david r. mandel

Search this site

Publications | Impersonal things I'll share with the world | Personal things I'll share with the world

AT LEAST I REPLIED: Reply to Simmons & Nelson's Data Colada blog post

In a recent blog post on Data Colada (hereafter, DC), entitled ""Exactly": The Most Famous Framing Effect Is Robust To Precise Wording," Joseph Simmons and Leif Nelson (S&N; with additional input from Uri Simonsohn) report the results of a replication study of Experiment 2 in my in-press JEP: General paper entitled "Do Framing Effects Reveal Irrational Choice?". They contacted me a few days prior to posting to give me advanced notice and to kindly offer me the chance to give up to a 150-word reply. Since 150-words isn't enough to reply properly, they have graciously agreed to link my full reply to their blog post.

I agree with them that this is a good paper for a replication and for the reasons they give – the original work by Amos Tversky and Daniel Kahneman is "foundational" and my critique is "novel and fundamental" (in their words). This we agree on and I welcome their attempt to replicate the findings.

There are a number of important issues to tackle in response to their blog post and broader issues they raise. I'll begin "local" and then branch out, so let me start with their experiment and the accompanying blog post.

Methods and replication quality

The experiment S&N replicated had a 2 (Frame: positive, negative) x 3 (Modifier: "at least", none, "exactly") factorial design. I used a framing problem similar to the Asian Disease Problem (ADP), which I had also used in an earlier paper. S&N said they focused on Experiment 2 because it was "the study most faithful to the original," meaning the Asian Disease Problem (ADP). However, it was in Experiment 1 that I actually used the ADP in a within-subjects design. I don't think this matters much, especially for comparing my results to S&N's, but again it is worth noting that I did report findings on the actual ADP.

The first thing I did was to review their materials and my own. I wanted to make sure that the terms "exactly" and "at least" weren't marked in my study. The methods in my paper don't indicate that they were, but it was worth double-checking. They weren't. So, that rules out one issue – unreported marking of linguistic modifiers. I've reviewed their materials and I agree with them that it is a precise methodological replication, except of course for sampling procedures.

Aside from their larger samples (a good thing), they used mTurk, while subjects in my experiment were from university samples. I am not sure whether this had an effect. I think it is worth noting, and it represents an uncertainty, but I will leave it at that. I am assuming that S&N use some sort of screening procedure to ensure that the subjects are fluent in English. It would be good to know what the inclusion/exclusion criteria were. And of course if they didn't, that would be important to know too.

As well, they asked subjects to recall how many people had their lives at stake, and they filtered out the roughly 50 cases that gave a wrong answer. I ran my initial tests on the comparability of our results with and without the filtering of these cases and it made little difference. I didn't use a similar filtering procedure (nor did Tversky and Kahneman), so it seems appropriate to include those cases.

Results

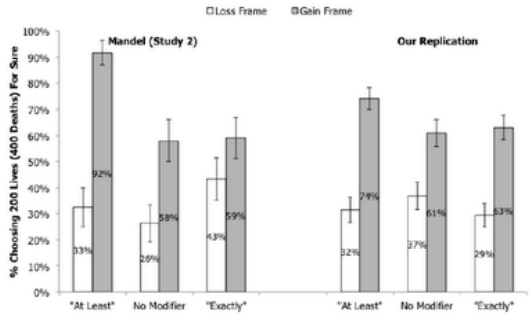
S&N present the results for the "no modifier" and "exactly" conditions on their blog post and link the full set of results with the "at least" condition for readers wanting to see the full set. They reported the results in the two formats I provided: percentages of subjects choosing the certain option and their choices weighted by strength of preference. I reported the percentages mainly because most earlier papers don't use the more sensitive weighting method, but it makes sense to do so because it seems worth knowing whether choosing a given option was the result of a strong preference for that option or little more that a coin toss (or somewhere in between).

In the analyses that follow, S&N and I have coded the choice of the certain option as 1 and the choice of the uncertain option as -1 and multiplied those by their preference rating (0 = equally preferable; 10 = much more preferable), so weighted choice can range

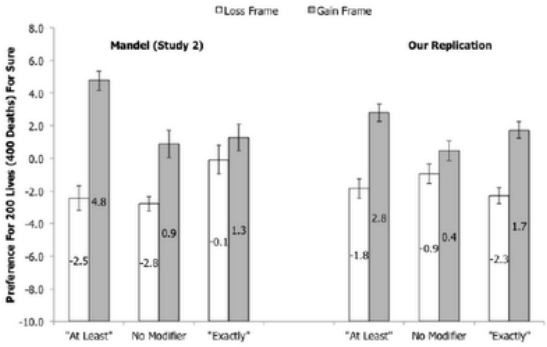
from -10 (strongest preference for the uncertain option) to 10 (strongest preference for the certain option), with 0 representing indifference.

Our analyses concur, and I'll present their filtered results for simplicity, although I'll later run tests on both the filtered and unfiltered samples.

Here are the percentages choosing the certain option in all conditions in both studies:



And here are the mean weighted choice findings:



S&N's Conclusions

S&N reported that "Unlike Mandel, we found a strong framing effect [in the unweighted choice data] even with the use of the word "exactly" ($p<.001$)." Likewise, they reported that the weighted choice data "also failed to replicate his result."

On the basis of this, S&N conclude:

"In sum, our replication suggests that Tversky and Kahneman's (1981) framing effect is not caused by this linguistic artifact."

They close the post by noting that, in then end, none of this was necessary since their colleague, Uri, runs the ADP in class every year with his students using exactly and he still gets the framing effect.

Case closed.

My Response

First, let me say that when Joe contacted me, he noted that his investigation with Leif was conducted entirely in the scientific spirit and not meant to prove me wrong. He said they planned to publish their results regardless of the outcome. I accept that that's so, and my own comments are likewise intended in the spirit of scientific debate.

Pure replication studies are important. I think we all agree that they add important information when done properly and when appropriate conclusions are drawn. I've already noted that I find this to be a pure replication, except for sampling variations, which may or may not account for any differences. I am not claiming that they do, nor denying that they might.

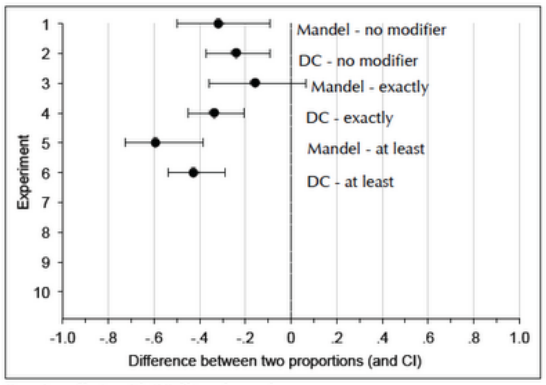
However, I don't think S&N's conclusions about my study's replicability, nor their conclusions about the theoretical significance of the results, follow quite as clearly from their findings as they suggest.

The replication

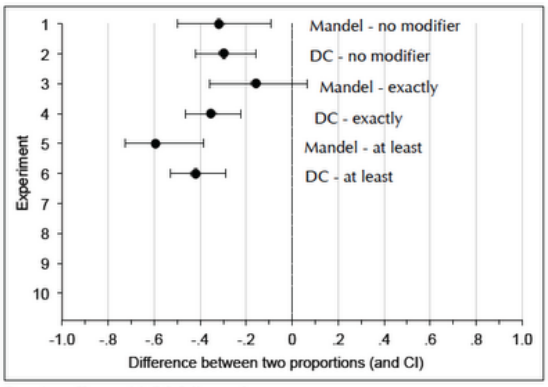
Let's begin with replicability since that is the proximate issue. Let's also be explicit about what replicability refers to here: Experiment 2 from my studies. I reported 3 studies (4 if you add in the demonstration experiment I reported in the General Discussion, which used the limited-vaccine version of the ADP adapted by Jou, Shanteau, & Harris, 1996). Each of those studies had a different design and tackled the issues in different ways. I have always regarded Experiment 2 as the weakest of the bunch, but I'll come back to that.

There were 6 experimental conditions in Experiment 2. Let's first take a look at the unweighted choice data. Here are the framing effects per condition in the original and replication experiments. Negative values indicate standard framing effects, i.e., more subjects choosing the certain option in the positive (gain) frame than in the negative (loss) frame. The first figure shows the results with S&N's filtering procedure and the second without filtering.

The error bars are 95% confidence intervals generated using Geoff Cummings' ESCI program for Mac, which I've used for much of the analysis that follows.



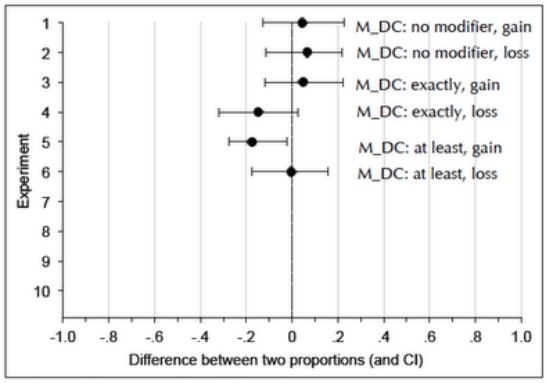
Framing effects with DC-filtered sample



Framing effects with DC-full sample

These figures show the effects. Nothing new here, but I find it useful to see all the 95% CIs, which show, for instance, that although my "exactly" condition was not significant, the effect was still in the direction of a standard framing effect. Also, it's evident that filtering didn't change much, so I'll move on with further analysis based on the full S&N dataset.

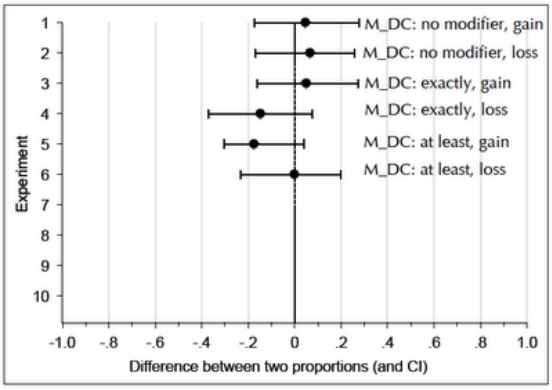
The next figure compares each of the 6 conditions across the two experiments.



Condition differences with DC-full sample

Five out of 6 conditions yielded percentages of certain option choosers that did not significantly differ. There was a significant difference between the percentages in the "at least" gain frame condition, where I found a higher proportion of sure-option choosers. However, in both my experiment and S&N's replication, that condition yielded the largest framing effect and, as we can see, the "at least" loss frame condition was virtually identical.

These six comparisons are uncorrected, and in no way theory driven. A replication study is just that – about statistical replicability of prior results. Obviously, replications that test many comparisons to the original study have a greater chance of yielding some evidence of non-replicability than those that have few comparisons. If we Bonferroni corrected, so that the overall Type I error rate was .05, then we'd need to use 99.12% CIs; i.e., (1 - .05/6) x 100. Here are the comparisons with 99% CIs (close enough):



With correction for alpha inflation, none of the 6 conditions differed significantly between my study and S&N's replication.

Let me say that again: using the standard metric for framing studies (percentage of subjects choosing the certain option), there was no statistically significant difference between any given condition in my Experiment 2 and S&N's replication experiment. In other words, each condition result was replicated by S&N.

Now that doesn't mean that the significance or nonsignificance of every reported effect will be preserved. S&N focus on the non-replicability of my nonsignificant framing effect in the "exactly" condition. But, as we saw earlier, that effect in my study was not zero. It was in the direction of a standard framing effect. With t(74) = 1.37, the probability of a same sign effect (p*rep*) is 83.2% (I use Bruno Lecoutre's LePrep to generate p*rep* and p*srep*). S&N got that, and a heck of a lot more. The probability of a significant same sign effect (p*srep*) at a two-tailed probability of 0.00000005 (which is the exact two-tailed probability corresponding to their t value of 5.64 with df = 225 yields) is 9.03E-04. So their result is indeed unexpected, given my result. Note, however, that a significant effect at alpha = .05, given my effect, is not that surprising: the p*srep* = 0.283—a 28.3% chance of a significant replication.

Note, however, that S&N's "exactly" framing effect is not all that probable even given their own "no modifier" framing effect obtained from the same study using the same subject base. The p*srep* = .17. That is, we would expect *not* to achieve their "exactly" significance level 83% of the time based on their "no modifier" result. Indeed, their effect in the critical "exactly" condition is very much like their "at least" condition. Given their "at least" framing effect, with t(213) = 6.77, p*srep* at a two-tailed probability of

0.00000005 is .78.

So, what S&N's results seem to bizarrely suggest is that subjects interpret the term "exactly" to mean "at least" (linguists, please keep calm!), or perhaps that explicit modification of quantifiers somehow (*how*?) affects framing effects in a similar way, amplifying them relative to no explicit modifier on the numeric quantifiers. These interpretations are more consistent with their findings than the explanation that the unmodified quantifiers are treated as meaning "exactly N lives saved/lost."

S&N don't attempt to explain why their "exactly" result is so strong and what that implies given the strength of their other two framing effects. Rather, they contrast my null (not nil!) effect with their very strong effect. Even without their replication, we knew there's close to a 3/10 chance of getting a significant effect at .05 given my null effect. So, a significant effect at that level in a replication isn't highly improbable. (To know this, they didn't even have to ask Uri.) The improbability of S&N's result, however, must be interpreted not only in light of my findings (as an improbable replication result), but also in terms of their own findings. I personally cannot make sense of why explicitly modifying the quantifiers in the certain option with "exactly" would strengthen the effect as much as "at least" did. It makes no sense to suggest that "exactly" means "at least" to most people. Prospect theory doesn't predict that effect. So, while S&N's "exactly" framing effect is an improbable result given mine, it is also a perplexing result, given their other results.

I don't want to belabour the point by repeating all of this with the weighted choice results. Let me just highlight a few observations. If one calculates Cohen's *d* for every framing effect in the two studies, as I've done in the ESCI Meta-Analysis of *d* Excel file here, you can test for effect heterogeneity (a significant Q indicates heterogeneity). What you'll find is that:

(a) S&N's "no modifier" and "exactly" conditions yield heterogeneous effects, Q = 4.15, p = .0418.

(b) S&N's "at least" and "exactly" conditions yield homogenous effects, Q = 0.01, p = .9231.

(c) The "no modifier" effects in Mandel and S&N are homogenous, Q = 1.02, p = .3121.

(d) The "exactly" effects in Mandel and S&N are heterogeneous, Q = 3.91, p = .048.

(e) the heterogeneity in (a) and (d) is comparable. For instance, consider I^2, which is a measure of the percentage of variation across studies or conditions that is due to heterogeneity rather than chance. I^2 = 75.9% for S&N's comparison of their "exactly" and "no modifier" weighted-choice framing effects, whereas I^2 = 74.4% for the comparison of the S&N's and my "exactly" weighted-choice framing effects.

So much for the proximate issue.

Broader issues

Among the broader issues that S&N's replication raises, an important one is how to make the most out of replication studies. Replications take time that could easily have been spent elsewhere. So, having conducted them, it makes sense to treat them carefully. This is not only to make the investment of time worthwhile, but, more importantly, so that they end up making a useful contribution to science.

The replication experiment may start out as a pure test of replicability, but replication researchers should be willing to triangulate the results of their findings with the body of evidence on the topic. That involves also considering the results they obtain not only in light of the original experiment's findings, but also in light of their other findings.

Ultimately, replication is merely one aspect of evidential triangulation. As scientists, we don't do enough of it, but replication researchers are not freed from the other burdens of triangulation: evaluating their findings in light of the larger body of evidence. For instance, when reading S&N say, "We could have just asked Uri", my immediate thought was "Well, I could have just pointed to Karl", since Karl Teigen and colleagues have shown in various published studies that people are prone to lower bounding numerical quantifiers. Those studies also include ADP adaptations (in particular, see Teigen & Nikolaisen, 2009). For instance, forecasters of outcomes such as "you will keep/lose $200" are seen to be less accurate when one ends up keeping or losing $180 (inconsistent with "at least") than when one ends up keeping or losing $220 (inconsistent with "at least"). This is so for positive and negative outcomes, so outcome desirability is ruled out as an alternative explanation. Or I could have pointed to the neo-Griceans and other work on scalar implicature. Or....

The point is that even a pure replication ought to be contextualized. Doing that well may end up taking more time than running the

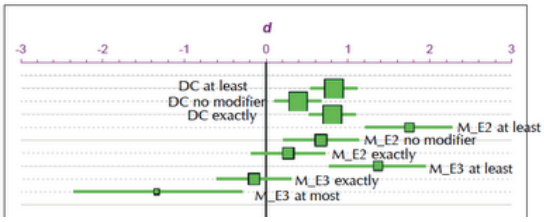
replication study itself and summarizing the findings. But without that deeper commitment, replications may not serve the greater aim of evidential triangulation. It is what you do with evidence that counts.

For instance, S&N seem to believe that their finding of a strong framing effect with "exactly" somehow reinstates prospect theory as a viable theoretical candidate for ADP-type risky choice framing effects. It doesn't. There is already strong evidence that ADP-style framing effects are not due to the psychophysics of valuation as captured in the value function model of prospect theory. Anton Kühberger, in his 1995 OBHDP paper, was the first to show that merely fleshing out the unexplicated part of the sure option within a given frame (so that the certain option matched the risky option in terms of stating explicitly that there were mixed outcomes for the sure option and a mixed set of extreme outcomes for the risky option) was sufficient to eliminate the framing effect. To be clear, if the gain frame is "If Plan A is adopted, 200 will be saved and 400 won't be saved" and the loss frame is "If Plan A is adopted, 400 will die and 200 won't die", the framing effect is eliminated. This modification actually strengthens the parallel construction of the framing manipulation across paired options, since now there are explicit mixed outcomes (i.e., positive and negative outcomes) noted in both options, and the methodological tweak in no way changes prospect theory's predictions, yet it obliterates the effect. Moreover, reversing the asymmetry, actually reverses the effect, yielding framing effects that are opposed to the directional effects predicted by prospect theory. Experiment 3 in my paper independently replicates those effects, and my earlier studies show that framing effects are eliminated when parallel explication is achieved by truncating the risky option so that the expected utility of *explicated* outcomes are matched. Mike Tombu and I have studies currently under review that show similar reversals based on manipulation of explication.

Now, one might be inclined to say that, even if those points were valid, and even if those studies showed what they did, there is still so much evidence in support of prospect theory, that it is people like me who have failed to triangulate the evidence. I mention this because I've already encountered it a number of times, not because it's a good argument. It isn't. It might be a good argument if I were claiming that prospect theory is wrong, but I don't claim that. What I actually concluded about prospect theory is "although prospect theory's value-function model has much evidence to support it from a wide range of experimental tasks (e.g., simple lotteries), it cannot accommodate the present findings." Not exactly a scathing critique. Indeed, I do not deny that prospect theory may do a good job of explaining certain types of framing effect. And, I am not "anti-framing" either, whatever that might mean. I mention it, strange as it sounds, because one reviewer seemed to think I was trying to prove that framing effects don't exist. Fortunately, I have done some work [e.g., here] showing that they do exist too, which helped my case, but imagine the fun I had.

I hope that if one reads my paper, they will be left with a clear sense that the core issue I was dealing with concerned the validity of inferences made by researchers about the linguistic interpretation of quantifiers that, in turn, affect how researchers characterize the logical coherence and rationality of the "decision-makers" they study. That's why I regard Experiment 3 as the most important. Experiment 3 took a hetero-phenomenological approach, while Experiments 1 and 2 did not.

That is, Experiment 3 explored how subjects interpreted the numeric quantifiers in the certain options as well the probabilities in the uncertain options. That allowed me to disaggregate the sample based on interpretational differences. When I did that, I found that subjects who said they had an "at least" interpretation of the certain option showed a standard framing effect, while those who said they had an "exactly" interpretation showed no significant effect, and a minority who said they had an "at most" interpretation showed a reversed framing effect. I have included these effects in the effect size file I linked earlier. They show stronger separability of effects than in Experiment 2, as shown here:



I also found that the likelihood of adopting an "at least" interpretation of the quantifier in the sure option was contingent on the extent to which that option was explicated. Subjects were more likely to adopt an "at least" interpretation when the sure option was not fully fleshed out, as in the ADP. However, even in the standard ADP-style condition, "only" 64% adopted the "at least" interpretation, a figure that plummeted to about 24% when the sure option was fleshed out. And, in the study I reported in the General Discussion, I showed how a variation to the ADP can make "at most" a likely interpretation of the sure option in the gain frame (namely, when only 200 can be saved because there are only 200 vaccines). I note all of this because reading the DC blog

post might leave readers believing that I had proposed that people *always* interpret numeric quantifiers as meaning "at least" n. In fact, I make a point of stating explicitly that I reject that view (what I've called naive unilateralism). I don't only make the point, I present evidence in the paper that does not support that view. And, I presented evidence a few years ago in this paper with Oshin Vartanian that showed that subjects interpret the quantities in DeMartino et al.'s fMRI framing study as exact values. For those who may be unfamiliar with that study, we asked 21 subjects (DeMartino et al had 20 subjects) in counterbalanced order:

Imagine you are in London England and encounter the following situation:

First, you receive £50. However, you must choose one of the following options:

Indicate your choice by checking the appropriate box:

☐ Lose £30

☐ Take the gamble

While contemplating the statement "Lose £30" did you consider it to mean:

☐ Lose at least £30.

☐ Lose precisely £30.

☐ Lose at most £30.

Imagine you are in London England and encounter the following situation:

First, you receive £50. However, you must choose one of the following options:

Indicate your choice by checking the appropriate box:

☐ Keep £20

☐ Take the gamble

While contemplating the statement "Keep £20" did you consider it to mean:

☐ Keep at least £20.

☐ Keep precisely £20.

☐ Keep at most £20.

We found that 17 out of the 21 said "precisely" in both conditions. That's 81% with 95% CI [60%, 92.3%]. There was, however, no framing effect evident in our impure replication.

Evolving standards for replication research, and research, more generally

The choice of what to replicate is not value-neutral, even if a decision is made to report whatever one finds. Had S&N regarded it as obvious that prospect theory were not a viable, if not the most viable, contender for explaining ADP-type problems, then would it have paid for them to replicate my study? Of course not. As S&N explained at the outset of their post, they regard the work they see my studies as challenging as "foundational," and they teach it to their students. Likewise, had the ADP been a framing task that nobody paid attention to, would it have made sense for me to run studies showing that it might actually not be a good manipulation of strict framing? Of course not, and I say as much in the paper.

The non-value-neutral basis of scientific inquiry is a good thing, and there's no need to hide it. Why shouldn't we be working on what we think is important or interesting? Or, testing the theories we sense are better on track than others? Or, attempting to replicate the ones we have our doubts about? Of course we should. But, we also should take precautions, and an important one, I think, is consistency in standards of assessment.

In the present case, S&N present a fairly dismissive assessment of my paper on the basis of their replication experiment. Yet, little over a month ago, they published a blog post entitled "Forthcoming in the American Economic Review: A Misdiagnosed Failure-to-Replicate", where they challenged the authors of a replication study that apparently claimed to have failed to replicate an anchoring effect reported by Ariely, Loewenstein, and Prelec (2003). One of their critiques is that the original paper was a multi-study paper, but the replicators only attempted to replicate one study. (Sounds familiar?) Another was that, although the effect size was smaller and the effect was nonsignificant in the replication, the two effect sizes had not been subjected to proper statistical comparisons with confidence intervals. When they did that, they found that the replicators' effect size had a 95% CI that included the original. These are all good points (buttressed in that case too by Uri's classroom data), and in many ways quite similar to the ones I've made in my own remarks on their replication study. It therefore strikes me that it would be of value if replication researchers clearly outlined what they were attempting to replicate, and what they regard as their criteria for concluding that a study was or was not successfully replicated. We need transparent standards and better adherence to them if we are to make the most of our science. I think such standards should specify ground rules for assessing not only comparative consistency (replication), but also internal coherence (do the findings across conditions in the replication make sense?). Rules that militate against selective focusing should resonate well with the authors, who have written about the topic more generally.

Criteria should be more transparent and consistently applied and, likewise, statistical reporting should be more precise. Simmons,

Nelson, and Simonsohn make a good case for their 21 words of disclosure, arguing that we devote far too time and space on standards for formatting documents, having clearer directives about such things as whether to put a period after a measure unit than we do about how to disclose our methods and analytical decisions. Likewise, I hope my preceding analysis highlights the lunacy of restricting statistical reporting to 3 decimal places and using imprecise (<.001) conventions. This does not aid clarity or lend itself to secondary analysis. I'd rather know that S&N got a significant framing effect in their "exactly" condition at a two-tailed p = 0.00000005 than to learn it was less than .001, even if it represents a failure to replicate the result of the same effect test in of my studies. Surely, we can save a bit of space in our papers to accommodate the kind of statistical precision that would facilitate better evidential triangulation over time. One percent less speculation should be more than enough to meet this requirement (and ample leave room for Simmons et al.'s 21 words).

Final thoughts on framing

All of this got me thinking about, and rereading, an email thread Daniel Kahneman and I had over this topic back in 2010. We both agreed that an exact interpretation of the quantifiers was an unlikely reading of the ADP. He said that he and Amos had the sense that the most probable interpretation was "about 200" or "about 400." I share that intuition. After all, in the ADP, we're dealing with a future epidemic. If subjects are asked their interpretation, and given the option to choose "at least", "at most", "exactly" or "about", we both agreed that "about" would be the most common response. However, we seem to have differed in terms of the inference we would draw from that finding. He viewed it as showing that I left the best interpretation out. However, I see "about" as an easy, noncommittal response to give, boosting its likely selection. I predict that if "about" interpreters were pressed to resolve their uncertainty by indicating whether *more* or *less* than 200 saved or 400 dying were more likely, they would be more likely to say *more*, consistent with a lower-bound "at least" interpretation. In other words, the resolution of "about" would not, according to my intuition, be symmetrically distributed around 200 or 400. For "about" not to matter in the ADP, that distribution would have to be symmetric. Otherwise, description invariance is violated and the problem ceases to be a strict framing problem -- again, the point of my paper. Incidentally, I had that intuition before learning about Teigen & Nikolaisen's studies, which have since strengthened my original intuition.

He also noted that it was subjects' sheer sense of surprise (or of being dumbfounded) when confronted with their inconsistency of choice (in within-subject studies) that made him doubt a linguistic interpretation of the effect. That's a wonderful point. If subjects interpret the values 200 and 400 as lower bounds, then why don't they simply say so when confronted with their inconsistency? One explanation is that it's because they don't. They have either precise ("exactly") or imprecise ("about") representations of quantity that are more or less symmetric. Another possibility, however, which I believe Karl Teigen had first suggested to me, is that those within-subject "insight" studies and our lower-bounding findings seem to indicate that numeric quantifier interpretations are labile. When frames are presented separately, the quantifiers may tend to be treated as lower bounds in the ADP, but when subjects see 200 out of the 600 right next to 400 out of the 600, that representation is cancelled and they now have an "exactly" interpretation. Their surprise would make sense if they were themselves unaware of how context had changed their representation. And, we know that people are often unaware of what influences their thinking and behavior.

I think these ideas suggest many studies worth pursuing. I find it amazing that a topic like risky choice framing has been so heavily focused on a psychophysical account of valuation for so long. Of course, the ADP was motivated by prospect theory, and if theories were bodies, it would have planetary gravitational force, but framing obviously also deals with mental representation, semantics, pragmatics, and how they affect judgment, evaluation, and decision making. I hope we will see a broadening of our focus on this topic in the years to come.

Comments