nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-020-0886-x

Replicating patterns of prospect theory for decision under risk

Kai Ruggeri[1,2] ✉, Sonia Alí[3], Mari Louise Berge[4], Giulia Bertoldo[5], Ludvig D. Bjørndal[6], Anna Cortijos-Bernabeu[7], Clair Davison[8], Emir Demić[9], Celia Esteban-Serna[10], Maja Friedemann[11], Shannon P. Gibson[12], Hannes Jarke[2], Ralitsa Karakasheva[13], Peggah R. Khorrami[14], Jakob Kveder[15], Thomas Lind Andersen[16], Ingvild S. Lofthus[6], Lucy McGill[17], Ana E. Nieto[18], Jacobo Pérez[18], Sahana K. Quail[19], Charlotte Rutherford[20], Felice L. Tavera[21], Nastja Tomat[15], Chiara Van Reyn[22], Bojana Većkalov[9,23], Keying Wang[10], Aleksandra Yosifova[24], Francesca Papa[25], Enrico Rubaltelli[26], Sander van der Linden[27] and Tomas Folke[1,2] ✉

Prospect theory is among the most influential frameworks in behavioural science, specifically in research on decision-making under risk. Kahneman and Tversky's 1979 study tested financial choices under risk, concluding that such judgements deviate significantly from the assumptions of expected utility theory, which had remarkable impacts on science, policy and industry. Though substantial evidence supports prospect theory, many presumed canonical theories have drawn scrutiny for recent replication failures. In response, we directly test the original methods in a multinational study (*n* = 4,098 participants, 19 countries, 13 languages), adjusting only for current and local currencies while requiring all participants to respond to all items. The results replicated for 94% of items, with some attenuation. Twelve of 13 theoretical contrasts replicated, with 100% replication in some countries. Heterogeneity between countries and intra-individual variation highlight meaningful avenues for future theorizing and applications. We conclude that the empirical foundations for prospect theory replicate beyond any reasonable thresholds.

One of the most influential papers across all of the behavioural sciences is the 1979 *Econometrica* article by Daniel Kahneman and Amos Tversky, entitled 'Prospect theory: an analysis of decision under risk'[1]. The study was conducted with university faculty and student participants from Israel, Sweden and the United States (with item sample sizes between 64 and 141). The items followed a typical structure in decision-making research: binary financial choices with probabilistic outcomes. Across 20 items, in which various choices ('prospects') were presented in terms of value and probability, the authors established that the resulting patterns diverged substantially from the predictions of expected utility theory, the dominant descriptive theory at the time.

Prior to prospect theory, there had been some exploration of deviations from the predictions of expected utility theory, such as the reflection effect[2] and observing inconsistent weighting of probabilities[3]. However, there was no general account of these deviations, and decision-making approaches still largely emphasized the prevailing traditions regarding choice, which comprised expected values, utility and the axioms of rational behaviour. As such, it was generally assumed that, when making decisions, rational individuals seek to optimize outcomes using stable algorithms tied to value, probability and cumulative wealth. Kahneman and Tversky used 20 binary choices organized into 13 contrasts (some items appeared in multiple contrasts) to challenge this model.

These contrasts highlighted six major deviations from the predictions of expected utility theory: the certainty effect, the reflection effect, the framing effect, the isolation effect, the overweighting of small probabilities and magnitude perception. The certainty effect

[1]Department of Health Policy and Management, Mailman School of Public Health, Columbia University, New York, NY, USA. [2]Policy Research Group, Centre for Business Research, Judge Business School, University of Cambridge, Cambridge, UK. [3]School of Psychology, University of Sussex, Brighton, UK. [4]Department of Personality & Health Psychology, Eötvös Loránd University, Budapest, Hungary. [5]School of Psychology, University of Padova, Padova, Italy. [6]Department of Psychology, University of Oslo, Oslo, Norway. [7]Department of Social Psychology and Quantitative Psychology, Faculty of Psychology, University of Barcelona, Barcelona, Spain. [8]School of Psychology and Neuroscience, University of St Andrews, St Andrews, UK. [9]Department of Psychology, Faculty of Philosophy, University of Belgrade, Belgrade, Serbia. [10]Division of Psychology & Language Sciences, University College London, London, UK. [11]Department of Experimental Psychology, University of Oxford, Oxford, UK. [12]Department of Psychology, Health & Professional Development, Faculty of Health & Life Sciences, Oxford Brookes University, Oxford, UK. [13]Division of Psychiatry and Applied Psychology, University of Nottingham, Nottingham, UK. [14]Department of Sociomedical Sciences, Mailman School of Public Health, Columbia University, New York, NY, USA. [15]Department of Psychology, Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia. [16]PPR Svendborg, Svendborg Municipality , Svendborg, Denmark. [17]Trinity College, Dublin, Ireland. [18]Department of Psychology, University Francisco de Vitoria, Madrid, Spain. [19]Department of Social Policy & Intervention, University of Oxford, Oxford, UK. [20]Department of Psychology, University of Cambridge, Cambridge, UK. [21]Department of Psychology, University of Cologne, Cologne, Germany. [22]Department of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium. [23]Faculty of Social and Behavioural Sciences, University of Amsterdam, Amsterdam, The Netherlands. [24]Department of Cognitive Science and Psychology, New Bulgarian University, Sofia, Bulgaria. [25]Organisation for Economic Cooperation and Development, Paris, France. [26]JDM Lab, Department of Developmental Psychology and Socialization, University of Padova, Padova, Italy. [27]SDM Lab, Department of Psychology, University of Cambridge, Cambridge, UK. ✉e-mail: kai.ruggeri@columbia.edu; nf2480@cbr.cam.ac.uk

refers to how small guaranteed outcomes are often preferred to larger risky outcomes when the expected values are equal. The reflection effect means that people tend to be risk seeking when maximizing gains, but risk averse when minimizing losses. The framing effect is an extension of the reflection effect, in that risk preferences change depending on whether a choice is presented in terms of gains or losses, even when the prospects of the options are held constant. Similarly, the isolation effect captures the fact that preferences for a choice may change depending on how it is structured sequentially. Finally, people are sensitive to relative as well as absolute magnitude. Applied to probabilities, this sensitivity leads to the overweighting of small probabilities, meaning that most find a difference between 1% and 2% more meaningful than a difference between 51% and 52%, even though the event is one percentage point more likely in either case. In the context of outcomes, it leads to shifts in magnitude perception, meaning that most people find the difference between $100 and $200 more meaningful than the difference between $1,100 and $1,200. In other words, the marginal value of the outcome generally scales with magnitude. There are few inferential statistics presented in the 1979 results, which are ostensibly single-item chi-squares; the bulk of the original argument focused on contrast pairs presented as descriptive statistics. However, as the effect sizes for these contrasts were generally large, their conclusions were broadly accepted.

It is difficult to overstate the level of influence that prospect theory has had on science, policy, management, financial services, government and beyond. The concept writ large has been cited as an explanatory framework for understanding a broad range of behaviours[4], including finance[5,6], investment[7,8], insurance[9,10] and political conflict[11]. This is only a limited selection of examples: shortly after Daniel Kahneman received the 2002 Nobel Prize for Economic Sciences, prospect theory was referred to as the most influential theoretical framework in all of the social sciences[12]. The 1979 paper that launched the theory has since become the most cited economic paper and is among the most cited in psychological science[13].

Affirmations and critiques of prospect theory have been available since almost immediately after the original study was published and picked up heavily through the 1980s and 1990s[14]. Much of this work investigated the initial claims, providing affirmations[15,16], adding details[17,18] and identifying gaps[19] in the original work. Studies covered a variety of methods, typically including similar experimental conditions to the original study, observing attitudes towards taxation and conducting meta-analyses of a battery of surveys across populations[20]. These studies were critical not only in advancing the science of decision-making under risk and uncertainty but also in establishing Tversky and Kahneman's revised cumulative prospect theory, published in 1992 (ref. [21]). While studies often made use of students[22] (many in master of business administration or other business-related degree programmes[19,23]), others studied individuals from specific professions[24].

In sum, these works have provided substantial evidence in support of the main tenets of prospect theory. While meta-analyses of those studies would probably show consistency in the direction of effects, direct and large-scale replications are much better suited to producing reliable effect size estimates[25]. There are many recent examples of effects in psychology that have been repeatedly replicated in concept and yet failed as large-scale direct replications[25]. Therefore, it is worthwhile to field a comprehensive, highly powered, direct replication of the original method across multiple countries.

Original research exploring prospect theory continues to the present day, and although it has generally shifted more towards testing applications, there remains considerable interest in the original study. For example, a recent large-scale replication[26] tested the behavioural paradoxes of prospect theory at the intra-individual level, concluding that while there can be a general effect for most people, individuals never exhibit all nine of the presumed irrational behaviours. However, though using some of the items from the original paper, that study focused on participants with low numeracy from one country using a single crowdsourcing platform.

We propose that a major cross-cultural replication of the original study is critical to determine what outcomes appear if tested again today. Given the widespread application of the behavioural sciences in government institutions around the world[27], this replication should involve participants from a number of countries (including those outside of North America and Western Europe) and backgrounds to determine whether the core tenets of prospect theory are consistent and broadly applicable, taking into account current, local and relevant income and financial standards.

Forty years have passed since the publication of the initial manuscript, and the conclusions of the original study are regularly used to interpret advances in the behavioural sciences as well as major world events. Perhaps the largest leap into such mainstream recognition occurred when prospect theory formed the core science behind Thaler and Sunstein's landmark writing on nudge theory, which popularized behavioural economics and revolutionized approaches to policy in organizations and governments around the world[28]. Given the time that has passed since the original study, its position among the behavioural science canon, its widespread influence in science and policy[29] and relevant critiques of its methods and conclusions, Kahneman and Tversky's 1979 study deserves an unbiased, robust reassessment, at a scale commensurate with its impact. This is particularly critical in light of concerns about replicability in behavioural science, as well as the translatability of findings between locations[30].

The original manuscript brought major change to the behavioural sciences. However, a number of ambiguities appear in the authors' methods, such as the precise sample characteristics and item responses and which currencies were used for participants from each of the three countries. Our study aims to bring fundamental observations leading to the formation of prospect theory in line with the critical standards of reproducibility in behavioural science in 2020 by attempting the replication of the full study from 1979. A failed replication would have seismic implications for behavioural science, whereas a successful replication would offer tremendous value and, undoubtedly, reassurance. Confirmation would clearly strengthen our confidence in core assumptions about decision-making. Additionally, a multinational replication would enable the empirical documentation of variability between locations and languages (or the lack thereof) in key aspects of financial decision-making under risk. This study attempts those steps by directly replicating well-established effects across a number of countries and languages.

Results

Data summary. In this study, we attempted a direct replication of 17 of the 20 items described in Kahneman and Tversky's 1979 paper proposing prospect theory, updating only the currency to present values and requiring all participants to complete all items. Our final sample consisted of 4,098 participants from 19 different countries covering 13 languages. Direct sampling accounted for 73.8% of the final sample (paid sample, 26.2%). The analyses include composite and separated samples, but with the wider view that if the tenets of prospect theory remain intact, the original 1979 results should be generally observed across populations (even with some variability).

The purpose of the analysis was twofold: first, to evaluate the reproducibility of the findings from one of the most influential papers in the behavioural sciences; and second, to unpack general themes of reproducibility based on sampling in multiple settings or languages, such as attenuation and the commutability of effects.

Demographics. Sixteen of the 19 countries recruited more than 141 participants (Table 1), which was the largest sample reported in the 1979 prospect theory study. All country samples are larger than

Table 1 | Samples per country by direct, paid and total *n*

Country	Language	Direct *n*	Paid *n*	Total *n*
Germany	German	186	141	327
Italy	Italian	157	144	301
United Kingdom	English	290	—	290
Australia	English	282	—	282
Mainland China	Simplified Chinese	259	—	259
Ireland	English	113	143	256
Serbia	Serbian	246	—	246
Hungary	Hungarian	101	142	243
United States	English	33	210	243
Norway	Norwegian	189	37	226
Slovenia	Slovenian	202	—	202
Spain	Spanish	199	—	199
Belgium	Dutch	127	65	192
Hong Kong	Traditional Chinese	160	—	160
Denmark	Danish	121	29	150
Chile	Spanish	89	56	145
Sweden	Swedish	106	33	139
Bulgaria	Bulgarian	98	29	127
Austria	German	70	41	111
Total		3,028	1,070	4,098

the original sample for all items apart from a single item (item 12), for which three of our country samples are smaller.

Of the final total sample ($n=4{,}098$), 50.7% were female. The median age was 29 years, and the ages ranged from 18 to 85 years. Sixty-seven per cent of the participants were university educated. For the country-specific demographics, see Table 2.

The preregistration included two exclusion criteria: participants who failed an attention check and participants whose completion time differed from the median completion time by more than three absolute deviations. We excluded 11 participants who were faster than three median absolute deviations[31] of the median completion time (86 s; the median completion time was 8 min). In the preregistration, we had planned to apply this criterion symmetrically to slow participants as well. However, given that 488 people failed this criterion, and we could assess data quality through the attention check, the slow participants were retained.

As we explored the data, we noted multiple additional indicators of poor data quality that led to further exclusions. Three participants were excluded for reporting an income as '99999' as we suspect these might be members of the research team testing the survey. Six participants were excluded for reporting being billionaires, which brought into question the validity of their responses. One participant was excluded for reporting a negative income. We also excluded five participants who reported being over 110 years old. To minimize the risk of participants mindlessly clicking through the questionnaire, we excluded participants who both (1) gave the same responses for more than 14 out of the 17 items and (2) completed the survey faster than one median absolute deviation below the median (6 min). These criteria led to 42 additional exclusions, making the final total sample size 4,098. The full annotated code used to clean and combine the data and make the exclusions is publicly available on the OSF platform.

To ensure that these deviations from the preregistration did not impact our conclusions, we created another dataset where we followed the preregistered exclusion criteria exactly. That is, we excluded only respondents who failed the attention check and who responded too quickly or too slowly (outside of three median absolute deviations[31] of the median completion time). This resulted in a final sample size of 3,666. All key analyses were repeated on this second dataset (Supplementary Results, section F6). All our results were qualitatively similar, and all of the conclusions reported in the results and discussion are consistent independent of the exclusion criteria.

Preplanned analyses. We planned four distinct sets of analyses in our pre-analysis plan (https://osf.io/wd4k5). The first of these involved replicating the (presumptively) chi-squared tests reported in the 1979 paper, which evaluated whether the response distribution for each item significantly differed from chance, on the country level. Our criterion for successful replication was detecting a significant effect in the same direction as the original study. We did not explicitly prespecify alpha thresholds, but since our power analysis for the country-level analysis was based on an alpha of 0.05, we applied it across the unpooled analyses. For the pooled data, we use a stricter 0.001 threshold.

For the pooled analyses, we ran a random effects meta-analysis to combine the information from all countries while respecting the hierarchical nature of the data. This was not preplanned but followed from recommendations from reviewers. We estimated the pooled effects in terms of log-odds, using maximum likelihood estimation, and then transformed the log-odds back into proportions before reporting them in Fig. 1. The overall replication rate for the pooled data was 16/17 (94.1%). All significant effects were in the same direction as in the original study, giving a 15/16 (93.8%) replication rate. Item 8 cannot be included in the second case, as it was not significantly different from chance in the original study. For the pooled sample, all items had response distributions that significantly differed from chance at the 0.001 threshold, apart from item 4 (response proportion for option A, 0.51; 95% confidence interval (CI), 0.48–0.54; $P=0.54$) and item 8 (response proportion for option A, 0.51; 95% CI, 0.48–0.53; $P=0.80$). The random effects meta-analysis suggests that the variation in effect sizes between countries cannot be accounted for by sampling variation alone (Q-tests were significant for all items), meaning that there is systematic variation in effect sizes across countries (Supplementary Results, section F5). We also tested what our results would have been if we had simply matched the sample sizes of the original study (Supplementary Results, section D1).

Next, we conducted unpooled analyses where we computed odds ratios for each item and country independently. Replication rates with these analyses showed that 247 out of 304 possible effects (81%) were significant in the same direction as in the original study. The unpooled analyses showed two important trends. First, while most effects replicated, there was a general attenuation of effects relative to the original study (Fig. 1). Specifically, 77% (95% CI, 72–82%) of the effects measured in this replication were smaller than those reported in the original study ($\chi^2=95.90$, d.f. $=1$, $P<0.0001$; this analysis was not preplanned).

There was much greater variation in replication rates between items than between countries. The replication rates between countries ranged from 69% to 94% (Fig. 2); the replication rates between items ranged from 15% to 100%. This means that individual items were more likely to have multiple failed replications than countries. It is also worth noting that there does not seem to be any pattern in terms of higher or lower replication rates based on values or income, as the four countries (Austria, Belgium, Germany and Ireland) with the same values as in 1979 are no more similar or more different from other countries. Additionally, there is no obvious pattern for higher or lower median national incomes.

The second set of preplanned analyses involved running bootstrapped analyses to evaluate to what extent our findings could be

Table 2 | Study demographics

Country	Language	*n*	% paid	Amount paid per participant (£)	% female	Age, median (IQR) (yr)	Income, median (IQR) (value reported in local currency as used in measure)	% university educated
Pooled	—	4,098	26.18	—	50.70	29 (24–38)	—	66.94
Australia	English	282	0	—	35.11	31 (26–37)	60,000 (33,000–90,000)	77.30
Austria	German	111	36.94	1.10	45.95	28 (24–36)	15,000 (5,000–30,000)	53.15
Belgium	Dutch	192	33.85	1.15	47.40	27 (23–38)	15,000 (1,860–25,000)	64.58
Bulgaria	Bulgarian	127	22.83		61.42	33 (26–42)	15,000 (7,000–25,100)	81.89
Chile	Spanish	145	38.62	0.84	51.03	27 (24–36)	2,700,000 (260,000–12,000,000)	64.14
Denmark	Danish	150	19.33	0.95	33.33	32 (26–40)	240,000 (74,000–300,000)	72.67
Germany	German	327	43.12	1.10	39.14	27 (24–33)	15,000 (4,250–30,000)	66.67
Hong Kong	Chinese (Traditional)	160	0	—	63.75	30 (24–43)	200,000 (30,000–425,000)	70.62
Hungary	Hungarian	243	58.44	0.70	43.21	29 (24–35)	2,000,000 (405,000–3,922,536)	60.08
Ireland	English	256	55.86	1.30	58.20	32 (24–41)	23,500 (10,000–35,000)	70.31
Italy	Italian	301	48.68	1.00	55.96	29 (23–43)	7000 (15–20,000)	45.03
Mainland China	Chinese (Simplified)	259	0	—	55.98	33 (27–41)	100,000 (60,000–200,000)	86.10
Norway	Norwegian	226	16.37	1.30	58.85	29.5 (25–38)	327,500 (131,500–469,250)	76.55
Serbia	Serbian	246	0	—	71.54	25 (23–35)	155,000 (20,500–600,000)	63.41
Slovenia	Slovenian	202	0	—	51.49	25.5 (23–35)	5,350 (2,000–15,000)	61.39
Spain	Spanish	199	0	—	57.29	27 (25–46.5)	8,000 (675–23,000)	77.39
Sweden	Swedish	139	23.74	1.30	27.34	28 (24–33)	250,000 (72,500–360,000)	41.73
United Kingdom	English	290	0	—	48.28	29 (24–38)	20,000 (9,069–32,375)	70.69
United States	English	243	86.42	0.70	54.32	30 (25–41)	26,000 (10,000–47,700)	62.14

IQR, interquartile range.

attributed to sampling variation. We found that sampling variation had a negligible impact on the findings, as would be expected given our sample size (Supplementary Results, section D2). The third set of planned analyses involved using hierarchical Bayesian models to test whether any demographic variables reliably influenced choices. There was no evidence that any demographic variable consistently predicted choice. For example, gender was the most common predictor to have posterior coefficients that did not cross zero, yet reliable differences between men and women were found in fewer than half of the items (Extended Data Fig. 1). For a more detailed explanation of the Bayesian analyses as well as posterior coefficient plots for all demographic variables for each item, see the Supplementary Results, section D3. In our preregistration, we had planned to allow both intercepts and slope coefficients to vary by country, but when we attempted these models, chains did not converge. We therefore simplified the models, treating the slope coefficients as fixed effects. This is further clarified in the Supplementary Results, section D3.

The fourth set of preplanned analyses involved evaluating the impact of demographics on the item contrasts, using a similar analysis strategy as for the third set of analyses. However, given the limited impact of the demographic variables in the third phase, we opted instead for a simpler approach, computing the odds ratios for the proportions of the contrasting items.

Unplanned analyses. When we used a random effects meta-analysis to explore the pooled effect sizes of the 1979 theoretical contrast pairs, 12 out of 13 contrasts replicated. The exception was the contrast between items 4 and 8 (log-odds, 0.03; 95% CI, −0.14–0.20; $P=0.76$), which tests for the presence of the reflection effect. However, since our sample was largely indifferent to options both in the gain domain (item 4) and in the loss domain (item 8), this observation does not challenge the presence of the effect itself (absence of evidence). Instead, it seems more likely that in this case there was simply no preference to reflect. Interestingly, there seems to be more homogeneity in the contrast effects than in the raw choices, as 4 out of the 13 contrasts (item 3 versus item 4, item 7 versus item 8, item 6 versus item 10 and item 4 versus item 11) did not show significant heterogeneity according to the *Q*-test (see Supplementary Results,

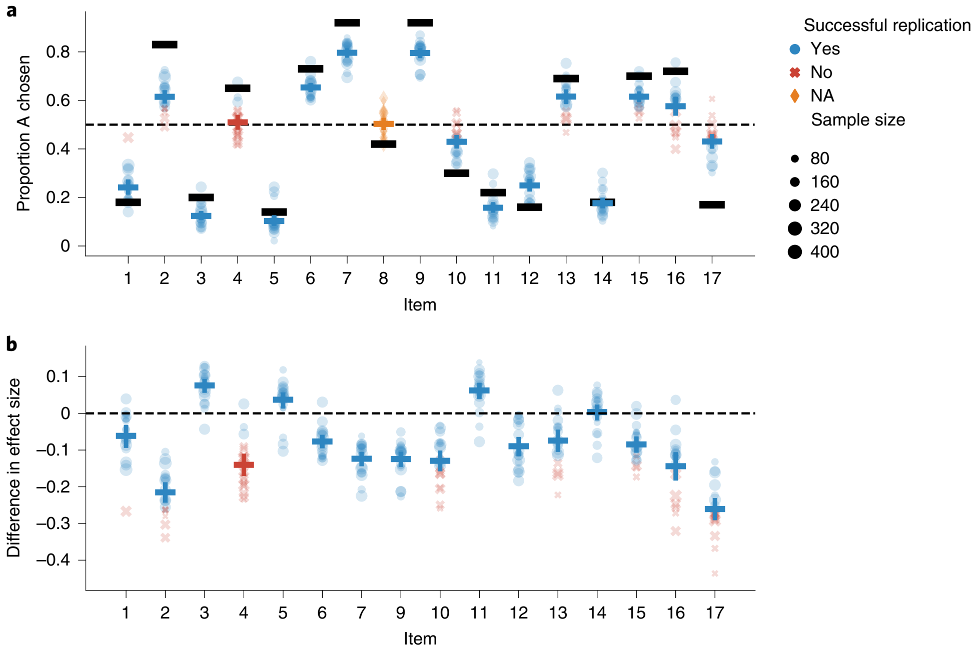


Fig. 1 | Effect sizes by item. a, Comparisons between the effects reported by Kahneman and Tversky in 1979 (black bars) and the current samples. A replication was deemed successful if the direction of the effect was the same as in the original study and significantly different from chance at an alpha threshold of 0.05. The dashed line indicates where participants would have picked option A 50% of the time for that item, indicating indifference relative to option B. **b**, The change in effect size compared with Kahneman and Tversky (1979). The dashed line represents the effect reported by Kahneman and Tversky. Values above the line suggest that participants in the current sample show a stronger preference for one of the options (are further from the 0.5 indifference point) than in the original study. Values below the line suggest a weaker preference (closer to the indifference point). There is an attenuation in effect sizes compared with 1979 for most items. In both **a** and **b**, the sample size is 4,098 for all items except item 14, where the sample size is 3,874, and item 15, where the sample size is 4,052. Each symbol represents one country; the blue, red and orange bars are the pooled results, with error bars indicating the 95% CIs. When the colours appear darker, this indicates a higher concentration of effects overlapping by country. NA: item 8 was not significant in the 1979 paper, and thus it is impossible to replicate.

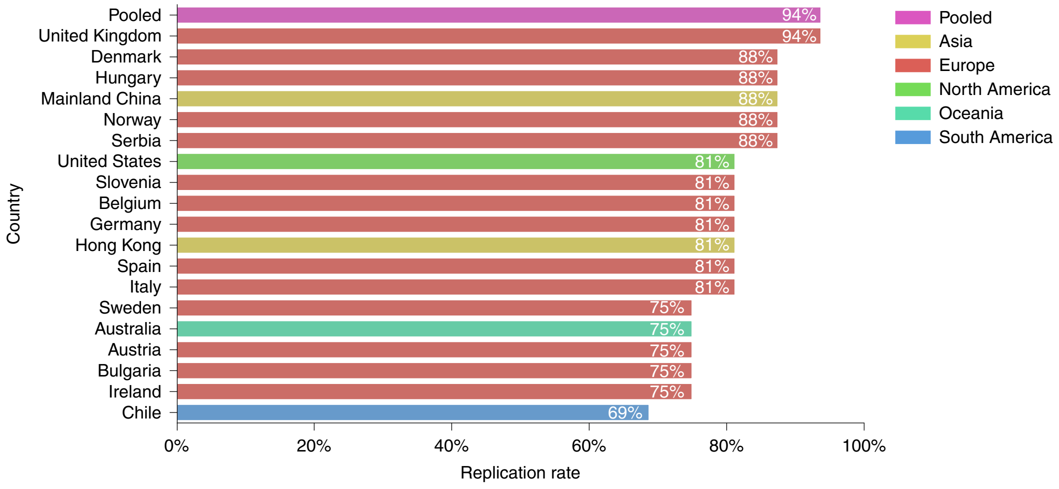


Fig. 2 | Item replication rates by country. A replication was deemed successful if the direction of the effect was the same as in the original study and significantly different from chance at an alpha threshold of 0.05. The sample size is 4,098.

section F5 for more details). Two of the three contrasts that tested the certainty effect and one (of one) contrast that tested the isolation effect did not show significant heterogeneity. These collectively suggest that these effects may not vary systematically between regions (at least for those countries evaluated here).

For the unpooled data, the contrasts replicated 89% of the time. As with the item-specific effects, there was a general attenuation of the contrast effects in the replication relative to the original study. The strength of this attenuation differed between contrast pairs (Fig. 3).

Most countries tested replicated at least 90% of the contrast effects, and the lowest replication rate recorded in any country was 77% (Fig. 4). As with the item-specific analyses, the replication rates varied more between contrasts than between countries.

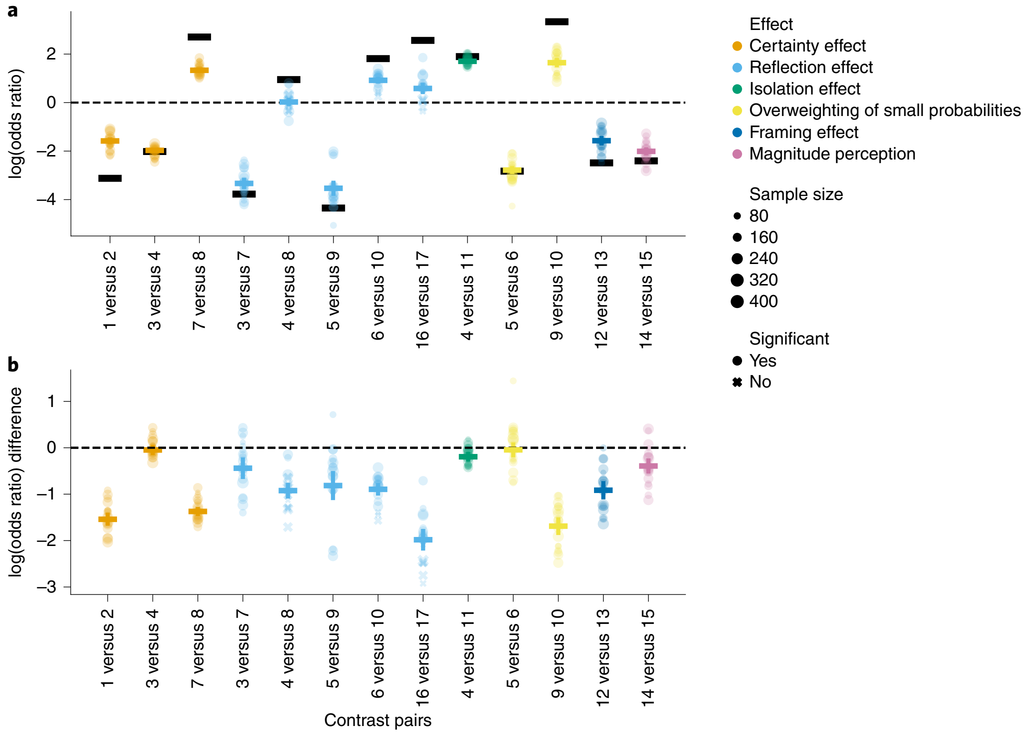


Fig. 3 | Effect sizes by contrast. a, Comparisons between the effects in Kahneman and Tversky in 1979 (black bars) and the current samples. A replication was deemed successful if the direction of the effect was the same as in the original study and significantly different from chance at an alpha threshold of 0.05. The dotted line represents the null hypothesis that participants show no difference in preferences for items in the contrast pair.**b**, The change in effect size compared with Kahneman and Tversky (1979). The dashed line represents the effect reported by Kahneman and Tversky. Values above the line suggest that participants in the current sample show stronger contrast effects between the items than in the original study, whereas values below the line suggest more similar choices between the contrasting items. There is an attenuation in effect sizes compared with 1979 for most contrasts. In both panels, the sample size is 4,098 for all contrasts, except 14 versus 15, where the sample size is 3,874. Each circle or × represents one country; the coloured bars are the pooled results (with error bars representing the 95% CIs).

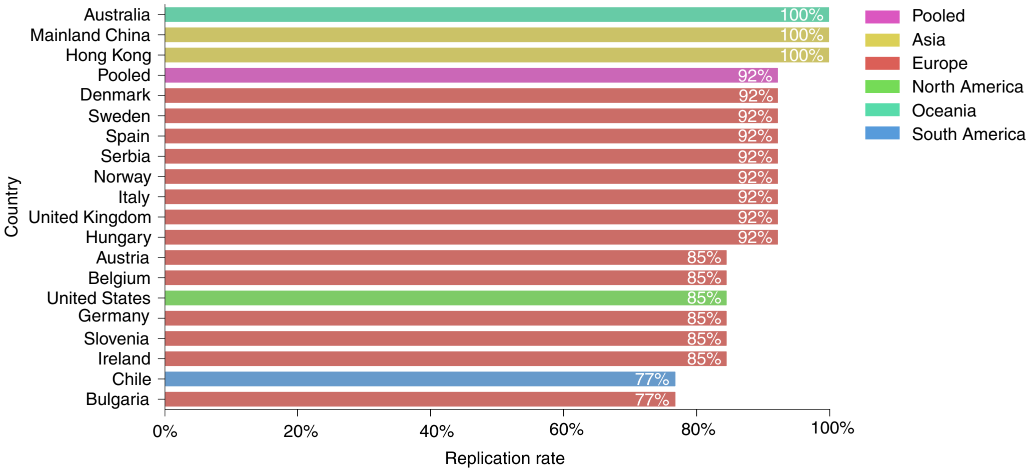


Fig. 4 | Contrast pair replication rates by country. A replication was deemed successful if the direction of the effect was the same as in the original study and significantly different from chance at an alpha threshold of 0.05. There was more variation in replication rates between contrasts than between countries. The sample size is 4,098.

Ten of 13 contrasts replicated consistently across all countries. All exceptions involved the reflection effect: the contrast between items 6 and 10 replicated in 84% of the countries, the contrast between items 16 and 17 replicated in 63% of the countries and the contrast between items 4 and 8 replicated in 16% of the countries. Though most effect sizes are attenuated compared with the original study, five out of six behavioural effects reported in 1979 consistently replicated across all 19 countries. The reflection effect had a combined replication rate of 73% across all items and countries.

Finally, we can evaluate prospect theory by studying choice patterns within respondents, as explained in the Methods. Choices that

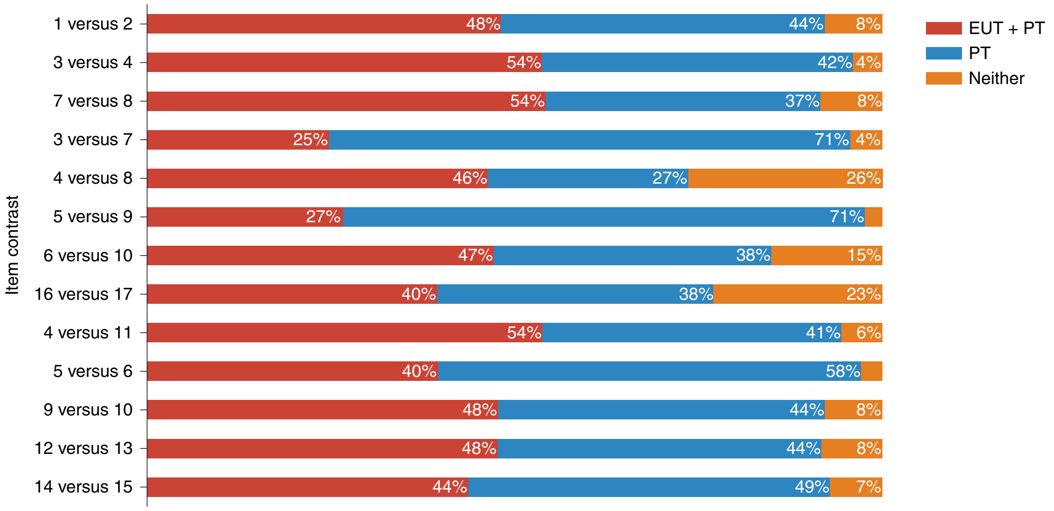


Fig. 5 | Choices congruent with prospect theory. This figure captures the proportions of choices that are congruent with expected utility theory and prospect theory, prospect theory only or neither theory. Expected utility theory can account for about 44% of the observed choices, whereas prospect theory can account for 91%. There is considerable variation between contrasts (the mean number of choices that could not be accounted for by either theory was 9%, but the standard deviation was 7.5%). However, prospect theory consistently provides a better explanatory framework for behaviour than expected utility theory in these cases. The sample size is 4,098 for all contrasts, except 14 versus 15, where the sample size is 3,874. EUT, expected utility theory; PT, prospect theory.

conform to expected utility theory also conform to some version of prospect theory, so we can code the choices for each item contrast as conforming to both theories, conforming only to prospect theory or conforming to neither (Fig. 5). We found that 91% of choices could be accounted for by prospect theory (but there was some variation between contrasts; s.d. = 7.5%), and 47% could be accounted for exclusively by prospect theory (s.d. = 12.95%). It is worth noting that the choices that go in the opposite direction of what is predicted by prospect theory are not uniformly distributed across contrasts, but are concentrated on the contrasts between 4 and 8, 16 and 17, and 6 and 10. All of these contrasts were intended to capture the reflection effect, again suggesting that the evidence for the reflection effect is weaker than that for the other behavioural effects.

To test whether the knowledge of loss aversion moderated any of the observed contrast effects, we used hierarchical logistic regression models to predict choices that could be explained only by prospect theory from whether people were aware of loss aversion or had correct intuitions about loss aversion. Between-country variation was accounted for by random intercepts. We found an effect of awareness of loss aversion for 2 out of 13 contrasts, both of which measured the reflection effect. For the contrast between items 5 and 9, people who were aware of loss aversion were slightly less likely to make choices that conformed to prospect theory (coefficient = −0.28(0.12), $z = -2.43$, $P = 0.01$). The same pattern applied for the contrast between items 16 and 17 (coefficient = −0.22(0.11), $z = -2.02$, $P = 0.04$). Intuitions about loss aversion had no significant effect for either contrast. Additionally, for the contrast between items 12 and 13, people who intuited loss aversion were slightly less likely to make choices that conformed to prospect theory (coefficient = −0.20(0.10), $z = -1.99$, $P = 0.05$). Actual awareness of loss aversion had an effect in the same direction, but this was not significant. Because of the high statistical power of these analyses, it seems as though the awareness of loss aversion has a very modest effect on choice, to the extent that such an effect exists at all. To see whether these effects may be present in specific countries, we present exploratory analyses in the Supplementary Results, section H.

One pattern that has consistently been ignored in research on decision-making under risk is the minority group outcome. Where a clear majority may consistently choose the certain, lower gain, there are always some that choose the risky option or accept a certain loss of lower value. We will analyse those patterns in future work while also making the data available, testing whether those individuals differ systematically from other participants. Still, it is critical to highlight this here to suggest that those deviations may be the archetype exception that proves the rule, and should be analysed in their own right. Given the attenuation we find, this is particularly meaningful in applications to public policy: while certain behaviours predicted by prospect theory may indeed represent modal decisions, the fact that they are not universal means that policymakers should also consider non-typical choice profiles (such as being risk seeking in all contexts).

The recent large-scale study testing decision paradoxes within individuals[26] concludes that the replication rate of prospect theory is poor, while our results indicate a general replication across items. We believe that any discrepancies are mostly illusory. Like us, Millroth et al. find attenuation effects for most items, as well as qualitatively similar patterns of effects as the 1979 original. Notably, they report non-significant effects for three items (2, 15 and 16; our numbering) where we find reliable effects. For these three items, Millroth et al. report effects substantially smaller than our pooled results, but within our observed between-country variation.

Given that the differences in effects fall within our own observed between-country variation, we do not want to overinterpret them, but one potential reason is that Millroth et al. suggest numeracy as a key moderator. Nearly 75% of their sample exhibits low numeracy, meaning that only a limited number of participants have high levels of numeracy. While we do not measure numeracy directly, our samples involve a wide range of educational attainment. As no fewer than 41% of participants within a country have university-level education in our study, it is likely that our samples have comparatively higher numeracy. Furthermore, as Millroth et al. only tested a single country, our disaggregated descriptive modelling based on multisite collection better demonstrates dispersion in the findings[25], highlighting the merits of the theory but also challenging the universality of application.

One point of agreement involves the distinction between population-level patterns and intra-individual patterns. Millroth et al. also counted the total number of prospect-theory-congruent but expected-utility-theory-incongruent choices per person, which they call prospect theory paradoxes. The mode in their sample was two prospect theory paradoxes per person. We run similar analyses in the Supplementary Results, section F4, and find more prospect-theory-congruent choices in our data (mode = 5).

Deviations from the preregistered analysis plan. Our final analyses deviate in multiple ways from our preregistered analysis plan. In the interest of transparency, we outline these differences here. The primary reason for this is that our thinking has matured since the preregistration, particularly including meta-analyses as a response to reviewer requests. Additionally, some of our planned analyses involved studying whether the contrast effects were moderated by demographic variables. In the end, we opted against these, as prior analyses suggested that the demographic variables had very limited predictive impacts on choice, and that the planned models were suboptimal to answer the questions of interest. Instead, we opted for a simpler analytic approach to establish the effect sizes of the contrasts, complemented with visualizations. The analyses that were originally planned but not included in the manuscript can be found in the Supplementary Results, section D, such as effects if using only the original sample sizes, bootstrapped analyses to assess variability and noise, and Bayesian models of demographic predictors of choice.

Additional analyses not part of the original plan include the chi-squared tests of contrast pairs, pooled analyses for items and contrast pairs, and assessing choice patterns within respondents. The last deviation was directly related to the recent study[26] published between completing the preregistration and completing the study. In it, the authors looked at intra-individual patterns of choices, rather than group-level patterns (which have been the norm). Applied at the intra-individual level, they find patterns other than what would be assumed under prospect theory, at least in terms of consistency. We wanted to briefly explore what these patterns looked like in our data, to make it easier for interested readers to compare the results and conclusions of the two studies.

Limitations. First, it is important to highlight that this study is testing whether the original 1979 work by Kahneman and Tversky replicated in a modern sample from multiple countries. This is in itself a critical goal, but does not address all criticisms linked to prospect theory or the effects that it seeks to explain, such as loss aversion. Replication is a critical first step in this process, in the sense that though a successful replication would not neutralize all possible criticisms, a failed replication would have suggested serious issues with the original theory. These findings cannot be interpreted as tantamount to saying that prospect theory is a fact, but the findings are generally consistent with the original conclusions.

We are also fully aware that arguments challenging prospect theory did not necessarily or categorically challenge the original method, but rather focus on context, interpretations and conclusions[32]. We also cannot assess whether one theoretical framework (for example, status quo bias) supersedes or dominates those associated with prospect theory (such as loss aversion), as has been argued. This could only be done experimentally. However, we do replicate features from the 1979 study such as reflection and framing effects, which provides insight to value and weighting functions, irrespective of loss aversion.

By randomizing item order and using an online survey rather than a printed booklet (as in the 1979 methods) for all participants, we were less concerned about the potential impact of order effects and more concerned about potential fatigue. Even so, there were no indications of such effects, as option A was selected around 45% of the time for all 17 items, regardless of which one was presented (Supplementary Results, section F2). There was no indication of measurement fatigue over the course of the survey, as the final items were not significantly different from those presented earlier. Sequential effects of contrast pair items will be tested in a separate study, but there is no indication that these have influenced the conclusions in our study.

The magnitudes of choice proportions vary somewhat between the direct and paid samples (Supplementary Results, section F3). These variations ultimately have little impact on the conclusions of the study. For items 4, 8 and 11, the paid samples show the opposite patterns (and therefore the opposite contrast for items 4 and 8) of what is predicted by prospect theory. The contrast between 16 and 17 is not significant in the paid sample. Both of these contrasts capture the reflection effect, in line with our finding that the reflection effect is less robust than the other contrast effects. However, looking at the direct sample alone, all of the 13 contrasts from the 1979 paper replicate. Decomposing the variance in responding caused by sample type from the variance caused by country is not trivial because we had different numbers of direct and paid participants in different countries (and some countries without one of the groups). Exploring these differences more carefully as well as testing potential explanations will be the focus of a follow-up paper.

We also note that the attention check itself may have resulted in a larger loss of sample than is indicative of actual problematic participants. The modification made in the US sample showed a slight improvement in passing rates on Prolific, meaning that future versions could be slightly less strict with this feature.

While we cannot compare the age distribution with that of the 1979 sample, we note that our sample is skewed towards a younger population. This might be one limitation on generalizability, given that risk aversion and other factors including computer literacy might vary systematically across age groups.

Finally, we acknowledge that a limitation of our study is that the power calculations were based on the effect sizes found in a single published study, which not is not necessarily an ideal approach[33]. Even though we selected the smallest effect size among those reported in the 1979 trial, this is just an estimate of the population effect. This could also be problematic in that the effect in the original study (which would probably be considered underpowered by current standards) may have been overestimated for any number of reasons. If this were the case, then our true statistical power would be lower. Therefore, we recommend that future replication attempts address this by using our meta-analytic effect sizes for power calculations (unless they focus on a specific setting included here), which are probably more accurate than the individual effects within countries or from the original study.

Discussion

With over 4,000 participants from 19 countries, we find that Kahneman and Tversky's 1979 findings replicate in the vast majority of analyses. To the extent possible, we used identical methods to those presented originally, modifying them only to make currency values relevant for a 2019 sample within each country. In doing so, we find a total replication of over 80% for individual analyses directly mirroring Kahneman and Tversky's methods. We also find 90% replication for directly testing the theoretical contrasts that were at the heart of their argument. Within all items and all contrast pairs testing specific theories, we find near-universal clustering of country results in the direction suggested by prospect theory. The replication rates are over 70% for both item-based and contrast analyses for all countries (except Chile, which had an item-based replication rate of 69%). In short, we find nothing to indicate failure to replicate or any fundamental flaws in the theory.

As would be expected with such a large sample (relative to the original), we did find evidence for the attenuation of some effects. In total, 77% of the effect sizes in our study were smaller than those reported in the original study. A third of our sample reported being aware of loss aversion, and an additional 50% who were unaware of loss aversion had an intuition that it was true. Though we lack data on awareness in 1979, it is plausible that the current numbers are higher given the popularity of behavioural economics. However, to the extent that we found an effect of the awareness of loss aversion on choice, it was weak, so it cannot fully explain the observed attenuation. We also note that some attenuation may be the result of methodological differences, such as lab-based versus internet-based data collection and the fact that the exact phrasing of the items is not always clear from the original paper. Both China and Hong Kong showed less of an attenuation effect than most other countries and showed a significant effect for the item 4 and item 8 contrasts. We currently do not know why this is the case, but it might be an avenue for further research. Combined, our findings provide further support for the effects of certainty, reflection, isolation, framing, magnitude perception and overweighting of small probabilities.

Attenuation in this case is perhaps a more relevant insight for reproducibility in general: since we regard the smallest effect from the original study as the benchmark for power calculations, we have a much larger sample than would be necessary for items or pairs with larger effects. As Kahneman and Tversky based their theory not on specific effect sizes but on distinct patterns of choice, we do not consider attenuation to be a major concern; instead, we consider it a meaningful insight for policy, as described under 'Limitations'. Simply put, the smaller effects mean that policy should be fully considerate of both the modal behavioural patterns and the non-trivial minority behavioural patterns.

While we do not suggest that our sample perfectly represents a global population, all analyses were sufficiently powered and show no indications of systematic bias that would undermine the findings. Any adjustments to the method from the preregistration have been highlighted explicitly in the Supplementary Results (see section E), all of which are relatively common in the course of conducting research. The unpaid or direct sampling approach produced enough participants in most countries that we did not need to include paid participants in all countries. The only procedural change was to collect data in the United States last; in the event that any flaws were found in other surveys, this was the easiest setting (language, accessibility of participants) for correcting, but this ended up not being necessary. Beyond this, there are no substantive changes to the procedure, thresholds or data collection, and changes from the preregistered analysis plan are mentioned alongside each analysis reported above.

Overall, our results are generally in line with the findings of the original study. For the pooled analyses, 2 out of 17 items (items 4 and 8) did not have response patterns that were significantly different from chance, one of which was not significant in the original study. The other was the reflected version of this same item (that is, the items had the same magnitudes and probabilities, but one item was in the gain domain and the other in the loss domain). As for specific theoretical contrast pairs, all but one replicated in the pooled sample. This was the contrast pair that tested the reflection effect for the two items that were not significantly different from chance. Because the other reflection contrasts did replicate, we interpret this not as a failure of the reflection effect but rather as a peculiarity of the items. Intuitively, if people are indifferent between the options in the gain frame, they are also indifferent in the loss frame, as there is no preference to reflect.

The results for the individual items and contrast pairs indicate that both the value and the weighting functions could be adequately approximated, resulting in the same conclusions as in the original article. In spite of some disparate results not indicating full replication, the threefold analysis process (individual items, contrast effects and comparison between prospect theory and expected utility theory) confirms that the main findings of the original study replicate on a general level. Hence, the effects reported in 1979 still remain a robust and widely applicable descriptive model of decision-making under risk and uncertainty.

We came into this study unbiased and without vested interests in the results of the trial. While we acknowledge prior commitment to this field of work and use of the theory under question, critiques of the 1979 study and concerns about reproducibility provided sufficient impetus to directly test long-held notions. Such challenges are critical for ensuring scientific quality, no matter how widely accepted the conclusions may be.

In the end, we find that Kahneman and Tversky's 1979 empirical foundation for proposing prospect theory broadly replicates. Some effects were less strong than in 1979, but this may be more a testament to the ease of accessing participants in 2019, rather than suggesting a flaw in the original study conclusions. Nothing in our results would indicate theoretical constructs (framing effects, overweighting of small probabilities and so on) from the original study to be unreliable. Rather, our results seemingly uphold those conclusions and, by default, some principles of loss aversion. Therefore, we consider this study compelling evidence for continuing to consider prospect theory as a viable explanation of individual behaviour, and therefore valuable for informing public policy around the world, in areas from financial decision-making to population well-being.

Methods

General summary. This trial involved—as closely as possible—the direct replication of the items used in the original paper on prospect theory[1]. We test original conclusions with contemporary analytical approaches through a combination of descriptive and inferential analyses, using the original outcomes as a reference point. Additional items were introduced to account for various demographic factors as well as knowledge of the hypothesized effects. Ethical approval was provided by the Centre for Business Research in the Judge Business School at the University of Cambridge, which was the staging location for the collaboration. All participants provided informed consent before beginning the study, which included being informed about the study and their rights as participants.

Participants. The participants were recruited from 19 countries, covering 13 languages. There was no systematic method for language or location inclusion beyond the collaborators that volunteered to participate. While there is a noted skew towards Euro-American regions, the generally random nature of inclusion is helpful for avoiding some level of systematic bias for participants. All data collection emanated from a single institutional account, with the data collected exclusively online; no in situ testing took place. The details on the other countries and languages considered are in the Supplementary Methods, section C.

There were two tracks for recruiting participants for the study. The first was direct contact with convenience samples for general testing of the procedure in a similar approach as in the original study, followed by participants recruited through Prolific, a paid online platform. For this study, we use 'direct sample' to refer to anyone not recruited in a paid sample. We use this robust approach intentionally to form insights about prospect theory specifically as well as about reproducibility through different platforms more generally.

In practice, direct sample participants were recruited through convenience samples, direct contact, online forums, social media posts, email circulars and various organizational membership channels.

Country-specific direct circulation, which was intentionally varied to buffer against bias, is further outlined in the Supplementary Methods, section C. This generally follows the replication approach used in the Many Labs trials[30,34], noting that we intentionally did not utilize psychology student participant pools[35]. It was important to have directly recruited participants to be similar to the 1979 study. Therefore, for each country, all project members targeted a minimum of 73 participants through direct collection, which was larger than the smallest sample (64) in the 1979 study and also in line with the smallest laboratory sample in the major 2014 multinational replication trial[30]. However, according to power calculations, the actual minimum for sufficient power for a chi-squared test to detect the smallest contrast reported by Kahneman and Tversky was 120 participants ('Power and error') and we also wanted to be over the highest sample size from the original study (141). Therefore, 218 was set as our ideal overall target, to be reached through a combination of direct recruitment (for at least 73 participants) and paid recruitment (for an additional 145 participants). This approach therefore also made it possible to assess any systemic differences between direct and paid samples.

Countries that exceeded the upper-bound desired threshold level of 218 through direct sampling (before exclusions) did not use paid samples. Though

Table 3 | Base version of the survey with 1979 values

2019	1979	Items	Response alternatives
1	1	Which option do you prefer?	A 33% chance at 2,500, a 66% chance at 2,400, and a 1% chance of 0 Guaranteed 2,400
2	2		A 33% chance of 2,500 (67% chance of 0) A 34% chance of 2,400 (66% chance of 0)
3	3		An 80% chance of 4,000 (20% chance of 0) 100% guarantee of 3,000
4	4		A 20% chance of 4,000 (80% chance of 0) 25% chance of 3,000 (75% chance of 0)
5	7		A 45% chance of 6,000 (55% chance of 0) 90% chance of 3,000 (10% chance of 0)
6	8		A 0.1% chance of 6,000 (99.9% chance of 0) 0.2% chance of 3,000 (99.8% chance of 0)
7	3′		An 80% chance of losing 4,000 (20% chance of losing 0) A 100% guarantee of losing 3,000
8	4′		A 20% chance of losing 4,000 (80% chance of losing 0) A 25% chance of losing 3,000 (75% chance of losing 0)
9	7′		A 45% chance of losing 6,000 (55% chance of losing 0) A 90% chance of losing 3,000 (10% chance of losing 0)
10	8′		A 0.1% chance of losing 6,000 (A 99.9% chance of losing 0) A 0.2% chance of losing 3,000 (A 99.8% chance of losing 0)
11	10	Imagine you are playing a game with two levels, but you have to make a choice about the second level before you know the outcome of the first. At the first level, there is a 75% chance that the game will end without you winning anything, and a 25% chance that you will advance to the second level. What would you choose in the second level?	An 80% chance of 4,000 (20% chance of 0) A 100% guarantee of 3,000
12	11	Imagine we gave you 1,000 right now to play a game. Which option would you prefer?	A 50% chance to gain an additional 1,000 (50% chance of gaining 0 beyond what you already have) A 100% guarantee of gaining an additional 500
13	12	Imagine we gave you 2,000 right now to play a game. Which option would you prefer?	A 50% chance you will lose 1,000 (50% chance of losing 0) A 100% chance you will lose 500
14	13	Which option do you prefer?	A 25% chance of 6,000 (75% chance of 0) A 25% chance of 4,000 (25% chance of 2,000, 50% chance of 0)
15	13′		A 25% chance of losing 6,000 (75% chance of losing nothing) A 25% chance of losing 4,000 (25% chance of 2,000, 50% chance of 0)
16	14		A 0.1% chance at 5,000 (99.9% chance of 0) A 100% guarantee of 5
17	14′		A 0.1% chance of losing 5,000 (99.9% chance of losing nothing) A 100% guarantee of losing 5

Most countries included 'gaining' for several items, which was necessary to distinguish from 'losing'. Items from 1979 with a prime symbol indicate a loss frame.

some participants were excluded later, these countries still exceeded the number targeted after exclusions. The paid sample was recruited via Prolific, which was why Chile was chosen for South America, as it was the only country in the platform pool for the continent. All participants received the equivalent of the minimum hourly wage for their country, prorated for the estimated time to complete the survey. This ranged from £0.70 to £1.30.

Instrument. To closely replicate the procedure, the same items as published in Kahneman and Tversky[1] were used, excluding the two verbal travel items and the verbal probabilistic insurance item (5, 6 and 9 in the original publication). The original travel items entailed a choice between having a chance to travel to England, France and Italy, and having a certain trip to England. Their subjective utility might differ markedly between countries, which adds a needless interpretive burden to this multicountry replication. Additionally, given the events of recent years, it is unclear whether a certain trip to England would reflect a positive utility to all participants. The probabilistic insurance item was excluded because the wording of the item was long and complex, creating concern that it more likely tests reading comprehension rather than theoretically relevant constructs. For posterity, we provide extensive detail on the method in the Supplementary Information, particularly for aspects not explicitly presented in the 1979 paper.

The financial values in each item were adjusted directly towards the median net household income in each location. The original study reported that the median net household income for Israel was about 3,000 Israeli pounds per month. Where possible, the median net household income in June 2019 was used as the relative value for each country in the replication. For example, an item that was 2,000 Israeli pounds in 1979 was 2/3 of the 3,000 reference value. For 2019 in the United States, the median net household income was about US$6,000 per month, so the same item would use US$4,000 for US participants in the replication. To put this explicitly, had the numerical values from 1979 been reused for dollars (that is, a reference of US$3,000), this would have meant that the values were worth about half what they were in the original study.

As some governments report mean income as the standard for the national average, the mean income was used in eight countries. The reference values were rounded to the nearest clean number to reflect the 1979 approach and to reduce complexity. All within-item prospects retained the same relative values as in the 1979 instrument. The details on this are included in the Supplementary Methods, sections B and C. This approach was ultimately decided on the basis of its being more important to expose participants to choices representing the same wealth as the original study, given that the specific numbers from the original study hold no theoretical value. Coincidentally, we were able to address this concern as Ireland, Austria, Germany and Belgium each had 3,000 as the reference value (Supplementary Results, section F1).

All items involve hypothetical monies only, in line with the original study. The lead and senior author have recently completed a multicountry study showing that the answers do not change substantially between hypothetical items and those involving real money[36], which has also been shown in other work[37–39] and reduces the need to validate with consequential choices.

Nine demographic measures were presented after the decision-making items to avoid stereotype threat influences: nationality, year of birth, gender, income, educational attainment and four measures of current financial circumstances and behaviours (strain, recent changes, investments and debts). We did not anticipate substantial differences between any groups, only moderate levels of variability.

An attention check item was included as the sixth item (preceding and following items were all randomized). This item gave the simple instruction

'Do not choose either option, just proceed to the next question.' Two options were presented, either of which, if answered, immediately excluded the participant by ending the survey. The options were between a guaranteed gain of 10,000 and a 99% chance of losing 5,000, which means that participants that were truly reading the options should immediately notice an obvious departure from the other items. During the early sampling stages, a minor flaw was identified in some versions of the survey for items 14 and 15. Those data were also excluded from the analyses, and the items were corrected before extensive sampling.

The full set of choice items is shown in Table 3, presenting the original problem numbers in the 1979 study and the corresponding item numbers used in the current study. Forward and back translation was used for all measures, with adjustments to the local currencies. The details of any specific issues within countries are also included in the Supplementary Methods, section C. None were deemed substantive enough to warrant highlighting here or requiring any parallel analyses to assess effect. The full set of all items in all languages is available with the preregistered material at osf.io/esxc4/.

Though the theoretical contrast pairs for constructs such as certainty effects, reflection effects and framing effects were not formally reported in 1979, the information in the paper makes it simple to compute odds ratios for all of the contrasts in question, both in aggregate and within individuals. This enabled us to use the same replication criterion as for the single item for the contrasts, namely by testing for significant effects (all countries pooled at 0.001; unpooled individual countries at 0.05) in the same direction as in the original study.

Procedure. All participants in each country completed identical surveys of 27 total items (US participants answered an additional item at the end of the survey on financial strain), including demographics. After providing informed consent, the participants responded to 17 choices under risk from the original study. The only difference in presentation relative to the original study was the language of the surveys (adapted to each country) and the monetary amounts used (adjusted to local purchasing power). The orders of the choice items were randomized. The 1979 paper presented the items in a pseudorandomized order to each participant, but the printed nature of the surveys limited them to a few different presentation orders, and no participants encountered all of the items. It is not clear from the original manuscript which items belonged to the same survey. In this replication, all participants encountered all items, but the randomized order should limit any confounds related to order.

At the end of the survey, the participants were asked whether they were familiar with the concept of loss aversion, as a proxy for general awareness of behavioural economics. This measure was not central to any hypothesis, but was included as a potential indicator in the event of systematic failures to replicate the 1979 results. All participants were tested over a 15-day window in July–August 2019; additional details on this timeline are included in the Supplementary Methods, section C.

Power and error. Given the likely heterogeneity in recruiting participants across multiple locations, we used three participant thresholds for each country: 73 (direct sample), 120 (country minimum) and 145 (target). We anticipated that getting direct sample participants would be difficult, but did not want to rely on an entirely paid sample with probably common participant pool demographics. The thresholds were based on the preregistered power calculation (https://osf.io/wd4k5), which indicated that we would need 120 participants to have 95% power to detect the smallest original contrast effect with an alpha threshold of 0.05.

Though the original paper ostensibly reports only chi-squared tests that compare the response distribution of each item with a balanced null distribution, Kahneman and Tversky's theoretical argument primarily relies on contrasting pairs ('theoretical contrast pairs') of response distributions (Supplementary Methods, section A). The smallest of these reported contrasts are between items 4 and 8, where 65% and 42% chose option A, respectively, which gives an odds ratio of 2.56. The smallest effect size is used as it requires the largest sample to validate at the highest level of power. Thus, to account for dropouts and exclusions, our working target was 145 participants per country via paid platforms, which is larger than the largest sample size reported in the original study (141) and allows for dropouts while remaining above the 120 threshold.

To more closely reflect the original study methods and to avoid relying only on paid participants, we targeted a minimum of 73 participants through direct sampling for each country to meet standards in replication[30] while being above the original study minimum. To maximize power and have the potential to address possible differences between samples, we used a combined ideal aim of 218 participants per country, in which 73 were from direct samples and 145 were paid. In this way, all countries and the total pool would be powered sufficiently beyond the minimum 120 necessary, even when applying conservative exclusion criteria. We met this criterion for all countries but one (Austria), which had a final sample size of 111. As direct sampling yielded substantially larger participation than had been anticipated in some countries, paid samples were sought in only 13 of the 19 countries.

Our approach ensured that the sample size for each country gave sufficient power for testing within locations as well as in composite. Because we aimed to collect data from a minimum of 15 countries, the total target sample was set at 2,910, which would have given us a power approaching 1 to detect the smallest anticipated effects at an alpha level of 0.0001.

All sample size calculations and power calculations are based on the bpower function in the Hmisc (v.4.2-0) package[40] in R. This also matched the approach presented by Many Labs[30] in replicating multiple psychological studies (one of which included gain–loss framing items published by Tversky and Kahneman in 1981 (ref. [41])). However, where Many Labs found that partial method replication attempts were not impacted by setting[42], we tested a single method comprehensively between locations (though such approaches are still likely to result in heterogeneity in replications[43]).

We first test whether the effects are similar to those in the 1979 trial across items by looking for significant deviations in the directions of effects in the replication study from the original findings. We assess this for all countries and groups, as well as in aggregate. Such descriptive approaches yield a large number of outputs, and we expected any substantial differences between the original trial and the replication to be spurious, with general clustering in the same direction as the 1979 findings. As such, we began with the assumption that if fewer than 5% of outcomes tested were in the opposite direction from anticipated, these would be assumed as noise in the form of type S error. As our sample would be substantially larger than in the original study, it was certainly likely to detect some evidence of decline effects, giving the overall impression of attenuation. However, we did not anticipate extreme declines as have been noted in other major replication attempts in the social sciences[44].

We note that our emphasis primarily considers type 1 and type S errors. We report, but do not focus heavily on, what could be considered type M errors (that is, the factor by which a statistically significant effect size overestimates the plausible effect size)[45]. These errors are presented in the analyses on attenuation and sample size.

Our data also allow us to evaluate the original theoretical argument by looking at individual choice patterns. For every theoretical contrast pair, choices conform to both expected utility theory and prospect theory, prospect theory only or neither. The original item set tested in 1979 was selected because expected utility theory predicted that choice for one item would perfectly predict preferences in the second items. Prospect theory, however, can account for violations of expected utility theory in one direction (for example, overweighting small probabilities) but not the other (for example, underweighting small probabilities). Prospect theory is a more general form of expected utility theory in that it adds a number of additional parameters to the traditional formalism, such as the weighting parameter that penalizes gains relative to losses. Depending on the values of these parameters, prospect theory may reduce to expected utility theory. Consequently, for each theoretical contrast pair, we can tally the proportion of choices that can be explained by expected utility theory, and compare it with the proportion of choices that can be explained by prospect theory as well as residual choices that can be explained by neither.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data are available with the preregistered material and code at osf.io/esxc4/.

Code availability

All code is available with the preregistered material and data at osf.io/esxc4/.

Received: 23 August 2019; Accepted: 17 April 2020;
Published online: 18 May 2020

References

1. Kahneman, D. & Tversky, A. Prospect theory: an analysis of decisions under risk. *Econometrica* **47**, 263–291 (1979).
2. Markowitz, H. Portfolio selection. *J. Financ.* **7**, 77–91 (1952).
3. Savage, L. J. *The Foundations of Statistics* (Wiley, 1954).
4. Barberis, N. C. Thirty years of prospect theory in economics: a review and assessment. *J. Econ. Perspect.* **27**, 173–196 (2013).
5. Altman, M. in *Behavioral Finance: Investors, Corporations, and Markets* Vol. 6 (eds Baker, H. K. & Nofsinger, J. R.) 191–209 (Wiley, 2010).
6. Odean, T. Are investors reluctant to realize their losses? *J. Financ.* **53**, 1775–1798 (1998).
7. Genesove, D. & Mayer, C. Loss aversion and seller behavior: evidence from the housing market. *Q. J. Econ.* **116**, 1233–1260 (2001).
8. Benartzi, S. & Thaler, R. H. Myopic loss aversion and the equity premium puzzle. *Q. J. Econ.* **110**, 73–92 (1995).
9. Johnson, E. J. et al. Can consumers make affordable care affordable? The value of choice architecture. *PLoS ONE* **8**, e81521 (2013).
10. Sydnor, J. (Over) insuring modest risks. *Am. Econ. J.* **2**, 177–199 (2010).
11. Levy, J. S. Loss aversion, framing, and bargaining: the implications of prospect theory for international conflict. *Int. Polit. Sci. Rev.* **17**, 179–195 (1996).
12. Mercer, J. Prospect theory and political science. *Annu. Rev. Polit. Sci.* **8**, 1–21 (2005).

13. Simonsohn, U. [15] Citing prospect theory. *Data Colada* http://datacolada.org/15 (2014).
14. Edwards, K. D. Prospect theory: a literature review. *Int. Rev. Financ. Anal.* **5**, 19–38 (1996).
15. Arkes, H. R. & Blumer, C. The psychology of sunk cost. *Organ. Behav. Hum. Decis. Process.* **35**, 124–140 (1985).
16. Uecker, W., Schepanski, A. & Shin, J. Toward a positive theory of information evaluation: relevant tests of competing models in a principal-agency setting. *Account. Rev.* **60**, 430–457 (1985).
17. Gregory, R. Interpreting measures of economic loss: evidence from contingent valuation and experimental studies. *J. Environ. Econ. Manage.* **13**, 325–337 (1986).
18. Loewenstein, G. F. Frames of mind in intertemporal choice. *Manage. Sci.* **34**, 200–214 (1988).
19. Newman, D. P. Prospect theory: implications for information evaluation. *Account. Organ. Soc.* **5**, 217–230 (1980).
20. Qualls, W. J. & Puto, C. P. Organizational climate and decision framing: an integrated approach to analyzing industrial buying decisions. *J. Mark. Res.* **26**, 179–192 (1989).
21. Tversky, A. & Kahneman, D. Advances in prospect theory: cumulative representation of uncertainty. *J. Risk Uncertain.* **5**, 297–323 (1992).
22. Diamond, W. D. The effect of probability and consequence levels on the focus of consumer judgments in risky situations. *J. Consum. Res.* **15**, 280–283 (1988).
23. Chang, O. H., Nichols, D. R. & Schultz, J. J. Taxpayer attitudes toward tax audit risk. *J. Econ. Psychol.* **8**, 299–309 (1987).
24. Payne, J. W., Laughhunn, D. J. & Crum, R. Multiattribute risky choice behavior: the editing of complex prospects. *Manage. Sci.* **30**, 1350–1361 (1984).
25. Kvarven, A., Strømland, E. & Johannesson, M. Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nat. Hum. Behav.* **4**, 423–434 (2020).
26. Millroth, P. et al. The decision paradoxes motivating prospect theory: the prevalence of the paradoxes increases with numerical ability. *Judgm. Decis. Mak.* **14**, 513–533 (2019).
27. *Behavioural Insights and Public Policy: Lessons from Around the World* (OECD, 2017); https://doi.org/10.1787/9789264270480-en
28. Thaler, R. H., & Sunstein, C. R. *Nudge: Improving Decisions about Health, Wealth, and Happiness* (Penguin, 2009).
29. McDermott, R. Prospect theory in political science: gains and losses from the first decade. *Polit. Psychol.* **25**, 289–312 (2004).
30. Klein, R. A. et al. Investigating variation in replicability. *Soc. Psychol.* **45**, 142–152 (2014).
31. Leys, C. et al. Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* **49**, 764–766 (2013).
32. Katsikopoulos, K. V. & Gigerenzer, G. One-reason decision-making: modeling violations of expected utility theory. *J. Risk Uncertain.* **37**, 35–56 (2008).
33. Simonsohn, U. Small telescopes: detectability and the evaluation of replication results. *Psychol. Sci.* **26**, 559–569 (2015).
34. Klein, R. A. et al. Many Labs 2: investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* **1**, 443–490 (2018).
35. Ebersole, C. R. et al. Many Labs 3: evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* **67**, 68–82 (2016).
36. Franklin, M., Folke, T. & Ruggeri, K. Optimising nudges and boosts for financial decisions under uncertainty. *Palgrave Commun.* **5**, 113 (2019).
37. Kühberger, A., Schulte-Mecklenbeck, M. & Perner, J. Framing decisions: hypothetical and real. *Organ. Behav. Hum. Decis. Process.* **89**, 1162–1175 (2002).
38. Beattie, J. & Loomes, G. The impact of incentives upon risky choice experiments. *J. Risk Uncertain.* **14**, 155–168 (1997).
39. Wiseman, D. B. & Levin, I. P. Comparing risky decision making under conditions of real and hypothetical consequences. *Organ. Behav. Hum. Decis. Process.* **66**, 241–250 (1996).
40. Harrell, F. E. Jr. Package 'Hmisc'. CRAN2018, 235-6 https://cran.r-project.org/package=Hmisc (CRAN, 2019).
41. Tversky, A. & Kahneman, D. The framing of decisions and the psychology of choice. *Science* **211**, 453–458 (1981).
42. Owens, B. Replication failures in psychology not due to differences in study populations. *Nature News* https://www.nature.com/articles/d41586-018-07474-y (19 November 2018).
43. Goldberg, M. & van der Linden, S. The importance of heterogeneity in large-scale replications. *J. Soc. Polit. Psychol.* **8**, 25–29 (2020).
44. Camerer, C. F. et al. Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637 (2018).
45. Gelman, A. & Carlin, J. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspect. Psychol. Sci.* **9**, 641–651 (2014).

Acknowledgements

We thank a number of colleagues and peers, including K. Kastelic, I. Sakelariev, T. Varkonyi, A. Víg, C. Saponaro, M. Frías and S. Deakin. We also thank Corpus Christi College Cambridge for support in hosting numerous researchers contributing to the study. We especially thank all team members from the Junior Researcher Programme. The authors received no specific funding for this work.

Author contributions

K.R. is the lead author and researcher responsible for all aspects of the manuscript. T.F. is a co-lead with primary responsibility for data management, analyses and visualization. S.A., M.L.B., G.B., L.D.B., A.C.-B., C.D., E.D., C.E.-S., M.F., S.P.G., H.J., R.K., P.R.K., J.K., T.L.A., I.S.L., L.M., A.E.N., J.P., S.K.Q., C.R., F.L.T., N.T., C.V.R., B.V., K.W. and A.Y. were part of the country-specific research teams who were responsible for data collection within each country, as well as country-specific supplementary details and general support of the writing. F.P., E.R. and S.v.d.L. were senior advisors on the study and provided input on the methods, analyses, writing and revisions.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41562-020-0886-x.

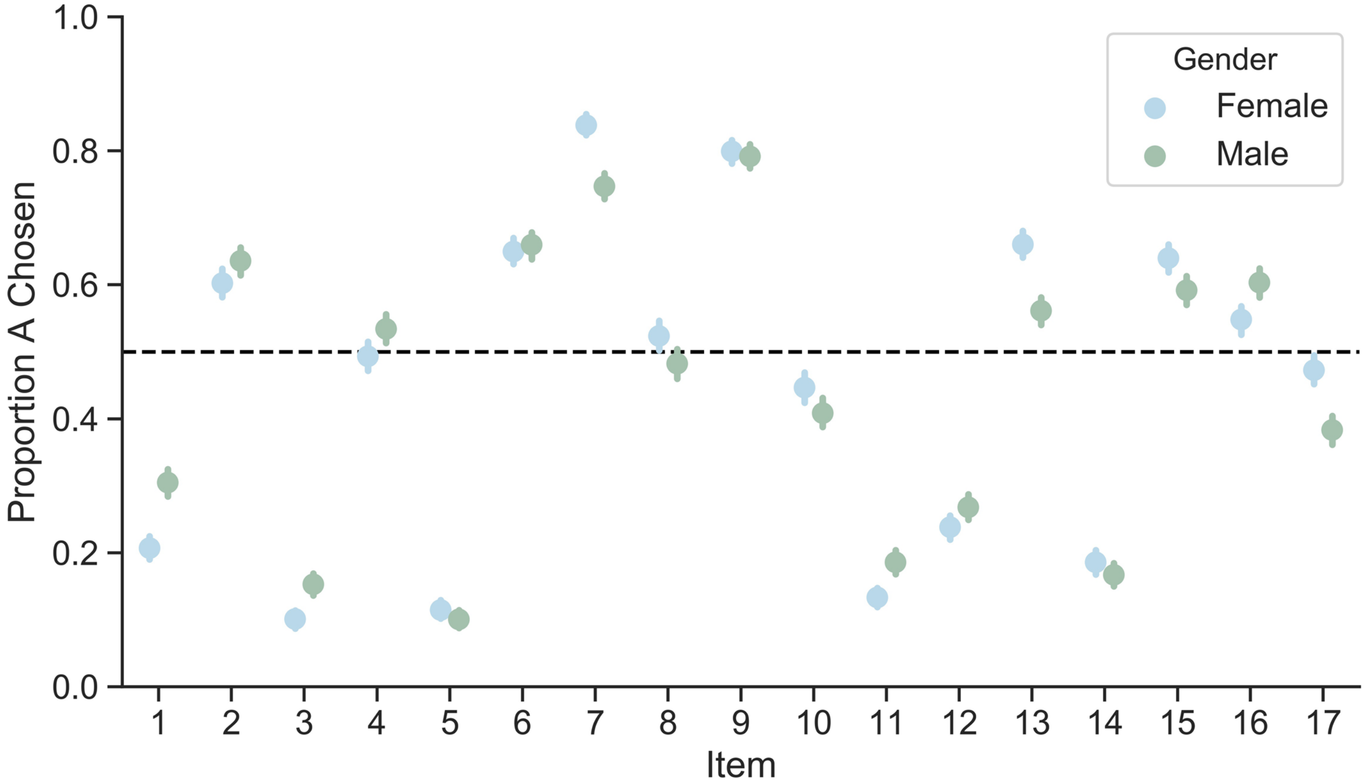
Supplementary information is available for this paper at https://doi.org/10.1038/s41562-020-0886-x.

Correspondence and requests for materials should be addressed to K.R. or T.F.

Peer review information Primary handling editor: Aisha Bradshaw.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | Choices by Gender. This figure captures the proportion of times participants chose option A as a function of their gender. Error-bars are bootstrapped 95% confidence intervals that respect the hierarchical structure of the data. There are clear gender differences for some items, but no general pattern. As this is the demographic variable with the most differences between groups, it is a meaningful indication of general consistency across the sample (that is, all other demographic indicators were even more similar).

naturereserach

Corresponding author(s): KAI RUGGERI, TOMAS FOLKE

Last updated by author(s): Mar 14, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

□ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

□ ☒ A description of all covariates tested

□ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

□ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

□ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

□ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

□ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All data were collected on the Qualtrics platform. Where specified, some participants were recruited via Prolific. Otherwise, all data were collected through circulation of the Qualtrics survey link.
Data analysis	All analyses were run in R statistical software and Python. Packages and versions are specified within the manuscript.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data are available with pre-registered material and code at osf.io/esxc4/

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

The study is a within participant survey design. Participants answered two alternative forced choice questions about their preferences about financial decisions under risk, and then provided demographic information (including information about their current financial circumstance). All data reported in the paper is quantitative, but some categorical demographic questions, had an "other" option that allowed participants to provide further information.

Research sample

Our final sample consisted of 4099 participants from 19 different countries covering 13 languages. For more information about each country sample see Table 3 in the main manuscript. Our samples are not perfectly representative of the general population in any given country (discussed in each country-specific appendix). Through online recruitment on a wide variety of platforms, we attempted to involve as a diverse sample as possible, because one of our main aims was to test the generality of the original prospect theory results.

Sampling strategy

We used a convenience sample, though we define this specifically in the manuscript. Of the total sample, 26% of participants were recruited through Prolific, with the remaining sample recruited directly. Because the central theoretical argument of the 1979 Prospect Theory paper rests on contrasts in choice preferences between paired items, we based our power calculations on the smallest of these contrast effects, which was an odds ratio of 2.56. We calculated that we required 120 participants to have 95% power to detect an effect of this size with an alpha threshold of .05. We therefore aimed to have a minimum of 120 participants per country, 18 out of 19 countries exceeded this sample size threshold (the exception was Austria which had a final sample of 111).

Data collection

All data were collected on the Qualtrics survey platform. No researchers were physically with participants when they completed the survey.

Timing

Data collection took place between the 23rd of July and the 4th of August 2019.

Data exclusions

We excluded 11 participants who were faster than three median absolute deviations (Leys, 2013) of the median completion time (86 seconds; The median completion time was 8 minutes). In the preregistration we had planned to apply this criterion symmetrically to slow participants as well. However, given that 488 people failed this criterion, and we could assess data quality through the attention check, the slow participants were retained. Three participants were excluded for reporting an income as “99999” as we suspect these might be members of the research team testing the survey. Six participants were excluded for reporting being billionaires, which brought into question the validity of their responses. One participant was excluded for reporting a negative income. We also excluded five participants who reported being over 110 years old. To minimise the risk of participants mindlessly clicking through the questionnaire, we excluded participants who both a) gave the same responses for more than 14 out of the 17 items and b) completed the survey faster than one median absolute deviation below the median (6 minutes). These criteria led to 42 additional exclusions, making the final total sample size 4099. The full annotated code used to clean and combine the data and make the exclusion will be made publicly available on the OSF platform.

Non-participation

Some participants did not complete the survey, which is mentioned briefly. However, as this is a single-setting, short survey, we do not track extensively details related to incomplete/withdrawal. The only exception for this relates to two countries where participants were sought - Kenya and Switzerland - where negative reactions to the survey, or simply a lack of response, occurred. These countries were then dropped from the full study.

Randomization

Participants were not allocated to experimental groups. All participants saw all items.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
□	☒	Human research participants
☒	□	Clinical data

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

We targeted adult participants in 19 countries: Australia, Austria, Belgium, Bulgaria, Chile, Denmark, Germany, Hong Kong, Hungary, Ireland, Italy, Mainland China, Norway, Serbia, Slovenia, Spain, Sweden, United Kingdom, USA.

Recruitment

There were two tracks for recruiting participants for the study. The first was direct contact with convenience samples for general testing of the procedure, followed by participants recruited through Prolific, a paid online platform. For this study, we use 'direct sample' to refer to anyone not recruited in a paid sample. We used this robust approach intentionally to form insights about Prospect Theory specifically as well as reproducibility through different platforms more generally.

More details about the recruitment strategies for specific countries, and comparisons of results between the two tracks are reported in the appendix.

Ethics oversight

This study received ethical approval from the Centre for Business Research at the University of Cambridge.

Note that full information on the approval of the study protocol must also be provided in the manuscript.