# STAGE 1 REVIEW

## Revisiting the "Belief in the law of small numbers": Conceptual replication and extensions

## Registered Report of problems reviewed in Tversky and Kahneman (1971)

Reviewed by: Kariyushi Rao
Behavioural Science Group, Warwick Business School
The University of Warwick, Coventry, UK

TABLE OF CONTENTS

## 1 Summary of the Research Plan

The authors plan to reproduce, replicate and extend seven "empirical demonstrations" of the belief in the law of small numbers (LOSN) presented in Tversky and Kahneman (1971). The authors make several improvements upon the original paper. First, the authors merge all seven demonstrations into one procedure, presented in random order to a large sample of Amazon Mechanical Turk Workers located in the United States. Second, the authors seek to minimize jargon and statistical terms in the experimental stimuli. Third, the authors plan to perform statistical tests of the hypothesized deviations (whereas Tversky and Kahneman only provided descriptive statistics).

For each demonstration, there is a purported "correct answer" (according to Tversky & Kahneman). The authors have done a commendable job clearly articulating what those correct answers might be, given the target article is often unclear and fails to state the "correct answer" directly. The authors will measure the deviation of participants' responses from the correct answer they've inferred from the target article. The results of the present experiment will be compared to the descriptive conclusions presented by the authors of the target article.

[*Note:* The authors have revised their original snapshot to exclude a scholar sample. I find their justification for this choice perfectly reasonable. In my opinion, the lay sample is more interesting and important than the scholar sample. I agree that the present research plan is sufficiently complex and challenging, and the present results (that exclude the scholar sample) will provide a significant contribution to the literature.]

### *1.A Is the research question scientifically valid?*

Yes. The present proposal meets the PCI standards. The research question is clearly defined. The research question is scientifically justifiable, and defined with sufficient precision as to be answerable through quantitative or qualitative research. The research question make sense in light of the extant theoretical and empirical literature in statistical reasoning, probability updating, and judgment and decision-making. The hypotheses are capable of answering the research question. The research question falls within established ethical norms. The authors have clearly distinguished work that has already been done from work yet to be done.

### *1.B Are the proposed hypotheses logical and plausible?*

Yes. The present proposal meets the PCI standards. A priori hypotheses are coherent and credible. Hypotheses follow directly from the research question. There is a sufficiently strong mapping between the theory, hypotheses, sampling plan, preregistered statistical tests, and possible interpretations given different outcomes. The authors have explained precisely which outcomes will confirm or disconfirm their predictions.

***1.C Is the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable) sound and feasible?***

The present proposal is of sufficient quality to merit IPA from PCI, but some improvements could be made. The study procedures and analyses are valid, for the most part. The authors have performed a statistical power analysis to the best of their ability given the lack of information in the target article, with appropriate (conservative) adjustments. The proposed sample size is sufficient to provide informative results. The authors clearly state their rules for randomization of experimental participants, and for data exclusion.

The authors do plan to rely on conventional null hypothesis significance testing. The authors also intend to interpret negative results from their one-sample t-tests and one-way ANOVAs as evidence that an effect is absent. The authors have not proposed Bayesian hypothesis testing or frequentist equivalence testing (inferential methods better capable of drawing conclusions about the implications of negative/null results). However, the proposed statistical methods are standard in the authors' field (psychology), and reviewers for the two "PCI-Interested" journals (JEP:G and JEP:LMC) will sometimes explicitly request traditional methods (e.g. ANOVA) be performed. So, I do not think the present analysis plan should preclude the authors from receiving an IPA from PCI.

The authors also plan to implement the paradigm suggested by LeBel and colleagues (2019) to judge the extent to which their experimental results replicate the original results in the target article.

*Suggestions*:

(1) Exclusion Criteria

Participants are incentivized to lie on each of the self-report measures proposed as exclusion criteria, so it is unlikely these criteria will serve their intended purpose. I recommend the authors run a qualification survey in advance of the focal study, and include the following substitutes for their first two self-report measures in the qualification survey.

*1. Participants indicating a low proficiency of English (self-report < 5, on a 1-7 scale).*

Ask participants the following two questions (or something similar):

1. What region of the United States do you live in currently? (Drop down list that includes "Prefer not to disclose")
2. What is your favorite thing about the region of the US where you live currently? Please respond with one complete, grammatically correct sentence. (Question should appear on a different page than the above. Question must be open response. Responses should each be read by the same human reviewer. Exclude all participants who do not provide a

complete, grammatically correct sentence. Exclude all participants who provide non-sequitur responses, e.g. "I love my television.")

*2. Participants who self-report not being serious about filling in the survey (self-report < 4, on a 1-5 scale).*
Restructure this question using Drazen Prelec's Bayesian Truth Serum.

Qualification survey responses should be checked by hand by the same person, to ensure that patterns are detected across responses (e.g. groups of participants colluding on the survey who paste copied text from internet sources instead of writing original responses). Qualified participants should be assigned an approval code, and the subsequent focal study should be restricted to those participants that have been assigned the approval code.

I don't see the purpose of requiring participants to confirm they are native American citizens born and raised in the United States. Participants are prone to lie on these types of questions, and this particular question does not necessarily indicate English proficiency or any particular (relevant) level of education or acculturation.

(2) Compensation
The hourly pay target for US-based Amazon Mechanical Turk Workers is too low. MTurk Workers in the US desire, expect, and actively seek out a pay rate equal to the most generous State minimum wage, which is $15.00/hour. The authors' target of $7.25/hour will result in selection issues, as more highly conscientious and experienced Workers are less likely to accept HITs at lower rates. The authors should also be aware that MTurk Workers can manually set hourly targets in their MTurk Dashboard that are perpetually displayed within the MTurk interface, so the $15.00/hour anchor will be salient for them.

(3) Comments on Table 1

Q3: The LOSN hypothesis should be rewritten as, " If a study reports that 0.8*X out of X infants preferred Toy A over Toy B, then people tend to perceive that as representative of the general population and therefore expect that 0.8*X out of X infants in the general population will prefer Toy A to Toy B. Regardless of what X is.

Q4: Given that you are trying to remove jargon and statistical language, shouldn't you avoid using the concept of a power analysis here? Instead, you should present a layman's explanation of what a "critical significance value" is. The power analysis version of the question might have been interesting when you were going to include a scholar sample, but without the scholar sample it doesn't seem as interesting or appropriate.

Q5: The generalized hypothesis should be rewritten as, "If a study reports a **surprising** phenomenon using any sample, then people tend to perceive their findings to be representative of the general population and therefore expect that the finding generally holds true for the general population." Same for the secondary hypotheses - the word "surprising" should be added in before "exploratory" in each case. The surprisingness of the phenomenon is really key here, especially from a Bayesian perspective. If a finding runs contrary to accumulated human knowledge (even lay knowledge), then our willingness to update in the direction of that finding should be smaller than if the finding does not run contrary to accumulated human knowledge.

Q6: For similar reasons to the above, the generalized hypothesis should be rewritten as, (1) "People do not differentiate between **exploratory studies that produce surprising results and confirmatory studies that seek to replicate those surprising results**," and separately, (2) "People ignore sample size. Participants perceive the **following to be equally representative: (1) an exploratory study with a sample size of X, and (2) a confirmatory study that seeks to replicate the results of the original exploratory study using a sample size of 0.5*X**."

Q8: The generalized hypothesis should be rewritten as, "People overestimate the likelihood **that a confirmatory study seeking to replicate several correlations found in an original exploratory study will produce support for at least 2/3 of those correlations, even if the confirmatory study has 2/5 the sample size of the original study**." And, the LOSN hypothesis should be rewritten as, "If an exploratory study with a sample of X found support for Y correlations, **then people overestimate the likelihood that a confirmatory study will replicate at least 2/3*Y correlations from the original study, even if the confirmatory study has a sample size of .4*X, regardless of what X is**."

(4) Comments on Table 3

In general it seems odd to use the word "experimenter" with a lay population. I suggest either using the generic "scientist" or "you" (as in Q3 of T&K, 1971) or "toy company executives," etc. Also, the notes on what it means to "find support" seem to commit a common error in the description of null hypothesis testing (that there is less than a 5% chance of obtaining X result if H1 is not true), and the language should be updated to avoid this error.

Q1: The phrase "a sample of X people" may be misinterpreted as a subset taken from a group of X people. E.g. "There were 100 people, and we took a sample of 10 out of that 100." The "clarification" note also commits an error in its description of null hypothesis testing. If you are really trying to get away from jargon and statistical language, consider changing the scenario and prompt to the following:

Scenario:
"You read a news report about [20/200/2000] people who participated in an experiment to test scientists' theory of X. The report indicates that the results of the experiment support the scientists' theory.

(Usually when scientists say they found support for a theory, it means they ran a statistical test that tells them if they ran the same experiment one hundred times, less than 5 of those experiments would produce the same results they got in the original experiment if their theory was false. Basically, they think it would be really hard to get the result they did if the theory wasn't true.)

The report also indicates that same scientists have just run the same experiment again with [10/100/1000] new people from the same population."

Prompt:
"How likely is it that the scientists will find support for their theory again in the experiment they just ran with [10/100/1000] new people? (Indicate your response as a percentage out of 100; e.g. 0% means there is absolutely no chance they will find support for their theory, 100% means they will definitely find support for their theory.)

Q2: There appears to be a typo (underlined and bold in red below), and I think there's a big difference between the (original) phrase "**known** to be 100" and the (new) phrase "**reported** to be 100." I suggest the following: "The average IQ of all eighth graders in a particular city is 100. A scientist randomly chose [50/500/5000] eighth graders from that city to test their IQ. The [first eighth grader / average IQ of the first **10/100** eighth graders] tested out of the [50/500/5000] chosen by the scientist [has an IQ of / is] 150.

*Important note:* I think the odds that the ***first*** child drawn from a sample of 50 having an IQ of 150 are higher than the odds of obtaining an average IQ of 150 from a contiguous sequence of 100 children drawn from a finite sample of 5000 children (I didn't work out the math on this one, so maybe the authors are right to assume that these odds are the same). If the odds are different then I don't think you are actually asking the same question when you increase the number from 1 to 10 to 100.

Q3: I'm not sure what it means for a "ratio of 80% of infants choosing Toy A over Toy B" to "persist." I think what's inferred by the grouping of the three questions is the following: "Suppose that [8/80/800] out of the [10/100/1000] infants in your *second* study *also* preferred Toy A over Toy B. If you were going to run one final study to conclude once and for all that

4 out of 5 infants in the world population prefer Toy A over Toy B, what is the minimum number of infants you would need to include in that final study? (Try your best to estimate.)"

Q4: The use of the words "positive association" and "expected association" are foreign to the lay population. Also, the meaning of 0.35 is ambiguous here, because the grade scale is not specified. Americans are used to a 4.0 GPA scale, or a letter-grade scale from A to F. Using a need for achievement scale that ranges from [-1, 1] with a GPA scale that ranges from 0 to 4 (and sometimes up to 5.0) makes the meaning of 0.35 difficult to understand for a lay person. I suggest using the following scenario instead:

Scenario:
"Psychologists who study two personality traits (Trait A and Trait B) expect there to be a positive relationship between these two traits in the general population. In other words, psychologists expect a people with higher ratings on Trait A to also have higher ratings on Trait B. Both traits are rated on a scale from 0 (does not exhibit the trait at all) to 10 (exhibits the highest level of this trait). Specifically, for each 1-point increase on the Trait A scale, psychologists expect people to exhibit a 0.35-point increase on the Trait B scale.

You read a news report about a study on the relationship between Trait A and Trait B. Before running that study, psychologists performed a statistical test using all of the existing evidence about the relationship between Trait A and Trait B. The test is supposed to determine how many people need to participate in the study in order to accurately detect the relationship between Trait A and Trait B. The result of the psychologists' test indicated that they need at least 79 people to participate in their study in order to accurately detect a 0.35- point increase in Trait B for each 1-point increase in Trait A.

(When scientists say "accurately detect" they usually mean that if they ran 100 experiments they would only detect the 0.35-point increase in less than 5 of those experiments if there wasn't really a positive relationship between Trait A and Trait B.) "

Prompt:
"If the psychologists ran their study with 79 people, how likely is it that they will find support for a 0.35-point increase in Trait B for each 1-point increase in Trait A?

(Usually when scientists say they found support for a relationship between two things, it means they ran a statistical test that tells them if they ran the same experiment one hundred times, they would find a relationship between Trait A and Trait B in less than 5 of those experiments if there wasn't really a relationship between those two things. Basically, they think it would be really hard to find a relationship there if it didn't really exist.)

Indicate your response as a percentage out of 100 (e.g. 0% means there is absolutely no chance they will find support for a 0.35-point increase in Trait B for each 1-point increase in Trait A, 100% means they will definitely find support for a 0.35-point increase in Trait B for each 1-point increase in Trait A)."

For all extension questions, I suggest using the verbose "What is the likelihood that in their sample of [79/790/7900] people, there will be a 0.35-point increase in Trait B for every 1-point increase in Trait A?" instead of the association language you have now.

Q5: Stripping out the words in bold here - "**you** completed a **difficult and time-consuming** experiment" - really changes the nature of this question. I don't think you're measuring the same thing anymore with the updated version. Also, the updated version still has a lot of unfriendly language for laypeople. Same thing with taking out "you" in the origininal "**you** were to run the same study again" changes the nature of the question. A person's confidence in their own ability to do the same thing twice is a very different judgment than a person's confidence that some stranger could do something they have no relevant information about twice. The "Reminder" note commits an error in its description of null hypothesis testing.

Q6: Laypeople don't really understand what it means for results to be "in the same direction," so it would be better to say something concrete here instead. The "Reminder" note commits an error in its description of null hypothesis testing.

Q8: Laypeople don't understand factors, associations, and correlations. The language should be updated to use words like "relationships" and phrases like "an increase or decrease in X tends to happen whenever there is an increase or decrease in Y." The "Clarification" note commits an error in its description of null hypothesis testing.

(5) If I understand correctly, the Conditions are defined by the magnitude of X. It seems that presenting participants with seven questions all having the same magnitude of X reinforces the validity of X as an appropriate magnitude (especially when you consistently refer to "experimenters" and "researchers" in the question prompts). It was not clear to me why all three versions of each question are not randomized across participants instead of having three different sets of participants each focus on a specific magnitude of X. E.g. It should be possible for a given participant to see one question with magnitude X, another with magnitude 10X, and another with magnitude 100X.

***1.D Is the clarity and degree of methodological detail sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses?***

Yes. The present proposal meets the PCI standards. The Stage 1 protocol contains sufficient detail to enable replication by an expert in the field and ensures protection against research bias, undisclosed procedural, or analytic flexibility. The protocol specifies sufficiently precise links between the research question, hypotheses, sampling plans, analysis plans, and contingent interpretations given different outcomes.

***1.E Have the authors considered sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained***

***results are able to test the stated hypotheses or answer the stated research question(s?.***
Yes. The present proposal meets the PCI standards. The authors include a statistical sampling plan that is sufficient in terms of statistical power and/or evidential strength. The authors have minimised discussion of post hoc exploratory analyses, apart from those that must be explained to justify specific design features. The authors describe attention checks. However, manipulation checks are not described.

*Suggestions:*

(1) Direct Replication
It would be informative to include the original forms of each stimulus, with *no adjustment* from the target article, in the procedure. I do not think every participant needs to see both versions of each stimulus. I can think of two ways to incorporate the original stimuli in your procedure that won't add a great deal of time to the procedure for each participant: (1) participants could be exposed to one duplicate (e.g. a given participant responds to both your version of Q1 and Tversky & Kahneman's version of Q1, and to your versions of Q2-Q8; another participant responds to your version of Q2 and to Tversky & Kahneman's version of Q2, and to your versions of Q1 and Q3-Q8); (2) for each participant, one of the seven stimuli could be the Tversky & Kahneman version, and the other six could be your version.