Judgment and Decision Making (2024), **Vol. 19**:e5 1–11
doi:10.1017/jdm.2023.48

EMPIRICAL ARTICLE

Choosing, rejecting, and closely replicating, 30 years later: A commentary on Chandrashekar et al.

Eldar Shafir[1] and Nathan N. Cheek[2]

[1]Department of Psychology, School of Public and International Affairs, Princeton University, Princeton, NJ, USA and [2]Department of Psychological Sciences, Purdue University, West Lafayette, IN, USA

Corresponding author: Eldar Shafir; Email: shafir@princeton.edu

Received: 22 October 2023; **Accepted:** 6 November 2023

Keywords: choice; rejection; compatibility; accentuation; replication

Abstract

In a 'very close replication' study using the same attributes as the original, Chandrashekar et al. (2021) report a failure to replicate some choose–reject problems documented in Shafir (1993). We find that several of the original attributes have changed their valence three decades later, and we compose new versions with updated attributes that fully replicate Shafir's (1993) original findings. Despite their apparent exactitude, 'very close replications' across contexts or time, when stimuli may have changed their meaning or valence, can be highly misleading, further exacerbating replication concerns.

In their paper, 'Accentuation and compatibility: Replication and extensions of Shafir (1993) to rethink choosing versus rejecting paradigms', Chandrashekar et al. (2021) report a 'very close replication' study of a series of choose–reject problems documented in Shafir (1993). In the original 1993 paper, Shafir reported several studies showing that choosing and rejecting, contrary to standard assumptions, often are not complementary. Across several choice problems, ranging from parental custody to vacation decisions, gambles, and ice cream choices, Shafir asked participants either to choose or to reject one of two options (the problems were not all binary, but the analysis remains similar). Shafir found that the *enriched* option, the one with more positive and negative attributes, which he interpreted as providing more compelling reasons for choice and rejection, had a greater share of being chosen and rejected than the impoverished option, which provided weaker reasons for choice or rejection. In Shafir's (1993) study, the enriched option's share of being chosen and rejected added up to significantly more, and the impoverished option to significantly less, than the expected 100%. In some cases, the enriched option was more likely to be both chosen and rejected. To explain this pattern, Shafir appealed to the notion of compatibility, suggesting that options' strengths weigh more heavily when people choose, whereas weaknesses matter more when people reject, and that enriched options, presenting both more positive and more negative attributes, thus receive more than their fair share of choice and rejection (Shafir, 1993; see also Shafir et al., 1993 for a review).

Chandrashekar et al. (2021) report a 'very close replication' study of a series of choose–reject problems reported in Shafir (1993) and conclude that, 'Taken together, the replication findings do not indicate consistent support for the original findings.' This echoes an earlier failure to replicate reported by Many Labs 2 (Klein et al., 2018; see also Shafir, 2018 for commentary), who presented Shafir's

(1993) original Custody Problem, about awarding or denying custody of a child to one of two parents, to thousands of participants in several countries.

In this brief commentary, we address Chandrashekar et al.'s (2021) conclusions regarding the failure to replicate the original problems. In doing so, we provide new data on the replicability of Shafir's (1993) studies and discuss challenges with the earlier "very close" replications. We focus our analysis on the Custody Problem because it is the one problem used in both recent replication projects, but our argument applies more generally (and is further discussed, along with several new choose–reject problems including a revised version of Shafir's (1993) Problem 2, the Vacation Problem, in Cheek and Shafir (2024)).

1. Replications and cultural changes between then and now

Let us begin with another, related study. A little over 20 years ago, Downs and Shafir (1999) investigated the role of compatibility and its impact on enriched versus impoverished options in the realm of social judgment. They presented participants with names of well-known personages with similar occupations, where, in each pair, respondents were more familiar with one personage than the other. For example, American participants were significantly more familiar with Ronald Reagan (an elder stateman at the time, who had been President of the United States a decade earlier) than with John Major (who had recently been the UK's Prime Minister). Similarly, American participants were more familiar with Woody Allen (an American director at the height of his career at the time) than with Federico Fellini (a great Italian director barely known in the United States).

Participants were presented with various adjectives and had to select which personage in a pair was better described by each adjective. As predicted, and consistent with the compatibility hypothesis, participants were more likely to select the more familiar over the less familiar personage across opposite adjectives. For example, Reagan was judged as more confident than John Major by 61% of respondents, and as more insecure by 71%, totaling 132% across these opposing adjectives (with John Major receiving a total of 68%). Woody Allen received a total of 119% across confident and insecure, compared to Fellini's 81%. José Canseco totaled 149% as compared to Tony Gwynn's 51%, David Letterman totaled 139% compared to Kathy Lee Gifford's 61%, and so forth. Not all judgments were quite this extreme, but the overall tendency of the more familiar personages, the ones offering more features compatible with the judgment, to be selected across opposite adjectives was pronounced and highly significant.

Now, how would a 'very close replication' of this study proceed? A 'very close replication' insists on using the same stimuli as in the original, with the claim that it is instructive to see how the phenomena persist through time. However, clearly the highly familiar items—Letterman, Canseco, and even Reagan—will not persist through time, or at least not to the same degree. The very hypothesis that generated these results in the late 20th century—namely, that participants will more easily find compatible instances in the familiar personages—predicts that those same stimuli will not replicate 30 or 40 or 50 years later, when those personages' renown will have waned. (Furthermore, it is unlikely, e.g., that having married one adopted daughter and allegedly abused another, Woody Allen would retain his 1990s 70% 'more moral' advantage over Fellini.) The judgmental compatibility phenomenon published by Downs and Shafir in 1999 should, of course, replicate, but it would almost certainly require new stimuli: instead of Woody Allen and José Canseco, we may need Kate McKinnon and LeBron James.

Clearly, some phenomena, such as classical optical illusions, which are the outcome of evolutionary trends (Gregory, 2009) will replicate over long periods of time, whereas other phenomena, like those that depend on people's attitudes toward Woody Allen or José Canseco, marriage, smoking, gender roles, or the environment, can change in just a few years. Where exactly to locate findings along this continuum is a theoretically interesting and nontrivial question. What is clear is that some findings are going to be time-sensitive in ways that optical illusions are not. And 'very close replications' need to

https://doi.org/10.1017/jdm.2023.48 Published online by Cambridge University Press

observe those distinctions, or they risk generating confusing failures to replicate outdated items, rather than contributing to our understanding of the phenomena that lie behind them (for similar points, see, e.g., Ferguson et al., 2014; McGuire, 2013). This brings us to the issue at hand: replicating the patterns of choosing and rejecting documented by Shafir (1993) three decades later.

2. Choosing and rejecting 30 years later

With the goal of conducting a 'very close replication,' Chandrashekar et al. (2021) chose to run the original materials used by Shafir (1993). Both replication projects (Chandrashekar et al., 2021; Klein et al., 2018) used Shafir's Custody Problem. The original problem read as follows (with the original results, from Shafir, 1993, p. 549, reproduced below):

Original Problem:

> Imagine that you serve on the jury of an only-child sole-custody case following a relatively messy divorce. The facts of the case are complicated by ambiguous economic, social, and emotional considerations, and you decide to base your decision entirely on the following few observations. [To which parent would you award sole custody of the child?/Which parent would you deny sole custody of the child?]

	Award	Deny
Parent A		
Average income Average health Average working hours Reasonable rapport with the child Relatively stable social life	36%	45%
Parent B		
Above-average income Very close relationship with the child Extremely active social life Lots of work-related travel Minor health problems	64%	55%

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. 'Above-average income' and 'very close relationship with the child' were highly positive (compatible with choice), whereas 'lots of work-related travel', and 'minor health problems' were much more negative (compatible with rejection; 'extremely active social life' in that context was close to neutral). In fact, Parent B's rates of being awarded and denied (119%) exceeded the total of 100% expected if choosing and rejecting were complementary, $z = 2.48, p < .02$.

This problem, however, is not a tidy optical illusion—it is built on attributes that may very well change over 30 years. 'Average working hours' might sound more positive at a time when 'more than ever, workers want to work fewer hours' (Lufkin and Mudditt, 2021), and 'average income' may sound more appealing in an era of rapidly increasing economic hardship that has left millions of Americans without adequate income to meet their basic needs. Similarly, as cultural norms change and perhaps grow more conservative, 'extremely active social life' may, three decades later, be judged more negatively, connoting a certain neglect of family life in favor of fun. 'Lots of work-related travel' was viewed, in the late 1980s, highly negatively. Subsequent decades, however, brought major changes, including the enormous increase in the frequency and popularity of work-related travel.[1] When we collected ratings for various parental attributes in late 2017, we found that 'lots of work-related travel' was rated neutral—perceived as neither negative nor positive. But when we collected ratings again in the fall of 2022, perhaps due to the coronavirus pandemic and a shift to remote work, 'lots of work-related travel' was perceived negatively again.[2] It is not always easy to predict how cultural change, unprecedented global crises like the pandemic, or, for that matter, some 'fashionable' associations, may change how people perceive certain stimuli three decades later. As a result, timely pilot testing will virtually always be a plus.

3. Pilot surveys

We conducted two preregistered pilot surveys to gauge attribute valence among online participants. Participants (Pilot Survey 1: n = 172 MTurkers, Pilot Survey 2: n = 164 MTurkers; recruited using CloudResearch; Litman et al., 2017) read the following instructions:

> Imagine a child custody case following a messy divorce. One of the parents must have custody of the child. The parents are described by various attributes below. For each attribute, please indicate on the provided scale how 'positive' (good) or 'negative' (bad) in your opinion it is for a parent to have that attribute.

Participants were then presented with the original 10 attributes as well as several new attributes, and they rated each attribute's valence on a scale from −5 (highly negative) to 0 (neutral) to 5 (highly positive). Both surveys followed this procedure; the second survey was used to rate additional attributes for new versions of the Custody Problem. We preregistered both Pilot Survey 1 (https://aspredicted.org/by4jy.pdf) and Pilot Survey 2 (https://aspredicted.org/my3ib.pdf) through AsPredicted.org. Data, materials, and analysis code for all studies are available on the Open Science Framework (https://osf.io/cxst6/). Ratings for all attributes from both surveys are presented in Table 1.

As was to be expected, we found that the valence of some of the original attributes had changed. Interestingly, several attributes deemed neutral 30 years ago—average working hours, reasonable rapport with the child, relatively stable social life—were now quite positive. In fact, the impoverished parent, viewed neutrally three decades earlier, was viewed quite positively in 2022.

4. New (properly normed) versions of the Custody Problem

Based on the updated attribute ratings, we composed two new versions of the Custody Problem. These were close variations on the original problem with updated attributes intended to correct for changes in attribute perceptions over time. In the first new version, we ensured that the impoverished parent's attributes (Parent A below, though actual order of presentation was counterbalanced) were

[1] US air traffic figures soared from 297 million in 1980 to 638 million by just 2000; there is been a 300% increase in the number of overseas trips taken since the early 1990s; US residents have logged 464.4 million person-trips for business purposes in 2019—about 1.3 million Americans travelling for business overnight every day (https://www.ustravel.org; https://www.centennialofflight.net/essay/Social/impact/SH3.htm; The Guardian, July 1, 2019).

[2] The original 1993 attribute ratings, collected on pieces of paper then, are, unfortunately, no longer available, and some 2017 rating data were lost with personnel transitions. However, all data for the studies reported in full in this paper were collected anew and are fully available on our OSF page.

https://doi.org/10.1017/jdm.2023.48 Published online by Cambridge University Press

Table 1. *Parental attribute valence ratings in Pilot Surveys 1 and 2.*

Attribute	Pilot Survey 1	Pilot Survey 2
Original impoverished parent		
Average income	1.24 (1.58)	
Average health	1.74 (1.72)	
Average working hours	2.03 (1.88)	
Reasonable rapport with the child	2.42 (1.90)	
Relatively stable social life	2.62 (1.65)	
Original enriched parent		
Above-average income	2.78 (1.68)	2.83 (1.63)
Very close relationship with the child	4.24 (1.42)	4.47 (1.14)
Extremely active social life	−.44 (2.40)	
Lots of work-related travel	−1.85 (2.28)	
Minor health problems	−.19 (1.63)	.07 (1.60)
New attributes		
High income	3.01 (1.78)	
Fairly close relationship with the child	3.19 (1.69)	
Moderately active social life	1.01 (1.67)	1.13 (1.67)
Somewhat risky financial behavior	−2.15 (2.18)	
High need for order and cleanliness	1.22 (2.23)	1.35 (2.18)
Has a couple drinks almost every night	−1.74 (2.24)	−2.40 (2.14)
Did not finish college	−.27 (1.53)	
Needs to travel occasionally	−.23 (1.75)	−.01 (1.84)
Tends not to get enough sleep	−1.16 (1.91)	
Goes out drinking occasionally	−.62 (2.14)	
A bit impatient at times	−.62 (1.74)	
A little overprotective	.74 (1.79)	
Doesn't read books often	−.72 (1.55)	
Lives on street with somewhat busy rush hour traffic	−1.16 (2.32)	−1.55 (2.17)
Moderately strict about following household rules	1.45 (1.65)	
Very strict about following household rules	.49 (2.24)	
Not very strict about following household rules	−.92 (2.03)	
Mild dietary restrictions		.53 (1.35)
Loses temper only infrequently		.84 (2.27)
Occasionally is slightly distracted during conversations		−.26 (1.44)
Minor disagreements with neighbors		−.60 (1.54)
Occasionally indulges in sugary desserts		.52 (1. 46)
Has minor bouts of insomnia once in a while		−.30 (1.52)
Not great at telling jokes		.01 (.97)
Slightly below-average income		−.48 (1.63)
Every once in a while, has disagreement with child		.45 (1.44)
Slightly unpredictable social life		−1.20 (1.69)
Goes out for a drink or two with friends occasionally		.32 (1.84)
Infrequently reads books		−.39 (1.81)
Fairly relaxed about household rules		.57 (1.87)
Only rarely makes a risky financial investment		1.66 (2.03)
Lots of work-related travel (during COVID times)		−2.31 (2.18)
Lots of work-related travel (during non-COVID times)		−1.99 (2.26)

https://doi.org/10.1017/jdm.2023.48 Published online by Cambridge University Press

Table 1. *(Continued).*

Attribute	Pilot Survey 1	Pilot Survey 2
Tends to be somewhat sleep deprived		−1.06 (1.81)
Not very strict about household rules		−.71 (1.86)
Frequently goes out drinking		−3.52 (1.98)
Light working hours		2.23 (1.88)
Good health		3.59 (1.71)

Note: Descriptive statistics—means (SDs)—from the two pilot surveys. Attributes were rated from −5 (highly negative) to 0 (neutral) to 5 (highly positive). The attribute 'average working hours' was rated twice in Pilot Study 1. The second time it was rated, the average rating was 1.98 (*SD* = 1.77).

all relatively neutral (absolute value of average valence ratings below .55). We further ensured that the enriched parent's attributes (Parent B below) were either highly positively rated (three attributes rated 2.23 or higher) or negatively rated (two attributes rated −.92 or lower) in a combination roughly approximating the original enriched parent's attribute ratings. The instructions were identical to those in Shafir (1993) and reproduced in the context of the *Original Problem* above. For both versions, we followed Simonsohn's (2015) guidelines for powering replications by recruiting enough participants to ensure at least 2.5 times the original sample size (170), plus an additional cushion to detect even smaller effects.

We administered the first new Custody Problem in Study 1, preregistered through AsPredicted.org (https://aspredicted.org/n94mv.pdf) and run on Prolific. It is shown below, along with the percentage of participants who selected each option in the choose and the reject conditions (n = 552)[3]:

New Custody Problem 1

	Award	Deny
Parent A	(n = 272)	(n = 280)
Not great at telling jokes Needs to travel occasionally Has minor bouts of insomnia once in a while Mild dietary restrictions Did not finish college	29%	43%
Parent B		
Fairly close relationship with child Above-average income Light working hours Has a couple of drinks almost every night Not very strict about following household rules	71%	57%

[3] For Study 1, we recruited 600 participants and, based on our preregistered plan, excluded, before data analysis, those who failed an attention check or did not confirm nonrandom responding, resulting in a final sample of 552. See Appendix for details and participant demographics.

We found that the majority (71%) of participants asked to award custody awarded custody to the enriched parent, and the majority (57%) of participants asked to deny custody denied custody to the enriched parent. Mirroring Shafir's (1993) results, the sum of percentages of participants selecting the enriched option in the two conditions exceeds the 100% that would be expected if choosing and rejecting were complementary, instead totaling 128%, $z = 6.64$, $p < .001$. (Chandrashekar et al., 2021, reported one-tailed tests; we follow their convention, but the p-values for this problem and the next problem are below .001 with both two-tailed and one-tailed tests.)

For the sake of robustness, we composed a second new version with a slightly different goal. Whereas the New Custody Problem 1 intentionally approximated the attribute valences in the original problem, it used a different set of attributes. In our second new version, we sought to more closely approximate the attributes used in the original, while still aiming to compose fairly balanced enriched and impoverished options. Accordingly, we assigned the impoverished parent (Parent A below) two of the original traits that were currently somewhat positively rated ('average income' and 'average health'), while including one slightly negatively rated attribute ('not very strict about household rules') and keeping the other attributes close to neutral. We then composed the enriched option using the same attributes as in the original, with two attributes highly positively rated and one attribute highly negatively rated, along with two slightly negative attributes. We reasoned that revising the original impoverished parent's currently highly positively rated attributes may allow us to retain the original enriched option and replicate a pattern similar to the original.

We administered the second new version in Study 2, preregistered through AsPredicted.org (https://aspredicted.org/fg7vn.pdf) and run on Prolific. The New Custody Problem 2 is shown below, along with the percentage of participants who selected each option in the choose and the reject conditions (n = 564)[4]:

New Custody Problem 2

	Award	Deny
Parent A	(n = 281)	(n = 283)
Average income Average health Needs to travel occasionally Not very strict about household rules Goes out for a drink or two with friends occasionally	33%	48%
Parent B		
Above-average income Very close relationship with the child Extremely active social life Lots of work-related travel Minor health problems	67%	52%

[4] For Study 2, we recruited 600 participants and, based on our preregistered plan, excluded those who failed an attention check question and those who did not confirm nonrandom responding, resulting in a final sample of 564. See Appendix for details and participant demographics.

https://doi.org/10.1017/jdm.2023.48 Published online by Cambridge University Press

As before, and similar to the original, the sum of the percentages of participants who selected the enriched option in the two conditions exceeds the 100% expected if choosing and rejecting were complementary, instead totaling 119%, $z = 4.40$, $p < .001$. Both new versions of the Custody Problem yielded results that are consistent with the original findings of Shafir (1993).

5. Concluding thoughts about 'very close replications'

We began by gauging the valence of possible attributes to be assigned to the impoverished and enriched options in the Custody Problem, and we then composed two new versions of the problem that yielded results consistent with Shafir's (1993) findings. We found that the attributes used in the original Custody Problem had different valences three decades later, sufficient to render the original problem inadequate for testing Shafir's original research question. In fact, the results of the 'very close replications' of Chandrashekar et al. (2021) and Klein et al. (2018) show that the original problems run 30 years ago do not reproduce the original findings at present. (We also ran the original later in 2021 and failed to obtain the 1993 results, as reported in Cheek and Shafir 2024.) Nonetheless, the original decision pattern documented in Shafir (1993), with an asymmetry between choice and rejection occurring due to the enriched option obtaining more than its fair share relative to the impoverished option, appears alive and well.

We are not the first to raise concerns about the practice of simply readministering the same experimental materials in a sufficiently different context and then interpreting the different results as a sign of the initial findings' unreliability (e.g., Crandall and Sherman, 2016; Gergen, 1973; Schwarz and Strack, 2014). Yet, the practice appears to persist without fully addressing these concerns. We view this as problematic and in need of refinement. When stimuli are context- or time- (or, obviously, language-) sensitive, running them 'as is', in different contexts, years later, in different cultural context, or in a language foreign to the respondents, without the appropriate updates or translations, is bound to mislead. From the Me Too movement to diets, commuting, bragging presidents, and family roles, enough has changed in the United States over the past 30 years to warrant a revision of the attributes used to describe relevant stimuli before certain replications ought to be explored. Otherwise, 'very close replications' do not seem, from a theoretical perspective, very close after all.

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). In addition, researchers attempting replications need to discuss their findings with greater subtlety. It is problematic, in our view, to conclude that a decades-old finding 'does not replicate' without sufficient consideration for the different contexts (time, culture, norms, associations, etc.) in which the studies occurred. In some circumstances, 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.

There is, furthermore, a concern that replications with little theoretical validity can impede research progress. Chandrashekar et al. (2021) used the original materials from Shafir (1993) not only in an attempt at a replication, but also to address some theoretical work around the proposed mechanism of 'accentuation' (Wedell, 1997). Because of their faulty stimuli, however, it may be difficult to know what to make of the rest of their theoretical treatment.

In his analysis of choice–reject discrepancies, Shafir (1993) attributed the observed patterns to the well-established principle of compatibility, where the weighting of inputs is enhanced by their compatibility with the response (see, e.g., Kornblum et al., 1990; Proctor and Reeve, 1990; Shafir, 1995; Slovic et al., 1990 for relevant discussions). Wedell (1997) presented data in tension with this interpretation, and argued for an 'accentuation' hypothesis, according to which a greater demand for justification in choice compared with rejection leads to accentuation of the differences between alternatives in the choice condition. Cheek and Shafir (2024) document several more choose–reject patterns and discuss the contributions of both compatibility and accentuation, as well as the significant role of simple response errors, in contributing to the emergence of the choose–reject discrepancy.

https://doi.org/10.1017/jdm.2023.48 Published online by Cambridge University Press

We hope that this brief reply contributes both to basic research on decision-making and to more careful design and interpretation of replication studies. For now, what is clear is that the choose–reject patterns documented more than 30 years ago are, given the necessary updates, real and replicable.

Data availability statement. The datasets for this article can be found at https://osf.io/cxst6/.

Funding statement. This research received no specific grant funding form any funding agency, commercial, or not-for-profit sectors.

Competing interest. The authors declare none.

References

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. N. & Shafir, E. (2024). *Further exploring choosing versus rejecting*.

Crandall, C. S. & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93–99.

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.

Ferguson, M. J., Carter, T. J., & Hassin, R. R. (2014). Commentary on the attempt to replicate the effect of the American flag on increased Republican attitudes. *Social Psychology*, *45*(4), 301–302.

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.

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, *26*(2), 309–320.

Gregory, R. L. (2009). *Seeing through illusions*. Oxford: Oxford University Press.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490.

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: cognitive basis for stimulus-response compatibility–a model and taxonomy. *Psychological Review*, *97*(2), 253.

Litman, L., Robinson, J., & Abbercock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavioral Research Methods*, *49*(2), 433–442.

Lufkin, B. & Mudditt, J. (2021). *The case for shorter work week.* BBC. https://www.bbc.com/worklife/article/20210819-the-case-for-a-shorter-workweek

McGuire, W. J. (2013). An additional future for psychological science. *Perspectives on Psychological Science*, *8*(4), 414–423.

Proctor, R. W., & Reeve, T. G. (1990). Research on stimulus-response compatibility: Toward a comprehensive account. In *Advances in psychology* (Vol. 65, pp. 483–494). North-Holland.

Schwarz, N. & Strack, F. (2014). Does merely going through the same moves make for a "direct" replication? Concepts, contexts, and operationalizations. *Social Psychology*, *45*(4), 305–306.

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

Shafir, E. (1995). Compatibility in cognition and decision. In *Psychology of learning and motivation* (Vol. *32*, pp. 247–274). Cambridge, MA: Academic Press.

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., Simonson, I., & Tversky, A. (1993). Reason-based choice. *Cognition*, *49*(1-2), 11–36.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569.

Slovic, P., Griffin, D., & Tversky, A. (1990). Compatibility effects in judgment and choice. In R. Hogarth (Ed.), *Insights in decision making: A tribute to Hillel J. Einhorn* (pp. 5–27). Chicago: The University of Chicago Press.

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

Appendix

Additional details about data collection and demographic characteristics of participants are reported for all studies below.

https://doi.org/10.1017/jdm.2023.48 Published online by Cambridge University Press

Pilot Survey 1

We recruited 200 participants from Mechanical Turk using CloudResearch (Litman et al., 2017). To be included in analyses, participants had to pass an instructional manipulation check and confirm that they did not respond randomly. A total of 172 participants met these criteria and were included in analyses. This sample included 82 women, 88 men, and 2 nonbinary participants. Then, 22 participants identified as Black, 11 identified as Asian, 10 identified as Latinx, 4 identified as Native American, 124 identified as White, 6 identified as Multiracial, and 2 identified with additional categories (numbers may exceed sample size because participants could select multiple categories). The average age of participants was 38.16 (*SD* = 11.57) and age ranged from 19 to 72.

Pilot Survey 2

As in Pilot Study 1, we used CloudResearch to recruit 200 participants from Mechanical Turk and excluded those who failed an instructional manipulation check and/or indicated that they responded randomly, leaving a final sample of 164. This sample included 82 women, 81 men, and 1 nonbinary participant. Then, 22 participants identified as Black, 14 identified as Asian, 14 identified as Latinx, 5 identified as Native American, 116 identified was White, 8 identified as Multiracial, and 2 identified with additional categories (numbers may exceed sample size because participants could select multiple categories). The average age of participants was 38.22 (*SD* = 11.90) and age ranged from 19 to 70.

Study 1

In Study 1, we aimed to recruit 600 participants through Prolific to ensure that we had more than 2.5× times (see Simonsohn, 2015 for details) the original sample size of Shafir (1993), along with additional power to further detect potentially smaller effect sizes within our budgetary constraints. In total, 600 participants completed the study, of whom 552 met the inclusion criteria (same as previous studies). This sample included 235 women, 299 men, 17 nonbinary participants, and 1 who did not report gender. Then, 47 participants identified as Black, 75 identified as Asian, 60 identified as Latinx, 5 identified as Native American, 376 identified as White, 30 identified as Multiracial, and 2 identified with additional categories (numbers may exceed sample size because participants could select multiple categories). The average age of participants was 36.42 (*SD* = 13.14) and age ranged from 18 to 93.

Three hundred sixty-six participants reported annual personal incomes ranging from $0K-$50K, 137 reported annual personal incomes ranging from $50K-$100K, and 48 participants reported annual personal incomes above $100K. Two hundred forty-four participants reported annual household incomes ranging from $0K-$50K, 180 reported annual household incomes ranging from $50K-$100K, and 127 reported annual household incomes above $100K. On a 10-point scale, participants reported an average subjective social status of 4.76 (*SD* = 1.78) and on a scale from 0 (completely liberal) to 100 (completely conservative) participants leaned liberal, with an average rating of 34.18 (*SD* = 28.05).

Study 2

Following the sample size planning of Study 1, we recruited 600 participants through Prolific, of whom 564 met the inclusion criteria (same as previous studies). This sample included 244 women, 301 men, and 19 nonbinary participants. 48 participants identified as Black, 56 identified as Asian, 45 identified as Latinx, 7 identified as Native American, 413 identified as White, 19 identified as Multiracial, and 2 identified with additional categories (numbers may exceed sample size because participants could select multiple categories). The average age of participants was 38.36 (*SD* = 14.41) and age ranged from 18 to 80.

Here, 359 participants reported annual personal incomes ranging from $0K to 50K, 147 reported annual personal incomes ranging from $50K to $100K, and 58 participants reported annual personal

https://doi.org/10.1017/jdm.2023.48 Published online by Cambridge University Press

incomes above $100K. Then, 206 participants reported annual household incomes ranging from $0K to $50K, 212 reported annual household incomes ranging from $50K to $100K, and 146 reported annual household incomes above $100K. On a 10-point scale, participants reported an average subjective social status of 5.11 (*SD* = 1.73) and on a scale from 0 (completely liberal) to 100 (completely conservative) participants leaned liberal, with an average rating of 33.27 (*SD* = 27.09).

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