

## **Not lost in translation: Successfully replicating Prospect Theory in 19 countries**

**Kai Ruggeri**, Sonia Alí, Mari Louise Berge, Giulia Bertoldo, Ludvig D Bjørndal, Anna Cortijos-Bernabeu, Clair Davison, Emir Demić, Celia Esteban-Serna, Maja Friedemann, Shannon P Gibson, Hannes Jarke, Ralitsa Karakasheva, Peggah R Khorrami, Jakob Kveder, Thomas Lind Andersen, Ingvild S Lofthus, Lucy McGill, Ana E Nieto, Jacobo Pérez, Sahana K Quail, Charlotte Rutherford, Felice L Tavera, Nastja Tomat, Chiara Van Reyn, Bojana Večkalov, Keying Wang, Aleksandra Yosifova, Francesca Papa, Enrico Rubaltelli, Sander van der Linden, Tomas Folke\*

**Corresponding author:** kai.ruggeri@columbia.edu | Department of Health Policy and Management, Mailman School of Public Health, Columbia University

\*Data manager and co-lead: t.folke@cbr.cam.ac.uk | Policy Research Group, Centre for Business Research, Judge Business School, University of Cambridge

*With special thanks to the Junior Researcher Programme*

### **Abstract**

Kahneman and Tversky's 1979 article on Prospect Theory is one of the most influential papers across all of the behavioural sciences. The study tested a series of binary financial (risky) choices, ultimately concluding that judgments formed under uncertainty deviate significantly from those presumed by expected utility theory, which was the prevailing theoretical construct at the time. In the forty years since publication, this study has had a remarkable impact on science, policy, and other real-world applications. At the same time, a number of critiques have been raised about its conclusions and subsequent constructs that were founded on it, such as loss aversion. In an era where such presumed canonical theories have increasingly drawn scrutiny for inability to replicate, we attempted a multinational study of N = 4,099 participants from 19 countries and 13 languages. The same methods and procedures were used as in the original paper, adjusting only currencies to make them relative to current values, and requiring all participants to respond to all items. Overall, we found that results replicated for 94% of the 17 choice items tested. At most, results from the 1979 study were attenuated in our findings, which is most likely due to a more robust sample. Twelve of the 13 theoretical contrasts presented by Kahneman and Tversky also replicated, with a further 89% replication rate of the total contrasts possible when separating by location, up to 100% replication in some countries. We conclude that the principles of Prospect Theory replicate beyond any reasonable thresholds, and provide a number of important insights about replications, attenuation, and implications for the study of human decision-making at population-level.

## Introduction

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* (Kahneman & Tversky, 1979). The study was conducted with university faculty and student participants from Israel, Sweden, and the US (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 resulting patterns diverted substantially from the predictions of Expected Utility Theory, the dominant descriptive theory at the time.

Prior to widespread acceptance of Prospect Theory, decision-making approaches largely emphasized 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 optimise outcomes using stable algorithms tied to value, probability and cumulative wealth. Based on the findings in their study, Kahneman and Tversky concluded that certainty carries greater weight for gains, whereas risks are preferable when losses are likely, and that small probabilities are often over-weighted. They argued that utility is more likely dependent on the change in assets relative to a reference point, rather than the absolute outcome asset, as previously believed. These decisions also appear to be influenced by how information is framed, meaning that the same values presented in different ways can elicit different choices from the same individuals, and therefore different patterns of behaviours across populations.

Critically, they convincingly contended that people were largely loss averse: that most people find losing a certain amount more aversive than gaining the same amount is appetitive. Thus, if given the choice between maintaining current wealth and accepting a 50-50 gamble of gaining or losing an equivalent

amount of money, few people would take the gamble. Furthermore, the larger the sums involved, the fewer people would accept it. Yet, when given the choice between a certain loss, and a gamble that could lead to no loss or a greater loss, most would opt to take the gamble, seemingly to avoid the loss altogether.

Though single item chi-square tests were the only statistics reported in 1979, the bulk of the original argument focused on theoretical contrast pairs. Specifically, Kahneman and Tversky devised a way to test the prevailing theory of decision-making at the time, which would have predicted that choice preferences would be consistent across the items, but where their sample did not. There are 13 such contrasts among the items in the replication, and these contrasts can be mapped onto 6 distinct behavioural effects that Prospect Theory sought to explain: the certainty effect, the reflection effect, the framing effect, the isolation effect, overweighting of small probabilities, and range adaptation.

### *Impact*

The loss aversion component of Prospect Theory has provided an explanatory framework for understanding a broad range of behaviours (Barberis, 2013) in finance (Altman, 2010; Odean, 1998), investment (Genesove & Mayer, 2001; Benartzi & Thaler, 1995), insurance (Johnson et al., 2013; Sydnor, 2010), and political conflict (Levy, 1996). Extensions of the work have provided foundational insights to explain behaviour under riskless circumstances. Those include the endowment effect (Thaler, 1980; Tversky & Kahneman, 1981; Köszegi & Rabin, 2006), status quo biases (Kahneman et al., 1991), and the hedonic impact ratings assigned to losses compared to gains (McGraw et al., 2010).

It is difficult to overstate the level of influence that Prospect Theory has had on science, policy, management, financial services, government, and beyond. Shortly after Daniel Kahneman received the 2002 Nobel Prize for Economic Sciences, it was identified as the most influential theoretical framework

in all of the social sciences (Mercer, 2005). It has since become the most cited economic paper and is among the most-cited in psychological science (Simonsohn, 2014).

Forty years have passed since the initial manuscript was published, and conclusions of the original study are increasingly used to understand advances in behavioural sciences as well as in 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 popularised behavioural economics and revolutionised approaches to policy in organisations and governments around the world (Thaler & Sunstein, 2009). We propose that a major replication of the original study is critical to determine what outcomes appear if tested again today. Given the widespread application of behavioural sciences in government institutions around the world (OECD, 2017), this replication should involve participants from a number of countries to determine if core tenets are broadly applicable, considering current, local, and relevant income and financial standards.

#### *Debates and basis for replication*

As should happen in science, with notoriety and acceptance has come a number of critiques of Prospect Theory. For example, Kahneman and Tversky themselves realised that Prospect Theory struggles with prospects that have many possible outcomes, and only deals with risky choices (where probabilities are known) as opposed to uncertain ones (where they are not). Consequently, they updated their theory to deal with these shortcomings (Tversky & Kahneman, 1992).

A recent criticism levied against one of the primary tenets from the 1979 study – loss aversion – claims that more rigorous evaluation of its findings (and those borne from it) portend the *loss of loss aversion*. In a narrative, non-systematic review, Gal and Rucker (2018) claim that current evidence does not support the notion that losses are more impactful than gains are, that the theory merely remains in strong support through persistent beliefs among scientists, and that most conclusions present

“inferential concerns” that can be reimagined (Gal & Rucker, 2018). For example, that people sell stocks that had early gains too quickly and hold longer those that lose value slowly. They contend this is not because of the disposition effect (based on loss aversion) as argued by Odean (1998), but simply because of perceived mean reversion on the part of the trader. While those arguments were largely refuted in scientific and mainstream literature (Ritholtz, 2018), and multiple global trials have generally confirmed normative patterns within countries (Rieger et al., 2011; Wang et al., 2017; Xie et al., 2018) relevant debates on loss aversion raised in behavioural science literature continue.

Given the time that has passed since the original study, its position among the behavioural science canon, its widespread influence in science and policy (McDermott, 2004), and relevant critiques of its methods and conclusions, Prospect Theory deserves an unbiased reassessment. This is particularly critical in light of concerns about replicability in behavioural science, as well as the translatability of findings between locations (Klein et al., 2014), ahead of attempting to resolve any current debate. There is also no question that methods have at times been designed with bias to aid researchers in obtaining the results and conclusions they desired (Ioannidis, 2012; Freedman, 2011). It has been convincingly argued that direct replication by laboratories independent from the original study is the best way to provide evidence for the robustness of psychological effects (Simons, 2014).

Since multiple theories that were presumed canon have failed to convincingly replicate (Klein et al., 2018), it is important to test those well-known beliefs even if (or perhaps especially if) they are generally accepted by the scientific community. In other words, attempted replications are invaluable for insights where effects are presumed robust, which certainly applies to the underpinning evidence for loss aversion (Walasek & Stewart, 2015). This may be specifically true in a “hot” scientific field such as behavioural economics, which was largely spurred by Prospect Theory. This concern raised as Corollary 6 by Ioannidis (2005): more popular topics are more likely to produce false results. Given that the behavioural economics ‘movement’ has been used as an unfortunate (and misguided) framework to

criticise fundamentals of economic theory (Bruni & Sugden, 2007), Prospect Theory certainly fits this condition.

Kahneman and Tversky's manuscript brought major change to behavioural sciences. However, a number of ambiguities appear in their methods, such as the precise sample characteristics, 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 critical standards of reproducibility in behavioural science in 2019 by attempting replication of the full study from 1979. A failed replication would have seismic implications for behavioural science, whereas a successful replication would still offer tremendous value and, undoubtedly, reassurance. Confirmation would clearly strengthen our confidence in core assumptions about decision-making. Additionally, a multi-national replication would enable empirical documentation of variability between locations and languages (or the lack thereof) in key aspects of financial decision-making with risk.

## **Methods**

### *General summary*

This trial involved - as closely as possible - the direct replication of items used in the original paper on Prospect Theory (Kahneman & Tversky, 1979). We test original conclusions with contemporary analytical approaches through a combination of descriptive and inferential analyses, using original outcomes as a reference point. Additional items were introduced to account for various demographic factors as well as knowledge of the hypothesised 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.

### *Participants*

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 data collected exclusively online; no *in situ* testing took place. Details on other countries and languages considered are in the appendix.

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 use this robust approach intentionally to form insights about Prospect Theory specifically as well as 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 organisational membership channels. All project members targeted a minimum of 73 participants through direct collection. This threshold was greater than half of the sample size needed to meet necessary power levels (see *Power Calculation*), 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 (Klein et al., 2014). Convenience sampling in this case was intentional to assess potential differences between an online sample and one with engrained, systemic bias. Country-specific direct circulation, which was intentionally varied to buffer against bias, is outlined in Appendix C. This generally follows the replication approach used in the ManyLabs trials (Klein et al., 2014; Klein et al., 2018), noting that we intentionally did not utilise psychology student participant pools (Ebersole et al., 2016).

Countries that exceeded the total desired power threshold of 218 through direct sampling did not make use of paid samples. Though some participants were excluded later, they still exceeded the number targeted following 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 GBP0.70 to GBP1.30.

### *Instrument*

In order to closely replicate the procedure, the same items as published in Kahneman and Tversky (1979) were used, excluding the two verbal travel items and the verbal probabilistic insurance item (5, 6, & 9 in the original publication). The original travel items entailed a choice between having a chance to travel to England, France, and Italy, or having a certain trip to England. Their subjective utility might differ markedly between countries, which adds a needless interpretive burden to this multi-country 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 convoluted, 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 appendices, particularly for aspects not explicitly presented in the 1979 paper.

Financial values in each item were adjusted directly toward median net household incomes 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  $\frac{2}{3}$  of the 3,000 reference value. For 2019 in the United States, median net household income is about \$6,000 per month, so the same item would use \$4,000 for US participants in the replication. To put explicitly, had the numerical values from 1979 been reused for dollars (i.e., \$3,000 reference), this would have meant values were worth about half what they were in the original study.

As some governments report mean income as the standard for national average, mean income was used in eight countries. 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. Details on this are included in Appendix B and C. This approach was ultimately decided on the basis of it being more important to expose participants to choices representing the same wealth as the original study, given 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 (see Appendix D).

All items involve hypothetical monies only, in line with the original study. The lead and senior author have recently completed a multi-country study showing answers do not change substantially between hypothetical items and those involving real money, which has also been shown in other work (Kühberger et al., 2002; Beattie & Loomes, 1997; Wiseman & Levin, 1996), and reduces the need to validate with consequential choices.

Nine demographic measures were presented after decision-making items to avoid stereotype threat influences: nationality, year of birth, gender, income, and educational attainment, as well as four measures of current financial circumstances and behaviours (strain, recent changes, investments,

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 randomised). 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 or a 99% chance of losing 5,000, which means participants that were truly reading the options should immediately notice an obvious departure from other items.

The full set of choice items is shown in Table 1, 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 local currency. Details of any specific issues within countries are also included in Appendix C. None were deemed substantive to warrant highlighting here or requiring any parallel analyses to assess effect. The full set of all items in all languages is in Appendix G.

2019	1979	Items	Response alternatives
1	1		A 33% chance at 2500, a 66% chance at 2400, and a 1% chance of 0 Guaranteed 2400
2	2		A 33% chance of 2500 (67% chance of 0) A 34% chance of 2400 (66% chance of 0)
3	3		An 80% chance of 4000 (20% chance of 0) 100% guarantee of 3000
4	4		A 20% chance of 4000 (80% chance of 0) 25% chance of 3000 (75% chance of 0)
5	7	Which option do you prefer?	A 45% chance of 6000 (55% chance of 0) 90% chance of 3000 (10% chance of 0)
6	8		A 0.1% chance of 6000 (99.9% chance of 0) 0.2% chance of 3000 (99.8% chance of 0)
7	3'		An 80% chance of losing 4000 (20% chance of losing 0) A 100% guarantee of losing 3000
8	4'		A 20% chance of losing 4000 (80% chance of losing 0) A 25% chance of losing 3000 (75% chance of losing 0)
9	7'		A 45% chance of losing 6000 (55% chance of losing 0) A 90% chance of losing 3000 (10% chance of losing 0)
10	8'		A 0.1% chance of losing 6000 (A 99.9% chance of losing 0) A 0.2% chance of losing 3000 (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 4000 (20% chance of 0) A 100% guarantee of 3000
12	11	Imagine we gave you 1000 right now to play a game. Which option would you prefer?	A 50% chance to gain an additional 1000 (50% chance of gaining 0 beyond what you already have) A 100% guarantee of gaining an additional 500
13	12	Imagine we gave you 2000 right now to play a game. Which option would you prefer?	A 50% chance you will lose 1000 (50% chance of losing 0) A 100% chance you will lose 500
14	13		A 25% chance of 6000 (75% chance of 0) A 25% chance of 4000 (25% chance of 2000, 50% chance of 0)
15	13'		A 25% chance of losing 6000 (75% chance of losing nothing) A 25% chance of losing 4000 (25% chance of 2000, 50% chance of 0)
16	14	Which option do you prefer?	A 0.1% chance at 5000 (99.9% chance of 0) A 100% guarantee of 5
17	14'		A 0.1% chance of losing 5000 (99.9% chance of losing nothing) A 100% guarantee of losing 5

Table 1. *Base Version of the Survey with 1979 Values*. Most countries included “gaining” for several items, which was necessary to distinguish from “losing”. Items from 1979 with ‘ indicate a loss frame. The incorrect response alternatives for items 14 and 15 are not listed in the table.

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 make 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 (pooled .0001 and unpooled .05, respectively) in the same direction as in the original study.

### *Procedure*

All participants within each country completed identical surveys of 27 total items<sup>1</sup>, including demographics. After providing informed consent, participants responded to 17 choices with 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 order of the choice items were randomised. The 1979 paper presented each participant items in a pseudo-randomised order, but the printed nature of 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 randomised order should limit any confounds related to order.

At the end of the survey, participants were asked if they were familiar with the concept of loss aversion, as a proxy for general awareness of behavioural economics. We had no hypothesis regarding how awareness of loss aversion may have biased participants to answer in line with expectations (i.e., similar

---

<sup>1</sup> US participants completed a 28th item on financial circumstances. This was added at the end and did not impact other items. The item was included only for the purposes of comparing with a standard question on financial strain.

patterns as the 1979 results) or counter to them. However, we test if any differences exist between those who claimed they were familiar with the concept and those who did not. All participants were tested over a 15-day window in July-August 2019, additional details on this timeline are included in Appendix C.

### *Power and error*

Based on a pre-registered power-calculation (<https://osf.io/esxc4/>), we estimated that we would need 120 participants to have a 95% power to detect the original effects with an alpha threshold of .05. Though the original paper only reports chi-square tests that compare the response distribution of each item relative to a balanced null distribution, Kahneman and Tversky's theoretical argument primarily relies on contrasting pairs ("theoretical contrast pairs") of response distributions (see Appendix 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, in order to account for dropouts and exclusions, we targeted a minimum of 145 participants per country, which is larger than the largest sample size reported in the original study (141), and allows for dropouts while remaining above the 120 threshold. We met this criterion for all countries but one (Austria), which had a final sample size of 111.

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 power approaching 1 to detect the smallest anticipated effects at an alpha level of .0001.

All sample size calculations and power calculations are based on the `bpower` function in the `Hmisc` (v. 4.2-0) package (Harrell, 2019) in R. This also matched the approach presented by ManyLabs (Klein, 2014) in replicating multiple psychological studies (one of which included gain-loss framing items published by Kahneman and Tversky in 1981). However, where ManyLabs found partial method replication attempts were not impacted by setting (Owens, 2018), we tested a single method comprehensively between locations.

We first test if effects are similar to the 1979 trial across items by looking for significant deviation between replication findings and 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 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 as anticipated, these would be assumed as noise in the form of Type S error.

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 (i.e., the factor by which a statistically significant effect size in the replication overestimates the plausible effect size) (Gelman & Carlin, 2014). These are instead 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 either 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 (e.g., overweighting small probabilities) but not the other (e.g., 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 to the proportion of choice that can be explained by Prospect Theory as well as “residual choices” that can be explained by neither.

## **Results**

### *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. Our final sample consisted of 4099 participants from 19 different countries covering 13 languages. Direct sampling accounted for 73.8% of the final sample (paid sample 26.2%). Analyses include composite and separated samples, but with the wider view that if loss aversion was indeed present, and the tenets of Prospect Theory intact, 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 findings from one of the most influential papers in behavioural sciences. 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 (see Table 2), which was the largest sample reported in Prospect Theory 1979. All country samples are larger than the original sample for all items apart from for a single item (item 12), where three of our country samples are smaller.

<b>Country</b>	<b>Language</b>	<b>Direct <i>n</i></b>	<b>Paid <i>n</i></b>	<b>Total <i>n</i></b>
Germany	German	186	141	327
Italy	Italian	155	147	302
UK	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
USA	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
<b>Total</b>		3026	1073	4099

Table 2. Samples per country by direct, paid, and total *n*.

The total final sample consisted of 4099 participants, of which 50.7% were female. The median age was 29 and ages ranged from 18 to 85. Sixty-seven per cent of the sample was university educated. For country-specific demographics see Table 3.

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.

Country	Language	N	% Paid	Amount Paid Per Participant (GB£)	% Female	Age, Median (IQR)	Income, Median (IQR) (Value reported in local currency as used in measure)	% University Educated
<b>Pooled</b>	-	4099	26.18	-	50.7	29 (24-38)	-	66.94
<b>Australia</b>	English	282	0	-	35.11	31 (26.25-37)	60,000 (33,000 - 90,000)	77.3
<b>Austria</b>	German	111	36.94	1.10	45.95	28 (23.5-36)	15,000 (5,000-30,000)	53.15
<b>Belgium</b>	Dutch	192	33.85	1.15	47.4	27 (23-38.25)	15,000 (1,860-25,000)	64.58
<b>Bulgaria</b>	Bulgarian	127	22.83	-	61.42	33 (26-41.5)	15,000 (7,000-25,100)	81.89
<b>Chile</b>	Spanish	145	38.62	0.84	51.03	27 (24-36)	2,700,000 (260,000-12,000,000)	64.14
<b>Denmark</b>	Danish	150	19.33	0.95	33.33	32 (26-40)	240,000 (74,000-300,000)	72.67
<b>Germany</b>	German	327	43.12	1.10	39.14	27 (24-33)	15,000 (4,250-30,000)	66.67
<b>Hong Kong</b>	Chinese (Traditional)	160	0	-	63.75	30 (24-43)	200,000 (30,000-425,000)	70.62
<b>Hungary</b>	Hungarian	243	58.44	0.70	43.21	29 (24-34.5)	2,000,000 (405,000-3,922,536)	60.08
<b>Ireland</b>	English	256	55.86	1.30	58.2	32 (24-41)	23,500 (10,000-35,000)	70.31
<b>Italy</b>	Italian	302	48.68	1.00	55.96	29 (23-43)	7000 (15.375-20,000)	45.03
<b>Mainland China</b>	Chinese (Simplified)	259	0	-	55.98	33 (26.5-41)	100,000 (60,000-200,000)	86.1
<b>Norway</b>	Norwegian	226	16.37	1.30	58.85	29.5 (25-38)	327,500 (131,500-469,250)	76.55
<b>Serbia</b>	Serbian	246	0	-	71.54	25 (23-35)	155,000 (20,500-600,000)	63.41
<b>Slovenia</b>	Slovenian	202	0	-	51.49	25.5 (23-34.75)	5,350 (2,000-15,000)	61.39
<b>Spain</b>	Spanish	199	0	-	57.29	27 (25-46.5)	8,000 (675-23,000)	77.39
<b>Sweden</b>	Swedish	139	23.74	1.30	27.34	28 (23.5-33)	250,000 (72,500-360,000)	41.73
<b>UK</b>	English	290	0	-	48.28	29 (24-38)	20,000 (9,068.75-32,375)	70.69
<b>USA</b>	English	243	86.42	0.70	54.32	30 (25-41)	26,000 (10,000-47,700)	62.14

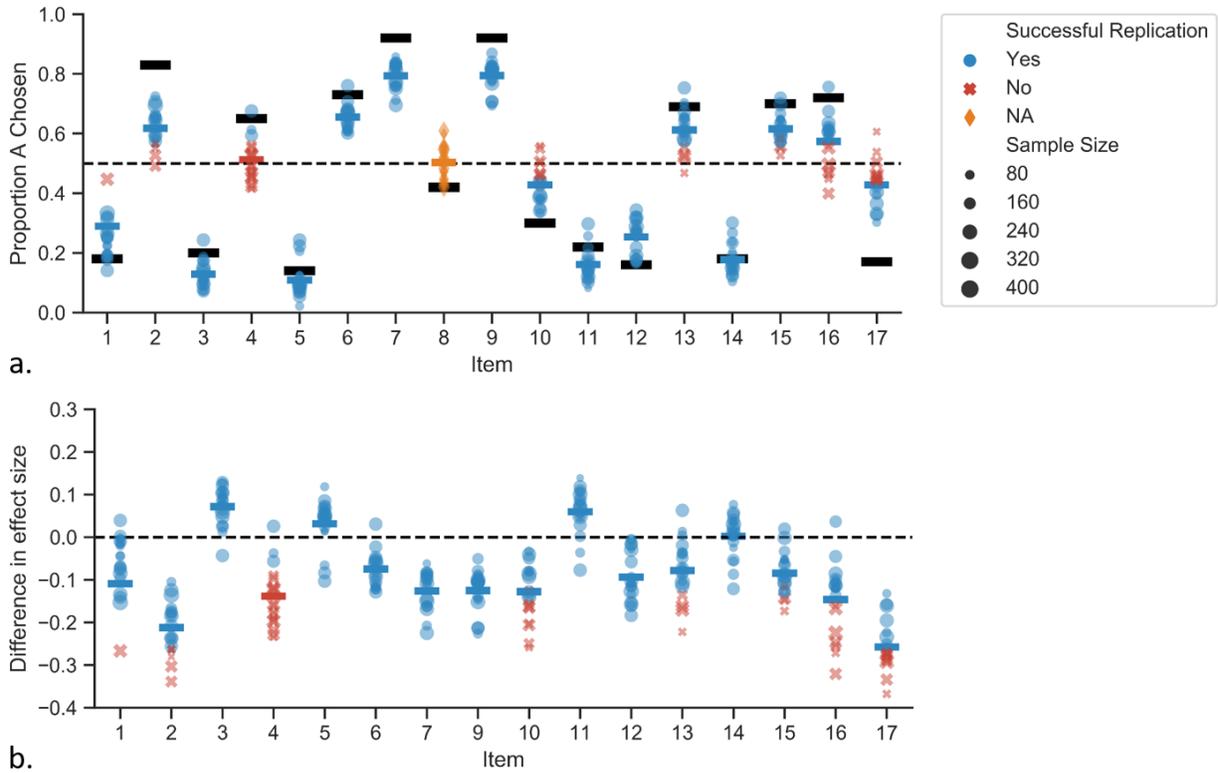
Table 3. Study demographics.

### *Replication analysis*

The original paper only reported chi-square tests that evaluated whether the response distribution for each item significantly differed from chance. Consequently, this is a natural starting point for our inquiry. Our criterion for successful replication was detecting a significant effect in the same direction as the original study. For the country specific analyses we use the conventional .05 alpha threshold, for the pooled data we use a stricter .0001 threshold.

The overall replication rate for 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 .0001 threshold, apart from item 4 ( $p=.14$ ) and item 8 ( $p=.67$ ).

Next we did unpooled analyses where we computed odds ratios for each item and country independently. Replication rates with these showed 247 out of 304 possible effects (81%) were significant in the same direction as the original study. The unpooled replications showed two important trends. First, while most effects replicated, there was a general attenuation of effects relative to the original study (see Figure 1). Specifically, 77% of the effects measured in this replication were smaller than those reported in the original study ( $\chi^2=96.99$ ,  $p<.0001$ ).



*Figure 1.* Effect-sizes by item. Each symbol represents one country; blue and red bars are the pooled results. a) Comparisons between the effects reported by Kahneman and Tversky in 1979 (the 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 .05. b) The change in effect size compared to Kahneman and Tversky 1979. The 0-lines represent the effect reported by Kahneman and Tversky. Values above the line suggest that the current sample show a stronger preference for one of the options (are further from the .05 indifference point) than the original study. Values below the line suggest a weaker preference (closer to the indifference point). There is an attenuation in effect sizes compared to 1979 for most items.

There was much greater variation in replication rates between items than between countries. The replication rate between countries ranged from 94% to 69% (see Figure 2); replication rates between

items ranged from 100% to 15%. This means individual items were more likely to have multiple failed replications than countries. It is also worth noting that there does not appear to be any pattern in terms of higher or lower replication based on values or income, as the four countries (Austria, Belgium, Germany, 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 high or lower median national incomes.

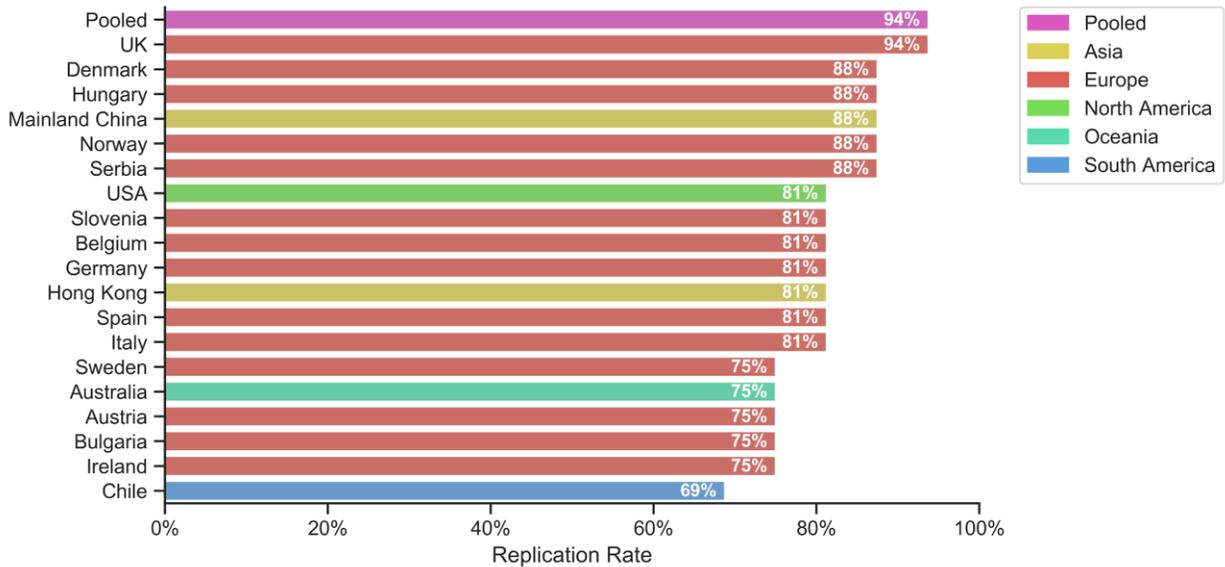


Figure 2. Item replication rates by country. A replication was deemed successful if the direction of the effect was the same as in the original study and significantly different from chance at an alpha threshold of .05.

Using the pooled data to retest the 1979 theoretical contrast pairs, 12 out of 13 contrasts replicated. The exception was the contrast between items 4 and 8 ( $p=.47$ ), 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. Instead, it seems more likely that in this case there was simply no effect to reflect.

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 (see Figure 3).

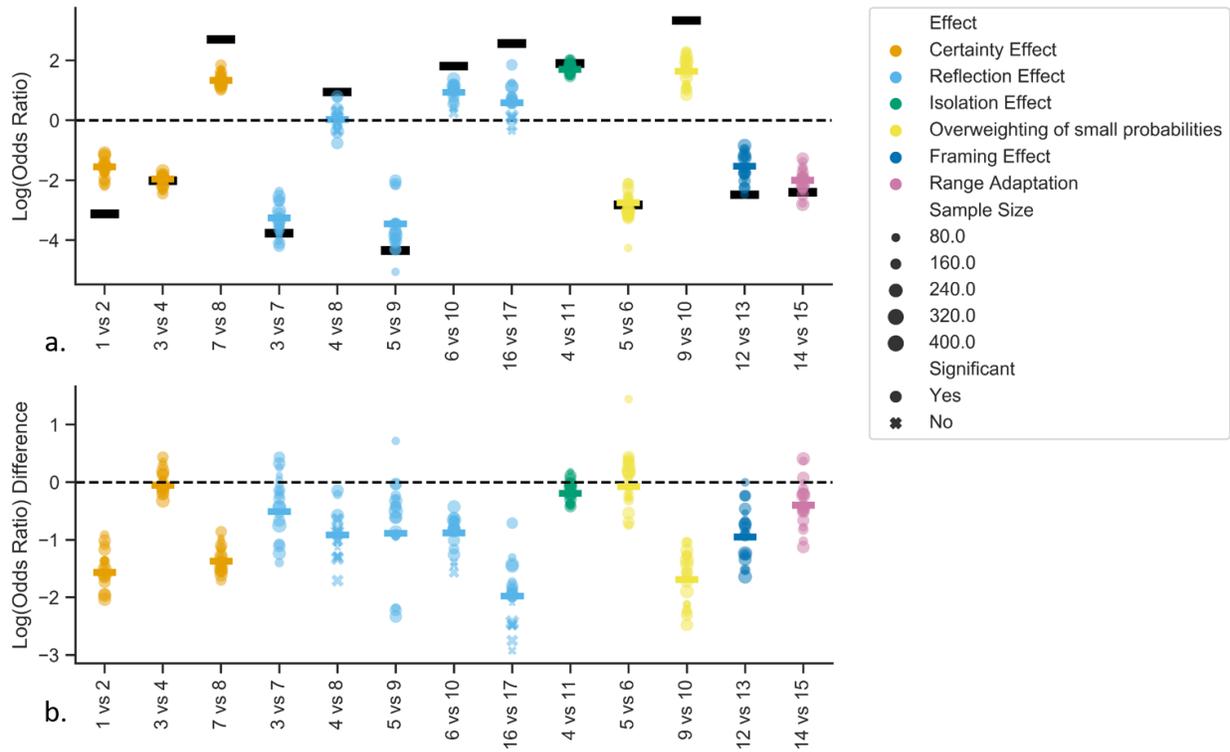
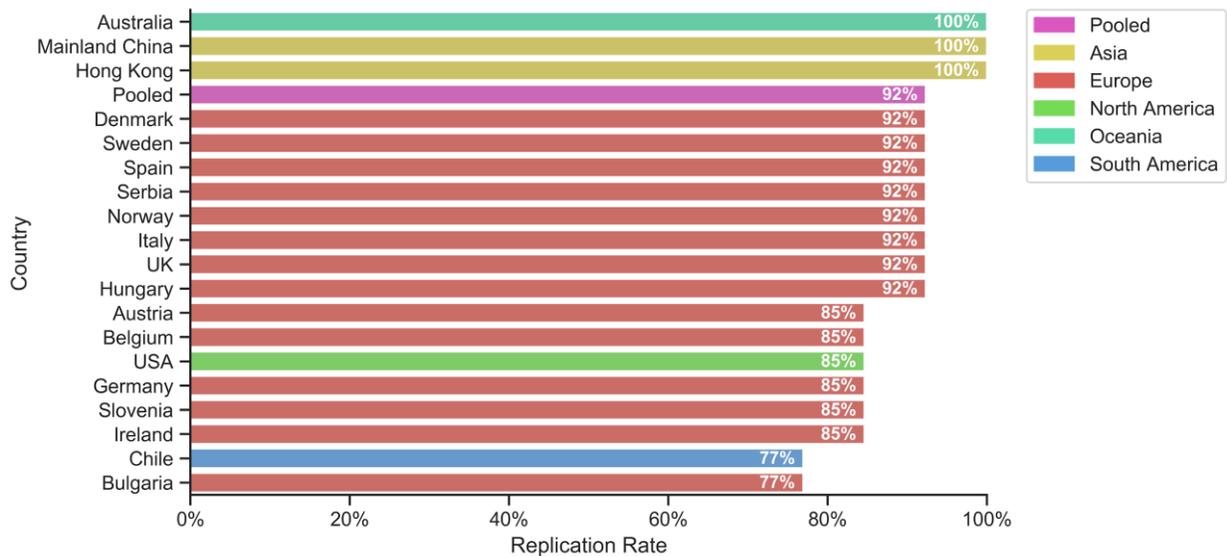


Figure 3. Effect sizes by contrast. Each circle or x represents one country; coloured bars are the pooled results. a) Comparisons between the effects in Kahneman and Tversky in 1979 (the black bars) and the current samples. A replication was deemed successful if the direction of the effect was the same as in the original study and significantly different from chance at an alpha threshold of .05. b) The change in effect size compared to Kahneman and Tversky 1979. The 0-line represents the effect reported by Kahneman and Tversky, values above the line suggest that the current sample show stronger contrast effects between the items than the original study, whereas values below the line suggest more similar effects between the contrasting items. There is an attenuation in effect sizes compared to 1979 for most contrasts.

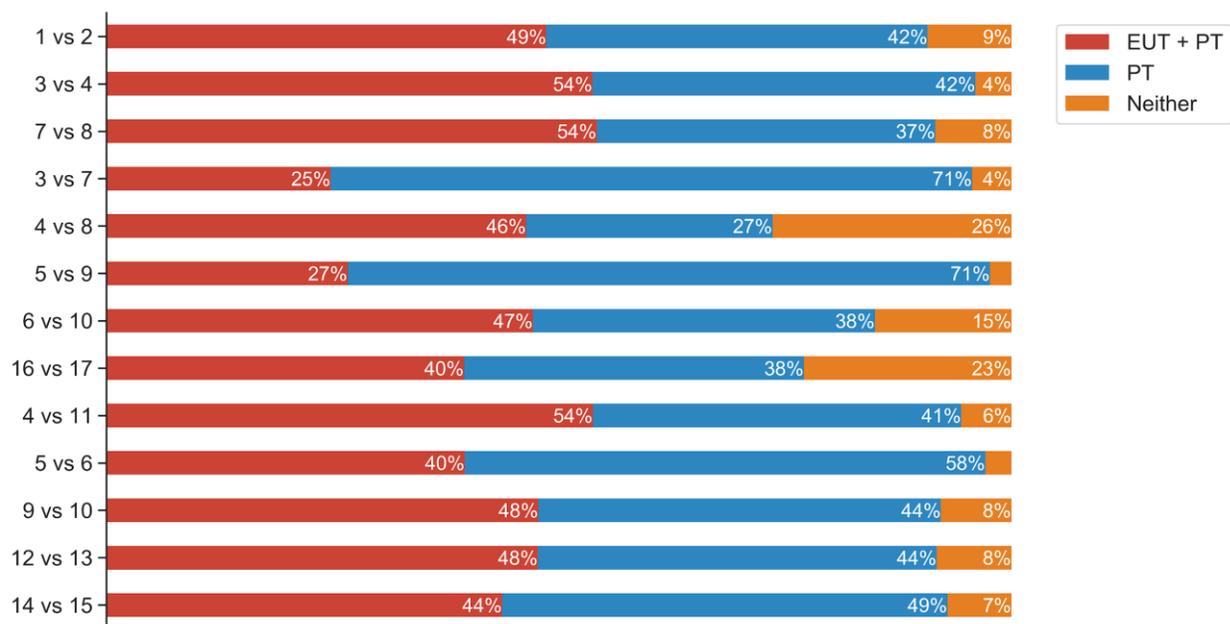
Most countries tested replicated at least 90% of the contrast effects, and the lowest replication rate recorded in any country was 77% (see Figure 4). As with the item-specific analyses, replication rates varied more between contrasts than between countries. Ten of 13 contrasts replicated consistently across all countries. All exceptions involved the reflection effect: the contrast between items 6 and 10 replicated in 84% of the countries, items 16 and 17 replicated in 63% of the countries, and items 4 and 8, which replicated in 16% of the countries. Though most effect sizes are attenuated compared to 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.



*Figure 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 .05. There was more variation in replication rates between contrasts than between countries.

Finally, we can evaluate prospect theory by studying choice patterns within respondents, as explained in the methods section. Choices that conform to Expected Utility Theory also conform to some version of Prospect Theory so we can code the choices for each item contrasts as either conforming to both

theories, only conforming to Prospect Theory or conforming to neither (See Figure 5). We found that 91% of choices could be accounted for by Prospect Theory (but there was some variation between items: SD = 7.5%) and 47% could be accounted for exclusively by Prospect Theory (SD = 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 contrast 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 evidence for the reflection effect is weaker than that of the others.



*Figure 5.* Choices congruent with Prospect Theory. This figure captures the proportion 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 were 9% but the standard deviation was 7.5%). However, Prospect Theory consistently provides a better explanatory framework for behaviour relative to EUT in these cases.

To test if knowledge of loss aversion moderated any of the observed contrast effects, we used hierarchical logistic regression models to predict choices that could only be explained 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 two out of thirteen 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 ( $z=-2.43$ ,  $p=.01$ ), the same pattern applied for the contrast between items 16 & 17 ( $z=-2.02$ ,  $p=.04$ ). Intuitions about loss aversion had no significant effect for either contrast. Additionally, for the contrast between items 12 & 13 people who intuited loss aversion were slightly less likely to make choices that conformed to prospect theory ( $z=-1.99$ ,  $p=.05$ ), 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 like awareness of loss aversion only has a very modest effect on choice, to the extent such an effect exists at all. To see if these effects may be present comparing between countries and rate of awareness, we present exploratory analyses in Appendix F (Hong Kong and China).

### **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 its subsequent constructs, such as loss aversion. Replication is a critical first step in this process, in the sense that though a successful replication would not neutralise all possible criticisms, a failed replication would have suggested serious issues with the original theory. We are also fully aware that arguments challenging Prospect Theory and loss aversion did not necessarily or categorically challenge the original method, but rather focus on

context, interpretations, and conclusions (e.g., Katsikopoulos & Gigerenzer, 2008). We also cannot assess if one theoretical framework (e.g., 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. By randomising all items for all participants, we were less concerned about the potential impact of order effects. Even still, there were no indications of such effects, as choice A was selected around 45% of the time for all 17 items, regardless of which one was presented. Sequential effects of contrast pair items will be tested in a separate study, but there is no indication that these have influenced conclusions in our study.

One pattern that has consistently been ignored in research on decision-making with 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 if those individuals differ systematically from other participants. We simply 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. A recent study testing decision paradoxes (Millroth et al., 2019) and numeracy raises this question for nine of the contrasts. As not in our pre-registered plan, we explored this (see Appendix D) in our own sample without hypothesis and find paradoxes appear for almost every participant, and most individuals demonstrate multiple paradoxes. As this is not a feature of the original 1979 study, we have not considered it further, but some details of these analyses are in Appendix D.

Magnitude of choice proportions vary somewhat between the direct and paid samples (see Appendix D). These variations ultimately have little impact on the conclusions of the study. For items 4, 8, and 11, paid samples show the opposite patterns (and therefore, 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 other contrast effects. However, looking at the direct sample alone, all of the thirteen 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 future versions could be slightly less strict with this feature.

While we cannot compare the age distribution to the 1979 sample, we note that our sample is skewed towards a younger population. This might be one limitation on generalisability, given that risk-aversion and other factors including computer literacy might vary systematically across age groups. As the focus of this particular study was on general reproducibility of a method, we have not explored in depth the potential effects of a number of demographic or financial indicators, specifically including age. Those will be reported in future dissemination.

## **Discussion**

With over 4,000 participants from 19 countries, we find 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 only to make currency values relevant for a 2019 sample within each country. In doing so, we find a pooled replication 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. Replication rates are over 70% for both item-based and contrast analyses for all countries (with the exception of Chile, which had an item-based replication rate of 69%). In short, we find nothing to indicate failure to replicate or any fundamental flaws to 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. One potential reason for this attenuation is that people are more aware of loss aversion today. 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 beliefs in 1979, it is plausible that current numbers are higher given the popularity of behavioural economics. However, to the extent we found an effect for 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 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 also 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, range adaption, 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

did not base their theory on specific effect sizes, but on distinct patterns of choice, we do not consider attenuation to be a major concern.

While we do not suggest our sample perfectly represents a global population, all analyses were sufficiently powered and show no indications of systematic bias that would undermine findings. Additionally, with only minor adjustments which are standard in conducting research, all critical methodological steps proposed in the pre-registered method were applied in practice. The only potential exception involved having a larger sample in most countries than targeted, which required fewer paid participants. Beyond this, we have no major issues to report for the method, and extensive details have been included in the appendices.

Overall, our results are generally in line with the findings of the original study. For the pooled analyses, two out of seventeen items (Item 4 and Item 8) did not have response patterns that were significantly different from chance, one of which were not significant in the original study. The other was the reflected version of this same item (i.e. the same magnitudes and probabilities, but one item 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 other reflection contrasts did replicate, we do not interpret this as a failure of the reflection effect, but rather 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.

Results for 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, comparison between PT and Expected Utility Theory) confirms that the main findings of the original study replicate on a general level. Hence, Prospect Theory still remains a robust and widely applicable descriptive model of decision-making under risk and uncertainty.

## **Conclusion**

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 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 conclusions may be.

In the end, we find that Kahneman and Tversky's 1979 empirical foundation of Prospect Theory replicates. Some effects were less strong than in 1979, but this may be more a testament to ease of accessing participants in 2019, rather than suggesting a flaw in the original study. Naturally, we did anticipate a full and comprehensive replication of findings, with some minor deviations. Nothing in our results would indicate loss aversion or other theoretical constructs (framing effects, over-weighting small probabilities, etc.) to be anything but reliable. Rather, our results seemingly uphold or magnify those conclusions.

Based on our findings, we roundly endorse the original conclusions of Prospect Theory - and much of the subsequent theory it has inspired. There is no question it has made an impact on individual behaviour as well as public policy around the world. We strongly encourage work to that end continue in areas from financial decision-making to population well-being.

## **Acknowledgments**

A number of colleagues, peers, and organisations will be thanked for their support of this work in the final publication. Some of these include Katarina Kastellic, Ivaylo Sakelariiev, Tamas Varkonyi, Amata Víg, Chiara Saponaro, Jacobo Perez, Manuel Frías, and Prof Simon Deakin. Authors confirm no conflict of interest in undertaking this study. We also thank Corpus Christi College for support in hosting numerous researching contributing to the study.

## References

- Altman, M. (2010). Prospect theory and behavioral finance. In H. K. Baker & J. R. Nofsinger (Eds.), *Behavioral Finance: Investors, corporations, and markets* (Vol. 6), 191-209. Wiley.
- Barberis, N. C. (2013). Thirty years of Prospect Theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27(1), 173-96.
- Beattie, J., & Loomes, G. (1997). The impact of incentives upon risky choice experiments. *Journal of Risk and Uncertainty*, 14(2), 155-168.
- Benartzi, S., & Thaler, R. H. (1995). Myopic loss aversion and the equity premium puzzle. *The Quarterly Journal of Economics*, 110(1), 73-92.
- Bruni, L., & Sugden, R. (2007). The road not taken: how psychology was removed from economics, and how it might be brought back. *The Economic Journal*, 117(516), 146-173.
- Ebersole, Charles R., Olivia E. Atherton, Aimee L. Belanger, Hayley M. Skulborstad, Jill M. Allen, Jonathan B. Banks, Erica Baranski et al. (2016) "Many Labs 3: Evaluating participant pool quality across the academic semester via replication." *Journal of Experimental Social Psychology*, 67, 68-82.
- Erev, I., Ert, E., and Yechiam, E. (2008). Loss aversion, diminishing sensitivity, and the effect of experience on repeated decisions. *Journal of Behavioral Decision Making*, 21, 575-597.
- Freedman, D. H. (2011). Lies, Damned Lies, and Medical Science. In M. Roach & T. Folger (Eds.), *The Best American Science and Nature Writing 2011: The Best American Series*, 112–126.
- Gal, D. (2006). A psychological law of inertia and the illusion of loss aversion. *Judgment and Decision Making*, 1, 23–32.
- Gal, D., & Rucker, D. D. (2018). The loss of loss aversion: Will it loom larger than its gain? *Journal of Consumer Psychology*, 28(3), 497-516.
- Gelman, A., & Carlin, J., (2014). Beyond power calculations: Assessing Type S (Sign) and Type M (Magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641-651.
- Gelman, A., & Hill, J. (2012). *Data analysis using regression and multilevel/hierarchical models* (Vol. 1). New York, NY, USA: Cambridge University Press.
- Genesove, D., & Mayer, C. (2001). Loss aversion and seller behavior: Evidence from the housing market. *Quarterly Journal of Economics*, 116(4), 1233-1260.
- Harrell Jr, F. E. (2019). Package 'Hmisc'. *CRAN2018*, 235-6.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, 7(6), 645-654.
- Johnson, E. J., Hassin, R., Baker, T., Bajger, A. T., & Treuer, G. (2013). Can consumers make affordable care affordable? The value of choice architecture. *PLoS One*, 8(12), e81521.  
doi:10.1371/journal.pone.0081521

- Kahneman, D., & Tversky, A. (1979). Prospect Theory: An analysis of decisions under risk. *Econometrica*, 47(2), 263-291.
- Katsikopoulos, K. V., & Gigerenzer, G. (2008). One-reason decision-making: Modeling violations of expected utility theory. *Journal of Risk and Uncertainty*, 37(1), 35-56.
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Cemalcilar, Z. (2014). Investigating variation in replicability. *Social Psychology*, 45, 142-152. <https://doi.org/10.1027/1864-9335/a000178>.
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, 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.
- Kőszegi, B., & Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4), 1133-1165.
- Kühberger, A., Schulte-Mecklenbeck, M., & Perner, J. (2002). Framing decisions: Hypothetical and real. *Organizational Behavior and Human Decision Processes*, 89(2), 1162-1175.
- Kuonen, D. (2018). An introduction to bootstrap methods and their application. *WBL in Angewandter Statistik ETHZ 2017*, 19, 1-143.
- Levy, J. S. (1996). Loss aversion, framing, and bargaining: The implications of prospect theory for international conflict. *International Political Science Review*, 17(2), 179-195.
- Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4), 764-766.
- McDermott, R. (2004). Prospect theory in political science: Gains and losses from the first decade. *Political Psychology*, 25(2), 289-312.
- McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.
- McGraw, A. P., Larsen, J. T., Kahneman, D., & Schkade, D. (2010). Comparing gains and losses. *Psychological Science*, 21(10), 1438-1445.
- Mercer, J. (2005). Prospect theory and political science. *Annual Review of Political Science*, 8, 1-21.
- Millroth, P., Nilsson, H., & Juslin, P. (2019). The decision paradoxes motivating Prospect Theory: The prevalence of the paradoxes increases with numerical ability. *Judgment and Decision Making*, 14(4), 513-533.
- Odean, T. (1998). Are investors reluctant to realize their losses? *Journal of Finance*, 53, 1775-1798.
- OECD. (2017). *Behavioural Insights and Public Policy: Lessons from Around the World*. OECD Publishing, Paris. <https://doi.org/10.1787/9789264270480-en>

- Owens, B. Replication failures in psychology not due to differences in study populations. *Nature News* (19 November, 2018).
- Rieger, M. O., Wang, M., & Hens, T. (2011). Prospect Theory around the world. *NHH Department of Finance & Management Science Discussion Paper* (2011/19).
- Ritholtz, B. (2018, August). A Challenge to the Biggest Idea in Behavioral Finance. *Bloomberg Opinion - Finance*.
- Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, 9(1), 76-80.
- Simonsohn, U. (2014). '[15] Citing Prospect Theory'. *Data Colada, February 10*. Available at <http://datacolada.org/15>.
- Sokol-Hessner, P., Hsu, M., Curley, N. G., Delgado, M. R., Camerer, C. F., & Phelps, E. A. (2009). Thinking like a trader selectively reduces individuals' loss aversion. *Proceedings of the National Academy of Sciences*, 106(13), 5035-5040.
- Sydnor, J. (2010). (Over) insuring modest risks. *American Economic Journal: Applied Economics*, 2(4), 177-99.
- Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization*, 1(1), 39-60.
- Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453-458.
- Tversky, A., & Kahneman, D. (1992). Advances in Prospect Theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4), 297-323.
- Walasek, L., & Stewart, N. (2015). How to make loss aversion disappear and reverse: Tests of the decision by sampling origin of loss aversion. *Journal of Experimental Psychology: General*, 144(1), 7.
- Wang, M., Rieger, M. O., & Hens, T. (2017). The impact of culture on loss aversion. *Journal of Behavioral Decision Making*, 30(2), 270-281
- Xie, Y., Hwang, S., & Pantelous, A. A. (2018). Loss aversion around the world: Empirical evidence from pension funds. *Journal of Banking & Finance*, 88, 52-62.