated the supplementary text in the "Participants" section. Ed & R1: We changed the inconsistency between the manuscript and data analysis in the Perceived wastefulness extension. R1: We explained "unified data collection." R2: We added Likelihood ratio tests in addition to Bayesian analyses.

Section	Actions taken in the current manuscript
Results	Ed & R1: We clarified the use of mixed ANOVA in Reason extension for Scenarios 2 and 3. We corrected the simulated mean “*M* = -0.04”. R1: We added descriptive data in Table 15.
Discussion	R2: We added planned discussions for Stage 2.

Note. Ed = Editor, R1/R2/R3 = Reviewer 1/2/3

[We note that we are not familiar with the titles and ranks of the reviewers, and looking for that information proves tricky. To try and err on the side of caution, we refer to all reviewers with the rank Dr./Prof. . We apologize for any possible misalignments and are happy to amend that in future correspondence.]

Reply to Editor: Dr./Prof. Doug Markant

> **Thank you for submitting your Stage 1 registered report entitled "Revisiting the Psychology of Waste: Replication and Extensions Registered Report of Arkes (1996)" to PCI: Registered Reports. I have now received comments from two expert reviewers and have also read the report myself. Overall, we're in agreement that your submission has several strengths, including a good justification for conducting a replication of the target article, well-documented plans for the study and clear criteria for evaluating the outcome of the replication, well-justified modifications to the original article, and a number of proposed extensions that improve upon the original study's methodological rigor.**
>
> **Based on my own reading and reviewers' comments, I've summarized below the main points that should be addressed in a revision prior to an IPA.**

Thank you for the reviews obtained, your feedback, and the invitation to revise and resubmit.

> **Major points**
>
> **1. Some aspects of the planned analyses need to be clarified. The description of the analyses starting on pg. 34 does not entirely match the simulated results. As reviewer TC notes, it's not clear how you'll use the ANOVA to draw conclusions about the role of the different reasons (relative to utility maximization) in wastefulness judgments.**
>
> **In general, the "Data analysis strategy" section would be improved by reiterating the hypotheses that are targeted by each set of tests (especially for the proposed extensions). The Bayesian analyses need more explanation about the approach and justification for the choice of prior (see also reviewer QA's suggestion about an alternative that doesn't rely on Bayes Factors). Reviewer TC also points out some inconsistencies between descriptions of the analyses of perceived wastefulness.**

Response: Thank you. We agree that reiterating the hypotheses will make the "Data analysis strategy" section easier to follow, and amended accordingly.

For ANOVA interpretation, please see the detailed reply #4 to Reviewer #1.

For Bayesian analyses, we employed the default Cauchy prior of 0.707. This choice represents a weakly informative prior, characterized by a heavy-tailed distribution. By adopting this approach, we intentionally avoid strong assumptions regarding the effect under investigation, while simultaneously accommodating the possibility of capturing both small and large effect sizes. We adopt a passive stance, refraining from assuming any specific effect or its absence, and deliberately selecting a Bayes Factor that lacks substantial informativeness. The Bayes Factor set 0.707 is considered a “neutral” Bayes Factor, in that it does not “expect or not expect” an effect. The replication - based on a well-powered Registered Report - can help shift that fairly neutral prior to a better prior. Furthermore, other researchers can easily substitute alternative Bayes Factors to explore their impact on the results.

Action: We made slight adjustments to our framing of the hypotheses, reflected in Table 1 and the Study Design table, and restated the hypotheses for each study in the “Data analysis strategy” and the “Results” sections to better align the stated hypotheses with our data analysis.

> **2. Given that the within-subjects design is a major deviation from the target article, I agree with reviewer TC’s suggestion that an evaluation of order effects should be carried out regardless of the outcome of any other tests. If there is an effect of order, examining the scenario that is presented first in a follow-up analysis seems sensible but will change the sensitivity to detect the target effect sizes and may leave you with a smaller sample than the minimum of 240, so an alternative analysis approach that accounts for order might be preferable.**

Thank you for the feedback. We understand and have acknowledged this concern.

We please ask that you reconsider this request, and will detail our reasons below. Yet, if you still feel unconvinced by these arguments we will gladly amend accordingly.

As noted in our manuscript, this is not the first time we are doing this kind of experimental design combining different studies, also implemented in more than dozen PCIRR replication projects. We have taken many steps to ensure that this would not impact things, such as in the randomization of order, and in our planned analyses. In our many previous replications, order has not been an issue, and this has been confirmed again with data collections in four PCIRR Stage 1 projects we collected data for this year (in a replication of seven problems reviewed in Read, 1999). For example, some of the most comprehensive demonstrations of that were with our replication of review articles where we ran many studies of the same phenomenon (e.g., 17 effects in our PCIRR replication of Thaler, 1999, in https://osf.io/4ca98, or our replication of nine problems in Kahneman and Tversky, 1972, in https://osf.io/nhqc4/, both concluded as successful), with no indication of order effects. It would be somewhat disappointing for robustness and generalizability of these phenomena if something like order or context would impact these findings. Overall, we (CORE team, 2024) have concluded over 120 replications of heuristics and biases in judgment in decision-making, with dozens using this design, all of them conducted online and with many of them combining many problems into a single unified design, summarizing very high replicability rates, and no indication of order effects. We mentioned some of the completed projects in our methods section: Petrov et al., 2023; Vonasch et al., 2023; Yeung & Feldman, 2022; Zhu & Feldman, 2023; two of those were Registered Reports with PCIRR, and there are many more.

Why not run those analyses anyways? Running all these analyses regardless of the results involves at least doubling the number of analyses, increasing decision flexibility, impacting readability and interpretability, when it is not clear what the added benefits are. It is not about effort/work in running those, these are very straightforward analyses to run, but rather it is about what to do with those once we run them. Consider the following as some of the examples:

- With 3 studies/scenarios in random order, there are many ways to analyze such a moderator: is it the positioning in the 3? Is it which of the specific studies came before it?
- What to conclude if a study is supported (with a signal) in the higher powered full sample but is not supported when only analyzing it as the first study? What to do in the opposite situation when it is supported as the first study but not with the better powered full sample?

To try and give some perspective to this issue, it might be helpful to compare those to other deviations in our replication from the original. One could argue that the participants' age, or education, or time (time of day, day of the week, month of the year, or season, etc.) are influencing factors, and that we should therefore always test age, education, or time as moderators, and then - if we see those as influencing - then rerun the main analyses focusing only on the time (whichever definition), education background, or age group as in the original. These might or might not be influencing factors, we do not know, we have no reason to suspect they would be, as these are factors that are not related to the core generalizable theory. Each possible analysis for each one of these factors adds another forking path of complexity and impacting ability to interpret the results. Especially given our experience with order - we saw no reason to suspect in advance that these would matter.

We therefore thought - given the strengths of the unified design - how to best reassure reviewers while minimizing flexibility and maximizing utility? The added benefits of running these analyses for replications are especially when the moderator may potentially cause the effect to go from signal to no-signal. If the effects replicate well, then there is little benefit in the additional analyses, these are not the core of the theory nor are these the main goals for the replication, just in the same way that the replication does not focus on age, education, or time. If someone were to later want to look at those factors, we make everything available and they are welcome to run additional analyses. However, when effects do not replicate well, we can conduct a series of exploratory analyses.

> **3. The purpose of the attention checks at the beginning of the survey is somewhat unclear. It sounds like participants will have to answer Yes to each of them in order to proceed. My impression is that if participants fail the attention checks they are given a chance to correct their response (although I couldn't quite tell from the survey flow). Perhaps this prompts people to pay closer attention at least momentarily, but it seems there's no plan to exclude people based on whether they initially fail these questions, correct?**

You and the reviewers have direct access to the survey's preview to examine how the survey would behave, and we welcome to try out and see what happens when participants fail to answer "yes" in the consent form. We wrote:

> [*For review: The Qualtrics survey .QSF file and an exported DOCX file are provided on the OSF folder. A preview link of the Qualtrics survey is provided on:* https://hku.au1.qualtrics.com/jfe/preview/previewId/1d9e5f02-121f-4083-a143-79ee31ad8687/SV_6QIXLrPc6cqrWYu?Q_CHL=preview&Q_SurveyVersionID=current]

To clarify the points raised in your specific concern here we would like to clarify that not answering "yes" does not allow the participants to proceed, the survey ends, and the participants are requested to return the task (meaning, that they do not take part). Qualtrics is set so that the same person does not allow a participant to retake (using cookie session, IP), or - put differently - participants are not allowed to correct their response.

> **4. A similar comment applies to the comprehension checks (see point by reviewer QA). It appears that the plan is not to exclude anyone based on these checks either, but rather to give them as many opportunities as they need to answer correctly in order to move on.**
>
> **Given the complexity of the scenarios I would be concerned that some participants are going to simply cycle through the options without understanding the scenario (and the risk of this might increase with later scenarios). This is a concern especially with the change from an in-person study in the target article to an online replication—if the replication fails, a natural question will be whether it's due to the online setting. In follow-up analyses it would be reasonable to use responses on the comprehension questions to divide the sample into groups. If you foresee going down that route, I'd recommend specifying at this point how you would approach that (e.g., deciding now what would be a reasonable number of incorrect responses where you would classify someone as an inattentive respondent).**

> **If not, please address how you will otherwise guard against low-effort or inattentive responses of the kind commonly seen in online samples.**

Thank you for the feedback, and we understand these concerns.

However, please consider that this is already a big improvement on the original survey, which did not validate anything, and participants could answer whatever they wanted. Participants in the target article's original surveys would not even have to worry about trying the different combinations. Therefore, this is already a far more conservative test of the target's hypotheses.

In our experience, most Prolific participants are very attentive and care about providing high quality work, as can be seen from our many successful replications using that platform, including from this year. In our manuscript, we cited many examples of our team's concluded high-quality replications conducted using MTurk (using CloudResearch) and Prolific participants. Many of those were successful, and many of those had comparable (or higher quality) results compared to in-person samples. We have completed over 120 of these replications and have had an overall very positive experience with these platforms. It is not only our experience, but there is a lot of data about Prolific showing that their participants, many who do tasks and take surveys for a living, take their work and answer surveys very seriously, arguably more so than participant-pool undergraduates and convenience samples. The comprehension checks are meant to help point their attention to important details that can be easily missed within the scenarios.

Many low-effort inattentive participants should be filtered in the first survey outline screen. These checks communicate a serious survey. Consider a participant who sees these attention and/or comprehension checks. You as the recommender reading our manuscript know what the outcome of getting those wrong would be, but the participants do not know that they would be able to try multiple times, they don't know what's on the next page, and they don't know how many of those are ahead. These participants are worried also about their quality reputation (as reflected by rejections over inattentiveness, a strategy which we do not employ). All they know is that there are attention and comprehension checks. Participants aiming to maximize profit and minimize time would therefore at this point either drop out or simply make the effort of answering correctly to minimize the time it would take them to complete the task.

> **Minor points**
>
> **5. The Introduction is generally well-written and clearly organized, but there are a few points that are a bit too terse and need some clarification.**
>
> **5.1 Pg. 8: "Our secondary goal was to build on the target's design and add extensions to refine the target's methods and gain further insights." — Although this information comes at a later point, it would be helpful to give**

> **a little more of a preview here of the extensions and the overall motivation for them.**

Thank you for the feedback. We revised and expanded to the following content:

> "Our secondary goal was to build on the target's design and add extensions to refine the target's methods and gain further insights. We added three extensions examining: 1) whether people indicate waste as a factor impacting their decisions in these situations, 2) a continuous measure of willingness to engage in behaviors to supplement the target's dichotomous choice measure, 3) the degree to which participants perceive the different options as wasteful, serving as the missing manipulation check ."

> **5.2 Pg. 15: "Given that the coding procedure was unclear and the process noisy…" — In what way? The preceding paragraph doesn't seem to explain these limitations, apart from reasons not being measured in Study 2.**

We appreciate the feedback. We revised to try and better explain what we meant:

> "Given that the coding procedure was not provided and the process involved a qualitative process with subjective ratings that may result in very different insights that would be challenging to compare to the original, we decided to instead build on the target's design and categorization, and switch from an open qualitative design to a fixed quantitative design. This also allowed us to implement this extension in all scenarios."

> **5.3 Pg. 17: "we were concerned about a possible discrepancy between the Arkes's (1996) conceptualization of the concept of wastefulness, and the laypersons' perspective of wastefulness." — What was this concern?**

Response: We mentioned this point mainly because the target article did not include pre-tests or manipulation checks, and so it is not clear whether the manipulations in the scenarios were indeed perceived by the participants as being about waste. Different people may have different definitions of wastefulness, and pre-tests and manipulation checks aim to ensure that what scholars perceive as waste is aligned with what laypersons perceive as waste.

We wrote in the same paragraph:

> "Given the inherent subjectivity of the concept of wastefulness and the potential for diverse interpretations, it is crucial to ensure that the different conditions manipulating wastefulness are indeed working as intended."

Action: We edited the content regarding Perceived wastefulness in the "Extensions" section as follows:

> "**Perceived wastefulness (needed manipulation check)**
>
> In this extension, we aimed to examine the extent to which individuals perceive wastefulness in behaviors presented in the study scenarios. Given the inherent subjectivity of the concept of wastefulness and the potential for diverse interpretations, it is crucial to ensure that the different conditions manipulating wastefulness are indeed working as intended. When reproducing these scenarios, we were concerned about a possible discrepancy between the Arkes's (1996) conceptualization of the concept of wastefulness, and the laypersons' perspective of wastefulness, given that there were no pre-tests reported and no included manipulation checks. As an exploratory direction, we also were interested in the differences in the strength of the wastefulness manipulations across the different scenarios, and the association between manipulation strength (as indicated by the manipulation checks) with the wastefulness avoidance effect."

6. **Clarify early on that the study will use a within-subjects design (see point by reviewer TC). I too found the phrasing about a "unified" design to be ambiguous and to cause some confusion at some points. In addition, I'd strongly recommend that for the planned experiment the authors describe the three tasks as separate "scenarios" (or something similar) rather than "studies" ("studies" is appropriate when describing the target article, but is confusing here given the within-subjects design).**

Thank you for the valuable feedback. We changed the description of our planned experiments to "scenarios.", and clarified that we are running the studies in a combined data collection in a within-subject design.

7. **Please address the question by reviewer TC about the "Likelihood" measure, where the simulated mean appears to fall outside the range of response.**

The reviewer caught an oversight, which did not match with our original coding. We changed it to "*M* = -0.04" in Table 9 under the "Results" section. Please see our reply to the reviewer below.

8. Assuming that the red text is meant to be deleted for the Stage 2 submission, I'd suggest including some or all of the text at the top of pg. 20 about pretesting and incentive pay as part of the main text (e.g., the planned pay rate and survey duration). The two sections also seem to repeat some of the same statements and could be consolidated.

Response: Thank you for the suggestion. The red text indeed serves a placeholder and guide for our Stage 2. We also have placeholders for more information in the supplementary, and all will be updated with all the details in Stage 2.

Action: We consolidated the placeholders at the beginning of the "Participants" section.

9. Comprehension checks are described in two places (first on pg. 26 then on pg. 28), but they appear to be describing the same questions. I recommend consolidating these sections so they are described in one place.

Thank you for the suggestion. We made some adjustments and expanded on our descriptions of the comprehension checks.

The two places serve different purposes depending on the section. We first mention comprehension checks in the Procedure section to give a sense of the overall survey flow. We then explained comprehension checks in detail in the Comprehension Check subsection.

In the Process section:

> We also added two multiple-choice comprehension checks presented after the scenario description which participants had to answer correctly before proceeding to the dependent measures. If answered incorrectly, participants are asked to re-examine their responses with as many attempts as needed until they answer correctly. This procedure was designed to signal the importance of carefully reading and comprehending the scenario, and to ensure that the participants read, processed, and understood the key piece of information in the scenarios. We note that this is a deviation from the target's procedure and was meant to ensure that participants understand the crucial scenario information and know what they are rating.

And then a separate, now expanded, subsection:

> **Comprehension checks**
>
> We added two multiple-choice questions for each scenario as comprehension checks to ensure participants understood the scenario content. One question was about the general scenario context, and the second was about the manipulation. Participants had to answer

these questions correctly before proceeding to the next stage to answer the dependent measures.

In Scenario 1, we asked - “Which is true for Mr. Munn?” and “Which is true for Mr. Fry?”, with options: 1) “Goes to the movies on Mondays, was offered a three-movie pack for $24, and has already watched two movies ($10 each, $20 overall)”; 2) “Goes to the movies on Fridays, was not offered a three-movie pack, and has already watched two movies ($10 each, $20 overall).”; 3) “Goes to the movies on Mondays, was not offered a three-movie pack for $24, and has already watched two movies ($10 each, $20 overall).”; and 4) “Goes to the movies on Fridays, was offered a three-movie pack, and has already watched two movies ($10 each, $20 overall). “.

In Scenario 2, we asked - “How much did you originally pay for the tax program you are no longer able to use?” ($0, $50, $100, $160), and “What happens to the old tax program you bought?” (“You can get a full refund for it”; “You can trade it in for a discount on the new package”; “It becomes completely useless”; “You can still use it for federal tax this year.”).

In Scenario 3, we asked - “What makes the tent developed by the other firm more competitive?” (“It's more waterproof.”; “It's easier to carry.”; “It's cheaper and lighter.”; “It's more customizable.”), and “What would happen if you abandon the project?” (“It would become useless and have no value.”; “You could sell it in smaller pieces for various applications.”; “You could sell it to the roofer for his tarp project.”; “You could sell it all as scrap for $5,000.”)

10. See the suggestions for improving the figures made by reviewer TC. NOTE: Although I agree with both points, given that the figures will be updated in the Stage 2 submission along with the rest of the results I don’t view these as necessary changes prior to IPA.

Indeed, the plots will look very different in Stage 2.

In Figure 10, the plot we generated used simulated data. Later, we will feed real Qualtrics data, which will produce a different pattern.

In Figure 11, we changed the labels in the manuscript to “Waste (no rebate) v.s. No waste (rebate)” to maintain consistency.

Reply to Reviewer #1: Dr./Prof. Travis Carter

> **I think the authors of this proposed replication and extension have are well prepared to produce a solid contribution. The proposed plan is a faithful replication of the original article, with well-articulated and well-thought out deviations from the original protocol to fit with the present (e.g., adjusting for inflation). Their proposed extensions are also well considered, intended to ameliorate clear deficits in the original articles method or reporting (e.g., manipulation check; continuous measures to complement the forced choice measures; more robust quantitative approaches to a measure that was originally purely qualitative). The proposed sample size is also a very nice improvement upon the original; the original article's samples were clearly insufficient to be very informative, even if they were normal at the time.**
>
> **I noticed a few small issues that I would suggest the authors address, but overall it appears to be a very solid plan to replicate and extend an important article that has so far not been revisited. Here are my suggestions:**

Thank you for the positive and supportive opening note and the constructive feedback.

> **.1. Recommend that you state very clearly much earlier in the article that you are having participants complete all three studies (the term "unified data collection" is a bit ambiguous---it could be taken to mean that they are randomly assigned to one of the three studies, rather than all three in a random order).**

Thanks for your suggestion. We edited the first paragraph in the “Design: Replication and Extension” section as follows:

> “**Design: Replication and Extension**
>
> We summarized the experimental designs in Tables 4, 5, and 6. Studies 1, 2, and 3 in the target article were conducted separately with independent samples. We ran the three scenarios together in a single unified data collection - Participants completed all three scenarios in random order. The display of scenarios and conditions was counterbalanced using the randomizer “evenly present” function in Qualtrics.”

> **.2. Relatedly, are there concerns about fatigue or bias being introduced by having them complete all three studies? The "unified" design is certainly efficient, and obviously you are doing the right thing by having the order counterbalanced, but you'll need to build in checks to see if the order matters (and *not* only if you fail to find support for the hypotheses, as stated in the note on pg 22), and if it does, how to handle that situation. Analyzing just the first scenario each person saw is one such solution, but that would reduce your power considerably.**

We understand and have acknowledged this concern. Please see our detailed reply to Editor on #2.

> **.3. Mean of "Likelihood" in Study 1 is simulated to be 1.97 (Table 9); is that meant to be the average of the three response options (coded as -1, 0 and +1), or did I misunderstand and that's a separate question? And if I did misunderstand, it's not clear which question that would be.**

Thank you so much for spotting that, much appreciated!

Indeed, it is coded as -1, 0, and 1. The 1.97 value was a typo that misrepresented our original coding results in our Rmarkdown. Given that it was simulated data for Stage 1 demonstration, it does not have any implications, but this does show the importance of simulated data, because if this were indeed the right number, then this would have indicated some issue with our coding.

We amended Table 9 to: "*M* = -0.04".

> **.4. Reasons: I'm not 100% sure I understand the conclusions you're aiming to draw from the repeated-measures ANOVA to analyze the Reasons ratings. That analysis will let you see if any of them are different than each other, but your language in interpreting the (simulated) results suggests you're hoping to do much more. How are you able to make an inference about whether their decisions were "influenced by considerations other than utility maximization" (p. 45)? If you are hoping to compare the other reasons to utility maximization, it seems like you'd need to ask about it explicitly. Plus, that particular analysis doesn't really lend itself to interpreting the absolute magnitude of those reasons. It's possible that participants could rate *all* of the listed reasons as being highly important to their decision, which would nonetheless show up as a non-significant F-test. You may wish to consider interpreting the reasons on an absolute scale (high vs. low) as well as relative to each other.**

Thank you for raising this issue. We agree that this does need better clarifying in the manuscript. This was not explained well in the target article, and we should aim to do better. We also realized that we should be more careful in framing the hypotheses and tests.

As a reminder, for the reasons extension, we included four reasons:

- Option chosen minimizes waste

- Option chosen minimizes negative emotions (regret, anger, sadness, shame, etc.)

- Option chosen maximizes value per money spent (benefit, enjoyment, convenience, etc.)

- Option chosen is more consistent with previous behavior and decisions.

Among these, the third reason emphasizes utility maximization, reflecting the pursuit of optimal outcomes based on benefit and satisfaction. This is what would be considered the neo-classical economics response, given that agents are perceived as utility maximizers. Using that approach, all other options are considered suboptimal, given that waste, emotions, or past behavior, should not take precedence over utility maximization. Among those reasons, the first reason emphasizes waste, which is the core idea in the target article.

Therefore, an ANOVA analysis allows us to compare all reasons to see whether one reason is emphasized over another. The neo-classical prediction would be for utility maximization to prevail. Therefore, we are looking at the comparison of utility maximization to all other reasons, especially the waste reason, given that this is the core and manipulated factor in the target article.

Moreover, in Scenarios 2 and 3 we have two conditions (waste versus no-waste), and then we could examine the interaction of waste conditions and reasons to see whether the manipulation of waste impacts the ratings of the different reasons.

We therefore changed from:

> Based on the brief description in that target article, we hypothesized that individuals' economic decisions would be associated with considerations not solely driven by utility maximization, including minimizing waste, minimizing negative emotions, maximizing value per money spent, and maintaining consistency with previous behavior and decisions. We planned an exploratory analysis comparing the different reasons, yet had no specific predictions as to which reasons would be the strongest in each of the scenarios.

To:

> According to the rational agent model in neo-classical economics the top reason would be to maximize utility. Given that the core argument of the target article is that people have considerations of waste that sometimes conflict with maximization of utility, this means that waste, emotions, or past behavior may be rated as higher priority than utility maximization. We therefore had competing hypotheses: The neo-classical hypothesis is that utility maximization would be the strongest reason, the target article's hypothesis given the emphasis on waste would be that ratings of waste reason would be higher than that of utility maximization, and two additional possibilities are hypotheses countering the neo-classical agent model that past behavior or emotions would be higher than utility maximization. We planned an exploratory analysis comparing the different reasons, and expected that (if waste indeed has an impact on decisions) people would rate waste as higher than utility.

We expanded on the competing hypotheses in Table 1:

> Exploratory competing hypotheses:
> Rational: People rate utility as the most important reason.
> Non rational: People rate utility similarly or lower than other reasons.
> Waste: People rate waste as higher than utility.
> Waste-top: People rate waste as the most important reason.
>
> Interaction: We expected bigger emphasis on waste in the waste condition.

We also revised the study design table, and improved on our plots in the "Results" section.

Ratings of all the reasons as similarly high is an acceptable result that would be considered no support for preference for either reasons of utility or waste.

.5. Wastefulness extension:

.5.1. inconsistency between descriptive statistics and analyses for Study 2 (only two means listed, but doing a mixed-model ANOVA).

.5.2. For studies 2 (p. 55) and 3 (p. 56), paragraph describes paired-samples t-tests instead of the analyses listed in Table 16 (p. 53).

Thanks for spotting these. We added descriptive data in Table 15 for the Scenario 2 tax program, and we adjusted and carefully rechecked all analyses.

.5.3 - Figure 10: This plot needs a bit more explanation. Perhaps this is a new type of plot that I'm unfamiliar with, but all of the dashed lines seem uninformative. They should at least be explained in the note.

Yes, we understand, right now it does not look very informative. These are lines that show per participant the change between the first measure and the second. It looks messy because the plot was generated based on simulated random data, so it goes in all directions. Once we obtain the real data, it should look very differently, hopefully with a clearer pattern. If it still looks very messy, we will remove the connecting dotted lines.

We also added the following to the figure notes:

> The dotted lines indicate the link between the two willingness responses for each participant.

.5.4. - Figure 11: Be consistent in labeling (include "rebate" vs. "no rebate" in addition to "waste" vs. "no waste")

Thank you for spotting it. We changed the labels in the manuscript to “Waste (no rebate) v.s. No waste (rebate).” This is now consistent with the Figure 11 labels.

Reply to Reviewer #2: Dr./Prof. Quentin Andre

> **My overall impression of the manuscript is very positive:**
>
> **• I agree with the authors' argument that Arkes (1996) has been an influential building block, and that the paper is an interesting target for replication.**
>
> **• The authors' familiarity with replications in general, and with the Registered Report format in particular, is evident: The hypothesis are very clearly laid out, the authors present simulated results (with appropriate conditional logic describing how they would describe significant vs. non-significant results), any differences between the original and the replication are very clearly laid out, and the extension that the authors are planning appears meaningful (if modest) in scope.**
>
> **• The sample size justification appears meaningful (using the small telescope approach offered by Simonsohn), and the authors' analytical strategy appears properly set-up for meaningful inferences, regardless of how the data turns out.**

Thank you for the positive and supportive opening note and the constructive feedback.

> **I only have minor comments and suggestions:**
>
> **.1. Unless I missed it, the authors do not discuss how they are planning to handle comprehension checks. Given the potential for misunderstanding the scenario (I must admit I found the movie scenario from Study 1 a bit hard to track when I first read it, something that the authors cannot be blamed for), it appears like an important aspect to discuss. In particular, in the between-subjects designs (Studies 2 and 3), it would be valuable to discuss how differential attrition across conditions will be handled (if, for instance, participants are less likely to pass the comprehension checks in one condition vs. the other, and they are excluded on this basis).**

Thanks to your comment we realized a potential misunderstanding and therefore the need to better explain our use of comprehension checks. In response to the feedback by the editor and your comment we expanded our original writing to the following:

> We also added two multiple-choice comprehension checks presented after the scenario description which participants had to answer correctly before proceeding to the

> dependent measures. If answered incorrectly, participants are asked to re-examine their responses with as many attempts as needed until they answer correctly. This procedure was designed to signal the importance of carefully reading and comprehending the scenario, and to ensure that the participants read, processed, and understood the key piece of information in the scenarios. We note that this is a deviation from the target's procedure and was meant to ensure that participants understand the crucial scenario information and know what they are rating.

We invite you and the editor/reviewer to go through our Qualtrics preview to better see what those comprehension checks are like. Participants cannot fail those comprehension checks, they must answer those correctly to be able to proceed, and there is therefore no attrition. Above, in our reply to the editor we explain why we feel this is a needed design and a needed improvement to the original.

> **.2. The authors have opted for a (mostly) exact replication of the findings of Arkes, only adjusting prices for inflation. I am well-aware of the "blame-if-you-do-blame-if-you-don't" aspect of conducting replications: Conceptuals replications are often dismissed on the grounds that the materials were insufficiently close to the original, while exact replications are often dismissed on the grounds that times have changed and that the materials need to be updated.**
> **Given this tension, I am wondering if there would be value in considering a scenario which would be a conceptual replication, either in addition (resource permitting) or in replacement of one of the original scenario.**

Thank you for the suggestion. We prefer to focus our investigation on direct replications, and given that we share all of materials, data, and code, hope that this reignite interest in these findings and will motivate and inspire further future direct and conceptual replications, with further adjustments, different contexts, and testing of potential moderating factors.

We added a planned discussion for Stage 2 in the "Discussion" section.

> [Following on Dr./Prof. Quentin Andre's comment, we will discuss the importance of conducting further direct and conceptual replications, with further adjustments, different contexts, and testing of potential moderating factors.]

> **.3. This is a matter of taste, but I do not find that Bayesian analysis based on Bayes factors are easy to interpret, given that they require a prior. If I may suggest a recommendation to interpret a null hypothesis, a Likelihood ratio test comparing the likelihood of the data under H0 to the likelihood of the data under a given H1 (which could either be the effect size of the original, or 33% of the original effect size following the "small telescopes approache". This statistics is directly interpretable as "How many times more likely is the data under H0 vs. H1".**

Thank you for the suggestion, we are happy to accommodate further reporting that would aid readers in making sense of our findings. We generally tried to focus on effect sizes and confidence intervals, and considered Bayesian analysis as aiding in the case of trying to quantify support for the null, and likelihood can be another tool for that.

We added Likelihood ratio tests to our Rmarkdown for Studies 1, 2, and 3 using Package 'likelihoodR' (Cahusac, 2023), and revised to the following:

> We pre-registered that in case we failed to find support for the hypothesis for any of the scenarios, we would run a complementary Bayesian analysis for that scenario (without outlier exclusions) using a prior of 0.707 and report likelihood ratio tests to quantify support for the null.

> **My first point is something that I would like to see addressed, while the second and third point are suggestions/matters of taste that the authors should feel free to ignore.**
>
> **Thank you for an enjoyable and very detailed read, and I wish you a smooth data collection process!**

Thank you, we appreciate the feedback and support.