Original Papers

Polish Psychological Bulletin
2022, vol. 53(3) 138–151
DOI: 10.24425/ppb.2022.141862

Agata Sobkow[*], *Marcin Surowski*[*], *Angelika Olszewska*[*], *Nina Antoniewska*[*], *Katarzyna Barcik*[*], *Urszula Bartkiewicz*[*], *Agnieszka Brzeska*[*], *Adrianna Brzozowska*[*], *Oliwia Budrewicz*[*], *Jakub Choja*[*], *Kamila Choma*[*], *Patrycja Chorbotowicz*[*], *Michalina Filimoniak*[*], *Łukasz Filip*[*], *Paweł Gambuś*[*], *Weronika Gierlik*[*], *Tomasz Gonczar*[*], *Katarzyna Goryczka*[*], *Maksymilian Góra*[*], *Marta Haczek*[*], *Weronika Hetmańczuk*[*], *Zuzanna Holka*[*], *Aneta Janosz*[*], *Nikola Kikowska*[*], *Joanna Kołcun*[*], *Zuzanna Kozłowska*[*], *Monika Kujawińska*[*], *Marcin Kuleszczyk*[*], *Aleksandra Lach-Galińska*[*], *Katarzyna Latacz*[*], *Adam Ławniczak*[*], *Katarzyna Majewska*[*], *Klaudia Makowska*[*], *Marta Mamzer*[*], *Iga Marciniszyn*[*], *Adam Masternak*[*], *Magdalena Matuszek*[*], *Jonasz Mehr*[*], *Ewelina Miela*[*], *Monika Mleczko*[*], *Paulina Morga*[*], *Magdalena Niemczyk*[*], *Damian Ostrowski*[*], *Jagoda Pełdiak*[*], *Kamil Piotrowicz*[*], *Antoni Płuciennik*[*], *Oskar Ryśkiewicz*[*], *Weronika Sekuła*[*], *Małgorzata Sikora*[*], *Natalia Sikora*[*], *Daria Sitko*[*], *Agata Sobczak*[*], *Julia Sosenko*[*], *Sonia Stando*[*], *Katarzyna Starek*[*], *Łukasz Ślak*[*], *Jagoda Świtała*[*], *Natalia Świtniewska*[*], *Agnieszka Tyc*[*], *Olga Urban*[*], *Natalia Wcisło*[*], *Katarzyna Wiśniewska*[*], *Joanna Wodzińska*[*], *Aleksandra Zabiełło*[*], *Monika Żygadło*[*], *Tomasz Zaleskiewicz*[*], *Jakub Traczyk*[*]

Conceptual replication study of fifteen JDM effects: Insights from the Polish sample

Abstract: We conducted pre-registered replications of 15 effects in the field of judgment and decision making (JDM). We aimed to test the generalizability of different classical and modern JDM effects, including, among others: less-is-better, anchoring, and framing to different languages, cultures, or current situations (COVID-19 pandemic). Replicated studies were selected and conducted by undergraduate psychology students enrolled in a decision-making course. Two hundred and two adult volunteers completed an online battery of replicated studies. With a classical significance criterion ($p < .05$), seven effects were successfully replicated (47%), five partially replicated (33%), and three did not replicate (20%). Even though research materials differed from the originals in several ways, the replication rate in our project is slightly above earlier reported findings in similar replication projects. We discuss factors that may underlie replication results (success vs. failure). We also stress the role of open science practices such as open data, open research materials, pre-registration, and registered reports in improving the replicability of results in the JDM field.

Keywords: *replication, decision making, reproducibility, pre-registration, risk perception*

INTRODUCTION

Since the seminal works by Popper (1934), reproducibility has been regarded as a cornerstone of modern science. Original findings that cannot be systematically obtained by other researchers who employ appropriate scientific methods in adequately powered studies are questionable for theory development and become futile in terms of their potential to be applied in different domains of everyday functioning (Simons, 2014). For example, if the significance of a promising effect of an intervention on health behaviors that fails to be replicated by other researchers diminishes, the intervention may be regarded as fruitless for policymakers responsible for health services.

A series of unsuccessful replications of notable psychological effects, such as ego depletion (Hagger et al., 2016), social priming (Doyen et al., 2012), or power posing (Ranehill et al., 2015) reported in papers published in prestigious peer-reviewed journals raised concerns about numerous scientific findings and initiated a debate commonly labeled as a replication crisis (see, for example, the open letter written by Kahneman; Yong, 2012).

Several factors were identified that might be responsible for the crisis, including selective data collection and reporting, data manipulations/transformations, treating exploratory analyses as confirmatory analyses, or drawing conclusions based on underpowered studies (Simmons et al., 2011). Failure to reproduce positive results has been observed in different disciplines from psychology, through neuroscience, and political sciences, to economics, indicating that various fields may struggle with a broad range of problems that prevent efficient reproducibility of scientific discovery. Interestingly, a survey conducted among researchers from various disciplines revealed that more than 70% of them experienced failure when trying to reproduce the results of another experiment (Baker, 2016).

[*] **SWPS University of Social Sciences and Humanities; Faculty of Psychology in Wroclaw, Poland.**
Corresponding author: Agata Sobkow - asobkow@swps.edu.pl

Funding source with grant number
This work was supported by the SWPS University of Social Sciences and Humanities, Faculty of Psychology in Wroclaw.

In response to this, investigators called for enhancing efforts aimed at verifying the robustness of effects and recommended running multi-lab pre-registered replication projects to establish replicability and boundary conditions essential to observing an effect of interest.

For instance, a replication project run by the Open Science Foundation (Open Science Collaboration, 2015) indicated that the successful replication rate of 100 effects published in three top psychological journals ranged from 36 to 47%, depending on the criterion used to assess successful replication (but see Gilbert et al., 2016). Klein et al. (2014), in their Many Labs, registered replication project (N = 6,344) conducted in 36 labs across the world, found that out of 13 psychological effects that were selected for the study, ten were successfully replicated (but with smaller effect sizes), two were not replicated, and there was weak evidence to conclude successful replication for one effect. In the Many Labs 2 project (Klein et al., 2018), only 15 of 28 psychological effects were successfully replicated (N = 15,305; 36 countries).

Replication projects were also conducted in the field of judgment and decision making (JDM). For example, it has been demonstrated that 94% of the effects predicted by prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992) were successfully replicated (Ruggeri et al., 2020). Also, other classical JDM effects such as money illusion (Ziano et al., 2021), decoy effect (Xiao et al., 2020), status quo bias (Xiao et al., 2021), framing (Zhou et al., 2021), or motivated numeracy (Persson et al., 2021) were subjects of replication and extension studies.

Despite the fact that the majority of JDM effects have been successfully reproduced, there are at least two potential concerns of replication studies in the JDM field that limit their ubiquitousness and generalizability. First, studies on decision making often incorporate simple monetary lotteries to draw conclusions about choices under risk/uncertainty and test cognitive models of choice (Fox et al., 2015; Wu et al., 2008). Parameters used in the process of lottery construction (e.g., payoffs) may exert an influence on choices. For example, depending on the currency used in the original study, investigators conducting replications can 1) use the same currency irrespective of the currency of the country where replication is carried out (e.g., $5 in the original study and $5 in the replication study), 2) change the original currency to the currency of the country where replication is carried out without changing values (e.g., $5 in the original study and 5 PLN – Polish Zloty – in the replication study), or 3) change the original currency to the currency of the country where replication is carried out along with changing values according to the present currency exchange rate (e.g., $5 in the original study and 19.10 PLN in the replication study). Finally, even when the same currency in the same country is used, researchers conducting replication could face the problem of differences in purchasing power. For example, when a replication study is conducted several years after the original study, because of the inflation, a fewer number of items could be purchased with a particular unit (e.g., $5). Based on the research on number sense and numerical cognition, such modifications are likely to modulate numeric information processing (Dehaene, 1999).

Second, language-, time-, and culture-related nuances observed in choice dilemmas used in JDM research may be critical for successful replications. For example, a person living in Europe will perceive a dilemma describing a lottery in which an individual can win a travel coupon covering the costs of going to Europe differently than a person living in the US (Rottenstreich & Hsee, 2001). Decisions in the classic Asian Disease Problem (Tversky & Kahneman, 1974) can be influenced by the current pandemic situation, and dilemmas describing an opportunity to purchase Barnes & Noble featured coupons can be difficult to understand in countries where the company does not operate (Gneezy et al., 2006).

In the present study, we attempt to make the initial step to address these concerns. We report the results of a project aimed at evaluating the reproducibility (and boundary conditions) of 15 classical and modern effects in the field of judgment and decision making. Nevertheless, our project differs from many other replication projects in psychology. In particular, all replicated JDM effects were adjusted to the Polish language, currency, and economic situation. Thus, our research could be considered conceptual replication rather than a direct replication (Hüffmeier et al., 2016). Additionally, replicated studies were selected, pre-registered, and conducted by undergraduate psychology students enrolled in a decision-making course, which seems to be of special importance because of educational benefits and the increasing methodological awareness of students (Jekel et al., 2020). Moreover, such a selection procedure allows a broad range of effects that are regarded as important by young people to be investigated, and not only effects that have gained an audience among experienced researchers.

METHOD & RESULTS

Participants

Two hundred and two Polish-speaking Prolific users (65 females, 133 males, and 4 persons who refused to answer the question about gender, M_{age} = 23.5, SD_{age} = 6.43) completed a series of online tasks. Participants were paid £2.50. Participation in the study was voluntary, and participants gave informed consent before the study. The study protocol was approved by the departmental Ethical Committee.

Procedure

During the decision-making course[1], undergraduate psychology students formed fifteen research teams and were asked to choose effects known from the judgment and decision making field that, in their opinion, had value for replication in Polish settings. These effects were selected and evaluated by the brevity of the procedure, feasibility to be conducted online, and on adults from the general

[1] The course syllabus was inspired by the work by Gilad Feldman https://mgto.org/teaching-courses/ and the Hagen Cumulative Science Project (Jekel et al., 2020).

population. Next, students prepared a pre-registration of replication of each effect on the Open Science Framework platform. We designed a simplified pre-registration form (in Polish) based on the "Replication Recipe" (Brandt et al., 2014) to make it more understandable for undergraduate students. Nevertheless, this simplified form contained all essential elements, such as the brief description of the replicated effect and its effect size, known differences between the original study and replication, required sample size, and plan of data analysis. We used G*Power (Faul et al., 2009) to calculate the required sample size for each effect based on the original effect size (we used effect sizes reported in original studies or estimated them based on available information). Conventionally, we assumed $\alpha = .05$ and $1 - \beta = .80$. Obtained sample ($N = 202$) was larger than the required sample size for each effect (however, see the description of the project S04 for an exception).

Research materials were translated and adapted to current settings (e.g., language, the COVID-19 pandemic). Then all tasks were coded in an online experimental procedure run under the Qualtrics. The order of presentation of tasks testing different effects was counterbalanced. The whole study lasted approximately 25 minutes.

THE DESCRIPTION OF 15 PSYCHOLOGICAL EFFECTS

For each of the tested effects, we provide a brief description of its nature, main findings based on the results reported in the original articles, differences between the original and replication study, and detailed results of the replication. We assumed that the effect is successfully replicated if it is statistically significant ($p < .05$) and its direction is consistent with the results of the original study. Additionally, in Figure 1, we present unified effects sizes estimated in the original studies (i.e., all effect sizes were converted to the Cohen's *d*) and effect sizes in the replication study with a 95% confidence interval.

Relationship between procrastination and dependent decision-making style (Geisler & Allwood, 2018) (S01)[2]

This original study investigated the relationship between decision-making styles and other individual differences. One hundred and eighteen Swedish students completed several questionnaires: General Decision-Making Style Inventory (GDMS; Scott & Bruce, 1995) containing five subscales (rational, intuitive, spontaneous, avoidant, and dependent), Self-Monitoring Scale (Lennox & Wolfe, 1984), Machiavellian Personality Scale (Dahling et al., 2008), Procrastination Scale–Student version (Lay, 1986), and Time-Style Scale (Usunier & Valette-Florence, 2007). This study demonstrated that different decision-making styles were related to specific differences in social orientation and time approach.

In the current study, we wanted to replicate the relationship between dependent decision style (the tendency to "seek advice and support of others or let others decide") and procrastination (the tendency to "use time resources by postponing the start or completion of tasks that need to be done"). In the original study, a small positive correlation was observed between these two variables ($r = .277$). Individuals who more often procrastinated also had a higher tendency to consult their decision with others.

We used the Polish version of the Pure Procrastination Scale (Stępień & Topolewska, 2014) and the translation of the dependent style subscale from the GDMS (Scott & Bruce, 1995). Both scales had excellent reliability (Cronbach's $\alpha_{\text{dependent style}} = 0.809$ and Cronbach's $\alpha_{\text{procrastination}} = 0.898$). We found a significant relationship between dependent decision style and procrastination ($r = .165$, $p = .019$). Thus the effect was replicated but smaller than in the original study.

[2] The codes refer to student projects on the OSF (https://osf.io/6sq8p/) and in Supplementary Table S1.

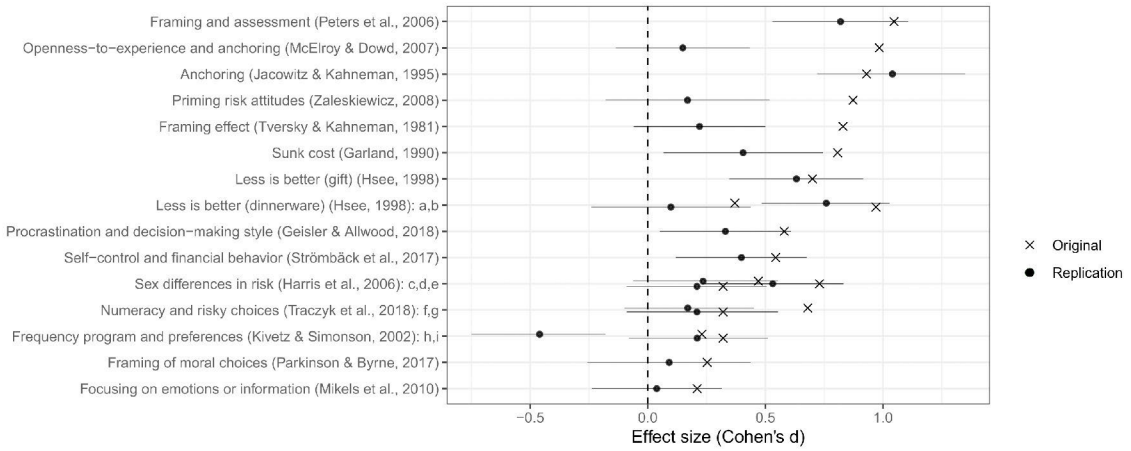

Figure 1. Effects sizes in the original („X") and replication (●) studies. Error bars represent a 95% confidence interval for the replicated effect sizes. To enable comparisons between effects sizes, we converted all estimates (e.g., Pearson's *r*, eta^2, etc.) to Cohen's *d*. The following letters denote specific effects tested in the replicated studies (see the description of effects in text for details): a) joint evaluation, b) separate evaluation, c) alcohol, d) gambling, e) risky recreational activities, f) the relationship between subjective numeracy and choices in high-payoff problems, g) the relationship between subjective numeracy and choices in low-payoff problems, h) cinema, i) clothing retailer.

The effect of sunk cost on the decision to escalate commitment to an ongoing project (Garland, 1990) (S02)

In the original study, 407 participants estimated the likelihood that they would authorize spending money on an ongoing investment project in the five different conditions when 1, 3, 5, 7, and 9 million dollars had already been spent on the project (with a total budget of 10 million dollars). Participants were assigned to one of three between-subjects conditions varying in the dependent variable measure: some participants estimated the probability of authorizing the next one million dollars for this project (1), others estimated the probability of authorizing remaining budget funds (2) or the perceived likelihood of the project's success (3).

Results showed a significant effect of sunk cost for authorizing the next one million dollars, $F(4, 145) = 6.67$, $p < .001$, $\eta_p^2 = 0.14$, and for authorizing all remaining budget funds, $F(4, 122) = 12.20$, $p < .001$ $\eta_p^2 = 0.28$. The subjective probability of charging all remaining funds and an additional one million dollars increased when the sunk cost was higher.

In our study, we intended to replicate the effect of investing an additional one million dollars for the project. However, we used only three between-subjects conditions, when 1 (10% of a project; $n = 69$), 5 (50%; $n = 67$), and 9 million (90%; $n = 66$) dollars had been invested in the project. The results showed a significant effect of sunk cost on the subjective likelihood of investing an additional one million dollars, $F(2, 199) = 14.51$, $p < .001$, $\eta_p^2 = 0.13$. Participants in the 10% condition estimated the probability of investing an additional million dollars for a project as $M = 42.26$ ($SD = 25.84$), in 50 % as $M = 52.79$ ($SD = 25.42$), and in 90% as $M = 66.33$ ($SD = 26.74$). The post-hoc comparisons with Bonferroni correction revealed that the difference between conditions 10% and 90% as well as between 50% and 90% were statistically significant ($t = -5.377$, $p < .001$, $d = -0.916$ and $t = -3.003$, $p = .009$, $d = -0.519$, respectively). While the difference between 10% and 50% conditions was only marginally significant ($t = -2.361$, $p = .058$, $d = -0.411$). Nevertheless, we argue that this effect was successfully replicated.

Priming risk attitudes and instrumental risk tendency (Zaleskiewicz, 2008) (S03)

In the original study, 200 participants were randomly assigned to one of the five conditions: four experimental and one control. In a 2 x 2 design of experimental conditions, participants were asked to solve a word search puzzle including words priming the attitude 2(encouraging, discouraging) toward risk 2(instrumental, stimulating). Sets of priming words were selected in pilot studies and contained, for example, *success* (encouraging instrumental), *excitement* (encouraging stimulating), *defeat* (discouraging instrumental), and *fear* (discouraging stimulating). Participants in the control condition did not solve any puzzle before the main task. Next, all the participants were shown six risk-related scenarios and declared their willingness to take these risks using a 7-point scale—three of these scenarios related to instrumental risk (e.g., taking a student loan) and three to stimulating risk (e.g., bungee jumping).

The results (the one-way ANOVA) showed that priming an instrumental risk attitude (encouraging or discouraging) influenced willingness to take instrumental risk, $F(4, 195) = 9.454$, $p < .001$, $\eta^2 = 0.16$. Participants who solved a puzzle containing words discouraging them from taking instrumental risk had the lowest tendency to engage in instrumental risk behaviors ($M = 3.33$). In contrast, those who solved a puzzle containing words encouraging them to take the instrumental risk were more willing to take such risk ($M = 5.10$). On the other hand, willingness to take stimulating risk was the highest among participants who solved a puzzle containing words encouraging them to take stimulating risk ($M = 5.01$), while the lowest tendency to take this type of risk was found in two groups solving puzzles with discouraging words ($M_{discouraging\_instrumental} = 3.12$, $M_{discouraging\_stimulating} = 3.33$), $F(4, 195) = 9.240$, $p < .001$, $\eta^2 = 0.16$.

In the present study, we wanted to replicate the effect of priming instrumental risk attitudes and willingness to take this type of risk. Similar to the original study, participants solved word search puzzles containing words encouraging or discouraging instrumental risk-taking. However, based on advice from the author of the original study, we decided to alter the control condition (in this replication study, participants solved a puzzle that contained only neutral words). After solving a puzzle, participants were shown three scenarios related to instrumental risk and declared their willingness to take such risks using 7-point scales. We hypothesized that participants primed with words encouraging to take instrumental risk would declare the highest willingness to take this risk. In contrast, those primed with discouraging words would have the lowest tendency to take this risk. Finally, participants completed the attention check task.

Because in the pre-registration form, we set eligibility criteria (e.g., finding at least eight words in a puzzle and correctly responding to the attention check questions), the analyses were performed on a smaller sample ($N = 127$: $n_{control} = 40$, $n_{discouraging} = 48$, $n_{encouraging} = 39$)[3]. The results showed that priming risk attitudes influenced willingness to take instrumental risk, $F(2, 124) = 4.721$, $p = .011$, $\eta^2 = 0.07$. Participants in the discouraging condition declared the lowest tendency to take risk ($M = 4.34$, $SD = 0.99$), while those in the encouraging condition declared the highest ($M = 4.93$, $SD = 0.98$). Post-hoc comparisons with Bonferroni correction revealed that the only significant difference was obtained between encouraging and discouraging conditions ($t = 2.913$, $p = .013$, $d = 0.592$), while the comparisons between experimental and control ($M = 4.78$, $SD = 0.79$) conditions were not significant ($t = -0.709$, $p = 1.000$ for comparison between control and encouraging) or marginally signifi-

[3] The results for a full sample ($N = 202$) were largely the same: $F(2, 199) = 5.558$, $p = .004$, $\eta^2 = 0.053$ with significant differences between discouraging and encouraging conditions ($t = 3.028$, $p = .008$, $d = 0.512$) and discouraging and control condition ($t = 2.764$, $p = .019$, $d = 0.471$).

cant (t = 2.188, p = .092 for comparison between control and discouraging). Therefore, we argue that our replication was partially successful.

Sex differences in risk preferences (Harris et al., 2006) (S04)

This study investigated the effects of sex differences on risk perception in four risk domains: gambling, recreation, health, and social. The authors also explored potential mediators of these relationships, such as perceived likelihood and severity of negative risk consequences or perceived benefits of risky activities.

In the present replication, we focused only on the effects of sex on declared past risk behaviors. In the original study, a sample of 657 US students (268 males and 389 females) responded to nine questions about their actual past risky activities using five-point scales. The results showed that males reported more risky behaviors when they were asked about smoking, $t(654) = 2.84$, $p = .013$, $d = 0.20$, drinking alcohol, $t(652) = 4.13$, $p < .001$, $d = 0.32$, frequency of being drunk, $t(654) = 2.90$, $p = .004$, $d = 0.22$, fast driving, $t(653) = 2.90$, $p = .004$, $d = 0.23$, breaking traffic laws, $t(654) = 3.30$, $p < .001$, $d = 0.25$, gambling, $t(653) = 9.53$, $p < .001$, $d = 0.73$, risky recreational activities, $t(655) = 5.99$, $p < .001$, $d = 0.47$, and public exposition during classes, $t(655) = 3.22$, $p < .001$, $d = 0.24$. The authors did not observe sex difference in the frequency of arguments with friends or family, $t(655) = 3.22$, $p = .115$, $d = 0.12$.

Because of budget constraints and the required sample size, we decided to pre-register only replication of the strongest effects: gambling ("How often do you gamble?"), engaging in risky recreational activities ("How often do you engage in risky recreational activities?"), and drinking alcoholic beverages ("How many alcoholic beverages do you typically drink in a week?")[4]. Nevertheless, we translated all nine risky behavior questions and used them in the procedure. We also slightly changed the question about public exposition during classes because a more representative sample took part in this study. We asked, "How often do you express your opinion in public?" instead of "How often do you raise your hand to answer or ask questions in class?".

Sixty-five females and 133 males participated in the replication study. Because four people declared another gender, they were excluded from the analyses (this eligibility criterion was declared in the pre-registration form). First, we performed t-tests for pre-registered effects[5]. We found that males engaged more often in gambling but not in risky recreational activities or frequently drinking alcoholic beverages (see Table 1). Thus the effects were partially replicated.

Additionally, we performed exploratory analyses for all other risk behavior questions (Table 1). We found that males declared that they engage in fast driving or breaking traffic laws more often than females. Moreover, they also had a higher tendency to drink heavily. Surprisingly, we found that females declared that they took more risk in a social domain: they got into arguments or expressed their opinion in public more often than males.

[4] The effect for this question was pre-registered, nevertheless because we did not achieve required sample size (N = 308) the analyses for this particular item could be underpowered.

[5] We decided to use t-tests similarly as in the original research. Nevetheless, because we found the response scale ordinal and there was a large disproportion between males and females in the replication study, we also checked for robustness of these effects using Mann-Whitney tests. These results yielded similar results to the t-tests: there was a significat difference between genders in gambling but not in frequent alcohol drinking and engaging in risky recreational activities.

Table 1. Gender differences in actual past risky behaviors

Risk Behavior Questions	Males *M* (*SD*)	Females *M* (*SD*)	Gender Difference	Pre-registered
Do you smoke?	1.59 (0.95)	1.46 (0.75)	$t(196) = -0.987$, $p = .325$, $d = -0.149$	no
How many alcoholic beverages do you typically drink in a week?	1.77 (0.86)	1.60 (0.66)	$t(196) = -1.379$, $p = .169$, $d = -0.209$	yes, but underpowered
How often have you had too much to drink or gotten drunk?	2.38 (1.20)	1.95 (1.07)	$t(196) = -2.445$, $p = .015$, $d = -0.370$	no
How often do you drive over the speed limit?	2.19 (1.19)	1.68 (1.02)	$t(196) = -2.963$, $p = .003$, $d = -0.448$	no
How often do you "bend" or break traffic laws?	2.39 (1.06)	1.92 (0.99)	$t(196) = -2.985$, $p = .003$, $d = -0.452$	no
How often do you gamble?	1.72 (0.98)	1.26 (0.57)	$t(196) = -3.504$, $p < .001$, $d = -0.531$	yes
How often do you engage in risky recreational activities?	1.75 (1.14)	1.51 (0.77)	$t(196) = -1.555$, $p = .122$, $d = -0.235$	yes
How often do you get into arguments with friends or family?	2.31 (0.89)	2.57 (0.85)	$t(196) = 1.969$, $p = .050$, $d = 0.298$	no
How often do you express your opinion in public?	2.88 (1.04)	3.20 (0.85)	$t(196) = 2.158$, $p = .032$, $d = 0.327$	no

Subjective numeracy and choices under risk (Traczyk et al., 2018) (S05)

In this study, 133 participants made twelve high-payoff choices and twelve low-payoff choices in binary two-outcome gambles framed as gains. Choice problems differed in their expected value (EV) ratios. When the EV ratio was high, choice problems could be regarded as meaningful because selecting an option with a higher EV could lead to a higher payoff. When the EV ratio was low, choice problems could be regarded as trivial because, irrespective of a chosen option, playing them repeatedly would lead to relatively small differences in payoffs. Individual differences in objective numeracy, subjective numeracy, fluid intelligence, and the need for cognition were measured to investigate their relationships with choices under risk and adaptive strategy selection.

Results showed that objective numeracy was related to adaptive strategy selection. That is, participants with higher objective numeracy, in comparison to participants with lower objective numeracy, maximized EV only when choice problems were meaningful (i.e., they were associated with high payoffs). When choice problems were trivial (i.e., choosing the option with a higher EV did not result in a large payoff), they switched to a heuristic strategy. Recently, this effect has been replicated successfully (Mondal, 2021). In contrast, people with higher subjective numeracy maximized EV in choice problems irrespective of the payoff. Subjective numeracy was related to more choices maximizing EV in both low-payoff problems (r = .200, p = .021, d = 0.408) and high-payoff problems (r = .323, p < .001, d = 0.683).

In the current study (N = 202), we wanted to replicate the effect of subjective numeracy in choices under risk. Participants completed a subjective numeracy test and made decisions in 24 choice problems. In contrast to the original study, an objective numeracy test, fluid intelligence test, and the need for cognition scale were not administered. We found that subjective numeracy was significantly related to the number of choices maximizing EV in high-payoff problems (r = .195, p = .005, d = 0.398) but it became non-significant in low-payoff problems (r = .086, p = .221, d = 0.173). Nevertheless, these two correlations did not differ significantly (Z = 1.558, p = .119). The effect was partially replicated and smaller than in the original study.

Framing of moral choices (Parkinson & Byrne, 2017) (S06)

This research focused on testing the moral echoing effect, which is a tendency to praise the decision-maker for a good outcome when a sure option is chosen but to blame the decision-maker for a bad outcome after a risky choice in a loss frame. Nevertheless, in the present study, we only aimed to replicate the effect of choice framing (gain vs. loss) on the moral acceptability of risky vs. safe options.

In the original study, a modified Asian Disease Problem was used. Two hundred and seven participants were randomized across a 2(a decision-maker's choice: sure vs. risk) x 2(frame: gain vs. loss) x 2(hypothetical outcome: good vs. bad) between-subjects experimental design. They were asked to indicate, among other things, whether the decision-maker (i.e., John) is morally responsible for people dying/surviving. Moreover, before the main task, participants were asked whether each choice (safe vs. risky) was morally acceptable using a 5-point scale (1–completely disagree, 5–completely agree). This was a control analysis to check whether participants tended to make the typical responses observed in framing studies (prefer a sure option in a gain frame and a risky option in a loss frame). Still, this effect was a target for our replication.

In the original research, a 2(frame: gain vs. loss; between-subjects) x 2(choice: sure vs risky option; within-subjects) ANOVA was conducted. The authors observed a main effect of choice, F(1, 199) = 4.11, p < .05, η_p^2 =.02, main effect of frame, F(1, 199) = 19.37, p < .001, η_p^2 = .09, and an interaction of choice and frame, F(1, 199) = 24.49, p < .001, η_p^2 = .11. The participants perceived risk choices as more morally acceptable in the loss frame, F(1, 96) = 32.31, p < .001, η_p^2=.25. Choosing the sure option was perceived as more morally acceptable in the gain frame than in the loss, F(1, 205) = 47.06, p < .001, η_p^2=.19.

Because our replication study was conducted during the COVID-19 pandemic, we decided to change the original "Asian Disease" into "novel and unknown disease in Ireland". All other details remained the same as the original scenario. Participants were randomly assigned to the gain vs. loss conditions (n_{gain} = 93; n_{loss} = 109). The results of our study showed a main effect of frame F(1, 200) = 8.12, p = .005, η_p^2 = .04, but not of choice F(1, 200) = 2.53, p = .113, η_p^2 = .01. However, we observed an interaction of choice and frame, F(1, 200) = 9.98, p = .003, η_p^2 = .04. Post-hoc tests with Bonferroni correction revealed that people perceived choosing a sure option in a gain frame as most morally acceptable (M = 3.63, SD = 1.06) while choosing this option in a loss frame was perceived as least morally acceptable (M = 3.00, SD = 1.01), t = 4.115, p < .001. Nevertheless, we did not observe a significant difference in moral acceptability of risky options depending on the frame (M_{gain} = 3.15, SD_{gain} = 1.03, M_{loss} = 3.15, SD_{loss} = 1.09), t = 0.024, p = 1.00. Generally, all the significant differences were found only between ratings of a safe option in a gain frame and other conditions (p = .013 for choosing a risky option in the context of gains; p = .010 for choosing a risky option in the context of losses). Thus, we argue that the effect of framing was only partially replicated.

Openness-to-experience influences response to the anchoring effect (McElroy & Dowd, 2007) (C11)

In this study, 197 undergraduate students completed the TIPI scale (a 10-item personality inventory based on the Big Five personality traits) and a traditional anchoring task. After completing the questionnaire, participants were asked to assess if the length of the Mississippi River is longer or shorter than 200 miles (low anchor condition) or 20,000 miles (high anchor condition). Then, they estimated the exact length of the Mississippi River.

Researchers performed a regression analysis with participants' openness-to-experience and anchor (low or high) as independent variables. Assessment of the river's length was treated as a dependent variable. The analysis showed a significant interaction between participants' openness-to-experience and the anchor (low vs. high), $F(1, 191) = 7.72$, $p < .007$. Higher levels of openness-to-experience were related to the anchoring effect. In the high anchor condition, higher openness-to-experience was associated with higher estimates, $F(1, 95) = 4.9$, $p = .03$. In the low anchor condition, higher openness-to-experience was associated with lower estimates: $F(1, 96) = 11.25$, $p = .002$.

In our replication, we used a Polish version of the TIPI scale (Sorokowska, et al., 2014) as a measure of openness-to-experience. Participants were asked to estimate the length of the Odra River (the second-longest Polish river). We calibrated anchors proportionally to the anchors from the original study. Additionally, miles were changed to kilometers.

To test the interaction effect, we performed a hierarchical regression analysis after exclusion of 2.5% of the highest and the lowest Odra estimations (because of outliers) and mean-centering of predictors ($n_{\text{low anchor}} = 105$, $n_{\text{high anchor}} = 87$). In the first model, we included openness-to-experience and anchors as predictors. The model was statistically significant: $F(2, 189) = 3.74$, $p = .026$. A high anchor led to higher estimates of the river's length, $b = 454.25$, $t = 2.490$, $p = .014$. There was no effect of openness-to-experience, $b = 34.95$, $t = 1.188$, $p = .235$. In the second model, we added an interaction term between openness-to-experience and the anchor as a third predictor. The change in model fit was statistically significant (R^2 change from 3.81% to 6.82%), $F(1, 188) = 6.09$, $p = .015$. Openness-to-experience moderated the effect of anchoring on numerical estimates, $b = -186.96$, $t = -2.467$, $p = .015$[6]. The effect was successfully replicated.

Self-control's impact on financial behavior (Strömbäck et al., 2017) (C12)

This study concerns the relationship between self-control and financial outcomes, such as financial behaviors and subjective financial well-being. A representative group of the Swedish population ($N = 2063$) completed the survey, including measures of financial behavior, well-being, self-control, deliberative thinking, and optimism.

The sample was split into two groups: high and low self-control based on the median level of all items from a shorter version of the Brief Self-Control Scale (Tangney et al., 2004) and the four items from the Short-Term Future Orientation Scale (Antonides et al., 2011). Two groups were compared in terms of self-reported financial behavior, measured as a mean score from the first twelve items (with a 5-point scale) of the Financial Management Behavior Scale (FMBS, Dew & Xiao, 2011). The mean FMBS score in the low self-control group was $M = 3.27$, and $M = 3.61$ in the high self-control group. The difference between groups was statistically significant, $t(2061) = -12.338$, $p < .001$.

In the replication study, we also used the first 12 items from FMBS to measure financial behaviors. We used five items from the Brief Self-Control Scale and four items from the Short-Term Future Orientation Scale to measure self-control. Items were translated into Polish by researchers. We split the sample at the median level of self-control and compared financial behavior between groups. Participants with high self-control ($M = 3.20$, $SD = 0.43$, $n = 97$) had higher FMBS scores than participants with low self-control ($M = 3.01$, $SD = 0.50$, $n = 105$). The difference was statistically significant, $t(200) = 2.826$, $p = .005$, $d = 0.398$. The effect was replicated.

Focusing on emotions or information in health-related decision dilemmas (Mikels et al., 2010) (C13)

In this study, 60 younger and 60 older adults were divided into three groups: focus on emotions, focus on information, and a control group. In each condition, they made decisions regarding health care plans. In the emotion-focus condition, participants were asked to focus on their emotional responses during the presentation of health care plans. In the information-focus condition, they were asked to focus on the details of each presented plan. In the control group, participants did not receive any instruction. The dependent variable was decision quality operationalized as the percentage of superior choices in the decision task.

There was a significant age group by condition interaction, $F(2, 114) = 6.83$, $p < .005$, $\eta_p^2 = .11$. For younger adults, the best decision quality was in the information-focus condition ($M = 84.4$, $SD = 8.2$). The difference between that group and the control condition ($M = 77.8$, $SD = 11.9$) was statistically significant, $t(38) = 2.08$, $p < .05$. Simultaneously, the difference between information-focus and emotions-focus ($M = 80.9$, $SD = 8.5$) groups was not statistically significant, $t(38) = 1.30$, $p > .20$, as well as the difference between emotions-focus and control groups, $t(38) = 1.11$, $p > .25$. For older adults, performance in the control group ($M = 79.9$, $SD = 9.0$) and emotion-focus condition ($M = 77.7$, $SD = 10.5$) was better than in the information-focus condition ($M = 70.1$, $SD = 11.2$), $t(38)$, $p < .005$, and $t(38) = 2.21$, $p < .05$, respectively.

In our study, we tried to replicate the effect on young adults. In contrast to the original study, older adults were not included in the sample, we did not verify cognitive functions at the beginning of the study, and there were no additional tasks except for healthcare plan dilemmas.

The difference between the information-focus condition ($M = 81.3$, $SD = 19.0$, $n = 63$) and control group ($M = 84.3$, $SD = 19.6$, $n = 70$), which we wanted to replicate, was not statistically significant, $t(199) = -0.905$, $p = .366$. Differences between the emotion-focus condition ($M = 85.5$, $SD = 17.4$, $n = 69$) and control group as well as emotion-focus and information-focus condition were not

[6] Analysis based on the full dataset (including outliers) revealed marginally significant effect, $b = -240.75$, $t = -1.914$, $p = .057$

significant, $t(199) = 0.385$, $p = .700$, and $t(199) = 1.277$, $p = .203$, respectively. The effect was not replicated, $F(2, 199) = 0.854$, $p = .427$, $\eta_p^2 = .009$.

Is a fewer amount of dinner plates better than more? (Hsee, 1998) (C14)

The less-is-better effect emerges when a less valuable option is more preferred than the objectively better option. In the original study, participants (104 college students) were asked to evaluate two dinnerware sets. The first set (set H) included every item from the second set (set L), but it also included additional pieces that were partly broken. The experimental manipulation was related to the way in which the two sets were presented: jointly (evaluation of H and L sets together) or separately (evaluation of H or L).

In the separate evaluation condition, set L was evaluated as more expensive (M = 32.69$) than set H ($M$ = 23.25$), $t = 3.91$, $p < .001$. In the case of the joint evaluation condition, the effect was opposite. That is, set L was perceived as less expensive (M = 29.70$) than set H ($M$ = 32.03$), $t = 2.15$, $p < .05$.

In our replication, we changed the currency from USD to PLN, extended the price range (proportional to the original study), and adjusted the names of elements in each set. Compositions of sets were the same as in the original study (the same number of broken and unbroken dinnerware elements). In joint evaluation (n = 67), according to the original effect, set L was evaluated as cheaper (M = 274.64 PLN) in comparison to set H (M = 332.87 PLN). The difference was statistically significant, $t(66) = 6.209$, $p < .001$, $d = 0.759$. In the separate evaluation condition, also according to the original effect, set L was evaluated as more expensive (M = 299.90 PLN, n = 72) than set H (M = 293.90 PLN, n = 63), but the difference was not statistically significant, $t(133) = 0.570$, $p = .569$, $d = 0.098$. The effect was only partially replicated.

The framing effect in decision making (Tversky & Kahneman, 1981) (C21)

In the original study (problem 10 drawn from Tversky & Kahneman, 1981), participants were randomly assigned to one of the two experimental conditions. In the first condition, they were instructed to imagine that they were considering purchasing a jacket for $125 and a calculator for $15. In the second condition, participants were asked to imagine that they were considering purchasing a jacket for $15 and a calculator for $125. In each condition, participants were asked to decide if they wanted to save $5 and buy a cheaper calculator on sale, but it involved a 30-minute walk from the current shop. Depending on the condition, such a decision would result in purchasing the calculator for $10 or $120, respectively. Saving $5 was attractive for 68% of participants when the calculator's price was low ($15). In the condition with a high price ($125), saving $5 was attractive only for 29% of participants.

In our replication, we changed the currency (from USD into PLN; $15 was changed into 15 PLN, etc.) and items to purchase (a jacket and calculator were replaced by a shirt and belt). For the attention check, participants were asked to enumerate items to purchase that were presented in the first stage of the task.

Firstly, we performed an analysis on the full dataset (without excluding participants with incorrect answers provided in the attention check task). Saving 5 PLN was attractive for 19.4% of participants when the belt's price was low (15 PLN, n = 98). In the condition with a high price (125 PLN, n = 104), saving 5 PLN was attractive only for 11.5%. The difference between groups was not statistically significant: $\chi^2(1, N = 202) = 2.39$, $p = .122$, $\varphi = 0.11$. Next, we compared groups after excluding 93 participants with incorrect answers in the attention check task. Saving 5 PLN was attractive for 20% of participants when the belt's price was low (15 PLN, n = 50). In the condition with a high price (125 PLN, n = 59), saving 5 PLN was attractive only for 13.5%. The difference between groups was not statistically significant: $\chi^2(1, N = 109) = 0.81$, $p = .367$, $\varphi = 0.09$. The effect was not replicated.

Students' assessment as a result of framing (Peters et al., 2006) (C22)

In this study, 100 participants were asked to assess students' quality of work using a 7-point scale (from -3 to +3). In the positive framing condition, students' results were framed as a percentage of correct answers. In the negative framing condition, scores were framed as a percentage of incorrect answers. For instance, the same student was described as a person who provided either 76% of correct answers or 24% of incorrect answers. Mean ratings in the positive framing condition were higher (M = 0.7) than in the negative framing condition (M = -0.1). The difference was statistically significant, $F(1, 96) = 26.3$, $p < .001$, $\eta^2 = 0.54$.

In our replication, participants were randomly assigned to one of the two conditions (i.e., positive vs. negative framing) and were asked to assess students' quality of work using the same scale as in the original study. In contrast to the original study, we did not include numerical skills as a moderator variable. The difference between conditions in our replication study was statistically significant, $F(1, 200) = 33.58$, $p < .001$, $\eta^2 = .144$. The positive frame led to higher ratings of the quality of students' work (M = 1.18, SD = 0.54, n = 103) as compared to the negative frame (M = 0.69, SD = 0.67, n = 99). The effect was successfully replicated.

Is a cheaper gift better? (Hsee, 1998) (C23)

In this study, 83 students were asked to imagine that they had received a gift from a friend. In the first scenario, it was a wool coat that cost $55 (with prices in the shop ranging from $50 to 500). In the second scenario, it was a wool scarf that cost $45 (with prices in the shop ranging from $5 to 50). Participants were asked to rate how generous that friend was using a 6-point scale.

Even though the gift was more expensive in the first scenario than the gift presented in the second scenario, participants rated that the friend described in the first scenario was less generous (M = 5.00) than the same

person described in the second scenario (M = 5.53). The difference was statistically significant, t = 3.13, p < .01.

We conducted replication with a change in currency (PLN instead of USD; we used 15 PLN, 130 PLN, 150 PLN, 200 PLN, and 600 PLN to replace $5, $45, $50, $55, and $500, respectively). The friend was rated as less generous in the first scenario (M = 4.80, SD = 1.15, n = 89) than in the second one (M = 5.44, SD = 0.91, n = 113), t(200) = -4.458, p < .001, d = -0.632. The effect was replicated.

Anchoring in estimation tasks (Jacowitz & Kahneman, 1995) (C25)

In this study, 103 students answered three consecutive questions in each of 15 problems (for instance, assessing the height of Mount Everest or the number of the United Nations members). Firstly, they respond if the right answer is bigger or smaller than the anchor. Secondly, they estimated the exact quantity. Finally, participants were asked to indicate their level of confidence. Values of anchors were based on answers from the calibration group (N = 53; high anchor: 85th percentile in each problem, low anchor: 15th percentile in each problem).

Jacowitz and Kahneman used an anchoring index (AI) to check the difference between groups. They calculated it as a quotient of the difference between medians in two conditions and the difference between two anchors. In this sense, AI measures the movement of the median estimate provided by participants toward the anchor they were presented with. The value of AI = 0 informs that there was no anchoring effect, while the value of AI = 1 suggests that median estimates provided by anchored participants are equal to the shown anchors. AI can be calculated separately for the low and high anchors. The overall mean of the AI in the original 15 problems was 0.49 (0.51 for the high anchors and 0.49 for the low anchors), suggesting that the median estimation moved approximately halfway toward the anchor. The mean value of the point-biserial correlation between participants' estimates and anchors over the 15 problems was r = .42.

We used 15 questions that were similar to the original study and adapted to Polish settings. Results are presented in Table 2. The value of mean AI was similar to the original study (0.47) and bigger in the high anchor condition (0.61) than in the low anchor condition (0.27). The mean value of the point-biserial correlation between participants' estimates and anchors over the 15 problems we used in the replication study was r = .46, p < .001. In addition to the original study, we performed a Student's t-test for independent groups for all 15 items separately. The difference between groups was statistically significant in 10/15 cases. The effect was successfully replicated.

The effect of frequency program requirements on preferences between rewards (Kivetz & Simonson, 2002) (C26)

In this study, participants were asked to imagine two frequency programs. The first program offered a luxury reward (i.e., pampering Swedish massage or two tickets for a concert at the San Francisco Symphony), while the second program offered a necessity reward (i.e., credit toward a future grocery bill). They had to choose one preferable program. In two studies, researchers manipulated the level of effort participants must invest to obtain the reward. For example, in the first study, participants (N = 159) were informed that they could get a reward in a car rental loyalty program if they rented a car 10 times (low effort and requirements) or 20 times (high effort and requirements). In the second study (N = 294), they were informed that they would get a reward in a loyalty program in Macy's shops after accumulating either $1,000 or $2,000 of purchases.

In the frequent car renter program, 26% of participants preferred a luxury reward when requirements were lower. The proportion was higher (41%) when requirements were also higher. The difference was statistically significant, χ^2 = 4.1, p < .05. In the frequent Macy's shopper program, 34% of participants preferred a luxury reward when requirements were low. Again, the proportion was higher (45%) when requirements were higher, and the difference was statistically significant, χ^2 = 3.9, p < .05.

In the original study, participants were recruited at the airport. In the replication study, we collected data using an online questionnaire. Participants were presented with two programs: a loyalty program for a cinema's clients and a loyalty program for a clothing retailer's clients. In the first scenario, participants had two options as a reward to choose from massage or credit toward a future discount store bill. In the second scenario, they were offered a voucher for a theatre show ticket or a discount bill.

In the cinema loyalty program condition (n = 184; 18 participants did not select any reward), 33% of participants preferred a luxury reward when requirements were lower (10 purchased tickets, n = 94). The proportion decreased (to 23%) when requirements were higher (20 purchased tickets, n = 90), but the difference was not statistically significant, χ^2(1, N = 184) = 2.11, p = .146, φ = 0.11.

In the case of the clothing retailer's loyalty program (n = 201; 1 participant did not answer a question regarding reward), 40% of participants preferred a luxury reward when requirements were lower (accumulating 1000 PLN of purchases, n = 101). The proportion decreased (to 19%) when requirements were higher (accumulating 2000 PLN of purchases, n = 100). The difference was statistically significant, but in the opposite direction to the original study: χ^2(1, N = 201) = 10.29, p < .001, φ = 0.23. The effect was not replicated.

DISCUSSION

The main goal of this project was to replicate a series of studies that were originally used to test effects related to judgment and decision making. Altogether, we aimed to conceptually replicate 15 effects concerning such issues as framing, anchoring, risk propensity, decision-making styles, etc. In 12 out of 15 cases, we succeeded in replicating the original effects, either completely or

Table 2 Comparison between high and low anchor conditions for all items separately. AI – anchoring index.

		low anchor (n = 98)		high anchor (n = 104)					AI		
	Calibration median	***M***	***SD***	***M***	***SD***	***t***	***p***	***d***	***high***	***low***	**AI**
Length of Wisla River (in km)	900	825.93	265.28	1147.04	278.70	-8.38	<.001	1.18	0.21	0.00	0.13
Height of Rysy (in meters above sea level)	2450	2279.37	329.72	2405.38	325.95	-2.73	.007	0.38	0.96	-0.08	0.01
Amount of meat eaten per year by average person in Poland (in km)	50	58.10	86.44	213.55	977.22	-1.57	.118	0.22	0.77	0.40	0.65
Distance from Gdansk to Rzeszow (in km)	615	553.05	167.41	1926.07	12002.17	-1.13	.259	0.16	0.24	0.24	0.24
Height of tallest redwood (in m)	88	81.41	121.23	242.98	128.04	-9.19	<.001	1.30	0.70	0.46	0.65
Number of the United Nations members	95	121.58	116.84	168.39	43.77	-3.73	<.001	0.54	0.98	-0.12	0.52
Number of professors at the Jagiellonian University	70	71608.93	707088.59	363.54	674.78	1.00	.321	0.14	1.00	0.40	0.83
Population of Warsaw	1500000	2270208.15	2054378.95	1352283.65	1566610.79	1.93	.059	0.51	0.30	-0.40	0.00
Year telephone was invented	1895	1869.40	41.04	1895.10	27.77	-5.24	<.001	0.74	-0.21	0.79	0.29
Average number of babies born per day in Poland	1000	1564.00	4508.95	14248.48	66566.77	-1.93	.056	0.27	0.52	0.78	0.60
Maximum speed of house cat (in km per hour)	20	22.89	13.54	34.13	12.97	-6.02	<.001	0.85	0.46	0.00	0.34
Amount of gas used per month by average person in Poland (in liter)	60	52.78	47.55	134.62	115.18	-6.67	<.001	0.92	0.83	0.50	0.70
Number of restaurants in Warsaw	600	1164.08	1800.75	2767.44	2841.52	-4.76	<.001	0.67	0.92	0.22	0.78
Number of colleges and universities in Masovian Voivodeship	30	37.22	66.46	91.65	63.29	-5.95	<.001	0.84	0.71	0.50	0.67
Number of Lincoln's presidency	5	8.16	7.65	13.34	6.07	-5.30	<.001	0.75	0.82	0.33	0.71

partially (see Supplementary Table S1 https://osf.io/svge5 for the summary of replicated effects).

Among seven successfully replicated effects, we can find well-established JDM effects: the sunk cost (Garland, 1990), the anchoring (Jacowitz & Kahneman, 1995), the framing (Peters et al., 2006), and the "less-is-better" (Hsee, 1998) effects. Notably, in the case of these effects, our results are consistent with the previous replication studies. For example, the anchoring effect (Jacowitz & Kahneman, 1995) was successfully replicated in the Many Labs project (Klein et al., 2014) and the "less-is-better" effect in the Many Labs 2 project (Klein et al., 2018). In our study, other successfully replicated effects regarded the role of individual differences in decision making, such as the role of openness to experience in susceptibility to anchoring (McElroy & Dowd, 2007), the positive relationship between procrastination and dependent decision-making style (Geisler & Allwood, 2018), and the self-control's impact on financial behavior (Strömbäck et al., 2017). Importantly, even though there were substantial differences in research materials used in the original and replication studies, most effects were successfully replicated. For example, we used the length of the Odra River as an anchor (the second-longest Polish river) instead of the Mississippi River, and we changed miles to kilometers (McElroy & Dowd, 2007). In the replication of the "less-is-better" effect, we changed the currency (PLN instead of USD) and provided more ecologically valid prices for products (coat and scarf) by replacing \$5, \$45, \$50, \$55, and \$500 with 15 PLN, 130 PLN, 150 PLN, 200 PLN, and 600 PLN.

In the case of five of the effects, our replication success was only partial. We defined the partial success when the main effect was significant, but some of the post-

hoc tests were not significant or when the effect was significant only for some items/conditions. We observed such a pattern in the case of the following effect: priming of instrumental risk attitude (Zaleśkiewicz, 2008), the effect of gender on risk taking (Harris et al., 2006), the relationship between subjective numeracy and risky choices (Traczyk et al., 2018), and the effect framing on the moral acceptability of risky vs. safe option (Parkinson & Byrne, 2017). Interestingly, the second operationalization of the "less-is-better" was only partially successful (Hsee, 1998): we found a significant effect in the joint evaluation condition but not when the items were evaluated separately.

In three studies that were conducted in our project, we did not find effects that were consistent with the original results. The first study in this set concerned the relationship between decision quality and participants' focus on either information or emotion, as compared to the control condition (Mikels et al., 2010). In the original study, the authors found that younger participants made better decisions when they focused on information (compared to controls), but no such difference was observed between the emotion focus condition and the control condition. In our project, the differences between all pairs of conditions were insignificant. One potential explanation was that our replication was carried out during the COVID-19 pandemic, and the decision was related to health issues. In the pandemic, people often faced information that later turned out to be incorrect (Van Bavel et al., 2020) which may have reduced people's general trust in medical knowledge, and even instructing participants to focus on information may not positively impact the quality of their decisions.

The second study, without successful replication, was related to the framing effect in decision making (Tversky & Kahneman, 1981). In the original study, saving $5 was attractive for the majority of participants when the product's price was low ($15) but not when the product's price was high ($125). Here, the reason behind the lack of replication may be due to the fact that we had to make several changes in the experimental stimuli. First, we changed the currency from $ to PLN (Polish Zloty). Second, we used different products than Tversky and Kahneman (1981) did in their study. Nevertheless, in the Many Labs 2 project (Klein et al., 2018), a similar approach for replicating this study was employed. Klein et al., in consultation with one of the authors of the original study, adjusted the currency and replaced items (a ceramic vase and a wall hanging) as well. Still, their replication was successful, but the effect was much smaller than in original study.

Finally, the third study that we were not able to replicate concerned the effect of frequency program requirements on preferences between rewards (Kivetz & Simonson, 2002). In the original experiment, the authors found that higher required effort makes consumers prefer luxury rather than necessity rewards. We did not observe such results in our replication. In this case, the lack of replication may also be related to the change in experimental stimuli. It is possible that the psychological difference between the rewards we presented as either a luxury reward or a necessity reward was too small, and participants did not perceive both rewards in a way we expected. Additionally, the original study was conducted in an airport where there are more luxurious rewards such as first-class lounges, duty-free shops, and fancy sitting arrangements as compared to home settings. This factor could also be a reason for unsuccessful replication.

To summarize, we successfully replicated 80% of the original studies that were included in our project. The replication rate we found in our research was higher than the one reported by the Open Science Foundation (Open Science Collaboration, 2015), which is less than 50%. However, it resembled the results found in the Many Labs registered replication project (Klein et al., 2014), which replicated approximately 77% of the original effects.

Importantly, our aim was not to conduct direct or close replications but rather to test the generalizability of these effects to other contexts, as well as investigate their validity and possible boundary conditions (Fabrigar et al., 2020; Hüffmeier et al., 2016). Thus, our procedures differed from the originals in several ways (see Table S1). If replication differs from the original study in too many details, we cannot be sure that the lack of support for the original effect is not caused by such changes. Since our study was exploratory in nature, future research should directly aim to investigate distinct differences: such as currency or products. Such an approach would allow putting forward specific hypotheses regarding replication success. It would also be valuable to conduct the prediction market among experts before collecting data (e.g., Dreber et al., 2015).

Moreover, as Hüffmeier et al. (2016) suggested, to ensure the testing of reproducibility of the effect, development of theory, and practical relevance, researchers should aim to conduct a series of replication studies varying in a scale of differences from the original research (exact replication, close replication, constructive replication, conceptual replication in the laboratory, conceptual replication in the field).

In our project, despite the obvious differences between original research and replication attempts, such as language/translation, currency, and current (epidemic) context, we also faced challenges with obtaining original research materials. Often, the exact wording was not available in original manuscripts or supplementary materials. Moreover, in many cases, we were not able to precisely calculate effect sizes for power analyses. Thus, we estimated them based on the available information. We argue that more JDM researchers should adopt Open Science practices such as open materials and open data to make this research more reproducible. For example, even among very recent (2015-2020) articles published in a well-known journal (i.e., *Organizational Behavior and Human Decision Processes*; Impact Factor: 4.9), only 22% had available research materials, and 32% provided access to the dataset (Logg & Dorison, 2021).

Finally, how can JDM researchers improve the replication rate in their field? We argue that adopting other solutions developed by the Open Science movement—pre-registrations and Registered Reports—may help. In the pre-registrations (on websites such as osf.io or aspredicted.org), researchers could describe their effects of interest, hypotheses, method, sample size, and detailed plan of analysis before the study is conducted. This approach helps to delineate between confirmatory analyses and exploratory analyses. Confirmatory analyses allow for testing predictions/hypotheses, and exploratory analyses help generate predictions for future research. Generally, the number of pre-registered studies has sharply increased in the last few years, especially in psychology journals (such as *Psychological Science* or *Journal of Experimental Psychology: General*, Simmons et al., 2021). Nevertheless, in some fields, e.g., consumer research journals (*Journal of Consumer Research*, *Journal of Consumer Psychology*), the pre-registration rate is still low (Simmons et al., 2021). In JDM journals, when we look at *Organizational Behavior and Human Decision Processes*, only 9% of papers provided pre-registration (Logg & Dorison, 2021). Thus, to ensure fewer false-positive results, more JDM scholars should pre-register their studies.

Second, to improve the quality of research, scientists may consider submitting their works as Registered Reports (it is possible in many journals such as *Judgement and Decision Making* or *Psychological Science*). In this format, reviewers evaluate the quality of the introduction, method, and plan of analyses before the study is conducted. If authors follow the accepted plan, their article is published even when their results do not support the hypotheses. The study by Soderberg et al. (2021) showed that Registered Reports were assessed by experts as better in all criteria (from the rigor of methodology and results to novelty and creativity) when compared with other (non-registered) papers.

To summarize, most of the JDM effects chosen by psychology students were replicated in a Polish setting; even our research materials differed from originals in many ways (e.g., language, culture, currency), and a study was conducted in an exceptional case context (during COVID-19 pandemic). Moreover, we showed that it is possible to engage undergraduate students in Poland to promote Open Science and accumulate evidence for the generalizability and validity of various JDM effects.

AUTHOR CONTRIBUTIONS

Agata Sobkow (AS) and Jakub Traczyk (JT) coordinated the project and acquired funding. Undergraduate students, AS, and JT chose effects for replications, prepared pre-registration, designed research procedures, translated and adapted research materials, as well as analyzed the data. AS, JT, Marcin Surowski, Angelika Olszewska, and Tomasz Zaleskiewicz drafted and reviewed the manuscript. All the authors commented on, edited, and approved the submitted manuscript.

We would like to thank Supratik Mondal for his insightful comments.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study, (simplified) pre-registration forms, and research materials are openly available on the Open Science Framework platform at https://osf.io/6sq8p/.

REFERENCES

Antonides, G., Manon de Groot, I., & Fred van Raaij, W. (2011). Mental budgeting and the management of household finance. *Journal of Economic Psychology*, *32*(4), 546–555. https://doi.org/10.1016/j.joep.2011.04.001

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van 't Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*(1), 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Dahling, J. J., Whitaker, B. G., & Levy, P. E. (2008). The development and validation of a new Machiavellianism scale. *Journal of Management*, 216–219. https://doi.org/10.1177%2F0149206308318618

Dehaene, S. (1999). *The number sense: How the mind creates mathematics*. Oxford University Press.

Dew, J., & Xiao, J. J. (2011). The Financial Management Behavior Scale: Development and validation. *Journal of Financial Counseling and Planning*, *22*(1), 43–59.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PloS One*, *7*(1), e29081. https://doi.org/10.1371/journal.pone.0029081

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., ... & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences, 112*(50), 15343-15347. https://doi.org/10.1073/pnas.1516179112

Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A Validity-Based Framework for Understanding Replication in Psychology. *Personality and Social Psychology Review*, *24*(4), 316–344. https://doi.org/10.1177/1088868320931366

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fox, C. R., Erner, C., & Walters, D. J. (2015). Decision Under Risk: From the Field to the Lab and Back. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making* (pp. 43–88). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118468333

Garland, H. (1990). Throwing good money after bad: The effect of sunk costs on the decision to escalate commitment to an ongoing project. *Journal of Applied Psychology*, *75*(6), 728–731. https://doi.org/10.1037/0021-9010.75.6.728

Geisler, M., & Allwood, C. M. (2018). Relating Decision-Making Styles to Social Orientation and Time Approach. *Journal of Behavioral Decision Making*, *31*(3), 415–429. https://doi.org/10.1002/bdm.2066

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science*, *351*(6277), 1037–1037. https://doi.org/10.1126/science.aad7243

Gneezy, U., List, J. A., & Wu, G. (2006). The Uncertainty Effect: When a Risky Prospect is Valued Less than its Worst Possible Outcome. *The Quarterly Journal of Economics*, *121*(4), 1283–1309. https://doi.org/10.1093/qje/121.4.1283

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D.

T. D., Dewitte, S., … Zwienenberg, M. (2016). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Harris, C. R., Jenkins, M., & Glaser, D. (2006). Gender Differences in Risk Assessment: Why do Women Take Fewer Risks than Men ? *Judgment and Decision Making*, *1*(1), 48–63.

Hsee, C. K. (1998). Less is better: When low-value options are valued more highly than high-value options. *Journal of Behavioral Decision Making*, *11*(2), 107–121. https://doi.org/10.1002/(SICI)1099-0771(199806)11:2<107::AID-BDM292>3.0.CO;2-Y

Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology*, *66*, 81–92. https://doi.org/10.1016/j.jesp.2015.09.009

Jacowitz, K. E., & Kahneman, D. (1995). Measures of Anchoring in Estimation Tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161–1166. https://doi.org/10.1177/01461672952111004

Jekel, M., Fiedler, S., Allstadt Torras, R., Mischkowski, D., Dorrough, A. R., & Glöckner, A. (2020). How to Teach Open Science Principles in the Undergraduate Curriculum—The Hagen Cumulative Science Project. *Psychology Learning & Teaching*, *19*(1), 91–106. https://doi.org/10.1177/1475725719868149

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–292.

Kivetz, R., & Simonson, I. (2002). Earning the Right to Indulge: Effort as a Determinant of Customer Preferences toward Frequency Program Rewards. *Journal of Marketing Research*, *39*(2), 155–170. https://doi.org/10.1509/jmkr.39.2.155.19084

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Vega, D., Aveyard, M., Axt, J. R., Babalola, M., Bahník, Š., Barlow, F., Berkics, M. M., Bernstein, M. J., Berry, D., Bialobrzeska, O., Bocian, K., Brandt, M. J., Busching, R., … Zeng, Z. (2018). Many Labs 2: Investigating Variation in Replicability Across Sample and Setting. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/doi.org/10.1177/2515245918810225

Lay, C. H. (1986). At last, my research article on procrastination. *Journal of Research in Personality*, *20*(4), 474–495. https://doi.org/10.1016/0092-6566(86)90127-3

Lennox, R. D., & Wolfe, R. N. (1984). Revision of the Self-Monitoring Scale. *Journal of Personality and Social Psychology*, *46*(6), 1349–1364.

Logg, J. M., & Dorison, C. A. (2021). Pre-registration: Weighing costs and benefits for researchers. *Organizational Behavior and Human Decision Processes*, *167*, 18–27. https://doi.org/10.1016/j.obhdp.2021.05.006

Mondal, S. (2021). Sensitivity of numerate individuals to large asymmetry in outcomes: A registered replication of Traczyk et al. (2018). *Decyzje*, 35, 5-26. DOI 10.7206/DEC.1733-0092.150a

McElroy, T., & Dowd, K. (2007). Susceptibility to anchoring effects: How openness-to-experience influences responses to anchoring cues. *Judgment and Decision Making*, *2*(1), 48–53.

Mikels, J. A., Löckenhoff, C. E., Maglio, S. J., Carstensen, L. L., Goldstein, M. K., & Garber, A. (2010). Following your heart or your head: Focusing on emotions versus information differentially influences the decisions of younger and older adults. *Journal of Experimental Psychology: Applied*, *16*(1), 87–95. https://doi.org/10.1037/a0018500

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

Parkinson, M., & Byrne, R. M. J. (2017). Moral judgments of risky choices: A moral echoing effect. *Judgment and Decision Making*, *12* (3), 236–252.

Persson, E., Andersson, D., Koppel, L., Västfjäll, D., & Tinghög, G. (2021). A preregistered replication of motivated numeracy. *Cognition*, *214*, 104768. https://doi.org/10.1016/j.cognition.2021.104768

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and Decision Making. *Psychological Science*, *17*(5), 407–413. https://doi.org/10.1111/j.1467-9280 .2006.01720.x

Popper, K. (1934). *The Logic of Scientific Discovery*. Routledge.

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the Robustness of Power Posing: No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women. *Psychological Science*, *26*(5), 653–656. https://doi.org/10.1177/0956797614553946

Rottenstreich, Y., & Hsee, C. K. (2001). Money, kisses, and electric shocks: On the affective psychology of risk. *Psychological Science*, *12*(3), 185–190. https://doi.org/10.1111/1467-9280.00334

Ruggeri, K., Alí, S., Berge, M. L., Bertoldo, G., Bjørndal, L. D., Cortijos-Bernabeu, A., Davison, C., Demić, E., Esteban-Serna, C., Friedemann, M., Gibson, S. P., Jarke, H., Karakasheva, R., Khorrami, P. R., Kveder, J., Andersen, T. L., Lofthus, I. S., McGill, L., Nieto, A. E., … Folke, T. (2020). Replicating patterns of prospect theory for decision under risk. *Nature Human Behaviour*, *4*(6), 622–633. https://doi.org/10.1038/s41562-020-0886-x

Scott, S. G., & Bruce, R. A. (1995). Decision-making style: The development and assessment of a new measure. *Educational and Psychological Measurement*, *55*(5), 818–831. https://doi.org/10.1177/0013164495055005017

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2021). Pre-registration: Why and How. *Journal of Consumer Psychology*, *31*(1), 151–162. https://doi.org/10.1002/jcpy.1208

Simons, D. J. (2014). The Value of Direct Replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Soderberg, C. K., Errington, T. M., Schiavone, S. R., Bottesini, J., Thorn, F. S., Vazire, S., Esterling, K. M., & Nosek, B. A. (2021). Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour*, 1–8. https://doi.org/10.1038/s41562-021-01142-4

Sorokowska, A., Słowińska A., Zbieg A., & Sorokowski, P. (2014). Polska adaptacja testu Ten Item Personality Inventory (TIPI) – TIPI-PL – wersja standardowa i internetowa. Wrocław: WrocLab.

Stępień, M., & Topolewska, E. (2014). Style Tożsamości w Ujęciu Berzonsky'ego a Prokrastynacja. In *Młoda psychologia* (Vol. 2, pp. 145–159). Liberi Libri.

Strömbäck, C., Lind, T., Skagerlund, K., Västfjäll, D., & Tinghög, G. (2017). Does self-control predict financial behavior and financial well-being? *Journal of Behavioral and Experimental Finance*, *14*, 30–38. https://doi.org/10.1016/j.jbef.2017.04.002

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, *72*(2), 271–324. https://doi.org/10.1111/j.0022-3506.2004.00263.x

Traczyk, J., Sobkow, A., Fulawka, K., Kus, J., Petrova, D., & Garcia-Retamero, R. (2018). Numerate decision makers don't use more effortful strategies unless it pays: A process tracing investigation of skilled and adaptive strategy selection in risky decision making. *Judgment and Decision Making*, *13*(4), 372–381.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

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. https://doi.org/10.1007/BF00122574

Usunier, J.-C., & Valette-Florence, P. (2007). The Time Styles Scale: A review of developments and replications over 15 years. *Time Society*, *16*, 333–366. https://doi.org/10.1177/0961463X07080272

Van Bavel, J. J., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., Drury, J., Dube, O., Ellemers, N., Finkel, E. J., Fowler, J. H., Gelfand, M., Han, S., Haslam, S. A., Jetten, J., … Willer, R. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, *4*(5), 460–471. https://doi.org/10.1038/s41562-020-0884-z

Wu, G., Gonzalez, R., Zhang, J., & Gonzalez, R. (2008). Decision Under Risk. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making* (pp. 399–423). Blackwell Publishing.

Xiao, Q., Lam, C. S., Piara, M., & Feldman, G. (2021). Revisiting status quo bias: Replication of Samuelson and Zeckhauser (1988). *Meta-Psychology*, *5*. https://doi.org/10.15626/MP.2020.2470

Xiao, Q., Zeng, S., & Feldman, G. (2020). Revisiting the decoy effect: Replication and extension of Ariely and Wallsten (1995) and Connolly, Reb, and Kausel (2013). *Comprehensive Results in Social Psychology*, *4*(2), 164–198. https://doi.org/10.1080/23743603.2021.1878340

Yong, E. (2012). Nobel laureate challenges psychologists to clean up their act. *Nature*. https://doi.org/10.1038/nature.2012.11535

Zaleskiewicz, T. (2008). Samoregulacja i podejmowanie ryzyka. Rola procesów automatycznych. *Czasopismo Psychologiczne*, *14*(2), 287–296.

Zhou, L., Liu, N., Liao, Y.-Q., & Li, A.-M. (2021). Risky choice framing with various problem descriptions: A replication and extension study. *Judgment and Decision Making*, *16*(2), 28.

Ziano, I., Li, J., Tsun, S. M., Lei, H. C., Kamath, A. A., Cheng, B. L., & Feldman, G. (2021). Revisiting "money illusion": Replication and extension of Shafir, Diamond, and Tversky (1997). *Journal of Economic Psychology*, *83*, 102349. https://doi.org/10.1016/j.joep.2020.102349