Judgment and Decision Making (2025), **vol. 20**:e34 1–17
doi:10.1017/jdm.2025.10011

COMMENTARY

On the process and value of direct close replications: A rejoinder to Shafir and Cheek's (2024) commentary on Chandrashekar et al. (2021)

Subramanya Prasad Chandrashekar [1] and Gilad Feldman [2]

[1]Department of Psychology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway and [2]Department of Psychology, University of Hong Kong, Hong Kong SAR

Corresponding author: Gilad Feldman; Email: gfeldman@hku.hk

Received: 15 September 2024; **Revised:** 1 July 2025; **Accepted:** 4 August 2025

Keywords: choosing; rejecting; replication; commentary; open science

Abstract

The studies in Shafir (1993, *Memory & Cognition* 21, 546–556) examined the impact of decision frames (choosing vs. rejecting) on decision-making. Our replication—Chandrashekar et al. (2021, *Judgment and Decision Making* 16, 36–56)—revealed mixed results with only partial support for the original findings, concluding a successful replication of only 2 out of 8 scenarios. Our data from an exploratory extension suggested a pattern in support of an alternative theoretical mechanism aligning with Wedell's (1997, *Memory & Cognition* 25, 873–887) accentuation hypothesis. Shafir and Cheek's (2024) commentary criticized our approach to replications, and the value and importance of direct close replications overall, and shared their views regarding the theory and scope of the phenomenon, with new information about what they consider to be needed steps to empirically test the phenomenon. In our response, we clarify misunderstandings and address empirical findings shared in the commentary. We discuss and defend the value and importance of direct replications and the necessity for full transparency regarding the theoretical assumptions and the process of empirical investigations. Finally, we call for the implementation of open science more broadly, in conducting more direct close replications, sharing of all protocols, materials, data, and code, and implementing outcome-blind reviewing and Registered Reports. These would allow for stronger theoretical and empirical foundations, and a more credible and robust psychological science.

1. Introduction

The Shafir and Cheek (2024) commentary discussed our Chandrashekar et al. (2021) article, which reported an attempted replication of the 8 studies reported by Shafir (1993). The studies in Shafir (1993) tested 2 decision frames—choosing versus rejecting—in a variety of contexts, and the findings challenged conventional assumptions by revealing that the 2 decision frames are not always complementary. In our replication report, we concluded a mostly unsuccessful replication, with mixed support for the findings reported in Shafir (1993), successfully replicating only 2 of the 8 scenarios. Using an extension, based on a suggestion in Shafir's commentary replying to a failed replication of one scenario from Shafir (1993) by Many Labs 2 (Klein et al., 2018), we provided an initial exploratory test of an

© The Author(s), 2025. Published by Cambridge University Press on behalf of Society for Judgment and Decision Making and European Association for Decision Making. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/jdm.2025.10011 Published online by Cambridge University Press

alternative theoretical account using the same scenarios, and concluded that our data were suggestive of supporting an accentuation theoretical account (Wedell, 1997) rather than the compatibility theoretical account suggested in Shafir (1993). The Shafir and Cheek (2024) commentary critiqued the definition and the value of direct replications, questioned our classification of the replication as a direct, very close replication, and argued that our process and evidence were an inadequate test of the phenomenon.

Going beyond the discussion about replications, the Shafir and Cheek (2024) commentary also shared a conceptual replication of 1 of the 8 scenarios Shafir (1993), and we tested a scenario that neither we nor Many Labs 2 (Klein et al., 2018) found support for. The commentary argued that to truly test the phenomenon, one must ensure that certain theory-driven valence properties are present in the tested context, potentially by using a pilot of the features used in the vignettes. Following this argument, they demonstrated empirically that the original scenarios used in Shafir (1993) do not align with their expected valence properties, and that with scenario modifications and item pretesting of 1 of the 8 scenarios, they were able to find support for the phenomenon.

We believe that this is an important dialogue and appreciate the opportunity to engage in a discussion regarding our replication attempt, our findings, and direct replications more broadly. We are encouraged by the commentary elaboration on their views regarding theory and generalizability limitations, the sharing of new suggested pretesting procedures, and the empirical demonstration of how the original stimuli can be updated. None of these are taken for granted or trivial, and they highlight a much-needed discussion about theories and empirical testing in judgment and decision-making that go far beyond our specific replication or the broader discussion on direct replications.

In this rejoinder, we clarify important aspects of our replication and process that we believe are relevant as a backdrop to the criticism raised by the commentary. We then address each of the issues raised and provide an in-depth discussion of the value and importance of direct replications and open science. Our summary of the rebuttal of the key points raised in the commentary is summarized in Table 1.

2. Chandrashekar et al. (2021): Background

In Chandrashekar et al. (2021), we reported a pre-registered replication and extension of a seminal paper by Shafir (1993) that documented several paradigms showing different behaviors when comparing decisions of 'choosing' versus decisions of 'rejecting'. Briefly, the studies in Shafir (1993) reported that a choice set with 'enriched' options, which had more positive and more negative attributes compared with an 'impoverished' option, was more likely to be both chosen and rejected. The suggested explanation for these effects was that of 'compatibility', in that choosing leads to an emphasis on strengths, whereas rejecting leads to an emphasis on weaknesses, resulting in the enriched option receiving a higher ratio of choice and rejection.

The Many Labs 2 mass collaboration (Klein et al., 2018) conducted—among many other replications—an initial partial replication of the work by Shafir (1993) focusing only on Problem 1, concluding a failure to replicate the effect. In a reply to the effort, the Shafir (2018) commentary raised several important limitations of that replication, arguing that the replication (1) was not a comprehensive replication of all scenarios, (2) lacked counterbalancing in the presentation of the alternatives, and (3) may have been unsuccessful due to the possibility that the meaning of some of these scenarios has changed over time. The commentary went further to mention Wedell (1997) as an alternative account with an 'accentuation' hypothesis, stating that: 'The experimental findings suggest that there are at least two distinct influences on simple binary choices of the kind considered here. Compatibility, the effect I attributed my findings to raises the weights of attributes compatible with the task at hand; accentuation, as proposed by Wedell (1997), leads to greater weighting of attribute differences in choice than in rejection' (p. 496).

Our replication addressed this exchange and was a step forward in helping the community make progress in understanding the phenomenon. We addressed the first 2 concerns raised by Shafir, as

https://doi.org/10.1017/jdm.2025.10011 Published online by Cambridge University Press

Table 1. *Summary and rebuttal of the main arguments made in Shafir and Cheek (2024).*

Issues noted in the commentary	Clarifications and notes	Recommendations
Interpretation of replication findings		
'We address Chandrashekar et al.'s (2021) conclusions regarding the *failure to replicate the original problems.*' (p. 2)	In Chandrashekar et al. (2021), we concluded that: 'We found no consistent support for the hypotheses across the eight problems: two had similar effects, two had opposite effects, and four showed no effects.' (abstract)	In Chandrashekar et al. (2021), we recommended that: 'Researchers conducting research in this domain and future research on this phenomenon can build on insights gained here to advance theory by defining the boundary conditions under which it operates and explore further ways on how it should be tested.' (p. 53)
'It is problematic, in our view, to conclude that a decades-old finding "does not replicate". . ..' (p. 2)	Therefore, in Chandrashekar et al. (2021), we did not conclude that the phenomenon 'does not replicate'. We reported support for 2 of 8 scenarios and found support for an alternative decades-old finding supporting a different theoretical account—Wedell's (1997) accentuation hypothesis.	
'All that replicators should be free to conclude is that the original materials, years later, or in different contexts, do not produce the same pattern of results.' (p. 8)	In Chandrashekar et al. (2021), we wrote exactly that: 'We successfully replicated the results of Problem 4 and Problem 5 of the original study. However, in Problems 7 and 8, we found effects in the direction opposite to the original findings, and our findings for Problems 1, 2, 3, and 6 indicated support for the null hypothesis.' (p. 51) And. . . 'Our replication does not rule out the compatibility account, only indicates that it is in need of further elaboration and specification, and further testing, and we see much promise in examining the interaction of the two accounts.' (p. 53)	In Chandrashekar et al. (2021), we recommended that: 'Other compatibility hypothesis stimuli be revisited with direct close replications or that new stimuli be developed before further expanding on the compatibility hypothesis.' (p. 53)
Replication closeness classification as 'very close'		
'Very close replications do not seem, from a theoretical perspective, very close after all.' (p. 8)	Chandrashekar et al.'s (2021) replication classification used LeBel et al.'s (2018) taxonomy, the most widely cited and used taxonomy for replications. The reference to a 'theoretical perspective' classification was	When evaluating replications, clearly define or refer to a clear taxonomy for classifying replication closeness. When referring to 'theoretical perspective' or 'theoretical closeness', define what that

(Continued)

https://doi.org/10.1017/jdm.2025.10011 Published online by Cambridge University Press

Table 1. *(Continued).*

Issues noted in the commentary	Clarifications and notes	Recommendations
	not sufficiently specified, leaving it unclear what theoretical closeness might look like, how that relates to the empirical demonstration included, and how that addresses the debate regarding the replication of Shafir (1993) and direct replications more broadly. Shafir and Cheek (2024) did not refer to or suggest an alternative taxonomy that defines or addresses very close 'theoretical perspective' replications. All piloting procedures suggested by the commentary are new, and were not previously shared nor specified in any of the decades-long follow-up literature.	means, what theory this refers to, and establish clear falsifiable theoretical criteria against which this is to be evaluated. Then, evaluate all replications conducted to show how they fare on the suggested criteria. Whenever possible, explain how the new taxonomy links with other taxonomies—complementary or supplementary, and which might be more relevant under what replication context or goals.
Value of and learning from direct replications		
'Very close replications' across contexts or time, when stimuli may have changed their meaning or valence, can be highly misleading, further exacerbating replication concerns.' (abstract) 'We also ran the original later in 2021 and failed to obtain the 1993 results, as reported in Cheek and Shafir (2024).' (p. 8)	The Shafir and Cheek (2024) and Shafir (2018) commentaries shared that they conducted multiple direct replications and have learned from those to inform their follow-up studies, and the possibility of needing to update the procedures using piloting. The Shafir and Cheek (2024) commentary referenced the direct replication in a citation to Cheek and Shafir (2018) and Cheek and Shafir (2024), yet with no further publicly available information. We are unable to assess or comment on unpublished findings. The community would benefit from sharing all failed and successful direct and conceptual replications, so that we can all learn from their process and findings.	Whenever possible and still relevant, first conduct direct replications with original items that are as close to the original items as possible to test whether original findings hold. If findings using original items do not hold, then use findings using original items as a baseline to determine how changes in items impact findings compared to the baseline. When adding to or adjusting procedures in the original article, explain in detail why those adjustments are needed, and outline in detail predictions regarding the impact of those adjustments and how those tie back to theory. Make all unpublished manuscripts that are referenced as evidence public. Make all replications public, successful and unsuccessful, conceptual and direct. Make materials, data, and code public for all data collection, including pretests/pilots.

(Continued)

https://doi.org/10.1017/jdm.2025.10011 Published online by Cambridge University Press

Table 1. *(Continued).*

Issues noted in the commentary	Clarifications and notes	Recommendations
'We are not the first to raise concerns about the practice of simply read-ministering the same experimental materials in a sufficiently different context and then interpreting the different results as a sign of the initial findings' unreliability.' (p. 8)	This view is unfortunately outdated, the citations that were used are old and have been addressed in detail in many follow-up rebuttals, some of which we cited (e.g., Nosek et al., 2022; Zwaan et al., 2018), and both are not aligned with the current best practices in psychological science. Replications allow the community to evaluate if the contextual factors affect an established theoretical assertion. The direct replication efforts reported by Many Labs 2, Chandrashekar et al. (2021), Shafir (2018), Cheek and Shafir (2018), Shafir and Cheek (2024), and Cheek and Shafir (2024) highlight the importance of replications, both direct and conceptual, for the process of understanding a phenomenon. Scholars who aim to build on the foundations of choosing versus rejecting paradigms now have a better understanding of how to think about and pursue such research, with better estimates of what works and what does not work. This reduces the investment of time and resources pursuing wrong methods or outdated items, and offers new insights regarding the possibility of the 2 accounts discussed in the literature.	Make direct replications mainstream, support and incentivize conducting, public sharing, and publishing direct replications. Clarify constraints on generality/generalizability of both their theory (level of universality and degree of precision; e.g., Glöckner, 2017; Glöckner and Betsch, 2011) and the measures, samples, validity, and methods (e.g., Simons et al., 2017). For generalizability claims not covered by the empirical demonstration, call for and promote direct replications in other contexts that would test those claims.
Piloting and pretesting items		
'Pilot testing when the Problem was first run had found that [. . .] were highly positive (compatible with choice), whereas [. . .] were much more negative (compatible with rejection) [. . .].'	That information and these piloting procedures were not fully documented and shared in Shafir (1993) or any of the follow-up literature. The pre-registrations did not detail thresholds for items to be classified as 'neutral', 'positive', and 'negative'.	In new manuscripts, share all details of all procedures relating to the investigation, including pilots and pretests. In commentaries sharing previously undisclosed information about completed published articles, make all information available about all procedures relating to all studies and all problems, including pilots and pretests. In both, explain how the piloting is linked back to theory and predictions, and define and pre-register clear thresholds for pretesting in devising new items.

https://doi.org/10.1017/jdm.2025.10011 Published online by Cambridge University Press

we attempted a replication of all 8 problems reported, and added several improvements, such as the counterbalancing of the display of the options. Furthermore, we added an exploratory extension that allowed testing for the possibility of the accentuation alternative account that Shafir suggested.

3. Chandrashekar et al. (2021): Findings

In our replication, we concluded that 'we found no consistent support for the hypotheses across the eight problems: two had similar effects, two had opposite effects, and four showed no effects' (underline not in original text). Although we consider the aggregated findings as a failure to replicate the findings of the target article as a whole, we tried to be very careful in how we framed our findings, and we note we deliberately framed those as a failure to find *consistent* support, rather than a failure to find any support, and failure to find consistent support specifically for the target article's findings, rather than a failure to find support for findings regarding the phenomenon all together. We also reiterate that our findings were not random noise, nor were they only representing null effects, and our findings were not a complete failure to find any support for the phenomenon. Quite the contrary, we found support for the phenomenon in 2 of the 8 scenarios, and using our added extension, we suggested support for an alternative account to Shafir's, one that Shafir (2018) specifically suggested in his commentary reply to Many Labs 2 (Klein et al., 2018). This means that using the original stimuli from the 1990s, with no additional modifications, we found support for the phenomenon in 2 of the scenarios, and overall concluded support for a pattern of results that was suggested in the literature following Shafir (Wedell, 1997), one which he himself pointed to when writing a commentary to Many Labs 2 (Shafir, 2018).

The commentary noted that we 'conclude[d] that a decades-old finding "does not replicate"' (p. 8) and, in their pushback, argued that 'all that replicators should be free to conclude is that the original materials, years later, or in different contexts, do not produce the same pattern of results, a finding that we find of limited insight by itself' (p. 8).

In fact, what they suggested we should conclude was exactly what we concluded in our replication. We opened our discussion section with the following: 'We successfully replicated the results of Problem 4 and Problem 5 of the original study. However, in Problems 7 and 8, we found effects in the direction opposite to the original findings, and our findings for Problems 1, 2, 3, and 6 indicated support for the null hypothesis'. We further wrote that 'our replication does not rule out the compatibility account, only indicates that it is in need of further elaboration and specification, and further testing, and we see much promise in examining the interaction of the two accounts'. And that 'we recommend that other compatibility hypothesis stimuli be revisited with direct close replications or that new stimuli be developed before further expanding on the compatibility hypothesis'.

4. Chandrashekar et al. (2021): Replication categorization

The commentary criticized our classification of our replication as a 'very close replication', arguing that 'very close replications do not seem, from a theoretical perspective, very close after all' (p. 8), emphasizing what seems like theoretical closeness over empirical closeness.

On that point, we reiterate what we wrote in our replication and clarify that our categorization used LeBel et al.'s (2018) taxonomy. To the best of our knowledge, this is the most common taxonomy of replications currently available, and we have employed it successfully throughout our many published replications. The commentary criticizes several aspects of the 'very close' categorization, especially regarding what they refer to as a 'theoretical perspective'. We feel that this is an important debate to have as a community, and we would welcome a broader debate to improve that taxonomy and further suggestions and examples on how to evaluate and pursue theoretical closeness. We come back to this point when we discuss the value of direct, very close replications.

https://doi.org/10.1017/jdm.2025.10011 Published online by Cambridge University Press

5. Chandrashekar et al. (2021): Process

The commentary criticized the value of direct replications, and so in this section, we aimed to chronologically outline the replications that took place in the process, since the publication of Shafir (1993), and through that to clarify the value that we see in direct replications and their contributions to the understanding of the focal findings. The main stages in this process are summarized in Table 2.

Many Labs 2 conducted the first published replication of the findings by Shafir (1993). A follow-up commentary by Shafir (2018) raised several methodological concerns, sharing that they also conducted replications of the findings by Shafir (1993) and wrote (p. 496):

> In follow-up work, my student Nathan Cheek and I have run several studies intended to further clarify these apparently conflicting patterns. In a working paper (Cheek and Shafir, 2018), we show that the original pattern is easy to replicate and that the pattern predicted by Wedell and observed in Many Labs 2 is also easy to document, and we further consider the cognitive mechanisms that might account for both.

Our Chandrashekar et al. (2021) replication addressed some of the points raised in that Shafir (1998) commentary by conducting an additional direct replication of all 8 scenarios reported in Shafir (1993) with the counterbalancing of options.

Table 2. *Timeline of scientific process and studies conducted following the publication of Shafir (1993).*

Time	Event
1993	Eight studies in Shafir (1993) demonstrated support for choosing versus rejection, and the article argued for a compatibility hypothesis.
2018	Many Labs 2 published the results of a mass replication (Klein et al., 2018), which reported a failed replication of one of the scenarios.
2018	A commentary by Shafir (2018) was published, critiquing Many Labs 2's failed replication of one of the scenarios, calling for replicating all scenarios, counterbalancing alternatives, and considering a possible change of context over time. The commentary noted unpublished failed direct replications: 'Propelled by the Many Labs 2 publication, [we] revisited the original study as well. Like the present authors, we, too, found that the original scenarios replicated poorly. Among other things, when we had them rated individually, several of the features used then now scored very differently, e.g., "lots of work-related travel" was hardly rated negative at all (−.28 on a −5 to 5 scale), etc.' The commentary also mentions unpublished evidence: 'We then piloted revised positive and negative features, drafted new items, and ran them. And we clearly documented the original choose-reject discrepancy. Which we then replicated with a pre-registered nationally representative sample. The choose-reject discrepancy of the original paper clearly replicates with revised items.'
2021	The replication by Chandrashekar et al. (2021) is published, reporting mixed support, finding support for 2 out of 8 scenarios, and initial evidence supporting an accentuation theoretical account (Wedell, 1997) over the compatibility theoretical account suggested in Shafir (1993).
2021	In a commentary to Chandrashekar et al. (2021), Shafir and Cheek (2024) noted additional unpublished failed direct replications: 'We also ran the original later in 2021 and failed to obtain the 1993 results, as reported in Cheek and Shafir (in preparation).'
2022–23	Shafir and Cheek's (2024) commentary reports pre-registering and running additional pretests and studies in 2022.

https://doi.org/10.1017/jdm.2025.10011 Published online by Cambridge University Press

In the commentary by Shafir and Cheek (2024) on our replication, the authors again shared that they conducted additional unpublished replications (p. 8):

> We also ran the original later in 2021 and failed to obtain the 1993 results, as reported in Cheek and Shafir (2024).

The published direct replications by Many Labs 2 and Chandrashekar et al. (2021) are important because they are the only available public reports of a replicability assessment of the original findings by Shafir (1993) as is. Those direct replications are also important because they set in motion the publishing of the commentaries by Shafir (2018) and Shafir and Cheek (2024), which have raised new contestations and understandings that were not fully articulated in Shafir (1993). Direct replications, therefore, help ensure that science can progress in studying a phenomenon by (1) assessing the replicability of the original findings using very similar conditions and stimuli, (2) eliciting public sharing of implicit knowledge and vital information not clearly communicated in the reporting of the original findings, and (3) motivating further public discussions regarding how to best further study the phenomenon. Without those, the public would only see the original studies by Shafir (1993), making it unclear whether the empirical findings reported in that article contain all that is needed for others to build on those findings, are replicable as is, and are up to date.

The commentaries by Shafir (2018) and Shafir and Cheek (2024) also briefly mentioned details and empirical evidence that would have helped progress our understanding of the phenomenon and the needed methods even further: (1) the process and pretests of the original findings from the target article, (2) the findings of the first reported failed direct replication and the procedures and materials of the adjusted scenario(s) replication from their response to the Many Labs 2 replication (Cheek and Shafir, 2018), and (3) the failed direct replications in response to our replication (Cheek and Shafir, 2024). It is unfortunate that neither of the commentaries further elaborated on their own failed direct replications nor on any of the details regarding the successful adjusted and conceptual replications. These would have been very helpful for the community and for us in our replication, so that we can all examine and learn from what the commentaries' authors did to make needed adjustments. We consider this a missed opportunity. Without access to further details about these studies, we as a community cannot verify, evaluate, or learn from the experience gained, and thus they remain implicit knowledge by the commentaries' authors. We kindly ask the commentaries' authors to follow up and share all the empirical evidence they have conducted regarding this phenomenon. This can be easily done through uploading all materials, data, and code to a publicly open and citable project on the Open Science Framework (osf.io) or Research Box (researchbox.org), and does not require a manuscript and going through a lengthy peer review.

To gain a better understanding of the phenomenon, it is crucial to publicly share all replications, direct and conceptual, successful and failed, by original authors and others in the community, even as preprints or public projects. The community needs to be able to evaluate claimed scientific evidence. For trust in science to be established and warranted, it is essential that scientific reporting is transparent, thereby enhancing the ability to assess the quality of research with greater certainty (Vazire, 2017).

6. Challenges in interpreting direct replication results: Hindsight, outcome, and confirmation biases

The commentary by Shafir and Cheek (2024) argued that replications 'can be highly misleading' and that our Chandrashekar et al. (2021) replication of Shafir (1993) is further 'exacerbating replication concerns'. Such arguments are often raised when the findings of the replication are not consistent with the findings reported by the original studies, resulting in criticism of direct replications as inherently flawed for not addressing any of the countless possible factors raised post hoc and therefore being of no value. In this section, we discuss the issues with post hoc outcome-based interpretations of direct replications and caution against possible hindsight, outcome, and confirmation biases in critiquing results of direct replications after the fact.

https://doi.org/10.1017/jdm.2025.10011 Published online by Cambridge University Press

One of the challenges with results-based arguments for problems with direct replications is that these arguments are commonly limited to failures, and yet some direct replications succeed in replicating original findings. Our Chandrashekar et al. (2021) was part of a mass collaboration effort in which we concluded over 120 replications of classic effects in the judgment and decision-making and social psychology literature (CORE, 2025). As part of this effort, in addition to the replication by Chandrashekar et al. (2021), we completed replications of many other classic articles from the same time as the studies reported by Shafir (1993). For example, in Ziano et al. (2021), we concluded a successful replication of a classic article by Shafir et al. (1997) regarding a phenomenon coined as the 'money illusion', the tendency to think about money in nominal rather than real terms. The stimuli from the 1990s were repeated as is in the replication, and included references to amounts of money that changed value over time due to inflation, economic terms that shifted over the 2 decades, and factors such as affect (e.g., happiness) and attitudes (e.g., job satisfaction) that may vary between samples and change over time. Despite the change in time, context, and samples, the results of the replication were very similar to those of the original findings, and out of the 4 problems, we concluded that 3 of the replication effects were consistent with the original findings (e.g., Problem 1; original: Cramer's $V = 0.26$, 95% CI [0.17, 0.37]; replication: $V = 0.28$ [0.21, 0.36]), and one of the replication effects was stronger than the effect reported in the original studies.

The commentary argued that 'what is clear is that some findings are going to be time-sensitive in ways that optical illusions are not', yet we argue that it is not at all clear what is time-sensitive, and many of the phenomena that we attempted to replicate with the CORE (2025) team showed remarkable consistency over time. To establish what is and what is not time-specific, we need to regularly run replications and specifically direct replications.

Furthermore, some phenomena may replicate well across contexts and cultures, and it is not clear if the context matters until we test it in that context. For example, a Brazilian team has followed up some of our CORE (2025) team replications, reusing our materials, translating those to Portuguese, and running those with Brazilian samples. One of these replications is de Moraes Ferreira et al. (2023), which repeated our Ziano et al. (2021) replication of Shafir et al. (1997). They, too, concluded remarkably similar results to our successful replication. Before seeing the results, scholars could have argued either way, with both reasons why this would replicate well and reasons why this is doomed to fail. The best way to establish whether the results hold in a different country using a different language is to first repeat it with direct close replications only varying the language and sample.

These successful replications of Shafir et al. (1997) are no coincidence. In the CORE (2025) team, we conducted many replications of classics dating back to the 1970s, and using different samples and contexts, our replication success rate has been relatively high (above 60%). In almost all of those replications, we reran the original stimuli and items with minimal to no adjustments. This goes to show that some of the JDM classics hold across time, context, and cultures, decades later.

We, therefore, need to be very careful with the criticism of direct replications only when results are negative and after the fact. Evaluating scientific outputs differently based on outcomes relates to the challenges of hindsight, outcome, and confirmation biases in the scientific process. For example, the publication of a successful replication is far less likely to result in further replications, conceptual adjustments, and public commentaries. If a replication's findings are not fully consistent with the original findings, then it tends to elicit stronger negative evaluations of the replication, or of the replicators and/or their motivations. This is despite replications often meeting much higher standards than those by the target articles in that they are often pre-registered, much better powered with larger, more diverse samples, fully transparent in all their materials, data, and code, and adding extensions that help better understand the effect or clarify the effect in case of a failure.

Are the studies showing hindsight, outcome, and confirmation biases replicable, hold across time and context, and impact evaluations of replicability? Our evidence from the CORE (2025) team suggests that the answer to all of those questions is 'yes'. Three of the target articles that we successfully replicated from the 1970s and the 1980s using very similar items with little to no adjustments were about these 3 biases. In Aiyer et al. (2023), we successfully replicated outcome bias demonstrated by

https://doi.org/10.1017/jdm.2025.10011 Published online by Cambridge University Press

Baron and Hershey (1988) showing that a decision resulting in a positive outcome was evaluated as better, more important, and more normative than the same decision resulting in a negative outcome, even when participants were explicitly told to ignore the outcome and have themselves indicated to have ignored the outcome. In Chen et al. (2021), we successfully replicated retrospective and prospective hindsight bias from Fischhoff (1975) and Slovic and Fischhoff (1977). In our third study in Chen et al. (2021), we showed hindsight and outcome bias regarding the replicability of hindsight bias. That is, we asked participants to predict the likelihood that a seminal study on hindsight bias (Fischhoff, 1975) would successfully replicate, in 3 conditions. In the 'foresight' condition, there was no indication of the outcome of our replication; in the 'hindsight positive' condition, we indicated that the replication outcome was successful and in support of the original findings; and in the 'hindsight negative' condition, we indicated that the replication outcome was negative and failed to find support for the original findings. In both hindsight conditions, participants were instructed to ignore the outcomes: 'Answer as if you do not know the outcome, estimating the probabilities at that time before the replication study was launched'. We found that when estimating the likelihood of a successful replication of hindsight bias, participants in the hindsight positive condition rated the likelihood as higher than those in the foresight condition, who in turn rated the likelihood as higher than those in the hindsight negative condition.

Seeing the results of a replication triggers outcome bias and hindsight bias—it is unavoidable, even for those who consider themselves unbiased, and even when explicitly told to ignore the outcomes. For those who are invested in a certain outcome, an outcome different from expectations may trigger confirmation bias, in the rejection of conflicting evidence and in seeking other evidence that would confirm held beliefs.

Replications are especially tricky. When replications succeed, a common reaction is along the lines of: 'Of course it would replicate, we knew it would all along, replicators conducted the replication well, but it is of no value, because we had the original findings.' When replications fail, a common reaction is along the lines of: 'of course it would fail, we knew it all along' either questioning the competency of the replicators, the quality of their replication decisions, and/or raising post-hoc reasons for factors that may have led to the failure. This is not to say that there is no merit in the suggested factors, or that failures are not associated with issues regarding replicators and the replication process. The commentary by Shafir and Cheek (2024) argued that their new evidence using a new pretesting procedure not documented in Shafir (1993) demonstrated on only 1 of 8 scenarios from Shafir (1993) shows that the evidence by Chandrashekar et al. (2021) with comprehensive replications of all 8 scenarios using the same terms as in Shafir (1993) is problematic and is of no value—'a finding that we find of limited insight by itself' (p. 8).

One of the key questions in assessing the methods of a replication is whether an expert examining the methods of the original studies against those of the replication, without seeing the replication outcomes, could have predicted the replication outcome and would have identified any critical flaws. Given that we followed the article by Shafir (1993) closely, that the article did not specify any generalizability criteria, context limitations, or pretests referenced in the commentary by Shafir and Cheek (2024), and that there were no publicly shared additional updates to the original findings, we thought one would expect the original findings to replicate successfully. Even if not, the community should welcome additional evidence, and indeed a second commentary on our replication by Ganzach (2025) concluded that 'I view Chandrashekar et al.'s (2021) replication as a worthwhile contribution to our understanding of this effect' (p. 3). It is unclear how replicators and researchers building on the work of Shafir (1993) were expected to guess procedures needed for the effect to emerge, and it is only thanks to the publication of the findings in the direct replications and the follow-up commentaries that we can now publicly discuss more appropriate methods that address missing information, unclear theory, or implicit assumptions in the original studies. It is unclear how the community or public are expected to know that the original findings do not replicate as is if there are no public updates to the original findings.

Are there any ways to address hindsight, outcome, and confirmation biases? Yes—Registered Reports and masked-outcome peer review are 2 simple procedures that can at least partly address these

https://doi.org/10.1017/jdm.2025.10011 Published online by Cambridge University Press

biases. In the last 3 years, we have increasingly shifted our replications to Registered Reports, mostly conducting our Registered Reports through the revolutionary *Peer Community in Registered Reports* initiative (Feldman, 2024; Lloyd and Chambers, 2024), which has an open peer-review process. In our experience, the Registered Reports process has greatly reduced these biases and post hoc outcome-based criticism. When everyone agrees on the methods before data collection, and publication is guaranteed regardless of outcomes, then there is less room for these biases to impact the peer-review process and influence outcomes.

Ideally, replications would be done as Registered Reports (Chambers and Tzavella, 2022; Nosek and Lakens, 2014; Soderberg et al., 2021), yet when this is not the case, such as when data and results already exist, another possible solution to address potential bias is to mask the editor and reviewers from the outcome, a process implemented by some journals like the *Royal Society of Open Science* ('results-blind track'; Chambers and Tzavella, 2022; Woznyj et al., 2018).

Shafir and Cheek's commentary further reflected on direct close replications using the original items as being 'problematic' and 'misleading'. We wholeheartedly disagree. Direct close replications are highly informative and greatly contribute to scientific progress and to the credibility of the scientific record. Replications have helped the original authors themselves revisit and reassess their own research, to share more from their previously unknown process of evaluating and choosing items, and now with new evidence using new items that demonstrate the phenomenon. Because of the new details we now have about the process, the public debate, and the evidence shared, the academic community and the public are all better off. Scholars that aim to build on the foundations of the choosing versus rejecting paradigms now have a better understanding of how to pursue such research, what works, and what does not, so that they do not invest more time and resources pursuing something using the wrong method or outdated items, and with new insights regarding the possibility of the 2 accounts discussed in the literature. This would not have come to light if it were not for the Many Labs 2 and our replications. Direct replications should be encouraged (Nosek et al., 2022) and should become mainstream and common (Zwaan et al., 2018).

7. Replicators working with original authors and adversarial collaborations

The Shafir and Cheek (2024) commentary argued that it would be preferable for replicators to work in collaboration with original authors:

> At a minimum, stimuli need to be revisited, in many circumstances preferably in collaboration with the original authors, to ensure that they afford a valid test of the original hypothesis (see also Fiedler et al., 2021 on manipulation checks). (p. 8)

We disagree. First, there is no evidence that involving original authors helps improve replications or the probability of a successful replication. If anything, the evidence we have so far from mass collaborations shows no benefits of involving original authors. There are many empirical demonstrations, of which we will mention those we consider to be the most impactful. Many Labs 5 (Ebersole et al., 2020) worked together with original authors to revisit 10 of the Reproducibility Project: Psychology (RP:P; Open Science Collaboration, 2015) over which original authors had expressed concerns regarding the replication designs before data collection. After working with the original authors, updating the protocols, and rerunning the studies, they concluded that 'we found that the revised protocols produced effect sizes similar to those of the RP:P protocols ($\Delta r = .002$ or .014, depending on analytic approach)'. In McCarthy et al. (2021), many labs conducted replications of Srull and Wyer's (1979) hostile priming effect, for the second time (the first failed mass-lab replication attempt was McCarthy et al., 2018), this time conducting both direct replications and conceptual replications involving the original authors. They concluded that 'the hostile priming effect for both the close replication ($d = 0.09$, 95% CI [−0.04, 0.22], $z = 1.34$, $p = .16$) and the conceptual replication with the involvement of the original authors ($d = 0.05$, 95% CI [−0.04, 0.15], $z = 1.15$, $p = .58$) were not significantly different from zero' and that

https://doi.org/10.1017/jdm.2025.10011 Published online by Cambridge University Press

'we suggest that researchers should not invest more resources into trying to detect a hostile priming effect using methods like those described in Srull and Wyer (1979)'. One of the original authors was involved in the entire process. Lastly, in one of social psychology's most remarkable replication stories, following a mass replication effort by ego-depletion scholars that failed to find support for ego depletion (Hagger et al., 2016), original authors pushed back on both the choice of study and methods of the mass replications (Baumeister and Vohs, 2016), and proceeded to conduct their own mass replication attempt on a study of their choice only to conclude similar very weak to no effects (Vohs et al., 2021).

Furthermore, the involvement of original authors in replications is sometimes (1) impossible, such as when authors have passed away, (2) unreasonable, such as when authors have left academia or retired, or (3) potentially extremely time-consuming or demanding, especially with highly impactful findings that elicit a lot of academic and public interest. With replications of old classics from decades ago, it is also sometimes the case that original authors do not have access to or remember much information regarding their published article or what took place, if it was not well-documented or kept.

Across our many replications, our experience in working with original authors has been very mixed. Some authors are a pleasure to engage and work with—they share everything they have, and provide support with constructive and objective feedback and suggestions. With others, there were issues that ranged from lack of responsiveness, hostility and/or long delays in replies, suspicion over motivations, deference to their authority and status, expectations of blind adherence to instructions given perceived expertise, and unreasonable to impossible requests and/or requirements.

For science, the best path forward is to ensure that anyone can conduct replications without the original authors. Original authors should publicly share as much as possible about their work to make replications possible and easy. As a community, we should encourage and potentially incentivize conducting and publishing external independent direct replications of published work. In Feldman (2022), we called on researchers to adopt a 'Check me, replicate me' pledge inviting others to check and replicate their work, with incentives for finding mistakes. For anyone to spend their scarce time and resources to revisit and follow up on others' work should be considered a valuable and important scientific pursuit. Original authors should consider replications as an opportunity for self-reflection—whether in improving the clarity, specificity, and falsifiability of their theories; enhancing the transparency of their methods and stimuli; or revisiting and possibly updating their own studies that would demonstrate stronger support for theories and evidence.

8. Pretesting items: Discussing Shafir and Cheek's empirical demonstration

The commentary by Shafir and Cheek (2024) included an empirical demonstration in which they reported a process of pretesting items in pilots that eventually led to what was concluded as a successful conceptual replication of 1 out of 8 scenarios from Shafir (1993). We are glad that our replication resulted in further empirical work that has the potential to improve theoretical clarity, methods, measurement, and generalizability. These procedures were not previously shared or explained in Shafir (1993) and—to the best of our knowledge—were not documented in any of the follow-up literature on this phenomenon; therefore, these are new procedures that serve as a good first step. Yet, to allow the community to learn from and build on that evidence, there are several gaps.

We outline remaining gaps regarding the evidence that was shared: (1) the demonstration was only conducted for 1 out of 8 scenarios, (2) it is unclear how the cutoff thresholds were determined, (3) it is unclear whether the measured valence in the commentary is different from the valence of the items in Shafir (1993), given that Shafir (1993) and the commentary did not provide the valence measurements for the items used in Shafir (1993), and (4) the process of choosing the items and the used cutoff threshold were not pre-registered, and so it is not clear whether these were determine a priori or post hoc, and whether and how these should generalize to other scenarios. Below, we elaborate on each of those gaps in detail.

https://doi.org/10.1017/jdm.2025.10011 Published online by Cambridge University Press

8.1. Empirical evidence: More information required

The Shafir and Cheek (2024) added pretesting of the valence for the attributes of 1 of 8 problems in Shafir (1993), noting that:

> Pilot testing when the Problem was first run had found that Parent A's attributes were perceived as neutral, essentially offering no compelling reason to award or deny custody, whereas Parent B's attributes were compatible with both choice and rejection. (p. 3)

The commentary then reported the findings of 2 pilot surveys that measured the valence of attributes used in the original scenario, as well as additional attributes, concluding that:

> We found that the valence of some of the original attributes had changed. (p. 4)

Changed how and from what? It is unclear what the Shafir and Cheek (2024) evidence is indicative of having changed from, given that Shafir (1993) did not report measuring any of those aspects, and so we have no way of comparing the two. The Shafir (1998) commentary on Many Labs 2 suggested the possibility that a similar procedure may have taken place in the studies reported in Shafir (1993). The Shafir (1998) commentary simply read (p. 496):

> When I devised the binary problems for my 1993 paper, I created enriched options that were preferred in choice, and then added negative attributes so that those options would also be rejected. This was a reasonable strategy to show compatibility, but it happened not to yield enriched options that were unpopular in choice, like those later created by Wedell, where a greater share of choice plus rejection can be obtained by the impoverished option.

Yet we have no access to the mentioned process or data, with the only mention of pretesting and manipulation checks in Shafir (1993) was that (p. 549):

> By way of manipulation checks, for each of the problems that follow, at least 85% of an independent group of subjects ($N = 26$) chose the enriched option as the option that has more reasons for and against it than the competing, impoverished option ($p < .001$ in all cases).

To evaluate the evidence and the claimed change in valence from 1993, it is essential that the data from Shafir (1993) and the procedure behind it are shared, and not just for the 1 scenario (out of 8) that they tried to replicate in the Shafir and Cheek (2024) commentary. Without that data, what the Shafir and Cheek (2024) commentary showed was only that the evaluations at the time they reran these studies in 2022–23 were not aligned with their unspecified valence expectations.

8.2. The need for fully specified, clearly testable, and falsifiable methods

Shafir and Cheek (2024)'s commentary argued that direct replications are of limited value because the original items should not have been used as is, and that for evidence supporting the phenomenon to be found, scholars need to follow a procedure of pretesting the valence of items and selecting specific items that meet unspecified criteria.

We find this argument puzzling, given that both commentaries by Shafir (2018) and Shafir and Cheek (2024) reported to have, at least twice, ran replications using the original items without going through any updating pretesting procedures. If the pretesting is taken for granted and there is no value in conducting direct replications, then why invest resources in conducting those? It is also unclear whether this procedure was part of the original studies in Shafir (1993) and if so—what exactly took place. It is a missed opportunity that both follow-up commentaries did not aim to fill that gap by sharing everything. Therefore, we call on the commentary's authors to further clarify and better specify under what circumstances they consider pretesting a crucial test of the theory and related phenomenon, and to share all that they have that is relevant for that process from their empirical investigations.

https://doi.org/10.1017/jdm.2025.10011 Published online by Cambridge University Press

The commentary seemed to argue that the value of the direct replications is limited because the replications did not use a piloting procedure, yet given that the suggested pretesting procedure is new, we are unsure how anyone was expected to follow it. Even with the provided evidence, given that it is not fully tested and underspecified, we are still unsure how anyone is expected to implement it for other empirical demonstrations. Replicators could not be expected to guess what needs to be done in order to test and demonstrate a phenomenon, or to guess the limitations on the generalizability of the stimuli. Our view is that, in science, everything should be clearly specified and should not involve unspecified procedures and assumptions, or require implicit knowledge. We are therefore grateful that, at least for one of the scenarios, the pretesting of the factors in the scenarios is a bit clearer. In future studies, we recommend making all procedures transparent and clear, sharing all that has taken place to result in the reported findings, clearly stating the procedures needed to conduct the investigation again in a different context and time, and setting a priori falsifiable criteria for the theory and hypotheses.

8.3. Constraints on generality/generalizability and implications for practical use

Articles should also make clear the constraints on generality/generalizability of both their theory (level of universality and degree of precision; e.g., Glöckner, 2017; Glöckner and Betsch, 2011) and their samples, validity, and methods (e.g., Simons et al., 2017). Given that the scope of generalization or context limitation of the items has not been clearly defined or established beforehand, this raises the question of falsifiability: How do we establish whether an empirical test of a theory or a phenomenon has been rejected or received empirical support? For falsifiability, there needs to be a clear criterion as to how to test items, and this needs to be defined a priori together with the theory, hypotheses, and empirical tests. We must avoid the unfalsifiable criteria by which evaluating whether items are relevant to the phenomenon is based solely on whether we find support for the phenomenon when testing the phenomenon using these items. If the items are not the criteria, but rather the criteria is the process by which the items are constructed, then both the theory and the empirical investigation demonstrating the phenomenon must be adjusted to include the procedure by which the items are created.

The commentary by Shafir and Cheek (2024) seems to claim that the phenomenon is so contextual that it must always be pretested and adjusted. This unfortunately severely limits the theory's generalizability and robustness, and therefore its importance. If the existence of the phenomenon always depends on the context and any failure to demonstrate the phenomenon can be put down to changes in context, then the natural question that arises is: Of what academic and practical use is the finding? In some of the commentary's claims, it appears as though the argument was that this specific empirical demonstration should only be expected to work as if studies are conducted at the same time, on the same sample, in the same context, and that each time we should test whether all stimuli have the same meaning. Yet the stimuli never have the same meaning. Even with the same participants, time passes, and people change and react differently at different times. In scientific follow-up studies, we use different samples and run them in different labs, at different times of the day, in different locations, with different background settings. The target article did not measure or report any of these factors, likely because they did not think that these should matter. If a phenomenon is truly robust and generalizable, to have enough meaning and to be of value and practical importance to the scientific community and stakeholders, then these small details better not matter. The factors that the authors expect to matter should be stated in advance and should be tied to the theory, and the empirical evidence should show that they have addressed those constraints. The argument that any other time, sample, or context requires a reformulation of the theory or empirical demonstration undermines its value. This is not just an empirical issue regarding replications; this is of importance to the broad audiences that rely on our findings in practice. Audiences need to know the scope of the phenomenon and how to adjust for implementation in their own setting.

Our team has shown that many classic judgment and decision-making studies from as early as the 1970s replicate well with original materials, even when these materials are very outdated, and even though much has changed over the past 5 decades. These have withstood the test of time, context,

https://doi.org/10.1017/jdm.2025.10011 Published online by Cambridge University Press

culture, language, and other factors, and it is what makes the decision-making phenomena so important in practice.

9. Conclusion

We responded to criticism by Shafir and Cheek's (2024) commentary regarding our replication and the value of direct replications. The commentary furthers our understanding of their views on the process required to empirically demonstrate the phenomenon. Yet more details are needed to clarify their theory and the scope and generalizability of the empirical process, to clearly specify how the cutoff thresholds are defined, and to further test the process rigorously for all problems from Shafir (1993). We addressed inaccuracies in reference to our work and responded in detail to their critiques of our replication and direct replications more broadly, reviewing the importance and value of direct replications, and calling for conducting and incentivizing more direct close replications.

Author biographies. S.P.C. is a postdoctoral researcher in the Department of Psychology at the Norwegian University of Science and Technology (NTNU). His research focuses on moral psychology, lay beliefs, and judgment and decision-making.

G.F. is an assistant professor in the Department of Psychology at the University of Hong Kong. His research focuses on judgment and decision-making.

Author contributions. Conceptualization, project administration, validation, writing—original draft, and writing—review and editing: S.P.C. and G.F. S.P.C. wrote the initial draft; G.F. revised and edited it. S.P.C. and G.F. finalized the manuscript for journal submission, with final edits by G.F.

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

Competing interests. We are the authors of the target article for the commentary. Other than that, we declare no potential competing interests with respect to the authorship and/or publication of this article.

References

Aiyer, S., Kam, H. C., Ng, K. Y., Young, N. A., Shi, J., & Feldman, G. (2023). Outcomes affect evaluations of decision quality: Replication and extensions of Baron and Hershey's (1988) Outcome Bias Experiment 1 *International Review of Social Psychology*, *36*(1), Article 12. https://doi.org/10.5334/irsp.751

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*(4), Article 569.

Baumeister, R. F., & Vohs, K. D. (2016). Misguided effort with elusive implications. *Perspectives on Psychological Science*, *11*(4), 574–575.

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour*, *6*(1), 29–42.

Chandrashekar, S. P., Weber, J., Chan, S. Y., Cho, W. Y., Chu, T. C. C., Cheng, B. L., & Feldman, G. (2021). Accentuation and compatibility: Replication and extensions of Shafir (1993) to rethink choosing versus rejecting paradigms. *Judgment and Decision Making*, *16*(1), 36–56.

Cheek, N., & Shafir, E. (2018). Further investigations into choosing versus rejecting. Unpublished manuscript, Department of Psychology, Princeton University.

Cheek, N. & Shafir, E. (2024). Further exploring choosing versus rejecting. Unpublished manuscript.

Chen, J., Kwan, L. C., Ma, L. Y., Choi, H. Y., Lo, Y. C., Au, S. Y., . . . & Feldman, G. (2021). Retrospective and prospective hindsight bias: Replications and extensions of Fischhoff (1975) and Slovic and Fischhoff (1977). *Journal of Experimental Social Psychology*, *96*, Article 104154. https://doi.org/10.1016/j.jesp.2021.104154

CORE Team (2025). Collaborative Open-science and meta REsearch. https://doi.org/10.17605/OSF.IO/5Z4A8; https://mgto.org/core-team/

de Moraes Ferreira, M., Santiago, M. Y. T., Seda, L., Batistuzzo, M. C., Bastos, R. V. S., Borborema, R. S., & Fatori, D. (2023). Replication of the 'money illusion' effect in a Brazilian sample. OSF. https://doi.org/10.31234/osf.io/fh597

Downs, J. S., & Shafir, E. (1999). Why some are perceived as more confident and more insecure, more reckless and more cautious, more trusting and more suspicious, than others: Enriched and impoverished options in social judgment. *Psychonomic Bulletin & Review*, *6*(4), 598–610.

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., . . . & Nosek, B. A. (2020). Many Labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309–331. https://doi.org/10.1177/2515245920958687

Feldman, G. (2022). 'Check me, replicate me' pledge—promoting collaborative replications, assessments, & corrections. https://doi.org/10.17605/OSF.IO/M3Z9G; https://mgto.org/check-me-replicate-me/

https://doi.org/10.1017/jdm.2025.10011 Published online by Cambridge University Press

Feldman, G. (2024). Registered reports and peer community in registered report are the future of science. https://doi.org/10.17605/OSF.IO/U7KJ5

Fiedler, K., McCaughey, L., & Prager, J. (2021). Quo Vadis, Methodology? The Key Role of Manipulation Checks for Validity Control and Quality of Science. *Perspectives on Psychological Science*, *16*(4), 816–826. doi: 10.1177/1745691620970602

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), Article 288.

Ganzach, Y. (2025). Accentuation explains the difference between choice and rejection better than compatibility: A commentary on Chandrashekar et al. (2021). *Judgment and Decision Making*, *20*, Article e1. https://doi.org/10.1017/jdm.2024.26

Glöckner, A. (2017). Research in judgment and decision making: Methodological challenges and theoretical developments, SPUDM 2017 presidential address. https://eadm.eu/wp-content/uploads/2017/08/Presidential_Address_Gloeckner_2017.pdf

Glöckner, A., & Betsch, T. (2011). The empirical content of theories in judgment and decision making: Shortcomings and remedies. *Judgment and Decision Making*, *6*(8), 711–721.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., . . . & Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573.

Klein RA, Vianello M, Hasselman F, et al. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225.

LeBel EP, McCarthy RJ, Earp BD, 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.

Lloyd, K. E., & Chambers, C. D. (2024). Registered reports: Benefits and challenges of implementing in medicine. *British Journal of General Practice*, *74*(739), 58–59.

McCarthy, R., Gervais, W., Aczel, B., Al-Kire, R. L., Aveyard, M., Marcella Baraldo, S., . . . & Zogmaister, C. (2021). A multisite collaborative study of the hostile priming effect. *Collabra: Psychology*, *7*(1), Article 18738. https://doi.org/10.1525/collabra.18738

McCarthy, R. J., Skowronski, J. J., Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., Acar, O. A., Aczel, B., Bakos, B. E., Barbosa, F., Baskin, E., Bègue, L., Ben-Shakhar, G., Birt, A. R., Blatz, L., Charman, S. D., Claesen, A., Clay, S. L., . . . Yıldız, E. (2018). Registered replication report: Srull and Wyer (1979). *Advances in Methods and Practices in Psychological Science*, *1*(3), 321–336. https://doi.org/10.1177/2515245918777487

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., . . . & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141.

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

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition*, *21*(4), 546–556.

Shafir, E. (2018). The workings of choosing and rejecting: Commentary on many Labs 2. *Advances in Methods and Practices in Psychological Science*, *1*(4), 495–496.

Shafir, E. & Cheek, N. N. (2024). Choosing, rejecting, and closely replicating, 30 years later: A commentary on Chandrashekar et al. *Judgment and Decision Making*, *19*, e5. doi: 10.1017/jdm.2023.48

Shafir, E., Diamond, P., & Tversky, A. (1997). Money illusion. *The Quarterly Journal of Economics*, *112*(2), 341–374. https://doi.org/10.1162/003355397555208

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Slovic, P., & Fischhoff, B. (1977). On the psychology of experimental surprises. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(4), Article 544.

Soderberg, C. K., Errington, T. M., Schiavone, S. R., Bottesini, J., Thorn, F. S., Vazire, S., . . . & Nosek, B. A. (2021). Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour*, *5*(8), 990–997.

Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology*, *37*(10), Article 1660.

Vazire, S. (2017). Quality uncertainty erodes trust in science. *Collabra: Psychology*, *3*(1), Article 1. https://doi.org/10.1525/collabra.74

Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A. J., Ainsworth, S. E., . . . & Albarracin, D. (2021). A multisite preregistered paradigmatic test of the ego-depletion effect. *Psychological Science*, *32*(10), 1566–1581.

Wedell, D. H. (1997). Another look at reasons for choosing and rejecting. *Memory & Cognition*, *25*(6), 873–887.

Woznyj, H. M., Grenier, K., Ross, R., Banks, G. C., & Rogelberg, S. G. (2018). Results-blind review: A masked crusader for science. *European Journal of Work and Organizational Psychology*, *27*:5, 561–576. https://doi.org/10.1080/1359432X.2018.1496081

https://doi.org/10.1017/jdm.2025.10011 Published online by Cambridge University Press

Ziano, I., Li, J., Tsun, S. M., Lei, H. C., Kamath, A. A., Cheng, B. L., & Feldman, G. (2021). Revisiting 'money illusion': Replication and extension of Shafir, Diamond, and Tversky (1997). *Journal of Economic Psychology*, *83*, Article 102349.
Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, Article e120.

Cite this article: Chandrashekar, S. P. and Feldman, G. (2025). On the process and value of direct close replications: A rejoinder to Shafir and Cheek's (2024) commentary on Chandrashekar et al. (2021). *Judgment and Decision Making*, e34. https://doi.org/10.1017/jdm.2025.10011

https://doi.org/10.1017/jdm.2025.10011 Published online by Cambridge University Press