UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Zhu, M., & Feldman, G. (2023). Revisiting the Links Between Numeracy and Decision Making: Replication Registered Report of Peters et al. (2006) With an Extension Examining Confidence. *Collabra: Psychology*, *9*(1). https://doi.org/10.1525/collabra.77608

Social Psychology

Revisiting the Links Between Numeracy and Decision Making: Replication Registered Report of Peters et al. (2006) With an Extension Examining Confidence

Minrui Zhu[1], Gilad Feldman[1][a]

[1] Psychology, University of Hong Kong

Keywords: Numeracy, judgment and decision making, registered report, replication, framing effect, confidence

https://doi.org/10.1525/collabra.77608

Collabra: Psychology

Vol. 9, Issue 1, 2023

Numeracy is individuals' capacity to understand and process basic probability and numerical information required to make decisions. We conducted a Replication Registered Report of Peters et al. (2006) examining numeracy as a predictor of positive-negative framing effect (Study 1), frequency-percentage effect (Study 2), ratio effect (Study 3), and bets effect (Study 4). With an online US American Amazon Mechanical Turk sample (N = 860), our replication using the target's dichotomizing of the numeracy measure found support for the original findings regarding interactions between numeracy and three decision-making effects. Numeracy was associated with weaker framing effect (η^2_p = 0.01, 90% CI [0.00, 0.02]), weaker ratio bias (*Cramer's V* = 0.17, 95% CI [0.10, 0.24]), and stronger bets effect (η^2_p = 0.02, 90% CI [0.01, 0.04]), yet we found no support for the frequency-percentage effect (η^2_p = 0.00, 90% CI [0.00, 0.01]). However, we found support for associations with all four studies when treating numeracy as a continuous variable. We extended the replication to examine confidence, yet the results were mixed with support found for only three conditions (Study 1 positive framing condition: r = -0.11, 95% CI [-0.20, -0.02]; Study 3: r = 0.15, 95% CI [0.08, 0.21]; Study 4 no-loss bet condition: r = 0.10, 95% CI [0.01, 0.20]), suggesting a much weaker and more complex relationship than anticipated. Materials, data, and code are available on: https://osf.io/4hjck/.

Numeracy

Decisions involving numbers, math, and statistics are common, and people rely heavily on their ability to accurately interpret, think about, and act on them. Numeracy is defined as the individuals' capacity to understand and process basic probability and numerical information required to make decisions. Research by Peters (2012) demonstrated that numeracy is a predictor of behavior in judgment and decision-making tasks.

We embarked on a direct replication of Peters et al. (2006) with two primary goals. Our first goal was to conduct an independent replication of the associations between numeracy and four decision-making paradigms. Our second goal was to examine an extension regarding the role of numeric confidence (or, subjective numeracy).

We begin by introducing the literature on numeracy and various decision making biases examined in the chosen article for replication - Peters et al. (2006). We provide a brief overview of the decision-making paradigms, in relation to numeracy. We then discuss our chosen target article, summarize its hypotheses and findings, and introduce our extension on the relationship between confidence, numeracy, and decision-making.

Attribute framing and numeracy

Framing effect is a well-established phenomenon in psychology and behavioral economics, in which decisions are influenced by the way information is presented, such as variations in valence - positive versus negative framing (Tversky & Kahneman, 1985).

Attribute framing is a type of framing effect and relates to the labeling of a particular attribute of an object or an event. For instance, ground beef with 75%-25% meat-fat ratio could be presented as "75% lean" or "25% fat". Levin and Gaelth (1988) found that people evaluated beef under the "% lean" framing more positively than in the "% fat" framing. Such framing effects have received empirical support by many follow-up studies (e.g., Freling et al., 2014; Piñon & Gambara, 2005).

[a] Corresponding author: Gilad Feldman, Department of Psychology, University of Hong Kong, Hong Kong SAR; gfeldman@hku.hk

PCIRR-Study Design Table

Question	Hypothesis	Analysis plan	Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis	Interpretation given different outcomes	Theory that could be shown wrong by the outcomes	**Observed outcome (Added in Stage 2)**
What is the relationship between numeracy and positive-negative framing effects	Higher numeracy is associated with weaker positive-negative framing effect	Mixed ANOVA Correlations	Our strategy for all replicated studies: 1. We keep the statistical method of the original paper as it treats numeracy as dichotomized. 2. We treat numeracy as a continuous variable, therefore adapt correlation.	Based on the criteria used by Lebel et al. (2019) We examine the replicability of the findings of Peters et al. (2006), and support for our suggested extensions.	Attribute framing effect	Numeracy was associated with weaker framing effect (η^2_p = 0.01, 90% CI [0.00, 0.02])
What is the relationship between numeracy and percentage and frequency effects	Higher numeracy is associated with weaker frequency-percentage effects	Factorial ANOVA Correlations			Frequency-percentage framing effect	We found mixed support for numeracy association with frequency-percentage effect. No support in replication using dichotomy (η^2_p = 0.00, 90% CI [0.00, 0.01]), but supported in an extension using continuous.
What is the relationship between numeracy and ratio bias What is the relationship between numeracy and affect precision	Higher numeracy is associated with more optimal choices in competing affective decisions. Higher numeracy is associated with higher affective precision, in competing affective decisions.	Chi-square test Independent t-test Correlations			Deliberate-experiential thinking modes Ratio bias	Numeracy was associated with weaker ratio bias (*Cramer's V* = 0.17, 95% CI [0.10, 0.24])
What is the relationship between numeracy and affective precision and affect in probabilities and numerical comparisons	Higher numeracy is associated with higher affective precision in probabilities and numerical comparisons. Higher numeracy is associated with greater affect in probabilities and numerical comparisons.	Factorial ANOVA Independent t-test Correlations			The highly numerate will focus more on details of numbers and draw more affective meanings. Bets effect	Numeracy was associated with stronger bets effect (η^2_p = 0.02, 90% CI [0.01, 0.04])
What is the relationship between objective numeracy and confidence under specific conditions?	The highly numerate is related to higher subjective confidence	Correlations			Associations with subjective confidence and objective numeracy	Mixed weak results.

Note. For the sampling plan please see power analysis in the methods section.

Attribute framing is related to how people understand and process numerical concepts, suggesting a possible link between numeracy and framing effects. Some studies found that less numerate people were more susceptible to framing effects (including attribute framing) (Choi et al., 2011; Gamliel et al., 2016; Gamliel & Kreiner, 2017). For instance, Gamliel and Kreiner (2017) demonstrated the relationship between numeracy and attribute framing bias: students with lower numeracy rated a university course higher if presented with success rates compared to failure rates. They suggested that decision makers with lower numeracy rely more heavily on "non-numerical information", whereas those with high numeracy pay more attention to numerical information and attain greater accuracy with numbers. Therefore, lower numeracy may be associated with stronger polarization due to the positive or negative valence of framing presentations.

Peters (2012) suggested that highly numerate individuals have the capacity to go beyond the specific numerical information and understand underlying relational information. For example, when a positive outcome is presented as 75% success rate, the highly numerate are more likely to also infer the complementary proportion of the failure rate of 25%, with similar logic for when the failure rates are presented and success rates are inferred. Therefore, the argument in relation to numeracy was that framing bias is attenuated when one is capable of grasping and processing both the positive and the negative information in a decision.

Frequency-percentage effect and numeracy

Frequency-percentage effect (or "frequency effect") is the phenomenon that decision making changes when the numbers are presented in forms of frequency (e.g., 10 out of 100) compared to percentage (e.g., 10%) (Gigerenzer, 1991; Hill & Brase, 2012).

Those higher on numeracy seem less likely affected by whether the number is represented in frequency or in percentage (Dickert et al., 2011; Hill & Brase, 2012; Peters et al., 2011). For instance, Peters et al. (2011) tested the relationship among patients. They informed patients of the side-effects of a medication in either frequency or percentage formats (i.e., 10 out 100 versus 10%) and then asked them to rate risk levels. They found that the less numerate were more likely to perceive the medication as less risky when presented in percentage format than in a frequency format. The possible mechanisms could be similar to those we previously discussed regarding the framing effect. Those higher on numeracy may be able to better understand the frequency and probability information as the same mathematical quantity (Hill & Brase, 2012; Peters, 2012).

Ratio bias and numeracy

Ratio bias (or numerosity effect) is the phenomenon that people tend to focus on absolute numbers rather than on probabilities (Peters et al., 2008; Reyna et al., 2009). For example, people are more likely to choose to select from a sample with a relatively large numerator/large denominator (e.g., 9 in 100) rather than the preferred odds yet relatively smaller numerator/small denominator (e.g., 1 in 10).

Reyna and Brainerd (2008) separated ratio bias into a heuristic ratio bias (i.e., identical probabilities in the two samples) and a non-optimal ratio bias (i.e., higher probabilities but smaller absolute numerator or lower probabilities but greater absolute numerator). One classic heuristic ratio bias example was from the study of Miller et al. (1989). Children randomly choose a cookie from one of two cookie jars, one containing 1 chocolate chip and 19 oatmeal cookies and the other containing 10 chocolate chips and 190 oatmeal cookies. The probabilities of having a chocolate chip are the same, yet Miller et al. (1989) found that children preferred to choose from the later one, with the larger numbers.

Peters et al. (2006) demonstrated that lower numeracy was associated with less optimal choices in ratio related decisions.

Affect, bets effect, and numeracy

Two modes of processing information appear to be affective-experiential and deliberative and are also known as the dual process model (Kahneman, 2003; Sloman, 1996). The model suggests that affective-experiential mode produces thoughts and feelings in a relatively effortless and spontaneous manner, whereas deliberative mode requires conscious reason-based and analytical thinking. Affect may provide information about the goodness and badness of an option and might as a consequence influence further choice processes.

Numeracy has been argued as moderating the association between affect and decision-making (Rottenstreich & Hsee, 2001; Traczyk & Fulawka, 2016), with the potential of aiding decision making, yet may sometimes lead to number overuse and worsen decisions (Pachur & Galesic, 2013; Peters & Bjalkebring, 2015). Those with higher numeracy seem to draw more precise affective information, then form relevant risk perception, and use that information in making related decisions.

In a demonstration of the possible advantages, Petrova et al. (2014) conducted a study about decision making regarding camera insurance. They found that participants with higher numeracy reported greater negative emotions to 90% chance of losing camera compared to 50% chance. In addition, they were willing to pay more on insurance against the loss when the loss probability was higher. By contrast, participants with lower numeracy seemed less sensitive to the two probabilities levels.

However, there are possible nuances and unintended side-effects to drawing precise numerical information, depending on the defined desired outcome. For example, Kleber et al. (2013) conducted research on donations and they found that numeracy was associated with donation behavior, with the more numerate focusing on projects with the greatest proportion of recipients, whereas those lower on numeracy tended to donate more with increases in both the number of recipients and the total number of people in need.

Choice of study for replication: Peters et al. (2006)

We chose the article by Peters et al. (2006) as the target for replication based on the following factors: its impact and potential for improvement on methodological limitations in the original studies.

The article has had much impact on scholarly research in the area of social psychology and judgment and decision making. At the time of writing (July, 2022), there were over 1400 Google Scholar citations of the article. In addition, Peters et al. (2006)'s work had important practical implications especially in the domains of medical decision making (Okamoto et al., 2012; Reyna et al., 2009) and financial decision making (Estrada-Mejía et al., 2016; Traczyk et al., 2018).

We reached out to the authors to request assistance with the original materials, and to try and assess any published and ongoing replication work. They indicated most of the original materials have been lost to time, yet kindly referred us to some of the extensive follow-up literature with conceptual replications and related materials, from which we were able to reconstruct most of the studies. We have also learned of other attempts at a replication of the broader numeracy literature in other languages (Polish) and have been in touch with their authors to coordinate efforts. To our knowledge, there are no published direct close replications of the target article's studies.

Examining the studies, we believe a direct replication is especially relevant given the low power and some of the statistical method choices. Their Studies 1-4 had 100, 46, 46, and 171 participants, respectively, which may seem low, especially given the interaction and supplementary analyses. Furthermore, the methods employed dichotomizing of the continuous numeracy scale, which we thought could be improved by analyzing as the intended continuous measure, and may allow for more accurate insights and conclusions.

We therefore aimed to revisit the classic phenomenon to examine the reproducibility and replicability of the findings with independent replications. We followed recent growing recognition of the importance of reproducibility and replicability in psychological science (e.g., Brandt et al., 2014; Open Science Collaboration, 2015; van't Veer & Giner-Sorolla, 2016; Zwaan et al., 2018) and embarked on a well-powered pre-registered very close replication of Peters et al. (2006).

Hypotheses and findings in target article

Peters et al. (2006) conducted four studies and we aimed to replicate all of them with needed adjustments and collected in a single data collection, with the experiments displayed in a random order (more on that in the methods section). Below we review the findings in each of the target's studies. We summarized the target's hypotheses and our extension's hypothesis in Table 1.

Study 1: Numeracy and Positive-negative Framing

Study 1 sought to examine the relationship between numeracy and attribute framing. They hypothesized that participants with low numeracy are more likely to be affected by attribute framing.

To test this, they recruited participants through campus newspapers. Participants first answered the numeracy scale developed by Lipkus et al. (2001). Then, they rated the quality of five psychology students' work. Participants were randomly assigned to positive or negative framing conditions. For instance, Emily received either 74% correct or 26% incorrect on her exam.

Peters et al. (2006) dichotomized numeracy to high numerate (9-11 items correct) and low numerate (2-8 items correct) with a median split. To test the hypothesis, they used a mixed ANOVA. They reported that higher numerate participants were less susceptible for framing bias (f = 0.25, 90% CI [0.00, 0.42]).

Study 2: Numeracy and Frequency-percentage Effect

Study 2 aimed to examine the relationship between numeracy and percentage-frequency framing effect. They hypothesized that participants with low numeracy are more likely to be affected by frequency-percentage effect. To test this, they recruited university students from a psychology course. Participants read the mental-patient scenario in either a frequency or percentage format and rate that the risk level of that patient who would harm someone. They ran a factorial ANOVA and found that low numerate rated lower in the percentage condition than frequency condition whereas the high numerate rated both conditions similarly (f = 0.31, 90% CI [0.00, 0.58]).

Study 3: Numeracy, Affect, and Ratio Bias

Study 3 intended to explore the association between numeracy and ratio bias as well as numeracy and the influence of affective information. They hypothesized that numeracy is associated with more optimal choices, evoking less affect and higher affective precision.

To test this, they recruited university students from a psychology course. Participants from Studies 2 and 3 were the same group. Participants read about a choice between two bowls, Bowl A-9-100 with affectively appealing description but less objectively favorable outcome (9 jellybeans of a bowl of 100) and Bowl B-1-10 with less appealing description but better results (1 jellybean of a bowl of 10). Participants rated their preference for a bowl and selected one. After indicating the preference and choice, they rated affect towards the Bowl A-9-100 option.

The authors used a chi-square test to examine participants' choices of two bowls and found that the less numerate were more likely to choose Bowl A-9-100 (φ = 0.77) and that the highly numerate showed higher preference for Bowl B-1-10 (d = -0.74, 95% CI [-1.33, -0.13]). In addition, the high numerate reported higher affect precision towards Bowl A-9-100 (d = 0.78, 95% CI [0.17, 1.36]). The study reported no support for differences in feelings (d = 0.46).

Table 1. Summary of hypotheses of replication and extension

Study	Hypothesis	Description of hypothesis
1	1(original)	The less numerate show a stronger framing effect than the highly numerate.
	1 (extension)	Higher numeracy is associated with weaker positive-negative framing effects.
2	1 (original)	The less numerate are affected more by the frequency-percentage effect than the highly numerate.
	1 (extension)	Higher numeracy is associated with weaker frequency-percentage effects.
3	1 (original)	The less numerate make more sub-optimal choices in competing affective decisions than the highly numerate.
	1 (extension)	Higher numeracy is associated with more optimal choices in competing affective decisions..
	2 (original)	The less numerate have the more positive affect about the affectively appealing bowl with less favorable objective probabilities in competing affective decisions than the highly numerate.
	2 (extension)	Higher numeracy is associated with more negative affect about the affectively appealing bowl with less favorable objective probabilities.
	3 (original)	The less numerate have lower affective precision about the affectively appealing bowl with less favorable objective probabilities than the highly numerate.
	3 (extension)	Higher numeracy is associated with higher affective precision about the affectively appealing bowl with less favorable objective probabilities.
4	1 (original)	The less numerate show smaller difference of rating of bets than the highly numerate.
	1 (extension)	Higher numeracy is associated with larger differences in the rating of bets.
	2 (original)	The less numerate draw less affective meaning in probabilities and numerical comparisons than the highly numerate.
	2 (extension)	Higher numeracy is associated with drawing more affective meaning in probabilities and numerical comparisons.
Extension: Confidence		
1, 2, 3, 4	1	Numeracy is positively associated with confidence.

Note. For each of the hypotheses we reframed the hypotheses deduced from the conclusions in the original article from a dichotomy (high numerate versus low numerate, labeled as “original”) to a continuous association (higher numeracy is associated with..., labeled as “extension”).

Study 4: Numeracy, Affect, and Bets Effect

Study 4 examined the relationship between numeracy and affect in probabilities and numerical comparisons. They hypothesized that numeracy is associated with affect arousal and affective precision.

To test this, they recruited volunteers from a subject’s pool of psychology department. Participants read the scenario about a bet with 7/36 chance to win $9 and 29/36 chance to win nothing or a bet with 7/36 chance to win $9 but 29/36 chance to lose 5 cents. The possible bets were visualized with a roulette wheel. Participants evaluated the attractiveness of the bet and their affect precision and affect, using the same scales as in Study 3.

They employed a factorial ANOVA and an independent samples *t*-test. They found that those high on numeracy rated the loss bet as more attractive in loss bet condition, whereas participants with low numeracy rated two conditions the same on average (*f* = 0.23, 90% [0.10, 0.35]). With respect to affect precision, participants with high numeracy had clearer feelings about the bets than those with low numeracy. The high numerate also reported more positive affect in the loss condition than in the no-loss condition, whereas there were weaker differences for the low numerate (*f* = 0.20, 90% [0.10,0.50]). Peters (2020) summarized such findings as “bets effect” in her book and we therefore also use this term.

We summarized the findings in the target article in Table 2.

Extension: Numeracy as a Continuous Measure

We added analyses to treat numeracy as a continuous variable instead of the dichotomy used in the target article. Methodologists have increasingly expressed concerns regarding the dichotomization of continuous variables as it might result in suboptimal interpretations (Altman & Royston, 2006; Fedorov et al., 2009; Lazic, 2018; Marimooryad & Busso, 2015). One of the primary limitations is the loss of information, and treating samples within the same group as having the same underlying properties.

Peters et al. (2006) conducted a median split of numeracy scores: Participants who achieved a score of 9 or more were categorized as highly numerate whereas those who achieved 8 or lower were categorized as less numerate. However, the differences between individuals who achieved 8 and 9 might be neglectable, and no different than the differences between individuals who achieved 9 compared to 10 or 7 compared to 8. In addition, dichotomization reduces the power of statistical tests and effect sizes (Bakhshi et al., 2012). Fedorov et al. (2009) argued that 100 continuous observations are statistically equivalent to 158 dichotomized observations. Thus, the aim of treating numeracy as a continuous variable is to obtain more accurate effects, max-

Table 2. Summary of original findings in the target article

S	Factors	E	Effect	% CIs	CIL	CIH
1	Numeracy and framing effect	*f*	0.25	90%	0.00	0.42
2	Numeracy and frequency-percentage effect	*f*	0.31	90%	0.00	0.58
3	Numeracy and bowl choice	*φ*	0.77	95%	/	/
	Numeracy and preference for bowls	*d*	-0.74	95%	-1.33	-0.13
	Numeracy and affect precision	*d*	0.78	95%	0.17	1.36
	Numeracy and affect	*d*	0.46	95%	/	/
4	Numeracy and attractiveness of bet	*f*	0.23	90%	0.10	0.35
	Numeracy and affect precision	*f*	/	/	/	/
	Numeracy and affect	*f*	0.20	90%	0.10	0.50

Note. CIL = lower bounds for CIs. CIH = higher bounds of CIs. We report 90% CIs for ANOVA eta-squared given the effect size is always positive.

imize power, and address potential misinterpretations resulting from dichotomization.

Extension: Confidence

We aimed to extend the replication by examining decision-making confidence. Confidence regarding a decision involving statistics may be considered as a measure of subjective numeracy or numeric self-efficacy, concerning how confident people are in their ability to understand numeric information and use mathematical concepts (Peters, 2020, p. 5). We discuss two rationales for this extension.

First, there are mixed findings regarding the association between subjective and objective numeracy. A body of research illustrates that subjective numeracy is positively associated with objective numeracy (Garcia-Retamero et al., 2015; Nelson et al., 2013; Peters, Fennema, et al., 2019). According to the Health Information National Trends Survey conducted by Nelson et al. (2013), participants who regarded themselves as high in subjective numeracy had higher correction rates of objective numeracy questions. Another recent study done by Rolison et al. (2020) illustrated that individuals with higher objective numeracy were more likely to have correct answers in health risk comprehension questions. However, some research found no support for such an association and people with low objective numeracy sometimes deem themselves as highly numerate (Gamliel et al., 2016; Liberali et al., 2012; Peters, Tompkins, et al., 2019). For instance, Peters et al. (2019) reported that the objective numeracy sometimes mismatches subjective confidence: 31% participants with high numeracy but low confidence and 44% participants with low numeracy but high confidence.

Second, most current studies measure trait subjective numeracy with self-report questionnaires and two frequently-used scales are Subjective Numeracy Scale developed by Fagerlin et al. (2007) and STAT-Confidence Scale developed by Woloshin et al. (2005). Self-report questionnaires target participants' traits or general impressions about their numeracy competence and preference for numbers. It may vary from specific numeric confidence regarding specific decision making paradigms. Very few studies directly ask participants to rate their confidence about their decisions and answers in response to specific scenarios. Therefore, this study intends to examine the relationship between objective numeracy and subjective confidence in four studies of Peters et al. (2006). We hypothesized that objective numeracy is positively associated with confidence in each study.

Pre-Registration and Open-science

We pre-registered the experiment on the Open Science Framework (OSF) and data collection was launched later that week. Pre-registrations, power analyses, and all materials used in these experiments are available in the supplementary materials. We provided all materials, data, code, and pre-registration on: https://osf.io/4hjck/. The IPA registration link is https://osf.io/r73fb.

We provided additional open-science details and disclosures in the supplementary materials under "Open Science disclosures" sub-section. All measures, manipulations, exclusions conducted for this investigation are reported, all studies were pre-registered with power analyses, and data collection was completed before analyses.

Methods

Power Analysis

We calculated effect sizes (ES) and power based on the statistics reported in the target article (see supplementary materials). We then conducted a power analysis using G*Power (Faul et al., 2007) for the statistical tests in each of the decision-making risk paradigms separately (i.e. framing effect, frequency-percentage effect, ratio bias and bets effect).

Power analyses were conducted on the results of the main findings in the original study that yielded significant effect and supported the hypotheses for Studies 1 to 4. The largest required sample size in all effects was a result from two-way between-subjective ANOVA, which when aiming for a power of 0.95 and alpha of 0.05 one-tail was $N = 314$. We provide further information regarding our calculations in the "Power analysis of original study effect to assess required sample for replication" section in the supplementary materials.

Table 3. Differences and similarities between original study and replication

	Peters et al. (2006) Study 1	Peters et al. (2006) Study 2 and 3	Peters et al. (2006) Study 4	US MTurk workers
Sample size	100	46	171	860
Geographic origin	US American			US American
Gender	55 males, 45 females	Not reported	79 males, 92 females	441 males, 415 females, 4 other/did not disclose
Median age (years)	Not reported	Not reported	Not reported	40
Average age (years)	26	Not reported	19	43.19
Standard deviation age (years)	Not reported	Not reported	Not reported	12.73
Age range (years)	Not reported	Not reported	Not reported	19-81
Medium (location)	Pencil and paper	Not reported	Not reported	Computer (online)
Compensation	$10	Not reported	Not reported	Nominal payment
Year	2005	2005	2005	2022

Given the possibility that the original effects are overestimated, we used the suggested Simonsohn (2015) rule of thumb, even if meant for other designs, and multiplied 314 by 2.5 resulting in 785 participants. Allowing for possible exclusions we summarized a total sample of 850 participants. Our sensitivity analysis indicated that a sample of 850 would allow the detection of f = 0.12 (one covariate, groups = 2, df = 1, 95% power, alpha = 5%, one-tail), an effect much weaker than any of the effects reported in the original, and the detection of r = 0.12 in our continuous measures extension, an effect considered weak in social psychology (Lovakov & Agadullina, 2021).

Participants

We recruited 919 participants, but 59 of them failed the verifications or consent at the beginning. They were not considered as participants and were filtered out. Therefore, we had 860 participants (*Mage* = 43.19, *SD* = 12.73; 415 (48.3%) females) from Amazon Mechanical Turk using the CloudResearch/Turkprime platform (Litman et al., 2017). We summarized sample demographics and details in Table 3.

Based on our extensive experience of running similar judgment and decision making replications on MTurk, to ensure high quality data collection, we employed the following CloudResearch options: Duplicate IP Block. Duplicate Geocode Block, Suspicious Geocode Block, Verify Worker Country Location, Enhanced Privacy, CloudResearch Approved Participants and Block Low Quality Participants. We also employed the Qualtrics fraud and spam prevention measures: reCAPTCHA, prevent multiple submission, prevent ballotstuffing, bot detection, security scan monitor and relevantID. We also reported the details of payment and duration of study in the "additional information in the study" section in the supplementary.

Design: Replication and Extension

We summarized the experimental design in Tables 4, 5, 6, and 7. To conduct a replication of the four studies in the original article, we ran the four studies together in a single data collection. The display of scenarios and conditions were counterbalanced using the randomizer "evenly present" function in Qualtrics. Scenarios were presented in random order and participants were randomly and evenly assigned into different conditions. This method was previously tested successfully in many of the replications and extensions conducted by our team (e.g., Adelina & Feldman, 2021; Vonasch et al., 2022; Yeung & Feldman, 2022). The methodology is especially powerful in addressing potential concerns about the target sample (e.g., naivety and attentiveness), such as when some studies in the target article replicate successfully whereas others from the same article do not, which suggests that it is likely the failed study that is the cause for the failure rather than the participants' characteristics. This methodology also allows for examining possible links between the different studies and the consistency in participants' responding to similar decision-making paradigms.

Procedures

Participants first read the consent form, study outline, and then acknowledged a warning about not looking up answers online. They were then randomly assigned to a condition in each of the four studies. The order of four studies and their conditions were randomized. After the completion of tasks of four scenarios, they completed two numeracy scales, in random order. Then, they verified not using external aids in answering the questionnaire. At the end, participants answered a number of funneling questions (seriousness towards the survey, study purpose conjecture, and feedback) and provided demographic information. We added a more comprehensive overview of the survey procedure in the "procedure" section in the supplementary.

Table 4. Study 1: Replication and extension experimental design

IV1: Numeracy [between subject/continuous] IV2: Positive-negative framing [between subject]	**IV1: Numeracy** Original numeracy scale. Extension numeracy scale.
IV2: Positive framing condition Scores framed positively “% correct”	**Dependent variable** Evaluation of students’ performance Please rate each student’s quality of work "*Very poor*" (-3) to "*Very good*" (3) (for each of the five students) **Extension dependent variable** Evaluation of subjective confidence level How confident are you that you made an accurate assessment of the five students? “*Not at all confident*” (0) to “*Very confident*” (6)
IV2: Negative framing condition Scores framed negatively “% incorrect”	

Table 5. Study 2: Replication and extension experimental design

IV1: Numeracy [between subject/continuous] IV2: Frequency-percentage description (risk format) [between subject]	**IV1: Numeracy** Original numeracy scale. Extension numeracy scale.
IV2: Frequency condition “Of every 100… 10 are estimated…”	**Dependent variable** Evaluation of risk level Please rate the level of risk that Mr. Jones would harm someone “*Low risk*” (1) to “*High risk*” (6) **Extension dependent variable** Evaluation of subjective confidence level How confident are you that made an accurate risk assessment? “*Not at all confident*” (0) to “*Very confident*” (6)
IV2: Percentage condition “Of every 100… 10% are estimated…	

Measures

We detailed the measures of the replications and extensions for each condition in Tables 4, 5, 6, and 7. We provided all materials, with all experimental manipulation and the scales used, in the supplementary materials.

Numeracy

Objective numeracy predictor was measured using the Numeracy Scale developed by Lipkus et al. (2001) (Cronbach’s α = 0.64). We refer to it as the “original numeracy scale”, and it has 11 items and the total mark is 11.

We added an additional numeracy measure as an extension: Numeracy Scale developed by Weller et al. (2013) (Cronbach’s α = 0.66). We refer to it as the “Rasch-based numeracy scale”, and it has eight items and the total mark is 8.

Manipulations

Study 1: Positive versus Negative Framing

Participants were randomly assigned to either positive framing or negative framing conditions. They were asked to rate the quality of five psychology students’ exam scores framed positively or negatively. The order of five exam scores was randomized.

Study 2: Frequency and Percentage Condition

Participants were randomly assigned to frequency or percentage conditions. Participants read the scenario of Mr. Jones, a mental patient with the potential to harm someone when released. Participants then rate the risk level of patients like Mr. Jones under either frequency framing (i.e., 10 out of 100 patients) or percentage framing (i.e., 10% of 100 patients).

Study 3: Ratio Bias

Participants first read a scenario describing two jellybean bowls. Bowl A-9-100 is the more attractive yet with less objectively favorable outcome than Bowl B-1-10. Participants rated their preference for Bowl A-9-100, and then chose one of the bowls. They then rated affect levels and affect precision of both bowls.

Study 4: No Loss versus Loss Condition

Participants were randomly assigned to loss versus no-loss conditions. Participants read the scenario on a bet with “a chance 7 out 36 chance to win $9 and 29 out 36 chance to win nothing” or with "a chance 7 out 36 chance to win $9 but 29 out 36 chance to lose 5 cents’'. The chance of bet was visualized using a picture of a roulette wheel. Participants evaluated the attractiveness of the bet, and then rated affect and affect precision towards two bets.

Table 6. Study 3: Replication and extension experimental design

IV: Numeracy [between subject/continuous] Original numeracy scale. Extension numeracy scale.
Dependent variables Preference of bowl Bowl A-9-100; 100 jellybeans, 9% colored (odds = 9 out of 100 = 9%) Bowl B-1-10: 10 jellybeans, 10% colored (odds = 1 out of 10 = 10%) "Imagine that if you select a colored bean, you will WIN $5. Would you prefer to pick from Bowl A or Bowl B?" "*Strong preference for Bowl A*" (6) to "*Strong preference for Bowl B*" (6) Affect precision for Bowl A-9-100 choice How clear a feeling do you have about the goodness or badness of Bowl A's 9% chance of winning? "*Completely unclear*" (0) to "*Completely clear*" (6) Affect for Bowl A-9-100 choice How good or bad does Bowl A's 9% chance of winning make you feel? "*Very bad*" (-3) to "*Very good*" (3) **[Added adjustment dependent variables]** Affect precision for Bowl B-1-10 choice How clear a feeling do you have about the goodness or badness of Bowl B's 10% chance of winning? "*Completely unclear*" (0) to "*Completely clear*" (6) Affect for Bowl B-1-10 choice How good or bad does Bowl B's 10% chance of winning make you feel? "*Very bad*" (-3) to "*Very good*" (3) Forced choice of bowls If you were forced to choose, which bowl would you prefer to choose from? "Bowl A" or "Bowl B" **Extension dependent variables** Evaluation of subjective confidence level How confident are you that made an optimal selection between Bowl A and Bowl B? "*Not at all confident*" (0) to "*Very confident*" (6)

Table 7. Study 4: Replication and extension experimental design

IV1: Numeracy [between subject/continuous] IV2: Bet type (loss vs. no-loss) [between subject]	**IV1: Numeracy** Original numeracy scale. Extension numeracy scale.
IV2: Bet - No loss condition "There is a 7/36 chance to win $9 and 29/36 chance to win nothing."	**Dependent variable** Evaluation of bet's attractiveness Please indicate your opinion of this bet's attractiveness "*Not at all attractive bet*" (0) to "*Extremely attractive bet*" (20) Affect precision for bet How clear a feeling do you have about the goodness or badness of the bet? "*Completely unclear*" (0) to "*Completely clear*" (6) Affect for bet How good or bad does the bet make you feel? "*Very bad*" (-3) to "*Very good*" (3) **Extension dependent variable** Evaluation of subjective confidence level How confident are you that you made an accurate assessment of the bet's attractiveness? "*Not at all confident*" (0) to "*Very confident*" (6)
IV2: Bet - Loss condition "There is a 7/36 chance to win $9 and 29/36 chance to lose 5 cents."	

Deviations

We note we made several adjustments that are deviations from the original's design. We summarize the details of the deviations with comparisons of the original paper and our replication in Table 8.

In terms of the measurement of numeracy, we added an objective numeracy scale developed by Weller et al. (2013).

Table 8. Classification of the replication, based on LeBel et al. (2018)

Design facet	Replication	Details of deviation
Hypothesis	Same+extension	We ran the original analyses and added a reframing of the hypotheses treating numeracy as a continuous variable.
IV construct	Same	
DV construct	Similar	We reconstructed our version of the scores of the four students described in Study 1, as the stimuli was not provided in the article.
IV operationalization	Similar	We randomized the order of the numeracy questions
DV operationalization	Similar	In Study 3, we added exploratory extra questions for more optimal choice on affect and affect precision (on Bowl B-1-10). We also added the question on compulsory bowl choices.
IV stimuli	Similar	Added an extra numeracy scale
DV stimuli	Same	
Procedural details	Similar+extensions	The dependent variables on the four studies were completed together, in random order Added a warning pledge before test Added a question confirming not using external aids to find answers Added familiarity questions in Studies 2, 3, and 4 We did not collect SAT scores
Physical settings	Different	Online questionnaire
Population (e.g., age)	Different	Online US American MTurk workers
Replication classification	Close replication	

Note. We summarized the replication as a close replication using the criteria by LeBel et al. (2018) criteria, summarized in the supplementary materials in section “Replication closeness”.

The rationale for this extension is that this scale has demonstrated sound psychometric properties based on Rasch analysis and is argued to have better predictive validity than previous scales. Several recent studies have adopted it and shown support for high internal consistency (Cheng, 2020; Dolan et al., 2016; Peters, Fennema, et al., 2019).

We added a warning pledge block at the beginning of the questionnaire to ask participants not to look for answers and added a question at the end asking participants whether they used any external aids to search answers after the completion of two numeracy scales.

We made minor visual adjustments in the original numeracy scale (Lipkus et al., 2001), we removed decimals in 10.00, and turned the 1,000 into 1000. Given that we asked for and validated the input of numbers without decimals and commas, these may confuse participants.

The target article paper used SAT scores as a proxy measure for intelligence as they demonstrated that intelligence is positively associated with objective numeracy. Collection of SAT scores is not applicable to our target sample, and is not a core component of the target article.

The target article ran data collection for each of the studies separately, and reported using pencil and paper in Study 1 (Studies 2, 3, and 4 not reported). We conducted data collection online in a unified design in which participants answer all the dependent variables of the four studies in random order.

In Study 1, the original paper did not report the specific scores of five psychology students, which only had one example. Therefore, we reconstructed our own version of the four students' scores with four percent increment or four percent decrement (i.e., 66%, 70%, 74%, 78% and 82%).

In Study 3, we added the questions of affect precision and affect for Bowl B-1-10. These were meant as exploratory measures to allow us to determine how participants feel about both options to allow for baseline comparisons. We considered the possibility that drawing conclusions from the ratings of only one of the two bowls may be lacking, whereas a comparison of the two options would be more accurate. In addition, we added the forced bowl choice after the preference of two bowls rating. The original paper conducted the chi-square test but did not report the process of categorization of Bowl A-9-100 and Bowl B-1-10. Therefore, we required an extra question to confirm the choices.

Data Analysis Strategy

Replication: As in the original

The original paper dichotomized the numeracy scores as high numerate and low numerate. They conducted a 2 between x 5 within mixed ANOVA in Study 1, a two-way ANOVA in Study 2, and in Study 3 a chi-square test on the question of choosing Bowl, and an independent *t*-test to test bowl preference, affect, and affect precision. In Study 4, they conducted a factorial ANOVA for main interaction effect (i.e., numeracy and attractiveness of bet, numeracy and affect, numeracy and affect precision) and independent *t*-test to compare the responses (i.e., rate of attractiveness, affect and affect precision) of the high numerate under two conditions.

Extension: Additional analyses

To our understanding, one of the major weaknesses of the target article is in their decision to dichotomize a continuous measure. In the replication, we supplemented the original analyses with additional analyses treating numeracy scores as intended - a continuous variable. Therefore, in Studies 1, 2, 3 and 4 we conducted correlational analyses and in Study 3 we conducted an extra independent *t*-test for bowl choice.

Extension: Confidence

We conducted correlational analyses for confidence level and numeracy scale score.

Results

In this section, we reported the results of the sample without exclusion. Our original plan for the exclusion was for the case in which the replication failed, and for the most part they did not, and we also realized that the large number of exclusions severely limited our power to detect the effects. Therefore, we focus our analyses on the full sample. We provided the results post-exclusions in the section "overview of post-exclusions" in the supplementary materials, and provided a table comparing the results with and without exclusions in Table S14.

Replication: Main effects

We first examined the main effect of each study, examining classic phenomena in judgment and decision-making.

In Study 1, we conducted an independent *t*-test and found the framing effect that participants rated the students' performance on exams higher when the results is positively framed than that is negatively framed (positive framing: *M* = 0.48, *SD* = 0.75; negative framing: *M* = -0.09, *SD* = 1.07; *t*(858) = 9.12, *p* < .001, *d* = 0.62, 95% CI [0.48, 0.76]).

In Study 2, we conducted an independent *t*-test to test the frequency-percentage effect. Participants rated the higher level of risk under frequency condition than percentage condition (frequency: *M* = 3.03, *SD* = 1.27; percentage: *M* = 2.58, *SD* = 1.17; *t*(858) = 9.12, *p* < .001, *d* = 0.37, 95% CI [0.23, 0.50])

In Study 3, we conducted a one-sample *t*-test on the preference of bowls and found that participants showed a stronger preference towards Bowl B-1-10 (*t*(859) = 20.04, *p* < .001, *d* = 0.68, 95% CI [0.61, 0.76]). We conclude that we failed to find support for previous ratio bias findings which showed stronger preference for the suboptimal choice, Bowl A-9-100.

In Study 4, we ran an independent *t*-test to examine the bets effect and found that participants rated higher attractiveness for the loss bet than the no-loss bet (no loss bet: *M* = 6.22, *SD* = 4.57; loss bet: *M* = 9.33, *SD* = 7.07; *t*(858) = 7.66, *p* < .001, *d* = 0.52, 95% [0.38, 0.66]), which supported the phenomenon.

Replication: Dichotomized numeracy

We first conducted statistical analyses that closely followed the methods used in the original article which dichotomized the continuous measure of numeracy into high numerate and low numerate via median split. The median of numeracy scores was 10 (mean = 9.69, range = 0-11). Therefore, participants whose overall score was 10 and above were classified as highly numerate and those whose overall score equal to or below 9 were classified as low numerate. We summarized the results of Studies 1, 2, and 4 in Table 9 and the results of Study 3 in Table 10. Further, we provided the descriptives of each subgroup analyzed of the following ANOVA tests in the supplementary materials (Table S17) to elaborate on the interaction effects.

In Study 1, we performed a mixed ANOVA and found an interaction effect of numeracy on framing effect (*F*(1, 855) = 5.02, *p* = .025, η^2_p = 0.01, 90% CI [0.00, 0.02]) (Figure 1). The effects were rather weak, barely below the pre-registered alpha threshold. We concluded support for the hypothesis that the less numerate show a stronger framing effect than the highly numerate, with weaker effects.

In Study 2, we conducted a 2-way ANOVA and failed to find support for an interaction between numeracy and the frequency-percentage effect, with weak effect just above the set alpha (*F*(1, 856) = 3.40, *p* = .065, η^2_p = 0.00, 90% CI [0.00, 0.01]) (Figure 2). The findings were in the right direction, and the differences in effect size and p-values between our replication of Study 1 and Study 2 are rather minor, so we hesitate to conclude one as supported and the other as failed, and yet the results did not meet our pre-set criteria, and therefore inconsistent with the hypothesis that the less numerate are affected more by the frequency-percentage effect than the highly numerate. We will return to this in our additional extension analyses.

In Study 3, we conducted two Chi-squared tests, one based on preference of bowl and the other based on the forced choice of bowl, to test the association between numeracy and bowl choice. Aforementioned, the original articles did not report the method of categorization of bowl choices based on the preferences of bowls. Then, we made an assumption that the original authors coded participants who indicated stronger preference for Bowl A-9-100 (i.e., 1-6) were coded as having selected Bowl A-9-100 and those who indicated preference for Bowl B-1-10 (i.e., 1-6) were coded as Bowl B-1-10, with the neutral value (0) neglected. We found support for an interaction between numeracy and ratio bias, and the result based on the forced choice of bowl (*χ2*(1, *N* = 860) = 24.91, *p* < .001, *Cramer's V* = 0.17, 95% CI [0.10, 0.24]) had a stronger effect than that based on preference of bowls (*χ2*(1, *N* = 789) = 12.53, *p* < .001, *Cramer's V* = 0.13, 95% CI [0.06, 0.20]).

In addition, we conducted independent t-tests on preferences of bowl and affective variables comparing the two numeracy groups. Participants with high numeracy (*M* = 3.01, *SD* = 3.50) showed greater preference for Bowl B-1-10 (over Bowl A-9-100) than participants with low numeracy (*M* = 1.56, *SD* = 3.89; *t*(858) = 5.51, *p* < .001, *d* = 0.40, 95% CI [0.25, 0.54]) (Figure 3). Those with lower numeracy (*M* =

Table 9. Studies 1, 2 and 4: Summary of statistical tests

	F	*df*	*p*	*η^2_p and CI*	Interpretation
Study 1 (Mixed ANOVA)					
Numeracy and framing effect	5.02	1, 855	.025	0.01 [0.00, 0.02]	signal inconsistent smaller
Study 2 (Factorial ANOVA)					
Numeracy and frequency-percentage effect	3.40	1, 856	.065	0.00 [0.00, 0.01]	no-signal inconsistent
Study 4 (Factorial ANOVA)					
Numeracy and attractiveness of bet	17.87	1, 856	< .001	0.02 [0.01, 0.04]	signal inconsistent smaller
Numeracy and affect of bet	9.27	1, 856	.002	0.01 [0.00, 0.03]	signal inconsistent smaller
Numeracy and affect precision of bet	0.02	1, 856	.890	0.00 [0.00, 0.00]	no-signal consistent

Note. CI = 90% confidence intervals. The interpretation of outcome is based on LeBel et al. (2019).

Table 10. Study 3: Summary of statistical tests

Low versus high numerate and bowl choice (Bowl A-9-100 and Bowl B-1-10)					
Chi-square test	*χ2*	*df*	*p*	*Cramer's V* and CI	
Dichotomized continuous numeracy	12.53	1	<.001	0.13 [0.06, 0.20]	signal
Dichotomized forced bowl choices	24.91	1	<.001	0.17 [0.10, 0.24]	signal
Low versus high numerate	*t*	*df*	*p*	*d* and CI	Interpretation
Preference of Bowls	5.51	858	<.001	0.40 [0.25, 0.54]	signal inconsistent smaller
Affect for Bowl A-9-100	-4.62	858	<.001	0.33 [0.19, 0.48]	signal consistent
Affect precision for Bowl A-9-100	3.00	858	.003	0.22 [0.07, 0.36]	signal inconsistent smaller

Note. CI = 95% confidence intervals. Independent *t*-test comparing the stated DVs between the high and low numerate split sub-samples. Dichotomized continuous numeracy is categorized bowl choices according to the preference of bowl. Dichotomized forced bowl choices is the adjusted DV. The interpretation of outcome is based on LeBel et al. (2019).

-0.48, *SD* = 1.48) showed higher affect than those with high numeracy (*M* = -0.94 , *SD* = 1.30; *t*(858) = -4.62, *p* < .001, *d* = 0.33, 95% CI [0.19, 0.48]). By contrast, the less numerate (*M* = 4.19, *SD* = 1.64) had less precise feelings towards Bowl A-9-100 than the highly numerate (*M* = 4.53 , *SD* = 1.53; *t*(858) = 3.00, *p* = .003, *d* = 0.22, 95% CI [0.07, 0.36]).

Therefore, our findings for Study 3 were consistent with the hypothesis that the less numerate make less optimal choices in competing affective decisions than the highly numerate, with lower affective precision.

In Study 4, high numerate participants rated the loss bet more attractively than no-loss bet (loss bet: *M* = 10.27, *SD* = 7.21; no-loss: *M* = 5.95 , *SD* = 4.47; *t*(570) = 8.63, *p* < .001, *d* = 0.72, 95% CI [0.55, 0.90]). By contrast, we found no support for difference in low numerate participants with much weaker effects(loss bet: *M* = 7.50, *SD* = 6.44; no-loss: *M* = 6.78 , *SD* = 4.74; *t*(286) = 1.09, *p* = .277, *d* = 0.13, 95% CI [-0.10, 0.36]). We conducted a two-way ANOVA and found support for an interaction between numeracy and bets effect with (*F*(1, 856) = 17.87, *p* < .001, η^2_p = 0.02, 90% CI [0.01, 0.04]) (Figure 4).

In addition, the highly numerate rated stronger affect towards bets in the loss condition (*M* = 0.16, *SD* = 1.83) than no-loss condition (*M* = -0.72 , *SD* = 1.35; *t*(570) = 6.62, *p* < .001, *d* = 0.55, 95% CI [0.38, 0.72]), with no support and weaker effects for the low numerate (loss bet: *M* = -0.18, *SD* = 1.61; no-loss: *M* = -0.38 , *SD* = 1.36; *t*(286) = 1.13, *p* = .260, *d* = 0.13, 95% CI [-0.10, 0.36]). We analyzed the interaction effect between numeracy and affect of bets and the results supported the hypothesis that the highly numerate experience stronger affect in probabilities and numerical comparisons than the low numerate (*F*(1, 856) = 17.87, *p* < .001, η^2_p = 0.02, 90% CI [0.01, 0.04]).

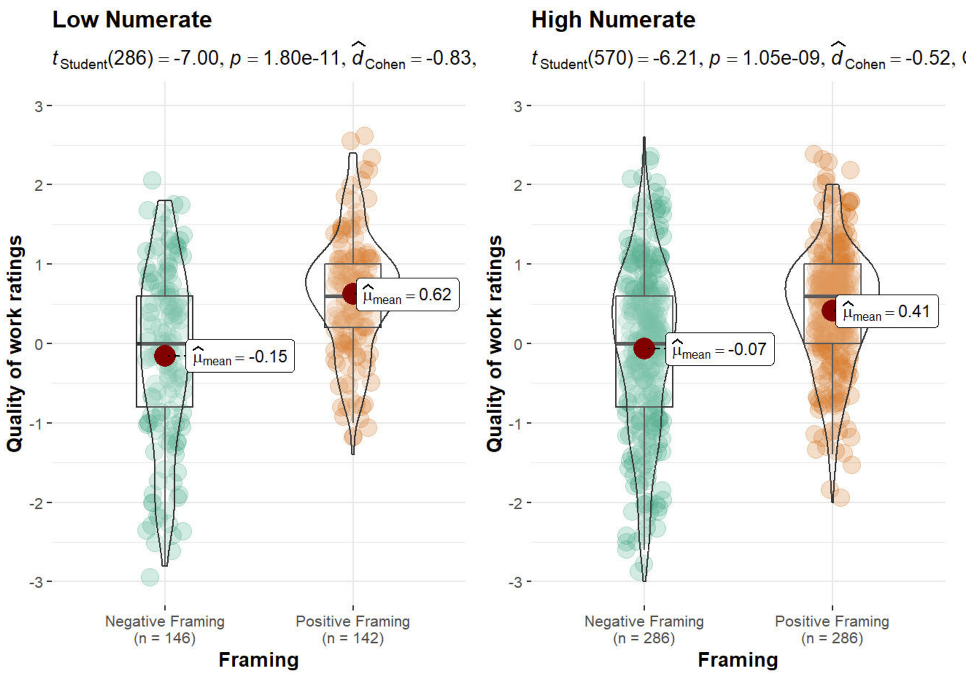


Figure 1. Study 1: Numeracy and attribute framing

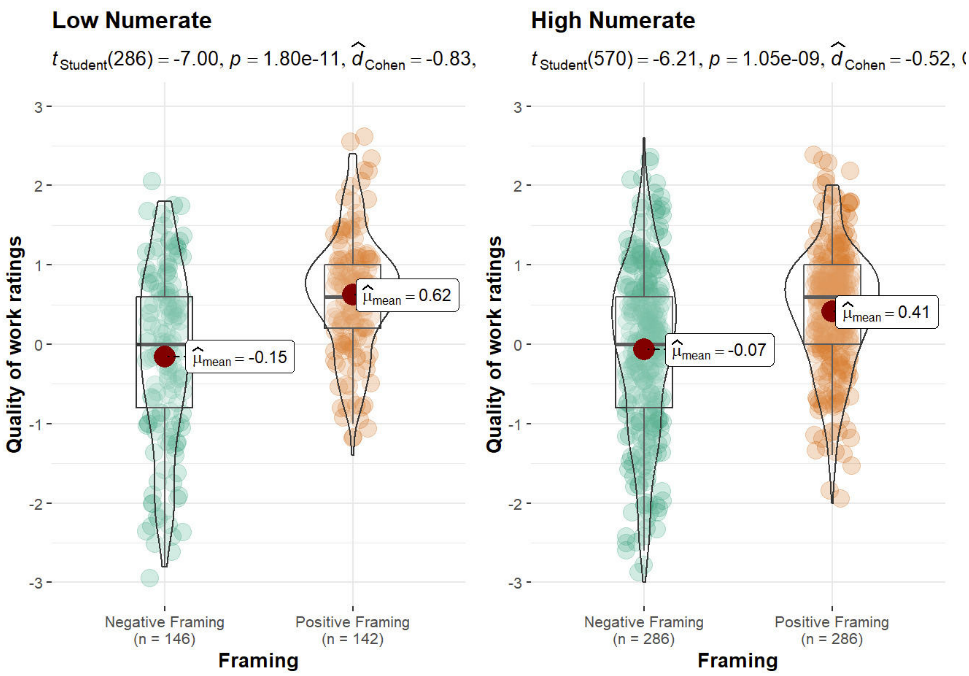


Figure 2. Study 2: Numeracy and percentage versus frequency representations of risk

Concerning the affect precision, both the low numerate and highly numerate showed greater affect precision towards loss bet than no-loss bet (low numeracy: $t(286)$ = 2.59, p < .001, d = 0.31, 95% CI [0.07, 0.54]); high numeracy: $t(570)$ = 4.36, p = .010, d = 0.36, 95% CI [0.20, 0.53]). Therefore, we found no support for an interaction between numeracy and affect precision of bets ($F(1, 856)$ = 0.02, p = .890, η^2_p = 0.00, 90% CI [0.00, 0.00]), consistent with the original findings of the target article.

Extension: Continuous numeracy

Original numeracy scale

In Study 1, we found support for stronger association between numeracy and ratings of students in the positive framing condition (r = -0.10, 95% CI [-0.19, -0.01], p = .036) than in the negative framing condition (r = 0.07, 95% CI [-0.03, 0.16], p = .177) (Figure 5). We compared the two

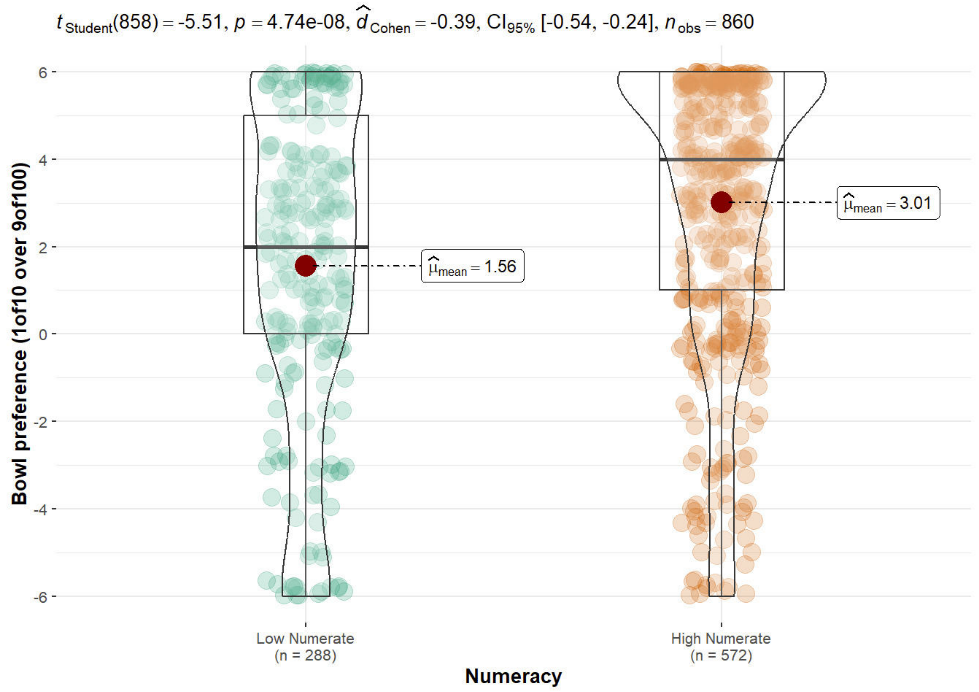


Figure 3. Study 3: Numeracy and bowl preference

Note. Higher score means stronger preference for Bowl B-1-10 over Bowl A-9-100 on a scale of -6 to 6.

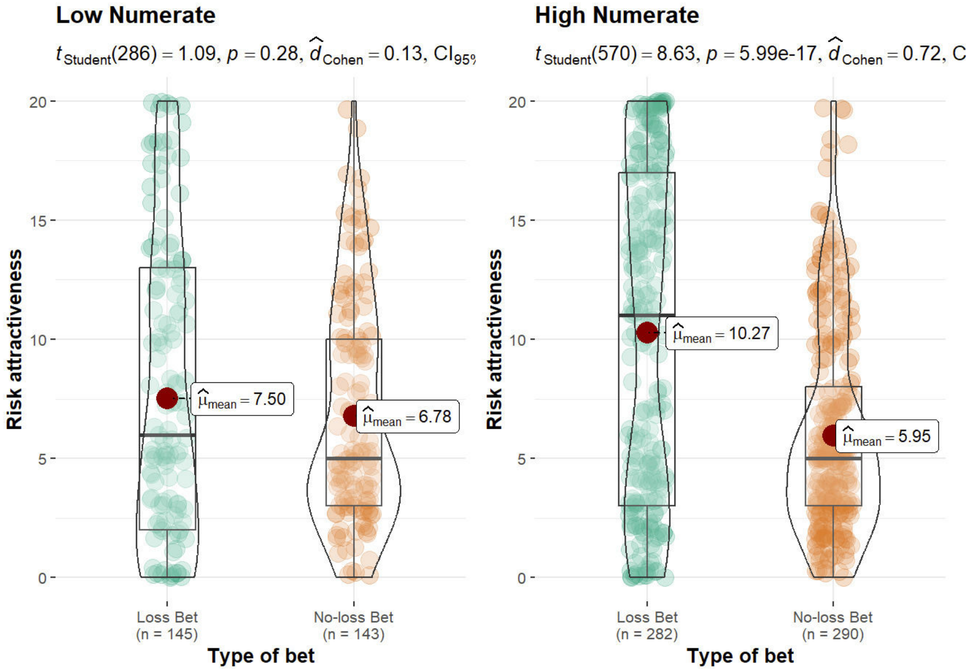


Figure 4. Study 4: Numeracy and rated attractiveness of a bet with and without loss

correlations with the tool "cocor" (Diedenhofen & Musch, 2015) and found support for differences in the strength of the two associations (z = -2.49, p = .013).

In Study 2, we found support for a stronger association between numeracy and ratings of risk level in frequency condition (r = -0.17, 95% CI [-0.26, -0.07], p < .001) than in the percentage condition (r = -0.03, 95% CI [-0.12, 0.07], p = .543) (Figure 6). We compared the two correlations using "cocor" and found support for differences in the strength of the two associations (z = -2.07, p = .039).

In Study 3, we conducted an independent t-test comparing the numeracy of the two bowl selections. The numeracy of participants who selected Bowl A-9-100 (M = 9.05, SD = 1.90) was lower than those who chose Bowl B-1-10

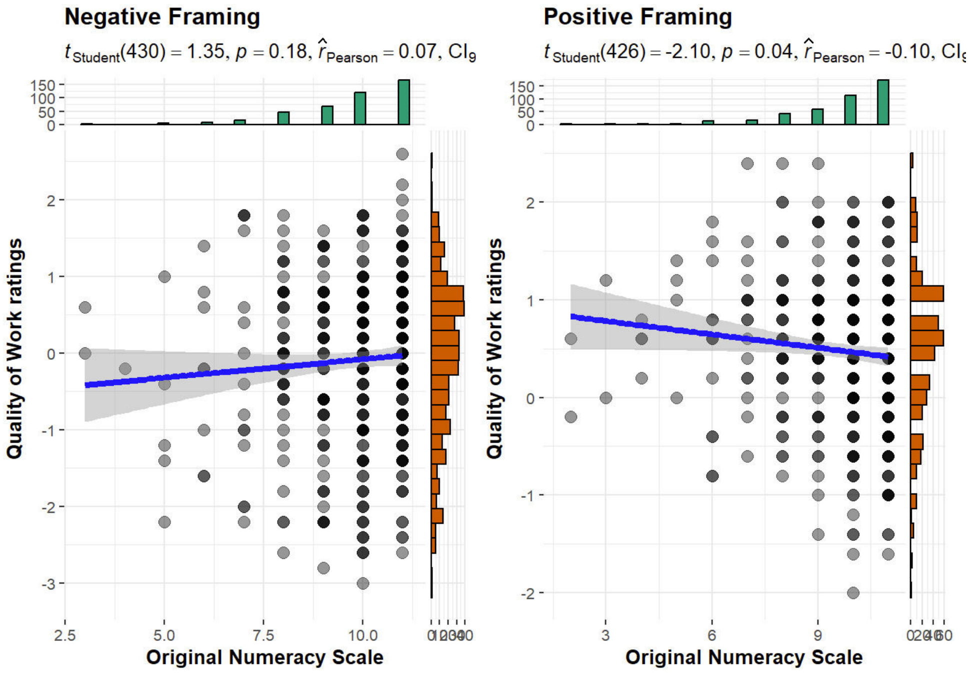


Figure 5. Study 1: Numeracy (original numeracy scale) and framing effect

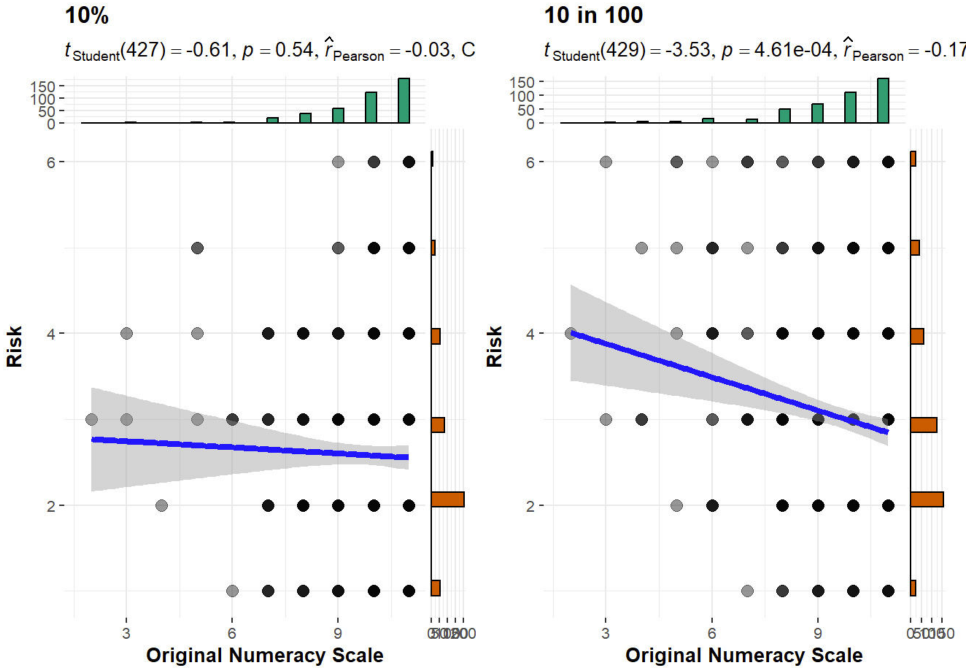


Figure 6. Study 2: Numeracy (original numeracy scale) and frequency-percentage effect

(M = 9.98, SD = 1.39; t(858) = 6.81, p < .001, d = 0.55, 95% CI [0.38, 0.72]). We also found support for associations between numeracy and bowl preference towards Bowl B-1-10 (r = 0.21, 95% CI [0.14, 0.27], p < .001) (Figure 7), lower affect for Bowl A-9-100 (r = -0.19, 95% CI [-0.25, -0.12], p < .001), and higher affect precision for Bowl A-9-100 (r = 0.14, 95% CI [0.07, 0.20], p < .001). These findings support the association between higher numeracy with the more rational choice of Bowl B-1-10, and less affect and higher affect precision in such a competing affective decision paradigm.

In Study 4, we found support for an association between numeracy and attractiveness of the two bets, a negative association with the no-loss condition (r = -0.13, 95% CI [-0.22, -0.04], p = .006), and a positive association with the loss condition (r = 0.21, 95% CI [0.11, 0.30], p < .001; comparison the associations: z = -5.03, p < .001) (Figure 8). We also found differences in associations between numeracy

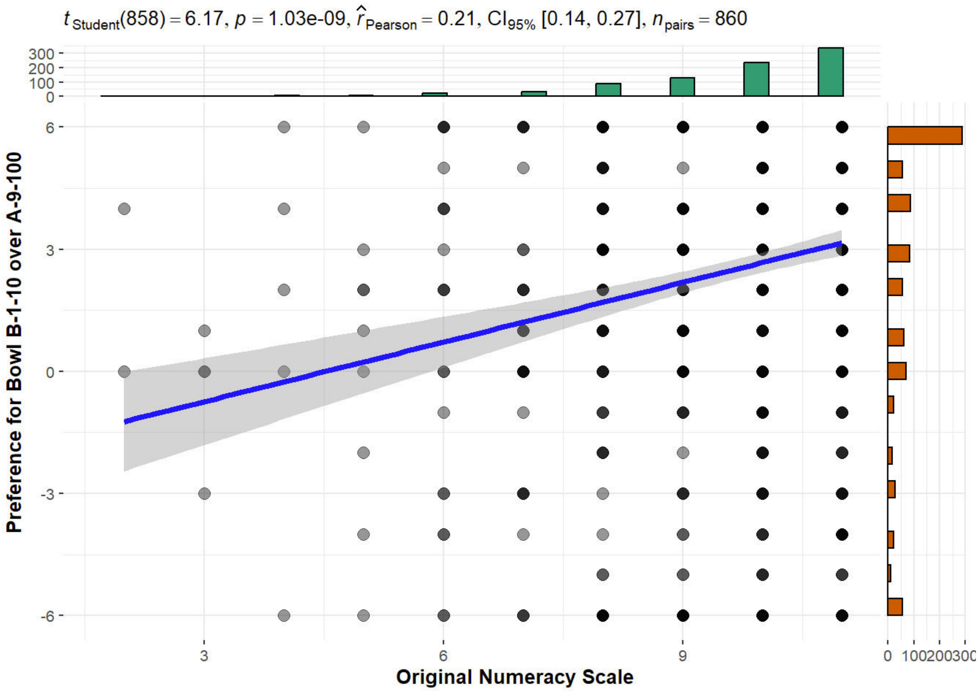


Figure 7. Study 3: Numeracy (original numeracy scale) and preference of bowls

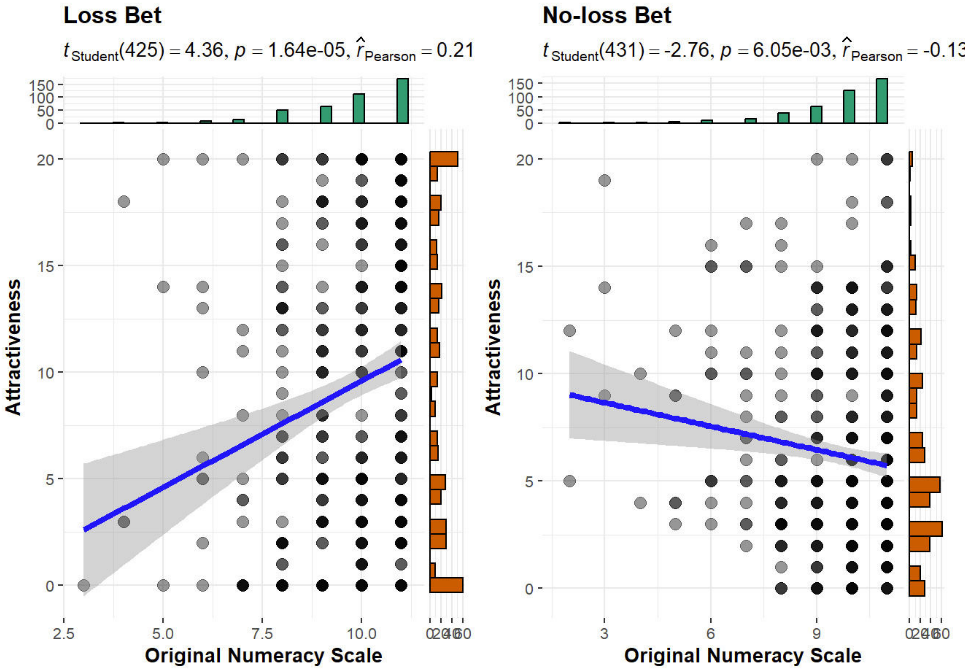


Figure 8. Study 4: Numeracy (original numeracy scale) and attractiveness of bets

and affect of bets (z = -3.82, p < .001), but no support for an effect regarding affect precision (z = 0.45, p = .654).

The results of four studies are summarized in Table 11, Table 12, and Table 13.

Extension: Rasch-based numeracy scale

The results based on Rasch-based numeracy were the same as those with original numeracy scale across four studies despite slight differences of effect size, and we provided those details in Tables 11, 12, and 13.

Extension: Confidence

We added an extension examining numeric confidence and tested the association between objective numeracy and numeric confidence with both original numeracy scale and rasch-based numeracy scale. The findings were mixed. We

Table 11. Studies 1, 2, 3, and 4: Summary of statistical tests

		Original scale			Rasch scale		
Study		*r* and CI	*p*	Spearman's rho	*r* and CI	*p*	Spearman's rho
1	Students rating in Positive framing	-0.10 [-0.19, -0.01]	.036	-0.11	-0.12 [-0.21, -0.02]	.017	-0.11
	Students rating in Negative framing	0.07 [-0.03, 0.16]	.177	0.07	0.07 [-0.02, 0.17]	.134	0.09
2	Risk rating in Frequency condition	-0.17 [-0.26, -0.07]	< .001	-0.12	-0.17 [-0.26, -0.08]	< .001	-0.15
	Risk rating in Percentage condition	-0.03 [-0.12, 0.07]	.543	-0.02	0.00 [-0.09, 0.10]	.919	-0.03
3	Bowl preference	0.21 [0.14, 0.27]	< .001	0.22	0.20 [0.14, 0.27]	< .001	0.00
	Affect for Bowl A-9-100	-0.19 [-0.25, -0.12]	< .001	-0.17	-0.20 [-0.26, -0.13]	< .001	-0.18
	Affect precision for Bowl A-9-100	0.14 [0.07, 0.20]	< .001	0.14	0.17 [0.11, 0.24]	< .001	0.18
4	No Loss condition						
	Attractiveness	-0.13 [-0.22, -0.04]	.006	-0.07	-0.07 [-0.16, 0.03]	.161	-0.03
	Affect	-0.16 [-0.25, -0.06]	.001	-0.12	-0.09 [-0.19, 0.00]	.053	-0.06
	Affect precision	0.16 [0.07, 0.25]	< .001	0.11	0.15 [0.05, 0.24]	.002	0.13
	Loss condition						
	Attractiveness	0.21 [0.11, 0.30]	< .001	0.23	0.21 [0.30, 0.12]	< .001	0.22
	Affect	0.10 [0.01, 0.20]	.032	0.14	0.11 [0.02, 0.21]	.020	0.14
	Affect precision	0.13 [0.04, 0.22]	.006	0.09	0.13 [0.03, 0.22]	.010	0.10

Note. CI = 95% confidence intervals.

Table 12. Studies 1, 2, and 4: Comparisons of correlations

		Fisher's *z*	*p*	Interpretation
Original numeracy scale				
Study 1	Numeracy and framing effect	-2.49	.013	signal
Study 2	Numeracy and frequency-percentage effect	-2.07	.039	signal
Study 4	Numeracy and attractiveness of bets	-5.03	< .001	signal
	Numeracy and affect	-3.82	< .001	signal
	Numeracy and affect precision	0.45	.654	no-signal consistent
Rasch-based numeracy scale				
Study 1	Numeracy and framing effect	-2.79	.005	signal
Study 2	Numeracy and frequency-percentage effect	-2.51	.012	signal
Study 4	Numeracy and attractiveness of bets	-4.14	< .001	signal
	Numeracy and affect	-2.93	.003	signal
	Numeracy and affect precision	0.30	.766	no-signal consistent

Table 13. Study 3: Numeracy and optimal bowl choice

Independent *t*-test	*t*	*df*	*p*	*d* and CI	Interpretation
Original numeracy scale					
Bowl Choice	6.81	858	< .001	0.55 [0.38, 0.72]	signal consistent
Rasch-based numeracy scale					
Bowl Choice	6.59	858	< .001	0.54 [0.37, 0.70]	/

Note. CI = 95% confidence intervals. Independent *t*-test comparing the numeracy between Bowl A-9-100 and Bowl B-1-10.

Table 14. Confidence: Summary of correlations with numeracy in Studies 1-4

Study		**Original**			**Rasch**		
		r and CI	*p*	Spearman's rho	*r* and CI	*p*	Spearman's rho
1	Positive framing condition	-0.11 [-0.20, -0.02]	.021	-0.12	-0.10 [-0.19, -0.01]	.038	-0.07
	Negative framing condition	-0.03 [-0.13, 0.06]	.474	-0.02	-0.04 [-0.14, 0.05]	.376	0.00
2	Frequency condition	-0.01 [-0.10, 0.09]	.868	0.02	-0.01 [-0.10, 0.09]	.852	0.03
	Percentage condition	-0.01 [-0.11, 0.08]	.801	-0.01	0.00 [-0.10, 0.09]	.952	0.03
3		0.15 [0.08, 0.21]	< .001	< .001	0.14 [0.08, 0.21]	< .001	0.19
4	No loss condition	0.10 [0.01, 0.20]	.030	0.08	0.11 [0.01, 0.20]	.027	0.10
	Loss condition	0.05 [-0.05, 0.14]	.325	0.03	0.06 [-0.03, 0.16]	.196	0.08

Note. CI = 95% confidence intervals

only found support for an association in the positive framing condition in Study 1 (r = -0.11, 95% CI [-0.20, -0.02], p = .021), in Study 3 (r = 0.15, 95% CI [0.08, 0.21], p < .001), and in the no-loss bet condition in Study 4 (r = 0.10, 95% CI [0.01, 0.20], p = .030) derived from original numeracy scale. We also conducted analyses using Rasch-based numeracy scale with similar results, detailed in Table 14.

Assumption checks and non-parametric tests

We used Levene's test to check the homogeneity of variances and the Shapiro-Wilks test to check the normality of variables for ANOVA and independent *t*-test. The homogeneity and normality were violated primarily because of the highly negative skewness of original numeracy scale and rasch-based numeracy scale. We first supplemented the analyses with a report of Spearman correlations, provided in the tables alongside correlations. We also conducted non-parametric tests: Aligned Rank Transform (ART) (Kay et al., 2021) to supplement the Mixed ANOVA (Study 1) and the factorial ANOVA (Study 2 and Study 4), and Mann-Whitney *U* test to supplement the independent *t*-test (Study 3). These robust tests showed similar results with comparable conclusions except for Study 2, which shifted from just above the threshold to just below the threshold (F(1, 856) = 3.40, p = .065, η^2_p = 0.00, 90% CI [0.00, 0.01]; non parametric: F(1,856) = 4.18, p = .04), which again shows the issues in the over-reliance on p-values threshold as a dichotomy of success/failure decisions.

We summarized results of robustness check in Table S18 in the supplementary materials.

Exploratory analyses: Affect precision for Bowl B-1-10 and numeracy scale associations

We added extra questions for affect and affect precision for Bowl B-1-10 in Study 3. We found that participants rated more positive affect towards Bowl B-1-10 than for Bowl A-9-100 (Bowl B-1-10: M = -0.22 , SD = 1.50; Bowl A-9-100: M = -0.78 , SD = 1.38; t(858) = 11.86, p < .001, d = 0.40, 95% CI [0.33, 0.47]). Participants also showed greater affect precision for Bowl B-1-10 than Bowl A-9-100 (Bowl B-1-10: M = 4.65 , SD = 1.47; Bowl A-9-100: M = 4.42 , SD = 1.58; t(858) = 6.09, p < .001, d = 0.21, 95% CI [0.14, 0.28]).

We examined the associations between the original numeracy scale and the extension rasch-based numeracy scale and found that they were strongly correlated (r = 0.83, 95% CI [0.81, 0.85], p < .001).

As we failed to find the support for Study 2 using dichotomized numeracy, we ran an additional analysis to examine possible order effects, with display order as a covariate, and found no support for the interaction (F(1, 855) =

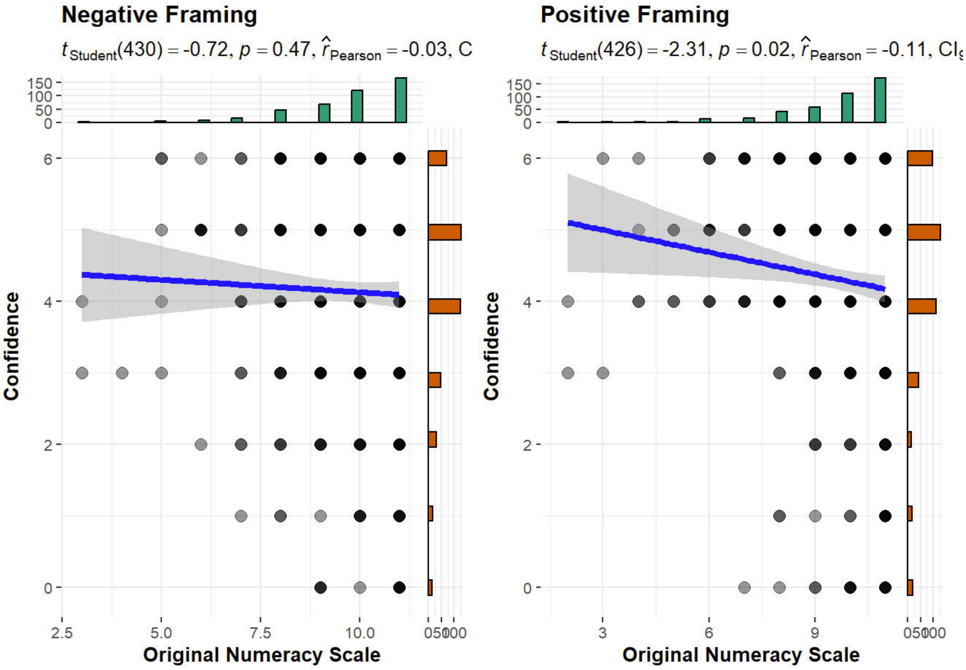


Figure 9. Study 1: Correlation between numeracy and confidence

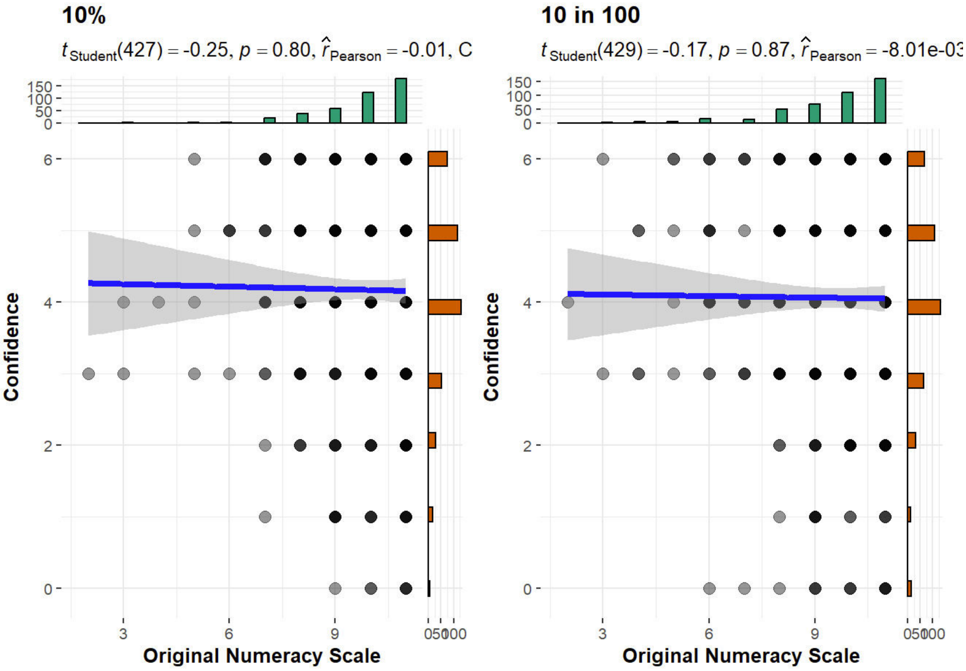


Figure 10. Study 2: Correlation between numeracy and confidence

3.29, p = .070, η^2_p = 0.00, 90% CI [0.00, 0.02]). In addition, we analyzed Study 2 only when it was the first study displayed to the participant, and also found no support for an interaction with an even weaker effect than for the whole sample (n = 224; F(1, 220) = 0.00, p = .986, η^2_p = 0.00, 90% CI [0.00, 0.00]). This suggests the order is not the reason for the failed replication using the dichotomous measure.

Comparing replication to original findings

Compared to the original findings (Table S1, S2, and S3 in the supplementary, Peters et al., 2006, p. 6), our replication findings based on dichotomous numeracy suggest support for numeracy as a predictor of framing effect (Study 1), ratio bias (Study 3), and bets effect (Study 4). When treating numeracy as a continuous variable, all four studies (in-

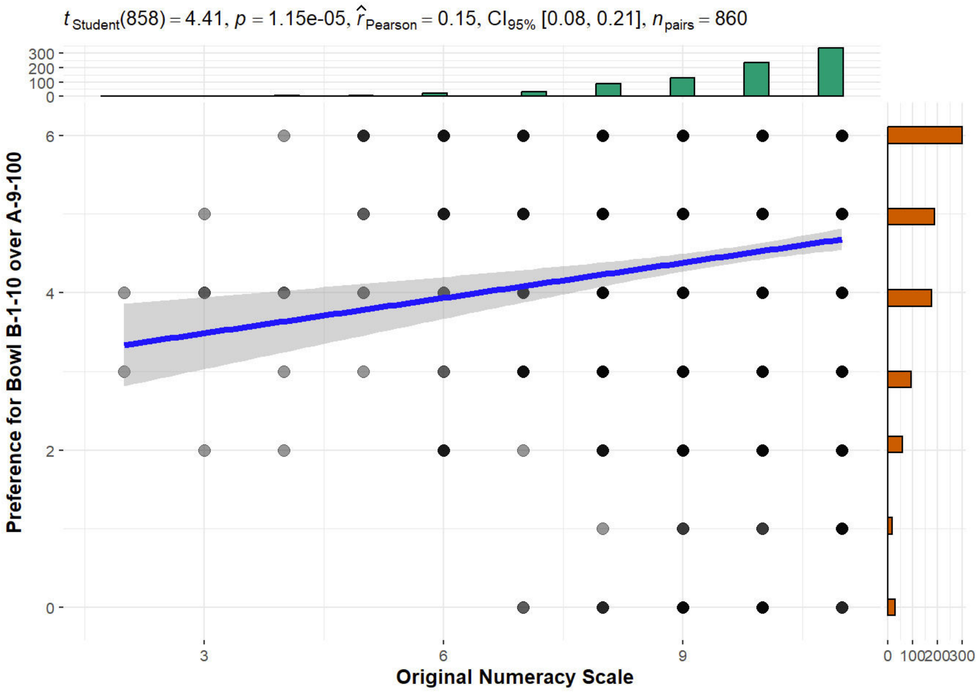


Figure 11. Study 3: Correlation between numeracy and confidence

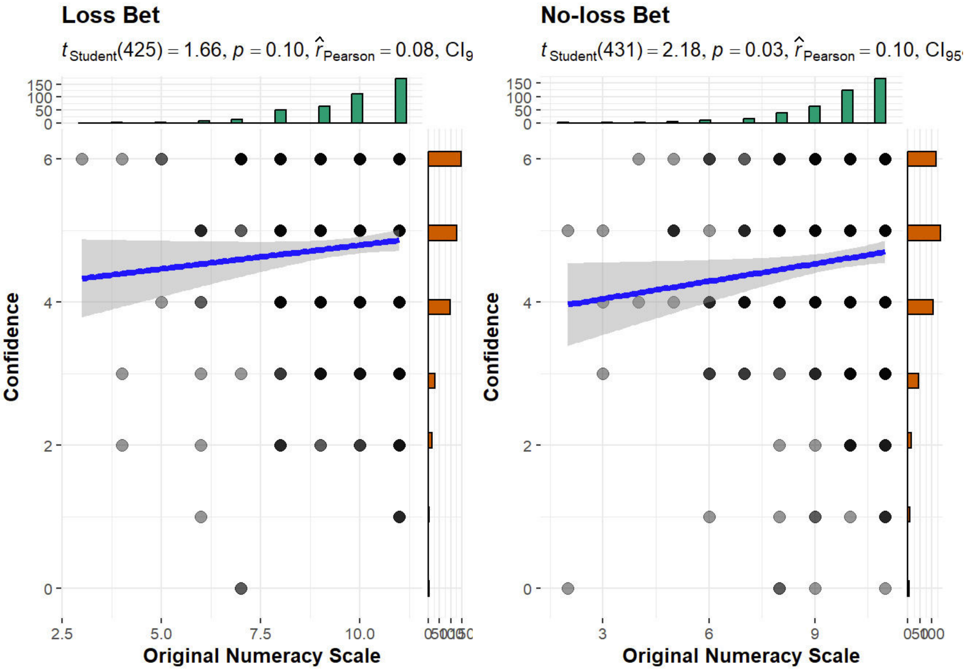


Figure 12. Study 4: Correlation between numeracy and confidence

cluding Study 2’s frequency-percentage effect) could be regarded as successful.

According to the criteria of LeBel et al. (2019) on the evaluation of replication results (see Figure S5 and S6), the replication effect sizes (i.e., Study 1, 2, and 4) showed signals and had inconsistent and smaller effects than those reported in the original. We summarize two minor discrepancies: (1) we found no support for numeracy as a predictor of frequency-percentage in Study 2 using the dichotomization method applied in the original, (2) the highly numerate felt more negative about the affectively appealing bowl with less favorable objective probabilities (i.e., Bowl A-9-100) compared to the low numerate.

Overall, we conclude this to be a successful replication of the target article, yet with much weaker effects than those reported by the original, and better aligned results when improving on the original’s methods using continuous measures rather than dichotomizing.

Discussion

We conducted a pre-registered replication and extension of Peters et al. (2006) with a larger, well-powered, and more diverse sample. Our findings mirroring the original's method of dichotomizing numeracy were mostly consistent with the original: (1) the highly numerate showed weaker framing effect (Study 1), (2) the low numerate participants showed stronger preference towards suboptimal choices, and showed more positive affect and low affect precision about their choices (Study 3), (3) the highly numerate showed a stronger bets effect (i.e., larger difference of rated attractiveness of bet under no-loss and loss conditions) and drew more affect from the less objectively favorable choices (Study 4). The findings for numeracy and frequency-percentage were weaker, though in the right direction and just below our pre-set threshold (Study 2). Our additional extension analyses using the continuous numeracy measure successfully replicated the results of all four original studies. Therefore, we conclude that our replication was mostly successful, with findings in the expected direction, yet with weaker effects. Concerning our added extension examining confidence, our findings regarding the association between objective numeracy and confidence were mixed.

Replication

The goal of the project was to assess the replicability of the research presented by Peters et al. (2006) in support of the interaction effects between numeracy and four decision-making paradigms. We first demonstrated support for three of the four classic effects: We showed a main effect for the framing effect, frequency-percentage effect, and bets effect, yet failed to show support for the ratio bias. That we were able to find numeracy as a predictor of ratio bias suggests that the bias is sensitive to the population and that factors such as sample numeracy impacted results. Our sample generally showed high numeracy, which may have resulted in weaker ratio bias effects. Also, our exploratory analysis of Bowl B-1-10 revealed that participants expressed more positive affect towards the optimal choice (Bowl B-1-10) rather than the non-optimal choices (Bowl A-9-100), as suggested by the dual process model. Bourdin and Vetschera (2018) discussed the situations that ratio bias phenomena may occur and found that it happens more frequently for low probabilities. However, the ratios they manipulated were more complex and required extra mental calculations compared to our target's paradigm (e.g., 1:9 vs. 9:90, and 1:9 vs. 8:91). It is possible that such ratios would show stronger effects, impacting the affective understanding of numbers. If that were the case, then those with lower numeracy would rely more on absolute numbers, which are more readily available. The scenario that we used might not be challenging enough, and therefore unable to serve as an affective hit to decision making.

That we failed to detect numeracy as a predictor of frequency-percentage effect using the original's method of dichotomizing was inconsistent with the target's findings and other previous studies (e.g., Dickert et al., 2011; Hill & Brase, 2012), though it is reassuring that we found support for the effect using the more accurate continuous method. One likely possible explanation is that the dichotomization of numeracy leads to the loss of power, as we noted in our discussion of the target's methodological weaknesses, and that given the weaker effects we needed a larger sample.

Extensions

Analyses Using Continuous Numeracy and the Rasch-based Numeracy Scale

We successfully replicated the original findings when we treated numeracy as a continuous variable, including Study 2 and affect for Bowl A-9-100 in Study 3. The results based on the rasch-based numeracy scale were consistent with those drawn on the original numeracy scale, which provided additional robust evidence to our findings. In future studies of numeracy and decision-making, we strongly recommend conducting continuous measure analyses. Given that the two numeracy scales showed comparable results, we consider either one or both to be good options.

Confidence

We ran extensions examining the relationship between objective numeracy and numeric confidence under specific conditions, and discovered three significant results with small effects. Therefore, we take our mixed findings as an indicator that such a relationship is not consistent or robust. It is possible that our single item question measuring confidence should be better validated or was not comprehensive enough to measure participants' confidence regarding engaging with and processing numeric information. We recommend more work to construct and test well-validated questions. For instance, Peters et al. (2019) selected the first four items from the subjective numeracy scale (Fagerlin et al., 2007) to measure numeric confidence (e.g., "How good are you at calculating 15% tip?"). In addition, Peters and Shoots-Reinhard (2022) suggested in their latest paper that numeric confidence is associated with persistence of choices being made and emotional reactions from experienced difficulty. Future studies could underline such measurement of variables. Another likely possibility is that numeracy and confidence are simply different constructs that capture different aspects of decision-making abilities which impact decision-making in different ways. Future studies can build on our data and initial investigation to examine the associations between confidence and decision-making biases and heuristics, and relate those to the literature on overconfidence, underconfidence, and the need for accuracy and calibration.

Limitations and Future Directions

As with all studies, several limitations should be addressed in future research. We initially set out to examine whether our participants were familiar with the very common decision-making paradigms and use that as an exclusion criteria. When analyzing the results, we realized the problem with this approach as the number of participants who indicated familiarity with the paradigm was much higher than we anticipated. Though we find support for the target's findings regardless of, it is possible that this resulted in the much weaker effects.

The first limitation is that our questions regarding familiarity (i.e., familiarity of scales and scenarios) were too ambiguous. Reviewing the feedback given by participants, some of them were confused about the meaning of familiarity. For instance, several participants perceived the understanding of questions as familiarity, rather than our intent in assessing whether they have seen those before, had experienced with similar decisions in real-life, or already know the right answer to the paradigm. This likely resulted in many of the participants being flagged for possible exclusion despite them not knowing the paradigms, which meant a severe loss of power and difficulty of detecting effect sizes for the post-exclusion analyses. However, when we compared the effects of pre and post exclusions, the effects were overall much stronger before exclusions, which may indicate that familiarity - at least in how we measured it - does not necessarily weaken the effects. This is an empirical question that should be examined in future studies using large samples, and we therefore hesitate against recommending excluding all participants who indicate knowing the paradigm. Instead, familiarity or experience with the paradigm can be considered as a possible moderator.

The high rate of indicated familiarity might also have to do with our target sample of highly experienced MTurk workers, a point that was raised in our Stage 1 review process. A large proportion of individuals indicated familiarity with the numeracy scales (56.2% for original numeracy scale and 63.4% for rasch-based numeracy scale), and given the popularity of these scales and MTurk workers experience with online studies, it is possible that they have indeed come across some of those scales before.

Moreover, the results of both numeracy scales were not normally distributed. The non-normal distribution of the original numeracy scale has been discussed by Weller et al. (2013), and they developed the rasch-based numeracy scale to avoid such statistical violation. However, the rasch-based scale appears to have been as easy as the original numeracy scale for our sample. Therefore, we conclude that future studies would need to take into account much a larger sample size and exclusion rates than we anticipated, as well as considering employing alternative numeracy scales, or to test for sample naivete.

Another methodological issue we faced was that we set our question validation for certain questions as too strict, without sufficient instructive instructions. For instance, the answer to Question 3 in the original numeracy scale should be 0.1% and we allowed only decimal input. However, several participants failed to input 0.001, which confused them about whether they gave a correct answer. Such issues should be noted in future studies with Qualtrics or other online questionnaire platforms, to be mindful of all the likely options of how participants may perceive the question or use the answer field.

Another obvious limitation with running studies online is the inability to completely prevent participants from using online shortcuts or calculators to answer numerical questions. We tried to address that best we could with a warning, and asking participants to pledge not doing so, and we relied on the participants' built-in incentives to finish the survey quickly to get paid, hoping that they would prefer to answer intuitively and fast rather than take the more lengthy and costly process of looking up answers. Similar studies done online may consider implementing extra measures with scripts that detect whether the participant has left the survey window, and take response time into account, to address possible issues.

To conclude, we were able to find support for the target's findings despite all those limitations, and this is an indication of robust findings. Our weaker effects may well be attributed to some of these limitations, and it is possible, and likely, that a more tightly controlled study would yield larger effects. Future studies can now use our materials to design stronger studies in the future.

Competing Interests

The author(s) declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Funding

This project has been supported by the Teaching Development Grant from the University of Hong Kong awarded to Gilad Feldman.

Authorship Declaration

Minrui Zhu conducted the replication as part of his thesis in psychology.

Gilad Feldman guided and supervised each step in the project, (later: conducted the pre-registrations, ran data collection), and edited the manuscript for submission.

Important Links and Information

Citation of the target research article:

> 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

In-principle Acceptance and Open-review

Provided on: https://rr.peercommunityin.org/articles/rec?id=165

Data Accessibility Statement

Materials, data, and code are available on: https://osf.io/4hjck/.

Contributor Roles Taxonomy

In the table, employ CRediT (Contributor Roles Taxonomy) to identify the contribution and roles played by the contributors in the current replication effort. Please refer to https://www.casrai.org/credit.html for details and definitions of each of the roles listed below.

Role	Minrui Zhu	Gilad Feldman
Conceptualization	X	X
Pre-registration	X	
Data curation		X
Formal analysis	X	
Funding acquisition		X
Investigation	X	
Pre-registration peer review / verification		X
Data analysis peer review / verification		X
Methodology	X	
Project administration		X
Resources		X
Software	X	
Supervision		X
Validation		X
Visualization	X	
Writing-original draft	X	
Writing-review and editing		X

Submitted: March 29, 2023 PDT, Accepted: April 23, 2023 PDT

References

Adelina, N., & Feldman, G. (2021). Are past and future selves perceived differently from present self? Replication and extension of Pronin and Ross (2006) temporal differences in trait self-ascriptions. *International Review of Social Psychology*, *34*(1), 29. https://doi.org/10.5334/irsp.571

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *Bmj*, *332*(7549), 1080. https://doi.org/10.1136/bmj.332.7549.1080

Bakhshi, E., McArdle, B., Mohammad, K., Seifi, B., & Biglarian, A. (2012). Let continuous outcome variables remain continuous. *Computational and Mathematical Methods in Medicine*, *2012*, 1–13. https://doi.org/10.1155/2012/639124

Bourdin, D., & Vetschera, R. (2018). Factors influencing the ratio bias. *EURO Journal on Decision Processes*, *6*(3–4), 321–342. https://doi.org/10.1007/s40070-018-0082-7

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*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Cheng, J. (2020). The role of numeracy and impulsivity in intertemporal choice and decision making. *Psychological Thought*, *13*(1), 254–272. https://doi.org/10.37708/psyct.v13i1.442

Choi, H., Wong, J. B., Mendiratta, A., Heiman, G. A., & Hamberger, M. J. (2011). Numeracy and framing bias in epilepsy. *Epilepsy & Behavior*, *20*(1), 29–33. https://doi.org/10.1016/j.yebeh.2010.10.005

Dickert, S., Kleber, J., Peters, E., & Slovic, P. (2011). Numeracy as a precursor to pro-social behavior: The impact of numeracy and presentation format on the cognitive mechanisms underlying donation decisions. *Judgment and Decision Making*, *6*(7), 638–650. https://doi.org/10.1017/s1930297500002679

Diedenhofen, B., & Musch, J. (2015). cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLoS ONE*, *10*(4), e0121945. https://doi.org/10.1371/journal.pone.0131499

Dolan, J. G., Cherkasky, O. A., Li, Q., Chin, N., & Veazie, P. J. (2016). Should health numeracy be assessed objectively or subjectively? *Medical Decision Making*, *36*(7), 868–875. https://doi.org/10.1177/0272989x15584332

Estrada-Mejía, C., de Vries, M., & Zeelenberg, M. (2016). Numeracy and wealth. *Journal of Economic Psychology*, *54*, 53–63. https://doi.org/10.1016/j.joep.2016.02.011

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: development of the Subjective Numeracy Scale. *Medical Decision Making*, *27*(5), 672–680. https://doi.org/10.1177/0272989x07304449

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Fedorov, V., Mannino, F., & Zhang, R. (2009). Consequences of dichotomization. *Pharmaceutical Statistics*, *8*(1), 50–61. https://doi.org/10.1002/pst.331

Freling, T. H., Vincent, L. H., & Henard, D. H. (2014). When not to accentuate the positive: Re-examining valence effects in attribute framing. *Organizational Behavior and Human Decision Processes*, *124*(2), 95–109. https://doi.org/10.1016/j.obhdp.2013.12.007

Gamliel, E., & Kreiner, H. (2017). Outcome proportions, numeracy, and attribute-framing bias. *Australian Journal of Psychology*, *69*(4), 283–292. https://doi.org/10.1111/ajpy.12151

Gamliel, E., Kreiner, H., & Garcia-Retamero, R. (2016). The moderating role of objective and subjective numeracy in attribute framing. *International Journal of Psychology*, *51*(2), 109–116. https://doi.org/10.1002/ijop.12138

Garcia-Retamero, R., Andrade, A., Sharit, J., & Ruiz, J. G. (2015). Is patients' numeracy related to physical and mental health? *Medical Decision Making*, *35*(4), 501–511. https://doi.org/10.1177/0272989x15578126

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases." *European Review of Social Psychology*, *2*(1), 83–115. https://doi.org/10.1080/14792779143000033

Hill, W. T., & Brase, G. L. (2012). When and for whom do frequencies facilitate performance? On the role of numerical literacy. *Quarterly Journal of Experimental Psychology*, *65*(12), 2343–2368. https://doi.org/10.1080/17470218.2012.687004

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697–720. https://doi.org/10.1037/0003-066x.58.9.697

Kay, M., Elkin, L. A., Higgins, J. J., & Wobbrock, J. O. (2021). *ARTool: Aligned Rank Transform for Nonparametric Factorial ANOVAs*. R package version 0.11.1. https://doi.org/10.5281/ZENODO.594511

Kleber, J., Dickert, S., Peters, E., & Florack, A. (2013). Same numbers, different meanings: How numeracy influences the importance of numbers for pro-social behavior. *Journal of Experimental Social Psychology*, *49*(4), 699–705. https://doi.org/10.1016/j.jesp.2013.02.009

Lazic, S. E. (2018). Four simple ways to increase power without increasing the sample size. *Laboratory Animals*, *52*(6), 621–629. https://doi.org/10.1177/0023677218767478

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402. https://doi.org/10.1177/2515245918787489

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2019). A brief guide to evaluate replications. *Meta-Psychology*, *3*, 1–9. https://doi.org/10.15626/mp.2018.843

Levin, I. P., & Gaeth, G. J. (1988). How consumers are affected by the framing of attribute information before and after consuming the product. *Journal of Consumer Research*, *15*(3), 374–378. https://doi.org/10.1086/20917

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making*, *25*(4), 361–381. https://doi.org/10.1002/bdm.752

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, *21*(1), 37–44. https://doi.org/10.1177/0272989x0102100105

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485–504. https://doi.org/10.1002/ejsp.2752

Mariooryad, S., & Busso, C. (2015). The cost of dichotomizing continuous labels for binary classification problems: Deriving a Bayesian-optimal classifier. *IEEE Transactions on Affective Computing*, *8*(1), 119–130. https://doi.org/10.1109/taffc.2015.2508454

Miller, D. T., Turnbull, W., & McFarland, C. (1989). When a coincidence is suspicious: The role of mental simulation. *Journal of Personality and Social Psychology*, *57*(4), 581–589. https://doi.org/10.1037/0022-3514.57.4.581

Nelson, W. L., Moser, R. P., & Han, P. K. J. (2013). Exploring objective and subjective numeracy at a population level: findings from the 2007 Health Information National Trends Survey (HINTS). *Journal of Health Communication*, *18*(2), 192–205. https://doi.org/10.1080/10810730.2012.688450

Okamoto, M., Kyutoku, Y., Sawada, M., Clowney, L., Watanabe, E., Dan, I., & Kawamoto, K. (2012). Health numeracy in Japan: Measures of basic numeracy account for framing bias in a highly numerate population. *BMC Medical Informatics and Decision Making*, *12*(1), 104. https://doi.org/10.1186/1472-6947-12-104

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

Pachur, T., & Galesic, M. (2013). Strategy selection in risky choice: The impact of numeracy, affect, and cross-cultural differences. *Journal of Behavioral Decision Making*, *26*(3), 260–271. https://doi.org/10.1002/bdm.1757

Peters, E. (2012). Beyond comprehension: The role of numeracy in judgments and decisions. *Current Directions in Psychological Science*, *21*(1), 31–35. https://doi.org/10.1177/0963721411429960

Peters, E. (2020). *Innumeracy in the wild: Misunderstanding and misusing numbers*. Oxford University Press. https://doi.org/10.1093/oso/9780190861094.001.0001

Peters, E., & Bjalkebring, P. (2015). Multiple numeric competencies: When a number is not just a number. *Journal of Personality and Social Psychology*, *108*(5), 802–822. https://doi.org/10.1037/pspp0000019

Peters, E., Fennema, M. G., & Tiede, K. E. (2019). The loss-bet paradox: Actuaries, accountants, and other numerate people rate numerically inferior gambles as superior. *Journal of Behavioral Decision Making*, *32*(1), 15–29. https://doi.org/10.1002/bdm.2085

Peters, E., Hart, P. S., & Fraenkel, L. (2011). Informing patients: The influence of numeracy, framing, and format of side effect information on risk perceptions. *Medical Decision Making*, *31*(3), 432–436. https://doi.org/10.1177/0272989x10391672

Peters, E., & Shoots-Reinhard, B. (2022). Numeracy and the Motivational Mind: The Power of Numeric Self-efficacy. *Medical Decision Making*, *42*(6), 729–740. https://doi.org/10.1177/0272989x221099904

Peters, E., Slovic, P., Västfjäll, D., & Mertz, C. K. (2008). Intuitive numbers guide decisions. *Judgment and Decision Making*, *3*(8), 619–635. https://doi.org/10.1017/s1930297500001571

Peters, E., Tompkins, M. K., Knoll, M. A. Z., Ardoin, S. P., Shoots-Reinhard, B., & Meara, A. S. (2019). Despite high objective numeracy, lower numeric confidence relates to worse financial and medical outcomes. *Proceedings of the National Academy of Sciences*, *116*(39), 19386–19391. https://doi.org/10.1073/pnas.1903126116

Petrova, D. G., van der Pligt, J., & Garcia-Retamero, R. (2014). Feeling the numbers: On the interplay between risk, affect, and numeracy. *Journal of Behavioral Decision Making*, *27*(3), 191–199. https://doi.org/10.1002/bdm.1803

Piñon, A., & Gambara, H. (2005). A meta-analytic review of framing effect: risky, attribute and goal framing. *Psicothema*, *17*(2), 325–331.

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, *18*(1), 89–107. https://doi.org/10.1016/j.lindif.2007.03.011

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin*, *135*(6), 943–973. https://doi.org/10.1037/a0017327

Rolison, J. J., Morsanyi, K., & Peters, E. (2020). Understanding health risk comprehension: The role of math anxiety, subjective numeracy, and objective numeracy. *Medical Decision Making*, *40*(2), 222–234. https://doi.org/10.1177/0272989x20904725

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

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Traczyk, J., & Fulawka, K. (2016). Numeracy moderates the influence of task-irrelevant affect on probability weighting. *Cognition*, *151*, 37–41. https://doi.org/10.1016/j.cognition.2016.03.002

Traczyk, J., Sobkow, A., Fulawka, K., Kus, J., Petrova, D., & García-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. https://doi.org/10.1017/s1930297500009244

Tversky, A., & Kahneman, D. (1985). The Framing of Decisions and the Psychology of Choice. *Behavioral Decision Making*, 25–41. https://doi.org/10.1007/978-1-4613-2391-4_2

van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2–12. https://doi.org/10.1016/j.jesp.2016.03.004

Vonasch, A., Yiu, H., Leung, W., Nguyen, A., Stephanie, C., Cheng, B., & Feldman, G. (2022). "Less is better" in separate evaluations versus "More is better" in joint evaluations: Mostly successful close replication and extension of Hsee (1998). *Open Science Framework*. https://doi.org/10.17605/OSF.IO/9UWNS

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., & Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *Journal of Behavioral Decision Making*, *26*(2), 198–212. https://doi.org/10.1002/bdm.1751

Woloshin, S., Schwartz, L. M., & Welch, H. G. (2005). Patients and medical statistics. *Journal of General Internal Medicine*, *20*(11), 996–1000. https://doi.org/10.1007/s11606-005-0245-7

Yeung, S. K., & Feldman, G. (2022). Revisiting the Temporal Pattern of Regret in Action Versus Inaction: Replication of Gilovich and Medvec (1994) With Extensions Examining Responsibility. *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/collabra.37122

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*. https://doi.org/10.1017/s0140525x17001972

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/77608-revisiting-the-links-between-numeracy-and-decision-making-replication-registered-report-of-peters-et-al-2006-with-an-extension-examining-confidenc/attachment/162531.docx?auth_token=jp0Y6zl45_OTHbm5rBRm

Supplemental Material

Download: https://collabra.scholasticahq.com/article/77608-revisiting-the-links-between-numeracy-and-decision-making-replication-registered-report-of-peters-et-al-2006-with-an-extension-examining-confidenc/attachment/162532.docx?auth_token=jp0Y6zl45_OTHbm5rBRm

Peer Review and Communication History

MS Title: Revisiting the Links Between Numeracy and Decision Making: Replication Registered Report of Peters et al. (2006) with an Extension Examining Confidence

Author Names: Minrui Zhu and Gilad Feldman

Submitted: Mar 29, 2023

In-principle Acceptance and Open-review
Provided on: https://rr.peercommunityin.org/articles/rec?id=165

Editor Final Decision: Accept
Apr 23, 2023

Dear Gilad ,

Thank you for submitting your Stage 2 Registered Report manuscript that underwent peer review at PCI-RR. I am happy to say that your paper is now officially accepted for publication in Collabra: Psychology. Congratulations on this excellent work, I think it will make an important contribution to the literature and I look forward to seeing it published! I hope your experiences with Collabra: Psychology have been positive and that you will continue to consider it as an outlet for your work.

As there are no further reviewer revisions to make, you do not have to complete any tasks at this point.

You will be receiving separate correspondence regarding any production and technical comments, data deposits, as well as publication charges. We work with the Copyright Clearance Center to process any applicable APC charges. Please note that your APC transaction must be completed before your article gets published.

You will have an opportunity to check the page proofs before we publish your article. Thank you again for publishing in Collabra: Psychology.

Sincerely,
Simine Vazire
Editor in Chief
Collabra: Psychology

Revisiting the Links Between Numeracy and Decision Making: Replication Registered Report of Peters et al. (2006) with an Extension Examining Confidence

Supplementary

Contents

Open Science disclosures

Data Collection

Data collection was completed before analyzing the data.

Conditions Reporting

All collected conditions are reported.

Data Exclusions

Details are reported in the exclusions section of this document

Variables Reporting

All variables collected for this study are reported and included in the provided data.

Analysis of the Original Article

Original Article Methods

Experimental Design

Study 1

Study 1 used 2 between (High numeracy vs. Low numeracy x 2 within (Positive framing vs. Negative framing) x 5 within (five students' scores) mixed-ANOVA. The study tested the association between numeracy and framing effect.

Study 2

Study 2 used factorial ANOVA, 2(High numeracy vs. Low numeracy) * 2(Frequency vs. Percentage) to examine the relationship between numeracy and frequency-percentage effect.

Study 3

Study 3 used the chi-square test for the bowl choice (Bowl A-9-100 or Bowl B-1-10) between high numeracy and low numeracy. Then it used the independent *t*-test to compare numeracy (high vs. low) and bowl preference, affect precision and affect respectively.

Study 4

Study 4 used the factorial ANOVA, 2(High numeracy vs. Low numeracy) * 2(No loss vs. Loss). It also used the independent *t*-test to compare both affect and attractiveness of bet of the high numerate under loss and no loss conditions. In addition, it also uses independent *t*-tests to compare both affect and affect precision of high numeracy and low numeracy under the no-loss condition.

Original Article Results

Sample Size

Please see the summary of sample size of the original article in Table 1 in the main manuscript. Sample exclusion was not reported in the original article.

Major Findings

Please see the summary of major findings of four studies in Table S1, S2 and S3.

Table S1

Studies 1 and 2: Summary of results

Statistical Tests	Conditions	*M*	*SD*	*df*	*F*	*p*	*η2*
Mixed ANOVA	Numeracy and frame effect	/	/	1, 96	5.6	< .05	0.11
Two-way ANOVA	Numeracy and frequency-percentage effect	/	/	1,42	4.0	< .05	0.42

Note. M = Mean, *SD* = standard deviation.

Table S2

Study 3: Summary of results

Statistical Tests	Conditions	*M*	*SD*	*df*	*χ²/ t*	*p*	*d*
Chi-square	Numeracy and Bowl choice	/	/	1	5.2	< .05	/
Independent *t*-test	Numeracy and Bowl preference	Low numerate: 1.7 High numerate: 4.1	/	44	-2.5	< .05	0.75
Independent *t*-test	Numeracy and affect	Low numerate: -0.5 High numerate: -1.1	/	44	N/A	.13	0.46
Independent *t*-test	Numeracy and affect precision	Low numerate: 3.7 High numerate: 5.0	/	44	-2.6	< .01	0.78

Note. M = Mean, *SD* = standard deviation.

Table S3

Study 4: Summary of results

Statistical Tests	Conditions	*M*	*SD*	*df*	*t/F*	*p*	*d*
Independent *t*-test	High numeracy and attractiveness under loss/no loss conditions	/	/	89	3.1	< .01	N/A
Independent *t*-test	Numeracy and affect under no-loss condition	Low numerate: -0.6 High numerate: 0.0	/	169	2.7	<.01	0.44
Independent *t*-test	Numeracy and affect precision under no-loss condition	Low numerate: 3.9 High numerate: 4.5	/	149	2.4	< .05	0.38
Independent *t*-test	High numeracy and affect under loss/no loss condition	Loss:1.No loss: 9 1.3	/	89	2.3	< .05	0.47
Two-way ANOVA	Numeracy and attractiveness	/	/	1, 169	8.0	< .01	/
Two-way ANOVA	Numeracy and affect	/	/	1, 169	7.2	< .01	/

Note. M = Mean, *SD* = standard deviation.

Effect Size Calculations of the Original Study Effects

Please see the Rmarkdown code and output provided in the OSF folder:

Peters et al 2006-rep-ext-es-power-analysis.Rmd/html

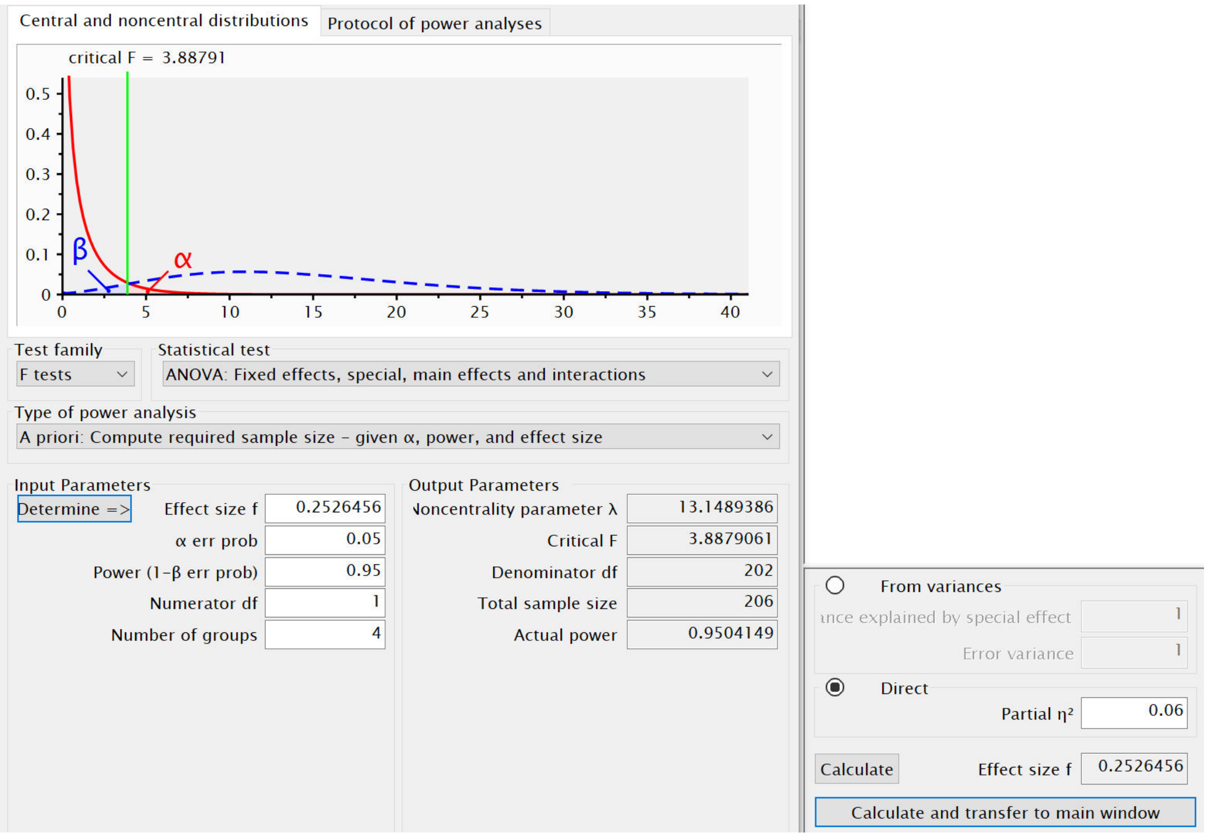
Power Analysis of Original Study Effect to Assess Required Sample for Replication

Please see the Rmarkdown code and output provided in the OSF folder:

Peters et al 2006-rep-ext-es-power-analysis.Rmd/html

Figure S1

The GPower input and output for the mixed ANOVA between numeracy and framing effect in Study 1


Central and noncentral distributions
Protocol of power analyses
critical F = 3.88791
0.5
0.4
0.3
0.2
0.1
0
β
α
0
5
10
15
20
25
30
35
40
Determine =>
Calculate
Calculate and transfer to main window

Test family: F tests

Statistical test: ANOVA: Fixed effects, special, main effects and interactions

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size f 0.2526456

α err prob 0.05

Power (1-β err prob) 0.95

Numerator df 1

Number of groups 4

Output Parameters

Noncentrality parameter λ 13.1489386

Critical F 3.8879061

Denominator df 202

Total sample size 206

Actual power 0.9504149

○ From variances

ınce explained by special effect 1

Error variance 1

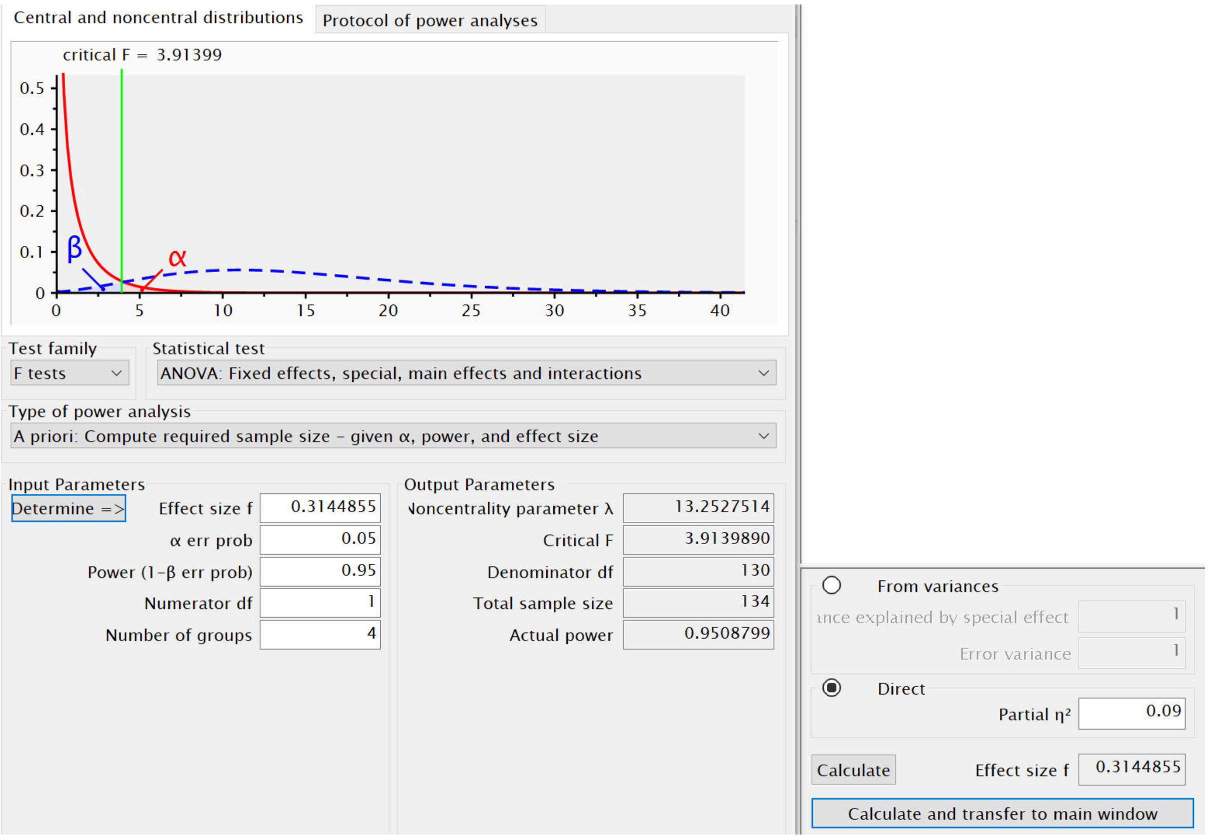
● Direct

Partial η² 0.06

Effect size f 0.2526456

Figure S2

The GPower input and output for the factorial ANOVA between numeracy and frequency-percentage effect in Study 2


Central and noncentral distributions
Protocol of power analyses
critical F = 3.91399
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
20
25
30
35
40
β
α
Determine =>
Calculate
Calculate and transfer to main window

Test family: F tests

Statistical test: ANOVA: Fixed effects, special, main effects and interactions

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size f: 0.3144855
α err prob: 0.05
Power (1-β err prob): 0.95
Numerator df: 1
Number of groups: 4

Output Parameters

Noncentrality parameter λ: 13.2527514
Critical F: 3.9139890
Denominator df: 130
Total sample size: 134
Actual power: 0.9508799

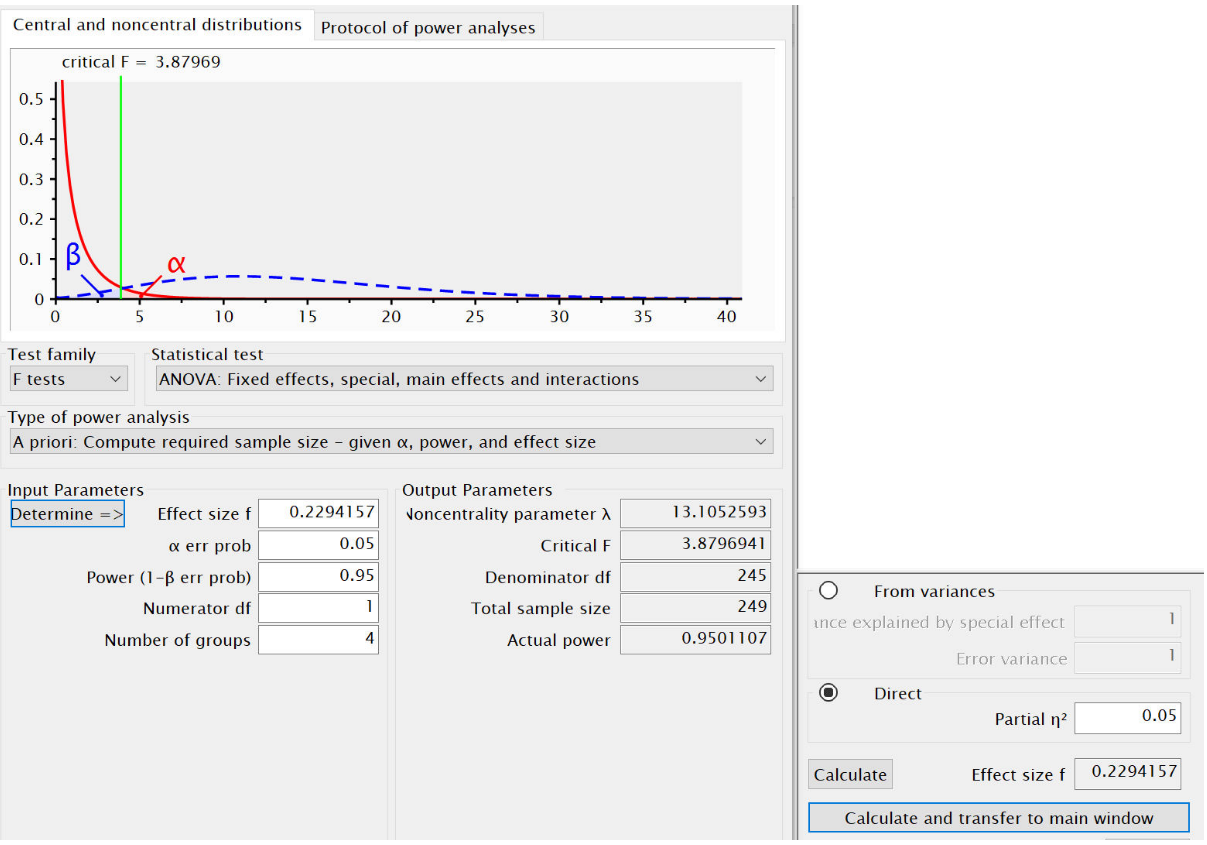
○ From variances
ınce explained by special effect: 1
Error variance: 1

● Direct
Partial η^2: 0.09

Effect size f: 0.3144855

Figure S3

The GPower input and output for the factorial ANOVA between numeracy and attractiveness of bet in Study 4


Central and noncentral distributions
Protocol of power analyses
critical F = 3.87969
0.5
0.4
0.3
0.2
0.1
0
β
α
0
5
10
15
20
25
30
35
40
Determine =>
Calculate
Calculate and transfer to main window

Test family: F tests

Statistical test: ANOVA: Fixed effects, special, main effects and interactions

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size f	0.2294157
α err prob	0.05
Power (1–β err prob)	0.95
Numerator df	1
Number of groups	4

Output Parameters

Noncentrality parameter λ	13.1052593
Critical F	3.8796941
Denominator df	245
Total sample size	249
Actual power	0.9501107

○ From variances

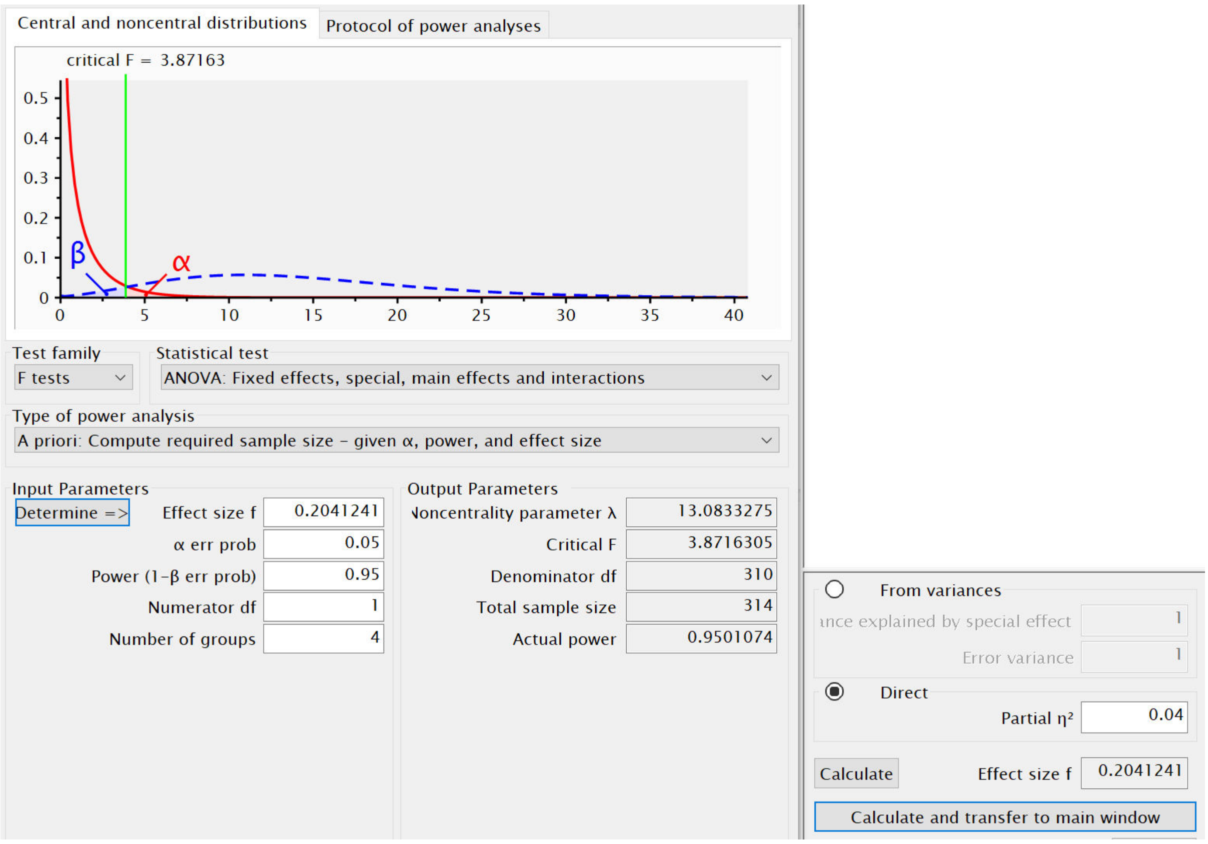
ınce explained by special effect	1
Error variance	1

● Direct

Partial η^2	0.05
Effect size f	0.2294157

Figure S4

The GPower input and output for the factorial ANOVA between numeracy and affect in Study 4


Central and noncentral distributions
Protocol of power analyses
critical F = 3.87163
0.5
0.4
0.3
0.2
0.1
0
β
α
0
5
10
15
20
25
30
35
40
Determine =>
Calculate
Calculate and transfer to main window

Test family: F tests

Statistical test: ANOVA: Fixed effects, special, main effects and interactions

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size f 0.2041241
α err prob 0.05
Power (1–β err prob) 0.95
Numerator df 1
Number of groups 4

Output Parameters

Noncentrality parameter λ 13.0833275
Critical F 3.8716305
Denominator df 310
Total sample size 314
Actual power 0.9501074

○ From variances
ınce explained by special effect 1
Error variance 1

● Direct
Partial η² 0.04

Effect size f 0.2041241

Comparison between original paper and materials authors provided

The authors indicated the materials were lost to time, but emailed us a list of materials of experiments from follow-up research. While using these for reproduction we realized that the contents deviated from the descriptions in the article. We therefore summarized the differences in Table below.

We are grateful for the authors' support of our project.

Table S4

Comparison between original paper and materials authors provided

Study	Descriptions from original paper	Materials from author	Deviations
1	N/A	Not sent	N/A
2	N/A	Not sent	N/A
3	Question of choice of bowl (inferred from Chi-square test) Question of bowl preference Question of affect and affect precision	Question of bowl preference Questions of affect and affect precision	No question of direct choice of Bowl
4	Scenario description: You will win nothing	Scenario description: You will lose nothing	Scenario description
	21-point scale on rating attractiveness from "Not at all attractive" (0) to "Extremely attractive" (21)	8-point scale on rating attractiveness from "Not at all attractive a bet" (1) to "Extremely attractive a bet" (7)	Scale on rating attractiveness

Materials and scales used in the replication + extension experiment

Procedure

1. Participants were recruited using the Amazon Mechanical Turk platform using CloudResearch.
2. Participants indicated agreement to consent form.
3. Participants read an overview of the experiment and warned about not looking for answers.
4. Participants answered the dependent variables of Studies 1, 2, 3, and 4 in random order.
5. Except for Study 3, participants were randomly assigned to one of two conditions and asked to complete the questions.
6. Participants completed two numeracy scales in a random order (original and extension) and answered a question whether they used external aids.
7. Participants answered three funneling questions about their seriousness towards the survey, conjecture of study purpose and provision of any feedback to the study.
8. Participants provided demographics (age, gender, place of birth, current residence, social class and engunder).
9. Participants rated the satisfaction with the pay offered for this task.
10. Finally, participants were debriefed.

Instructions and experimental material

Consent form

This study is conducted by Gilad Feldman of the psychology department at University of Hong Kong and colleagues. If you have questions or concerns regarding this project, please do not hesitate to contact Gilad Feldman gfeldman@hku.hk at any time.

Purpose of the study

To understand how people think, feel, make decisions, and act in various types of situations. Preferences and individual differences between people, as well as both internal and external factors, may affect these types of responses and this research intends to uncover and/or understand these processes.

Procedures.

This study will ask you to complete a set of questionnaires requiring decision making in various scenarios. The duration of this study has been indicated on the task that you accepted.

Potential risks.

This procedure has no known risks greater than those of ordinary daily life.

Potential benefits.

This study aims to add to existing research lines in the field of social-cognitive-personality psychology. We also hope that this study can provide you with a learning experience of participating in psychological research and possibly learning more about yourself and your beliefs, evaluations, preferences, personality, etc..

Compensation.
Compensation is offered through the online platform. The level of compensation has been indicated on the task that you accepted.

Confidentiality.
Your questionnaire responses are anonymous and strictly confidential. No personal identifiers are kept. Information obtained will only be used as aggregates for research purposes.

Participation and withdrawal.
Your participation is voluntary. This means that you can choose to stop at any time without negative consequences. If at any time you wish to withdraw, please simply indicate eight zeros as your completion code, and you will receive compensation regardless.

Questions and concerns
If you have any questions about the research, please feel free to contact Gilad Feldman at the University of Hong Kong (gfeldman@hku.hk). If you have questions about your rights as a research participant, contact the Human Research Ethics Committee, HKU (+852 2241-5267). EA210265

Please print a copy of this consent form for your records, if you so desire.

Study outline
This survey involves decision-making tasks. You will be asked to complete four tasks, then two short scales. After that, there are brief feedback and demographics questions.

For opinion related questions - There are no “right” or “wrong” opinion answers, so please state your opinion as honestly as possible. Thank you for your cooperation.

Warning about not looking for answers
Your statistical gut intuitions

In this study we will present you with various decisions and problems that require your statistical gut intuitions.

Important: This study only aims to test your statistical intuitions, not to test accuracy. We therefore ask that you please do not look for the answers to any of these problems, but rather answer based on what you know and think right now.

All answers are unidentified and anonymous, and are only used as aggregates for research to try and understand people's statistical intuitions.

To ensure that you understand the guidelines of not looking up answers to these questions, we ask that you please write down (or copy-paste) the following sentence to the text field below (not case sensitive)

I pledge to not search for answers to presented questions, and only answer based on my own knowledge and intuitions

Study 1 Experimental condition: Positive framing condition

Instructions:

Evaluating exam scores
Below are the exam scores of five psychology students.
Please rate each student's quality of work on a 7-point scale from "Very poor" (-3) to "Very good" (3)

Dependent variables:

a. Emily received 74% correct on her exam.
b. Jack received 78% correct on his exam.
c. Emma received 82% correct on her exam.
d. Oliver received 70% correct on his exam.
e. Sophia received 66% correct on her exam.

Extension question :

a. You gave the following ratings:
(scale: -3 = Very poor, 0 = Neutral, 3 = Very good)

[Selected Choice] for Emily (74% correct)
[Selected Choice] for Jack (78% correct)
[Selected Choice] for Emma (82% correct)
[Selected Choice] for Oliver (70% correct)
[Selected Choice] for Sophia (66% correct)

How confident are you that you made an accurate assessment of the five students?

Study 1 Experimental condition: Negative framing condition
Instructions:
Same as positive framing condition

Dependent variables :

a. Emily received 26% incorrect on her exam.
b. Jack received 22% incorrect on his exam.
c. Emma received 18% incorrect on her exam.
d. Oliver received 30% incorrect on his exam.
e. Sophia received 34% incorrect on her exam.

Extension question:

a. You gave the following ratings:
(scale: -3 = Very poor, 0 = Neutral, 3 = Very good)

[Selected Choice] for Emily (26% incorrect)
[Selected Choice] for Jack (22% incorrect)
[Selected Choice] for Emma (18% incorrect)
[Selected Choice] for Oliver (30% incorrect)

[Selected Choice] for Sophia (34% incorrect)

How confident are you that you made an accurate assessment of the five students?

Study 2 Experimental condition: Frequency condition

Instructions:
Estimating harm

In this study, please imagine the below scenario:

A patient – Mr. James Jones – has been evaluated for discharge from an acute civil mental health facility where he has been treated for the past several weeks.

A psychologist whose professional opinion you respect has done a state-of-the-art assessment of Mr. Jones. Among the conclusions reached in the psychologist's assessment is the following:

Of every 100 patients similar to Mr. Jones, 10 are estimated to commit an act of violence to others during the first several months after discharge.

Dependent variable :

a. Please rate the level of risk that Mr. Jones would harm someone.

Extension questions:

You rated the risk level of Mr. Jones: [Selected Choice]
(scale: 1 = Low risk, 6 = High risk).

a. How confident are you that made an accurate risk assessment?
b. Are you familiar with this scenario?

Study 2 Experimental condition: Percentage condition

Instructions:
Estimating harm

In this study, please imagine the below scenario:

A patient – Mr. James Jones – has been evaluated for discharge from an acute civil mental health facility where he has been treated for the past several weeks.

A psychologist whose professional opinion you respect has done a state-of-the-art assessment of Mr. Jones. Among the conclusions reached in the psychologist's assessment is the following:

Of every 100 patients similar to Mr. Jones, 10% are estimated to commit an act of violence to others during the first several months after discharge.

Dependent variable:

a. Please rate the level of risk that Mr. Jones would harm someone.

Additional questions:

a. You rated the risk level of Mr. Jones: [Selected Choice]
(scale: 1 = Low risk, 6 = High risk).

How confident are you that made an accurate risk assessment?

b. Are you familiar with this scenario?

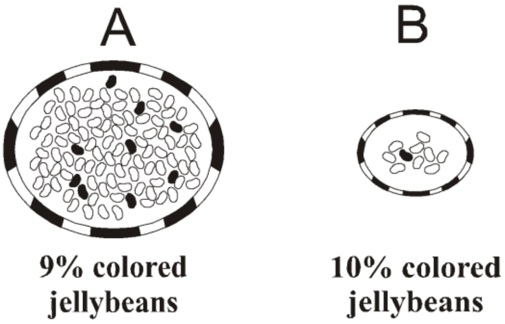
Study 3 Experiment condition

Instructions:
Jellybean bowls

Bowl A has 100 jellybeans, and Bowl B has 10 jellybeans.

Please imagine that once you have selected a bowl, it will be placed behind a screen, the experimenter will mix up the jellybeans randomly, and then you will reach around the screen (without looking at the bowl) and select a bean.

Imagine that if you select a colored bean, you will WIN $5.
Would you prefer to pick from bowl A or bowl B?



Dependent variables:

a. Which bowl would you prefer to choose from?
b. If you were forced to choose, which bowl would you prefer to choose from?
c. How clear a feeling do you have about the goodness or badness of Bowl A’s 9% chance of winning?
d. How good or bad does Bowl A’s 9% chance of winning make you feel?

Additional questions:

a. How clear a feeling do you have about the goodness or badness of Bowl B’s 10% chance of winning?
b. How good or bad does Bowl B’s 10% chance of winning make you feel?

c. You selected [Selected Choices], and gave the following ratings:
[Selected Choices] for the preference of Bowl A or Bowl B (scale: -6 = Strong preference about Bowl A, 6 = Strong preference about Bowl B)
[Selected Choices] for how clear the feeling about Bowl A's 9% chance of winning (scale: 0 = Completely unclear, 3 = Neutral, 6 = Completely clear)
[Selected Choices] for how clear the feeling about Bowl B 10% chance of winning (scale: 0 = Completely unclear, 3 = Neutral, 6 = Completely clear)

[Selected Choices] for how good or bad about Bowl A's 9% chance of winning (scale: -3 = Very bad, 0 = Neutral, 3 = Very good)
[Selected Choices] for how good or bad about Bowl B's 10% chance of winning (scale: -3 = Very bad, 0 = Neutral, 3 = Very good)

How confident are you that made an optimal selection between Bowl A and Bowl B?

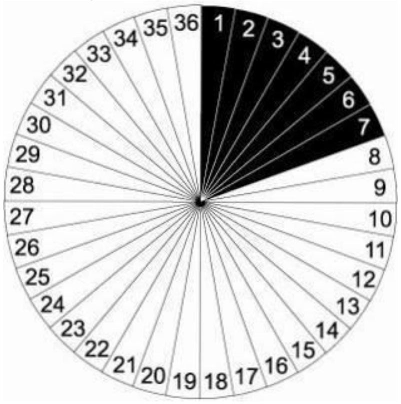
d. Are you familiar with this question?

Study 4 Experiment condition: No loss Condition
Instructions:
Will you take the bet?

We are interested in how attractive the prospect of playing the following bet is to you.

7/36 to win $9:
This means that there is a 7 out of 36 chance that you will win the bet and receive $9 and there is a 29 out of 36 chance that you will win nothing.

Here is a visualized roulette wheel with 36 numbers along the circumference. If a ball lands on any of the 7 numbers between 1 and 7 inclusive, you win $9. If it lands on 8-36, you win nothing.



Dependent variables:

a. Please indicate your opinion of this bet’s attractiveness. There is no right or wrong answer, we are interested only in your opinion about the attractiveness of playing the bet.
b. How clear a feeling do you have about the goodness or badness of the bet?
c. How good or bad does the bet make you feel?

Additional questions:

a. You gave the following ratings:

[Selected Choice] for bet's attractiveness (scale: 0 = Not at all attractive bet, 20= Extremely attractive bet)
[Selected Choice] for how clear a feeling about the bet (scale: 0 = Completely unclear, 3 = Neutral, 6 = Completely clear)
[Selected Choice] for how good or bad feel the the bet (scale: -3 = Very bad, 0 = Neutral, 3 = Very good)

How confident are you that made an accurate assessment of the bet's attractiveness?

b. Are you familiar with this scenario?

Study 4 Experiment condition: Loss Condition

Instructions

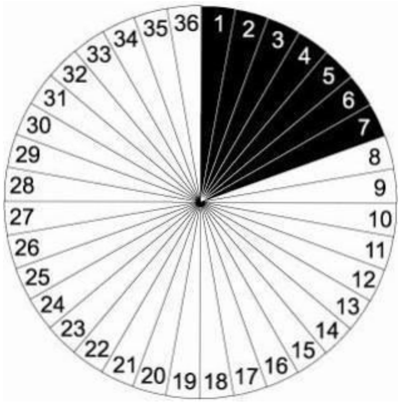
Will you take the bet?

We are interested in how attractive the prospect of playing the following bet is to you.

7/36 to win $9 29/36 to lose 5¢:

This means that there is a 7 out of 36 chance that you will win the bet and receive $9 and 29 out of 36 chance that you will lose 5¢.

Here is a visualized roulette wheel with 36 numbers along the circumference. If a ball lands on any of the 7 numbers between 1 and 7 inclusive, you win $9. If it lands on 8-36, you lose 5¢.



Dependent variables:

a. Please indicate your opinion of this bet’s attractiveness. There is no right or wrong answer, we are interested only in your opinion about the attractiveness of playing the bet.
b. How clear a feeling do you have about the goodness or badness of the bet?
c. How good or bad does the bet make you feel?

Additional questions:

a. You gave the following ratings:

   [Selected Choice] for bet's attractiveness (scale: 0 = Not at all attractive bet, 20= Extremely attractive bet)
   [Selected Choice] for how clear a feeling about the bet (scale: 0 = Completely unclear, 3 = Neutral, 6 = Completely clear)
   [Selected Choice] for how good or bad feel the the bet (scale: -3 = Very bad, 0 = Neutral, 3 = Very good)

   How confident are you that made an accurate assessment of the bet's attractiveness?
b. Are you familiar with this scenario?

Numeracy scale from original paper (Lipkus et al., 2001)
Instructions:
Statistical intuitions

Below are 14 questions presenting various decisions and problems that require your statistical gut intuitions. Please answer the question with a number (i.e., integer or decimal) where required.

Dependent variables:

a. Q1: Imagine that we roll a fair, six-sided die 1000 times. Out of 1000 rolls, how many times do you think the die would come up even (2, 4, or 6)? (Please input an integer)

b. Q2: In the BIG BUCKS LOTTERY, the chances of winning a $10 prize are 1%. What is your best guess about how many people would win a $10 prize if 1000 people each buy a single ticket from BIG BUCKS? (Please input an integer)

c. Q3: In the ACME PUBLISHING SWEEPSTAKES, the chance of winning a car is 1 in 1000. What percent of tickets of ACME PUBLISHING SWEEPSTAKES win a car? (Please input a decimal)

d. Q4: Which of the following numbers represents the biggest risk of getting a disease?
   1 in 100
   1 in 1000
   1 in 10

e. Q5: Which of the following represents the biggest risk of getting a disease?
   1%
   10%
   5%

f. Q6: If Person A's chance of getting a disease is 1% in ten years, and Person B's risk is double that of A, what is B's risk of getting a disease in ten years? (Please input an integer)

g. Q7: If Person A's chance of getting a disease is 1 in 100 in ten years, and Person B's risk is double that of A, what is B's risk (i.e, X in 100) of getting a disease in ten years? (Please input an integer)

h. Q8A: If the chance of getting a disease is 10%, how many people would be expected to get the disease out of 100? (Please input an integer)

i. A8B: If the chance of getting a disease is 10%, how many people would be expected to get the disease out of 1000? (Please input an integer)

j. Q9: If the chance of getting a disease is 20 out of 100, this would be the same as having a ____% chance of getting the disease. (Please input an integer)

k. Q10: The chance of getting a viral infection is .0005. Out of 10,000 people, about how many of them are expected to get infected? (Please input an integer)

Additional questions:

a. Have you ever come across any of the questions presented to you on this page?

If yes, You indicated you came across these questions before. Please briefly indicate where...

Rasch-based numeracy scale developed by Weller et al. (2011)

Instructions: Same as Numeracy scale in original paper

Dependent variables :

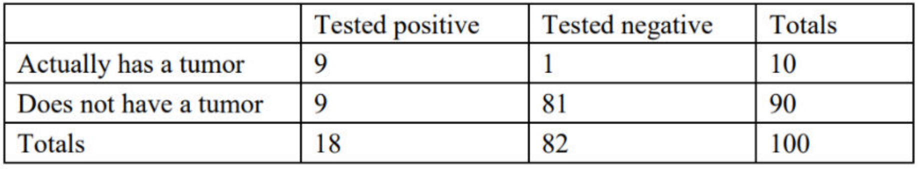
a. Q6: A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? (in dollars) (Please input a decimal)

b. Q7: In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? (Please input an integer)

c. Q8: Suppose you have a close friend who has a lump in her breast and must have a mammogram.

Of 100 women like her, 10 of them actually have a malignant tumor and 90 of them do not. Of the 10 women who actually have a tumor, the mammogram indicates correctly that 9 of them have a tumor and indicates correctly that 1 of them does not have a tumor. Of the 90 women who do not have a tumor, the mammogram indicates correctly that 81 of them do not have a tumor and indicates incorrectly that 9 of them do have a tumor.

The table below summarizes all of this information. Please Imagine that your friend tests positive (as if she had a tumor), what is the likelihood (i.e., X %) that she actually has a tumor? (Please input an integer)



	Tested positive	Tested negative	Totals
Actually has a tumor	9	1	10
Does not have a tumor	9	81	90
Totals	18	82	100

Additional questions :

a. Have you ever come across any of the questions presented to you on this page?

If yes, You indicated you came across these questions before. Please briefly indicate where...

Note: The questions: Q1, Q2, Q3, Q4, Q5 of rasch-based numeracy scale are the same as the questions Q1, Q2, Q3, Q8B, Q9 of numeracy scale in the original paper.

Exclusion for numeracy measurement

Instructions:

Our research depends on you using your intuition to answer our questions, so it is very important for us to know: Did you look up any questions? Did you use any aid to answer these questions?

You will be paid regardless, and there is no penalty, but for the sake of the accuracy of our research, we need to know.

Question:

a. I did NOT use any aids in answering this survey.
b. I DID use external aids to answer this survey

Funnelling section and Demographic section

Instructions:

Thank you, you completed the survey. A few quick final questions (3 on this page) and demographics (6 questions in the next page)...

Three funneling questions:

a. How serious were you in filling out this questionnaire? (1 = Not at all, 5 = Very much)
b. What do you think the purpose of the study was? (one sentence)
c. Help us improve for the next studies - Did you spot any errors? Anything missing or wrong? Something we should pay attention to in next runs? (briefly)

Six Demographic questions:

a. How old are you?
b. Please indicate your gender (Male/Female/Other/Rather not disclose)
c. Which country are you originally from? (country of birth)
d. In which country are you currently residing?
e. Please estimate your family's social class (Lower class/Working class/Lower middle class/Middle class/Upper middle class/Upper class)
f. How would you generally rate your understanding of the English used in this study? (Very bad/Bad/Poor/Neither good nor bad/Fair/Good/Very good)

Debriefing section

Instructions:

We would like to thank you for taking part and hope you found it interesting.

The experiments in which you participated today were designed to examine how personal and environmental factors may affect human cognition and decision making. In psychology, it has been known that information can affect person's behavior to a certain extent and that individual differences affect behavior. The purpose of the study was to know how exposure to stimuli and certain individual differences affect decision making and behavior.

It's important to note that all of the information that was collected today will be kept in complete confidentiality and there will be no attempt or interest in connecting your provided personal information with your responses. This data will be used for research purposes alone and not shared or reported to anyone. We are not interested in any one participant's responses by themselves. Rather, we are interested in the general responses of all participants when they are combined together.

We ask that you please do not share the details of this study with anyone because they may be potential participants and knowing the purpose of the study beforehand may affect the results. Thank you very much for your participation.

If you would like information about the results, or have further questions for us, please contact Gilad Feldman gfeldman@hku.hk at any time. You can also read more about Gilad's research in his website.

Scales used in the experiments

Study 1:

a. Rating for students' performance from "very poor" (-3) to "very good" (3)

Study 2:

a. Rating for risk level ranges from "low risk" (1) to "high risk" (6)

Study 3:

a. Rating for preferences of bowls ranges from "Strong preference for Bowl A"(6) to "Strong preference for Bowl B"(6)
b. Rating for affect precision ("how clear a feeling…") ranges from "completely unclear" (0) to completely clear (6)
c. Rating for affect ("how good or bad…") ranges from "very bad"(-3) to "very good"(+3)

Study 4:

a. Rating for attractiveness of bet ranges from "Not at all attractive bet"(0) to "Extremely attractive"(20)
b. Rating for affect precision ("how clear a feeling…") ranges from "completely unclear" (0) to completely clear (6)
c. Rating for affect ("how good or bad…") ranges from "very bad"(-3) to "very good"(+3)

Extension: Confidence questions in all four studies range from "not at all confidence"(0) to "Very confidence"(6)

Answer of Numeracy scales

Numeracy Scale in original article (Lipkus et al., 2001)

Q1: **500** out of 1000

Q2: **10** people out of 1000

Q3: 0.1%

Q4: **1** in 10

Q5: 10%

Q6: 2%

Q7: **2** out of 100

Q8A: 10

Q8B: 100

Q9: 20

Q10: 5

Rasch-based numeracy Scale (Weller et al., 2011)

Answers of Q1(Q1), Q2(Q2), Q3(Q3), Q4(Q8B), Q5(Q9) are shown above.

Q6: **5** cents

Q7: **47** days

Q8: **9** out of 100

Exclusion criteria

Generalized exclusion criteria

General criteria:

1. Participants indicating a low proficiency of English (self-report < 5, on a 1-7 scale)
2. Participants who self-report not being serious about filling in the survey (self-report < 4, on a 1-5 scale).
3. Participants who failed to complete the survey. (duration = 0, leave question blank)
4. Participants not from the US.

Specific criteria

1. Participants answer “yes” in “Have you ever come across any of the questions presented to you on this page?” at the end of original numeracy scale. The whole responses will be excluded.
2. Participants answer “I DID use external aids to answer this survey” after the completion of two numeracy scales. The whole responses will be excluded.
3. Participants who answer “yes” in “Have you ever come across any of the questions presented to you on this page?” at the end of the Rasch-based numeracy scale. The responses of this numeracy scale will be excluded.
4. Participants answer “yes” in familiarity questions in study 2. The responses of Study 2 will be excluded.
5. Participants answer “yes” in familiarity questions in study 3. The responses of Study 3 will be excluded.

6. Participants answer “yes” in familiarity questions in study 4. The responses of Study 4 will be excluded.

Handling outliers: Strategy

Outlier handling strategy followed the recommendations by Leys et al. (2019). The median absolute deviation (MAD) was used to detect univariate outliers. After the detection, we found 41 outliers in the original numeracy scale and 20 outliers in the rasch-based numeracy scale. In detail, the score of the original numeracy scale smaller than 6.08 were outliers and that of the rasch-based numeracy scale smaller than 4.2 were outliers. However, we are determined to keep them as they rightfully belong to the distribution of interest despite the increase of variances and decrease in statistical power. In addition, it is informative that a small part of participants would achieve relatively low numeracy scores.

Comparisons and deviations

Overview of post-exclusions

The number of excluded participants for each exclusion criterion is summarized in Table Table S5 and the number of participants after exclusion in each study is summarized in Table S6.

Table S5

Summary of participants fulfilling the exclusion criteria

Exclusion criteria	Number of excluded participants
General criteria 1 (Low proficiency of English)	6
General criteria 2 (Low seriousness of the study)	6
General criteria 3 (Failed completion of study)	0
General criteria 4 (Not US participants)	0
Specific criteria 1 (Familiarity of original numeracy scale)	483
Specific criteria 2 (Usage of external aids)	9
Specific criteria 3 (Familiarity of rasch-based numeracy scale)	545
Specific criteria 4 (Familiarity of scenario in Study 2)	225
Specific criteria 5 (Familiarity of scenario in Study 3)	318
Specific criteria 6 (Familiarity of scenario in Study 4)	305

Table S6

Summary of post-exclusions in four studies

	Number of post-excluded participants (analyzed with original numeracy scale)	Number of post-excluded participants (analyzed with rasch-based numeracy scale)
Study 1	339	234
Study 2	253	173
Study 3	221	151
Study 4	224	153

Replication: Main effect

Table S7

Studies 1-4: Descriptive statistics and main effect

Study	Conditions		Main effect
1	Positive framing	Negative framing	Framing effect
	M = 0.49, *SD* = 0.71, *n* = 169	*M* = -0.07, *SD* = 1.11, *n* = 198	*t*(363) = 5.56, *p* < .001, *d* = 0.58, 95% CI [0.37, 0.80]
2	Frequency condition	Percentage condition	Frequency-percentage effect
	M = 3.04, *SD* = 1.30, *n = 125*	*M* = 2.64, *SD* = 1.07, *n* = 141	*t*(264) = 2.76, *p* = .006, *d* = 0.34, 95% CI [0.09, 0.58]
3	Choice of Bowls		Preference of Bowls
	/	/	*t*(235) = 9.81, *p* < .001, *d* = 0.64, 95% CI [0.50, 0.78]
4	Bet - No loss Condition	Bet - Loss condition	
	Attractiveness		Affect of Bet
	M = 6.25, *SD* = 4.56, *n* = 122	*M* = 9.08, *SD* = 7.03, *n* = 116	*t*(236) = 3.70, *p* < .001 *d* = 0.48, 95% CI [0.22, 0.74]
	Affect		
	M = -0.65, *SD* = 1.36, *n* = 124	*M* = -0.25, *SD* = 1.70, *n* = 116	/
	Affect precision		
	M = 4.11, *SD* = 1.43, *n* = 124	*M* = 4.65, *SD* = 1.45, *n* = 116	/

Note. *n* = number of participants, *M* = mean, *SD* = standard deviation.

Replication: Original's analyses with dichotomized numeracy

Table S8

Studies 1, 2 and 4: Summary of statistical tests

	F	*df*	*p*	η^2_p and CI	Interpretation
Study 1 (Mixed ANOVA)					
Numeracy and framing effect	1.49	1, 362	.223	0.00 [0.00, 0.02]	no-signal inconsistent
Study 2 (Factorial ANOVA)					
Numeracy and frequency-percentage effect	0.07	1, 262	.794	0.00 [0.00, 0.01]	no-signal inconsistent
Study 4 (Factorial ANOVA)					
Numeracy and attractiveness of bet	2.09	1, 234	.149	0.01 [0.00, 0.04]	no-signal inconsistent
Numeracy and affect of bet	1.05	1, 234	.307	0.00 [0.00, 0.03]	no-signal inconsistent
Numeracy and affect precision of bet	2.53	1, 234	.113	0.01 [0.00, 0.04]	no-signal consistent

Note. CI = 90% confidence intervals. The interpretation of outcome is based on LeBel et al. (2019).

Table S9

Study 3: Summary of statistical tests

Low versus High numerate and Bowl Choice					
Chi-square test	*χ2*	*df*	*p*	*Cramer's V* and CI	
Numeracy as dichotomized continuous variable	0.36	1	.548	0.04 [0.00, 0.17]	no-signal
Dichotomized forced choices	3.00	1	.083	0.11 [0.00, 0.24]	no-signal
Low versus high numerate	*t*	*df*	*p*	*d* and CI	Interpretation
Preference of Bowls	-1.59	234	.112	0.21 [0.05, 0.47]	no-signal inconsistent
Affect for Bowl A-9-100	2.02	234	.045	0.27 [0.00, 0.53]	signal inconsistent
Affect precision for Bowl A-9-100	0.16	234	.873	0.02 [-0.24, 0.28]	no-signal inconsistent

Note. CI = 95% confidence intervals. Independent *t*-test comparing the stated DVs between the high and low numerate split sub-samples. Dichotomized continuous numeracy is categorized bowl choices according to the preference of bowl. Dichotomized forced bowl choices is the adjusted DV. The interpretation of outcome is based on LeBel et al. (2019).

<u>*Extension: Analyses using continuous numeracy*</u>

Table S10

Studies 1, 2, 3, and 4: Summary of statistical tests

Correlation		*r* and CI	*p*	Spearman's rho
Original numeracy scale				
Study 1	Rating of students in Positive framing condition	-0.05 [-0.20, 0.10]	.492	-0.06
	Rating of students in Negative framing condition	0.09 [-0.05, 0.22]	.224	0.09
Study 2	Risk rating in Frequency condition	-0.15 [-0.32, 0.02]	.088	-0.11
	Risk rating in Percentage condition	0.00 [-0.16, 0.17]	.986	-0.01
Study 3	Bowl preference	0.09 [-0.04, 0.22]	.154	0.09
	Affect for Bowl A-9-100	-0.17 [-0.29, -0.04]	.011	-0.11
	Affect precision for Bowl A-9-100	0.04 [-0.09, 0.17]	.520	0.02
Study 4	No Loss condition			
	Attractiveness	0.01 [-0.17, -0.19]	.933	0.02
	Affect	-0.07 [-0.24, 0.11]	.453	-0.03
	Affect precision	0.09 [-0.09, 0.27]	.312	-0.02
	Loss condition			
	Attractiveness	0.20 [0.02, 0.37]	.028	0.20
	Affect	0.09 [-0.09, 0.27]	.331	0.11

	Affect precision	0.29 [0.12, 0.45]	.001	0.14
Rasch-based numeracy scale				
Study 1	Rating of students in Positive framing condition	-0.01 [-0.19, 0.18]	.941	-0.03
	Rating of students in Negative framing condition	-0.01 [-0.17, 0.16]	.926	0.00
Study 2	Risk rating in Frequency condition	-0.20 [-0.40, 0.01]	.057	-0.18
	Risk rating in Percentage condition	0.02 [-0.18, 0.22]	.855	0.00
Study 3	Bowl preference	0.16 [0.01, 0.31]	.043	0.18
	Affect for Bowl A-9-100	-0.16 [-0.30, 0.00]	.048	-0.14
	Affect precision for Bowl A-9-100	0.16 [0.01, 0.31]	.043	0.12
Study 4	No Loss condition			
	Attractiveness	0.04 [-0.18, 0.26]	.720	0.09
	Affect	0.01 [-0.21, 0.23]	.924	0.06
	Affect precision	0.07 [-0.15, 0.28]	.558	0.00
	Loss condition			
	Attractiveness	0.11 [-0.11, 0.32]	.337	0.09
	Affect	0.01 [-0.21, 0.23]	.929	0.02
	Affect precision	0.43 [0.24, 0.60]	< .001	0.28

Note. CI = 95% confidence intervals. Interpretation is using the LeBel et al. (2019) criteria.

Table S11

Studies 1, 2, and 4: Comparisons of correlations

		Fisher's *z*	*p*	Interpretation
Original numeracy scale				
Study 1	Numeracy and framing effect	-1.33	.184	no-signal inconsistent
Study 2	Numeracy and frequency-percentage effect	-1.22	.224	no-signal inconsistent
Study 4	Numeracy and attractiveness of bets	-1.47	.142	no-signal inconsistent
	Numeracy and affect	-1.22	.222	no-signal consistent
	Numeracy and affect precision	-1.59	.113	no-signal inconsistent
Rasch-based numeracy scale				
Study 1	Numeracy and framing effect	0.00	1.000	no-signal inconsistent
Study 2	Numeracy and frequency-percentage effect	-1.49	.139	no-signal inconsistent
Study 4	Numeracy and attractiveness of bets	-0.44	.659	no-signal inconsistent
	Numeracy and affect	0.00	1.000	no-signal inconsistent
	Numeracy and affect precision	-2.44	.015	signal inconsistent

Note. The interpretation of outcome is based on LeBel et al. (2019).

Table S12

Study 3: Numeracy and optimal bowl choice

Independent *t*-test	*t*	*df*	*p*	*d* and CI	Interpretation
Original numeracy scale					
Bowl Choice	2.21	234	.028	0.33 [0.03, 0.62]	signal consistent
Rasch-based numeracy scale					
Bowl Choice	2.06	234	.040	0.31 [0.01, 0.60]	signal consistent

Note. CI = 95% confidence intervals. Independent *t*-test comparing the numeracy between Bowl A and Bowl B. The interpretation of outcome is based on LeBel et al. (2019).

Extension: Confidence

Table S13

Confidence: Summary of statistical tests in Studies 1-4

Correlation		*r* and CI	*p*	Spearman's rho
Original numeracy scale and Confidence level				
Study 1	Positive framing condition	-0.01 [-0.16, 0.14]	.909	0.00
	Negative framing condition	-0.07 [-0.20, 0.07]	.361	-0.05
Study 2	Frequency condition	-0.01 [-0.18, 0.17]	.955	0.01
	Percentage condition	0.03 [-0.14, 0.19]	.764	0.01
Study 3		0.16 [0.03, 0.28]	.014	0.19
Study 4	No loss condition	0.18 [0.00, 0.35]	.048	0.00
	Loss condition	0.11 [-0.07, 0.29]	.229	0.06
Rasch-based numeracy scale and Confidence level				
Study 1	Positive framing condition	0.04 [-0.14, 0.23]	.638	0.05
	Negative framing condition	0.03 [-0.13, 0.20]	.692	0.07
Study 2	Frequency condition	-0.01 [-0.22, 0.20]	.919	0.00
	Percentage condition	0.04 [-0.16, 0.24]	.695	0.09
Study 3		0.17	.034	0.17

		[0.01, 0.31]		
Study 4	No loss condition	0.17 [-0.05, 0.38]	.123	0.04
	Loss condition	0.23 [0.01, 0.42]	.041	0.15

Note. CI = 95% confidence intervals.

Pre-exclusions versus post-exclusions

Table S14

Summary for pre-exclusions and post-exclusions

Study	Main effects/ Interaction effects with numeracy	Replication methods	Pre-exclusion (cohen's *d/η²ₚ/Cramer's V*/Fishers' *z* and CI)	Post-exclusion (cohen's *d/η²ₚ/Cramer's V/r* and CI)
Original numeracy scale				
1	Framing effect	Independent *t*-test	0.62 [0.48, 0.76]	0.58 [0.37, 0.80]
	Numeracy and framing effect	Mixed ANOVA	0.01 [0.00, 0.02]	0.00 [0.00, 0.02]
	Numeracy and framing effect	Correlation comparison	-2.49	-1.33
2	Frequency-percentage effect	Independent *t*-test	0.37 [0.23, 0.50]	0.34 [0.09, 0.58]
	Numeracy and frequency-percentage effect	Factorial ANOVA	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]
	Numeracy and frequency-percentage effect	Correlation comparison	-2.07	-1.22
3	Bowl preference	Independent *t*-test	0.68 [0.61, 0.76]	0.64 [0.50, 0.78]
	Numeracy and bowl choice (Original)	Chi-square test	0.13 [0.06, 0.20]	0.04 [0.00, 0.17]
	Numeracy and bowl choice (Adjusted)	Chi-square test	0.17 [0.10, 0.24]	0.11 [0.00, 0.24]
	Numeracy and Bowl preference	Independent *t*-test	-0.35 [-0.53, -0.17]	0.21 [0.05, 0.47]
	Numeracy and affect precision for Bowl A-9-100	Independent *t*-test	-0.21 [-0.29, 0.06]	0.02 [-0.24, 0.28]
	Numeracy and affect for Bowl A-	Independent *t*-test	0.33 [0.09, 0.45]	0.27 [0.00, 0.53]

	9-100			
	Numeracy and bowl choice (Adjusted)	Independent *t*-test	0.55 [0.38, 0.72]	0.33 [0.03, 0.62]
	Numeracy and bowl preference	Correlation	0.20 [0.14, 0.27]	0.16 [0.01, 0.31]
	Numeracy and affect precision for Bowl A-9-100	Correlation	0.17 [0.11, 0.24]	0.16 [0.01, 0.31]
	Numeracy and affect for Bowl A-9-100	Correlation	-0.20 [-0.26, -0.13]	-0.16 [-0.30, 0.00]
4	Attractiveness of bet	Independent *t*-test	0.52 [0.38, 0.65]	0.48 [0.22, 0.74]
	Numeracy and attractiveness of bet	Factorial ANOVA	0.02 [0.01, 0.04]	0.01 [0.00, 0.04]
	Numeracy and attractiveness of bet	Correlation Comparison	-5.03	-1.47
	Numeracy and affect precision of bet	Factorial ANOVA	0.00 [0.00, 0.00]	0.01 [0.00, 0.04]
	Numeracy and affect precision of bet	Correlation Comparison	0.45	-1.59
	Numeracy and affect of bet	Factorial ANOVA	0.01 [0.00, 0.03]	0.00 [0.00, 0.03]
	Numeracy and affect of bet	Correlation Comparison	-3.82	-1.22
Rasch-based numeracy scale				
Study 1	Numeracy and framing effect	Correlation comparison	-2.79	0.00
Study 2	Numeracy and frequency-percentage effect	Correlation comparison	-2.51	-1.49
Study 3	Numeracy and bowl preference	Correlation	0.20 [0.14, 0.27]	0.16 [0.01, 0.31]

	Numeracy and Bowl Choice (Extension)	Independent *t*-test	0.54 [0.37, 0.70]	0.31 [0.01, 0.60]
	Numeracy and affect precision for Bowl A-9-100	Correlation	0.17 [0.11, 0.24]	0.16 [0.01, 0.31]
	Numeracy and affect for Bowl A-9-100	Correlation	-0.20 [-0.26, -0.13]	-0.16 [-0.30, 0.00]
Study 4	Numeracy and attractiveness of bet	Correlation Comparison	-4.14	-0.44
	Numeracy and affect precision of bet	Correlation Comparison	0.30	-2.44
	Numeracy and affect of bet	Correlation Comparison	-2.93	0.00

Note. CI = 95% Confidence Interval for independent *t*-test, chi-square test and correlation. CI = 90% Confidence Interval for mixed ANOVA and factorial ANOVA.

Table S15

Summary for pre-exclusions and post-exclusions

Original numeracy scale			Pre-exclusion *r* and CI	Post-exclusion *r* and CI
Study 1	Confidence and numeracy under positive framing	Correlation	-0.11 [-0.20, -0.02]	-0.01 [-0.16, 0.14]
	Confidence and numeracy under negative framing	Correlation	-0.03 [-0.13, 0.06]	-0.07 [-0.20, 0.07]
Study 2	Confidence and numeracy under frequency condition	Correlation	-0.01 [-0.10, 0.09]	-0.01 [-0.18, 0.17]
	Confidence and numeracy under percentage condition	Correlation	-0.01 [-0.11, 0.08]	0.03 [-0.14, 0.19]
Study 3	Confidence and numeracy	Correlation	0.15 [0.08, 0.21]	0.16 [0.03, 0.28]
Study 4	Confidence and numeracy under bet no loss condition	Correlation	0.10 [0.01, 0.20]	0.18 [0.00, 0.35]
	Confidence and numeracy under bet loss condition	Correlation	0.05 [-0.05, 0.14]	0.11 [-0.07, 0.29]
Rasch-based numeracy scale			Pre-exclusion *r* and CI	Post-exclusion *r* and CI
Study 1	Confidence and numeracy under positive framing	Correlation	-0.10 [-0.19, -0.01]	0.04 [-0.14, 0.23]
	Confidence and numeracy under negative framing	Correlation	-0.04 [-0.14, 0.05]	0.03 [-0.13, 0.20]
Study 2	Confidence and numeracy under frequency condition	Correlation	-0.01 [-0.10, 0.09]	-0.01 [-0.22, 0.20]
	Confidence and numeracy under	Correlation	0.00 [-0.10, 0.09]	0.04 [-0.16, 0.24]

	percentage condition			
Study 3	Confidence and numeracy	Correlation	0.14 [0.08, 0.21]	0.17 [0.01, 0.31]
Study 4	Confidence and numeracy under bet no loss condition	Correlation	0.11 [0.01, 0.20]	0.17 [-0.05, 0.38]
	Confidence and numeracy under bet loss condition	Correlation	0.06 [-0.03, 0.16]	0.23 [0.01, 0.42]

Note. CI = 95% Confidence Interval. The report of pre-exclusions and post-exclusions will be complete after data collection.

Pre-registration plan versus final report

Table S16

Preregistration planning and deviation documentation

Components in pre-registration	**Location of 1) pre-registered decision/plan and 2) rationale for decision/plan** **[Location / link]**	**Were there deviations? What type?** **[no / minor / major]***	**If yes - describe details of deviation(s)** **[brief description / location / link]**	**Rationale for deviation** **[brief description / location / link]**	**How might the results be different if you had/had not deviated** **[brief description / location / link]**	**Date/time of decision for deviation + stage**
Hypotheses	1) Page 13 Table 1 “Summary of hypotheses of replication and extension” 2) We reframed the hypothesis in Study 4 to align it with the original article.	No	/	/	No effect	Stage 2
Measured variables	1) Page 27 Table 6 “Study 3: Replication and extension experimental design” 2) We corrected the typo: the rating of strong preference of Bowl A from “-6” to “6”. 6 was shown on the questionnaire.	No	/	/	No effect	Stage 2
Measured variables	In our manuscript, we changed the description “Bowl A” to “Bowl A-9-100” and “Bowl B” to “Bowl B-1-10” to decrease cognitive load on readers and make it clearer what A and B stand for.	No	/	/	No effect	Stage 2

Exploratory analysis	1) Page 46 (Exploratory analysis) Paragraph 4 2) We removed the comparison of the original effect and extension effect as exploratory analysis. We realized that effects conversions are not helpful and can be misleading.	No	/	/	No effect	Stage 2
Exploratory analysis	1) Page 46 (Exploratory analysis) Paragraph 5 2) We did not report timer completion as planned exploratory analysis. The time of completion of scales didn’t give us too much information about numeracy scales and subsequent analysis. In addition, we could not explain why some participants took longer to complete the questionnaire.	No	./	/	No effect	Stage 2

*Notes.**Categories for deviations: Minor - Change probably did not affect results or interpretations; Major - Change likely affected results or interpretations.

Additional analyses and results

Descriptives of original analyses with dichotomous numeracy

Table S17

Descriptives for the original analyses with dichotomous numeracy

Study	Conditions		High numerate	Low numerate
Study 1	Positive framing		*M* = 0.41, *SD* = 0.75	*M* = 0.62, *SD* = 0.73
	Negative framing		*M* = -0.07, *SD* = 1.06	*M* = -0.15, *SD* = 1.10
Study 2	Frequency		*M* = 2.93, *SD* = 1.23	*M* = 3.21, *SD* = 1.31
	Percentage		*M* = 2.59, *SD* = 1.22	*M* = 2.55, *SD* = 1.06
Study 4	Bet effect	No-loss bet	*M* = 5.95, *SD* = 4.47	*M* = 6.78, *SD* = 4.74
		Loss bet	*M* = 10.27, *SD* = 7.21	*M* = 7.50, *SD*= 6.44
	Affect precision	No-loss bet	*M* = 4.42, *SD* = 1.35	*M* = 4.17, *SD* = 1.43
		Loss bet	*M* = 4.89, *SD* = 1.20	*M* = 4.61, *SD* = 1.44
	Affect	No-loss bet	*M* = -0.72, *SD* = 1.35	*M* = -0.38, *SD* = 1.36
		Loss bet	*M* = 0.16, *SD* = 1.83	*M* = -0.18, *SD* = 1.61

Note. M = mean, *SD* = standard deviation.

Assumption Checks

Table S18

Comparisons between parametric tests and non-parametric tests among four studies before exclusion

	Main effects/Interaction effect	Results of Parametric tests (Mixed ANOVA/ Factorial ANOVA/Independent t-test)	Non-parametric tests (Aligned Rank Transform/ Mann-Whitney U test)
Study 1	Framing effect	$t(858) = 9.12, p < .001, d = 0.62$, 95% CI [0.48, 0.76]	$U = 63253, p < .001$, $r_{rb} = 0.32$
	Numeracy and framing effect	$F(1, 855) = 4.42, p = .025, \eta^2_p = 0.01$, 90% CI [0.00, 0.02]	$F(1, 856) = 3.22, p = .07$
Study 2	Frequency-percentage effect	$t(853) = 9.12, p < .001, d = 0.37$, 95% CI [0.23, 0.50]	$U = 72873, p < .001$, $r_{rb} = 0.21$
	Numeracy and frequency-percentage effect	$F(1, 856) = 3.40, p = .065, \eta^2_p = 0.00$, 90% CI [0.00, 0.01]	$F(1,856) = 4.18, p = .04$
Study 3	Numeracy and preferences of bowls	$t(859) = 20.04, p < .001, d = 0.68$, 95% CI [0.61, 0.76]	$U = 63546, p < .001$, $r_{rb} = 0.23$
	Numeracy and affect of Bowl A-9-100	$t(858) = -4.62, p < .001, d = 0.33$, 95% CI [0.19, 0.48]	$U = 68273, p < .001$, $r_{rb} = 0.17$
	Numeracy and affect precision of Bowl A-9-100	$t(858) = 3.00, p = .003, d = 0.22$, 95% CI [0.07, 0.36]	$U = 72679, p = .004$, $r_{rb} = 0.12$
Study 4	Bets effect	$t(858) = 7.66, p < .001, d = 0.52$, 95% [0.38, 0.66]	$U = 72338, p < .001$, $r_{rb} = 0.22$
	Numeracy and bets effect	$F(1, 856) = 17.87, p < .001, \eta^2_p = 0.02$,	$F(1, 856) = 16.54, p < .001$

		90% CI [0.01, 0.04])	
	Numeracy and affect of bets	$F(1, 856) = 17.87$, $p < .001$, $\eta^2_p = 0.02$, 90% CI [0.01, 0.04]	$F(1, 856) = 7.26$, $p < .001$
	Numeracy and affect precision of bets	$F(1, 856) = 0.02$, $p = .890$, $\eta^2_p = 0.00$, 90% CI [0.00, 0.00]	$F(1, 856) = 0.25$, $p = .616$

Exploratory analyses

Table S19

Comparisons of affect precision and affect towards Bowl A-9-100 *and Bowl B*-1-10

Before exclusion			
	Bowl A-9-100	Bowl B-1-10	Main effect
Affect precision	*M* = 4.42*, SD* = 1.58, *N* = 860	*M* = 4.65*, SD* = 1.47, *N* = 860	*t*(859) = 6.09, *p* < .001, *d* = 0.21, 95% CI [0.14, 0.28]
Affect	*M* = -0.78*, SD* = 1.38*, N* = 860	*M* = -0.22*, SD* = 1.50, *N* = 860	*t*(859) = 11.86, *p* < .001, *d* = 0.40, 95% CI [0.33, 0.47]
After exclusion			
Affect precision	*M* = 4.29*, SD* = 1.59, *N* = 236	*M* = 4.53*, SD* = 1.46, *N* = 236	*t*(235) = 3.28, *p* = .001, *d* = 0.21, 95% CI [0.08, 0.34]
Affect	*M* = -0.81, *SD* = 1.40*, N* = 236	*M* = -0.09*, SD* = 1.60, *N* = 236	*t*(235) = 7.13, *p* < .001, *d* = 0.46, 95% CI [0.33, 0.60]

Note. *N* = number of participants, *M* = mean, *SD* = standard deviation.

Additional information about the study

1. Setting: The study was conducted online via an online questionnaire using Qualtrics. There was no fixed physical setting in which the study was conducted. In addition, we did not disallow participation using any specific devices.
2. Duration of Study Sessions: Participants were expected 10 minutes to complete all study materials, sessions would be ended earlier if participants completed study earlier. The average time was 14.8 minutes.
3. Time of Day: As questionnaires are conducted online, there is no limit to what time of the day the participants should complete the questionnaire. They could do it at any time of their convenience.
4. Data collection dates: Data collection started on May 16, 2022, and ended on May 17, 2022.
5. Participant Recruitment: Participants were recruited using Amazon Mechanical Turk.

Data collection procedures:

This study was conducted on Amazon Mechanical Turk with American participants. We imposed the following settings in recruiting our participants:

1. Participants were paid $1.25 as a fixed participation reward. This amount was determined by multiplying the expected completion time (in mins.) with the minimal federal wage in the U.S. (i.e., $7.25 per hour).
2. The expected completion time was set at 10 minutes in advance.
3. The most time we allowed each worker to complete the study was 30 minutes.
4. We limited all workers' HIT Approval Rate to be between 95% and 100%.
5. We limited each worker's number of HITs approved to be between 5,000 and 100,000.
6. We blocked Suspicious Geocode Locations and Universal Exclude List Workers.
7. We blocked duplicate IP addresses and duplicate geolocation.
8. We enabled HyperBatch so that all eligible workers were able to participate in our HIT immediately after the survey was launched.
9. We restricted workers' location to be in the U.S.

Replication evaluation

Replication closeness

Given provided details on the classification of the replications using the criteria by LeBel et al., (2018) criteria in Table below and details of deviation in Table 8 in the manuscript. We summarized the replication as a "very close" replication.

A classification of relative methodological similarity of a replication study to an original study. "Same" ("different") indicates the design facet in question is the same (different) compared to an original study. IV = independent variable. DV = dependent variable. "Everything controllable" indicates design facets over which a researcher has control. Procedural details involve minor experimental particulars (e.g., task instruction wording, font, font size, etc.).

"Similar" category was added to the Lebel et al. (2018) typology to refer to minor deviations or extensions aimed to adjust the study to the target sample that are not expected to have major implications on replication success. See Olsson-Collentine, van Assen, and Wicherts (2020) on meta analysis showing minor to no expected impact due to variations in sample population or setting.

Table S20

Criteria for evaluation of replications by LeBel et al. (2018)

Target similarity	Highly similar				Highly dissimilar
Category	Direct replication			Conceptual replication	
Design facet	**Exact replication**	**Very close replication**	**Close replication**	**Far replication**	**Very far replication**
Effect/hypothesis	Same/similar	Same/similar	Same/similar	Same/similar	Same/similar
IV construct	Same/similar	Same/similar	Same/similar	Same/similar	Different
DV construct	Same/similar	Same/similar	Same/similar	Same/similar	Different
IV operationalization	Same/similar	Same/similar	Same/similar	Different	
DV operationalization	Same/similar	Same/similar	Same/similar	Different	
Population (e.g. age)	Same/similar	Same/similar	Same/similar	Different	
IV stimuli	Same/similar	Same/similar	Different		
DV stimuli	Same/similar	Same/similar	Different		
Procedural details	Same/similar	Different			
Physical setting	Same/similar	Different			
Contextual variables	Different				

Note. “Same” (“different”) indicates the design facet in question is the same (different) compared to an original study. "Similar" category was added to the Lebel et al. (2018) typology to refer to minor deviations or extensions aimed to adjust the study to the target sample that are not expected to have major implications on replication success. See Olsson-Collentine, van Assen, and Wicherts (2020) on meta analysis showing minor to no expected impact due to variations in sample population or setting.

Replication versus the original

Figure S5

Interpretation criteria for evaluation of replications outcomes by LeBel et al. (2019), if the original study detected a signal. A simplified replication taxonomy for comparing replication effects confidence intervals to target article original effect sizes.

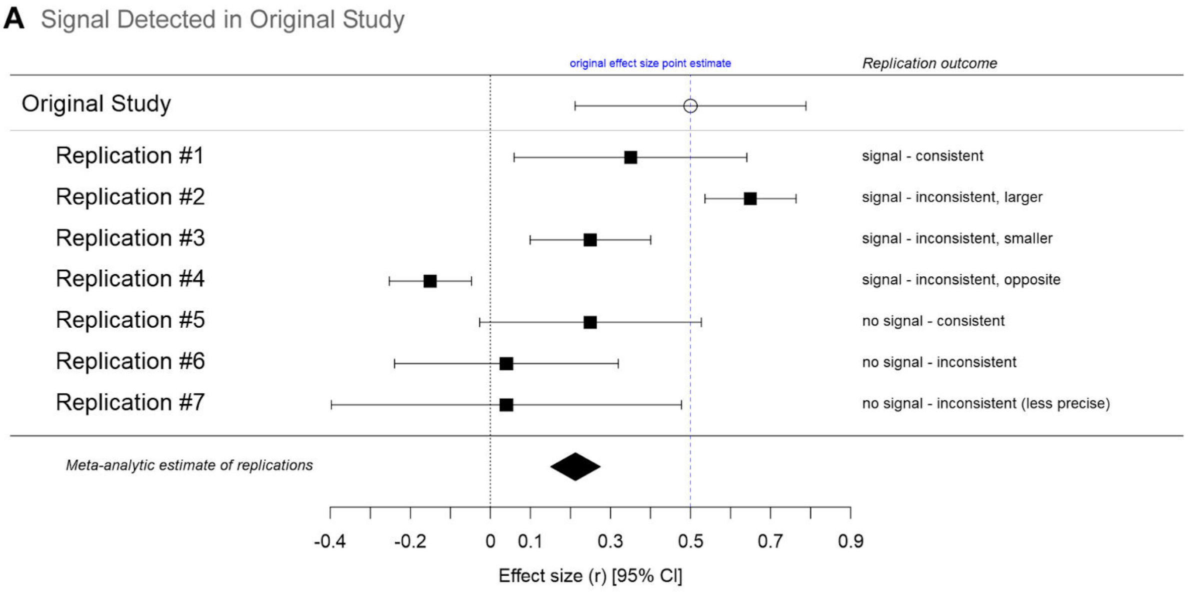
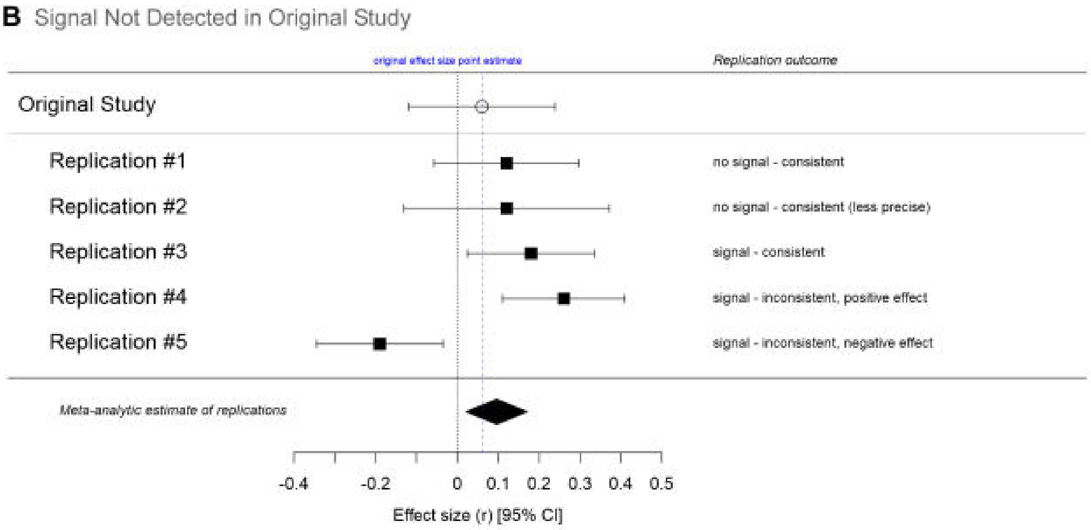


Figure S6

Interpretation criteria for evaluation of replications outcomes by (LeBel et al., 2019), if the original study failed to detect a signal (null finding)



References

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*, 389-402. https://doi.org/10.1177/2515245918787489

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2019). A brief guide to evaluate replications. *Meta-Psychology*, 3. https://doi.org/10.15626/MP.2018.843

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology*, 32(1). https://doi.org/10.5334/irsp.289