This version: 2026/02/12

Don't Ask Don't Learn: Participant Explanations Can Identify Confounds and Mechanisms

Daniel Banki
Esade Business School
bankidaniel@gmail.com

Uri Simonsohn
Esade Business School
urisohn@gmail.com

Abstract
We propose that psychologists should, by default, ask study participants to explain why they did what they did. We illustrate the benefits by revisiting three research paradigms that had sparked published debates. We ran direct replications, adding a question to collect participant explanations. Example 1 involves subjective evaluations of the numbers "9" vs "221". We confirm the presence of a suspected confound and discover a second unsuspected one. Example 2 involves the gambler's fallacy. We find that it arises from wrong probability beliefs, refuting recently published claims. Example 3 involves the choosing vs rejecting paradigm. We use it to illustrate a case where a hypothesized mechanism would not be expected to surface in participant explanations, but where collecting explanations is nevertheless useful for detecting possible unexpected confounds. As it happens, we found one; it accounts for about half the effect.

Reproduction Package: https://ResearchBox.org/4605 (Use Code: CWIFTJ)

In contrast to chemistry, astronomy, and most other sciences, the objects of study in psychology are generally capable of having and expressing thoughts. We propose taking advantage of this privilege by asking participants in our studies, by default, why they did what they did. Participant explanations have been seen with suspicion in psychology at least since the work by Nisbett and Wilson (1977). They proposed that participant reports will only be accurate "when influential stimuli are salient and [constitute] plausible causes of [behavior]" (p. 231). In the interim five decades, our field has largely acted as if these conditions are never attained. But, in fact, they routinely are.

Even if every hypothesis that psychologists study involves imperceptible stimuli that have effects that participants would consider implausible (a very extreme premise), it would still be valuable to collect participant explanations, because our psychology experiments require operationalizations, those operationalizations can introduce unexpected confounds, and those confounds can be salient to our participants. We shouldn't expect a participant to report that a 50-millisecond prime of a banana made them think of a monkey when writing their short essay. But asking "why did you write about a monkey?" may reveal that the lab is decorated with a photograph of Jane Goodall petting a toddler chimp.

The optimal strategy, it seems to us, consists of always collecting information about participants' thoughts. We can always ignore information we consider irrelevant, but we can never use information we did not collect.

When as researchers we choose not to collect participant explanations, we implicitly assume that participants are so unlikely to be right about why they did what they did that there is no point in even asking them; we are also implicitly assuming we are so unlikely to have an overlooked confound in our experiment that there is no point in collecting data that may flag it. If participants in an experiment systematically attribute their behavior to a cause other than what the researcher claims is at play, the burden should be on the researcher to persuade readers that the participants are wrong. The current approach, not collecting participant explanations, hides rather than prevents the misinterpretation of experimental results.

It is important not to let the pendulum swing back further than it should. We should indeed not expect participants to attribute their behavior to non-salient causes. For example, in many between-subjects designs, the cause of the behavior is not salient because participants do not observe the counterfactual condition they could have been assigned to. Participants may also shy away from answering truthfully for social desirability reasons (e.g., they might be reluctant to mention that a target's race impacted their evaluation). Also, participants may simply find it difficult to articulate clearly what was behind their decision. Absence of evidence for a hypothesized mechanism should not be treated as evidence of absence of that mechanism. But even when we do not expect participants to be able to articulate the true cause of their behavior, it is still valuable to collect participant explanations. While they may not confirm the presence of suspected mechanisms, they may reveal the presence of an unsuspected confound. (The logic is similar to the Wason (1968) task, where one ought to seek information that may disconfirm, not only confirm expectations). If nothing else, participant explanations are a cost-effective catch-all strategy against unexpected confounds.

We present three demonstrations of the power of participant explanations to enrich how much we learn from psychology experiments. We looked for studies that had drawn above-average scrutiny in the form of published debates, be it commentaries or replication attempts.

Our examples are thus a stringent test of the ability of participant explanations to be valuable; all the low-hanging fruit should already have been picked by multiple teams engaged in the debate. If participant explanations can add to our understanding of phenomena that adversarial teams have already tackled from different angles, participant explanations are likely to add to our understanding of many phenomena.

When working on each of the three examples, we were surprised by the findings. In all three cases, we found something fundamental in the original study or phenomenon that neither the original authors, nor the researchers who published comments on the original work, nor we, expected to find.

Example 1. Detecting a Suspected Confound: 9>221

Birnbaum (1999) published a well-known study where, in a between-subjects design, participants rated the number 9 as larger than the number 221. Birnbaum was arguing against between-subjects designs and explained his finding as follows: "when different groups judge the subjective size of numbers, . . . 9 brings to mind a context of small numbers . . . [while] 221 invokes a context of 3-digit numbers" (p. 243, abstract).

This interpretation was challenged by Leong et al. (2019). They propose that the effect is artifactual, caused by the scale Birnbaum opted for to measure the perceived magnitude of the numbers. Specifically, Birnbaum asked participants to judge the size of 9 and 221 using a 10-point scale anchored at 1 (*very very small*) and 10 (*very very large*). See footnote for a critical discussion of how the scale was reported by Birnbaum.[1] Leong et al. propose that "some participants

[1] The article by Birnbaum does not mention that the 10-point scale he used had numeric anchors assigned to the extremes. The description of the scale in the original paper reads "Judgments were made on a 10-point scale, ranging from very very small to very very large." (p. 245). Numbers were not mentioned. This omission is remarkable for three reasons. First, it is unusual. We searched Google Scholar for articles published in 1999 containing the phrase

[mistook] the response scale (rating from 1 to 10) for the intended reference set (numbers from 1 to 10)." (p. 648). They found that the "9>221" finding does not replicate with different scales.

Intuitively, the problem is that participants in the "9" condition were not comparing the number 9 to other numbers that spontaneously came to mind, but instead, to the other numbers the experimenter had placed in front of them, the 1–10 scale. Or, as one of our participants put it, "On a scale from 1 to 10 the number 9 is the number 9."

It took the research community 20 years to discover this artifact. The original paper was published during the Clinton administration, the artifact was documented during the Trump administration.[2] We wondered whether asking participants to explain their responses would have flagged the scale confound from the very beginning.

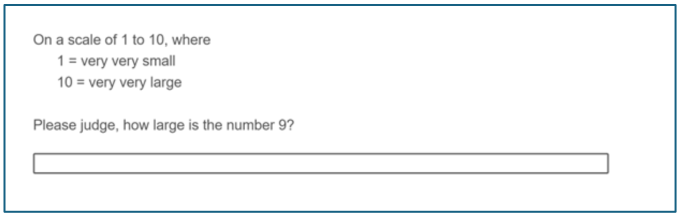
Method

Our sample, collected on 2025/10/24, consists of 200 US CloudResearch participants who were allowed to begin the study after passing an attention check (50.5% women, M_{age} = 37.15). As pre-registered, we excluded 3 participants because they copy-pasted text into the Qualtrics survey.

Participants were randomly assigned to evaluate the largeness of the number 9 or the number 221, using the original scale by Birnbaum. Figure 1 has a screenshot from our survey.

"10 point scale ranging from", and all ten results in the first page were papers that included the numeric anchors in the paper's description (we have a PDF of the Google results in our ResearchBox). Second, it makes the discovery by Leong et al. (2019) more remarkable. Readers of the Birnbaum paper cannot easily identify the scale confound because the scale was described incompletely, omitting said confound. And yet, Leong et al. did identify it (even if 20 years later). Third, while this specific omission may be rare, we suspect it is common that study descriptions fail to include all the design details that may produce confounds, making it difficult (or outright impossible) for readers and reviewers to spot them. Collecting participant explanations will help authors and readers spot confounds.

[2] According to Google Scholar, 189 articles published before 2020 cited Birnbaum (1999).



On a scale of 1 to 10, where
1 = very very small
10 = very very large

Please judge, how large is the number 9?

Figure 1. *Scale used to evaluate "9" and "221", same as Birnbaum (1999).*[3]

We then elicited participant explanations: “Please explain how you approached answering the question about how large the number [9/221] is and why you said <participant’s response>”.

Results

Replicating the original finding, "9" was rated as larger, M = 7.82, than was "221", M = 5.62, t(194.9) = 5.48, p < .001.[4]

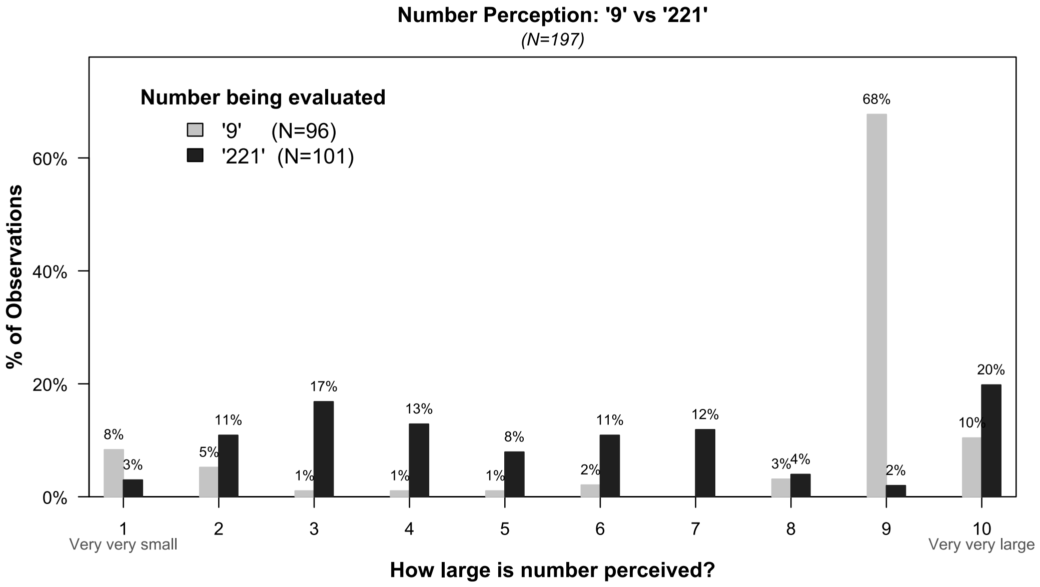


Figure 2. *Distribution of responses to the question "How large is the number [9/221]?"*

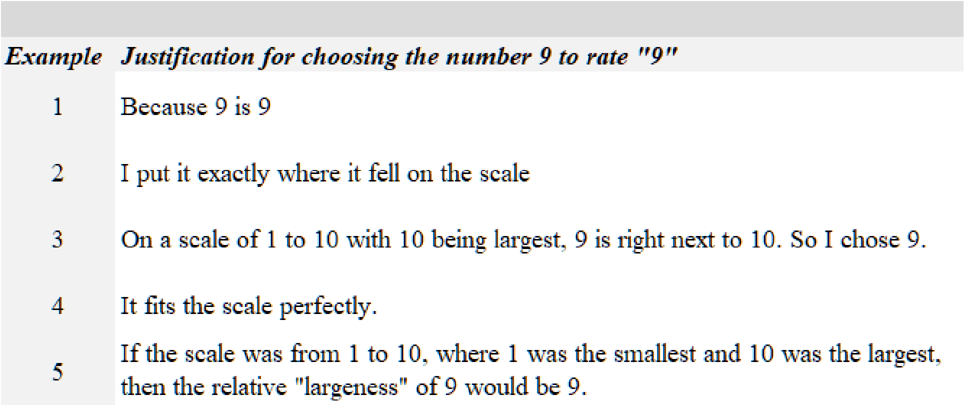
[3] The original article by Birnbaum (1999) links to the online form he used: https://web.archive.org/web/20160501000103/http://psych.fullerton.edu/mbirnbaum/number9.htm

[4] The means in the original were lower, M=5.13 and M=3.10 for "9" and "221", respectively.

The distribution of responses in Figure 2 (which neither previous paper reports) already suggests a problem. There is a glaring difference between conditions in the tendency to enter 9 as a response. To understand what drives this pattern, we turn to participant explanations.

We analyzed participant explanations in two steps. In the first step, we asked Claude 4.5 to create 5 post-hoc categories of explanations, separately for each condition. In the second step, we manually reviewed the explanations. We sorted them within condition & LLM category and read the first few at different levels of the dependent variable.[5]

The most relevant finding is that one of the 5 categories for the condition where participants evaluate the number 9 was "direct scale mapping" (with N=22 responses). Claude 4.5 describes this category as follows: "Participants directly map the number 9 to position 9 on the 1-10 scale". Figure 3 contains the first 5 answers in that category. They appear contiguously, they were not individually cherry-picked by us.



Example	*Justification for choosing the number 9 to rate "9"*
1	Because 9 is 9
2	I put it exactly where it fell on the scale
3	On a scale of 1 to 10 with 10 being largest, 9 is right next to 10. So I chose 9.
4	It fits the scale perfectly.
5	If the scale was from 1 to 10, where 1 was the smallest and 10 was the largest, then the relative "largeness" of 9 would be 9.

Figure 3. *First five participant explanations in the "direct scale mapping" category*

[5] We posted to ResearchBox the LLM-classified dataset, which also includes the manual annotations we made while reading the participant explanations.

When we inspected explanations in the other four categories, we found other cases clearly identifying a scale confound. For example: "The rating system was 0 to 10, and 9 is within that system." and "I felt like it was self explanatory, 9 is equal to 9".

We designed the study expecting that participant explanations would reveal the scale confound in the "9" condition. But we did not expect a confound in the "221" condition. Neither did Leong et al. (2019), who wrote: "note that this confusion could only affect responses in the 9 condition, because the 1–10 response scale can be mistaken for a comparative context for 9 but not for 221" (p. 648). The participant explanations revealed that we were all wrong. This design does not have one confound, it has two: the scale also produced artifactual responses in the "221" condition.

Specifically, some participants compared the number 221 to the highest number in the scale, 10, and correctly observed that 221>10. For example, consider these participant explanations we collected: "[221] is a lot higher than the numbers 1 through 10", "For reference, the 10 was labeled as very very large, so 221 being much larger should also be a 10.", and "I said 10 because 221 is very large on the scale". This probably explains an interesting pattern in Figure 2 we had initially missed. While most of the data for the "221" condition is to the left of the "9" condition, there are more participants at the highest possible value (10) in the "221" condition than in the "9" condition. Also, there are ten times as many 10s as there are 9s in the "221" condition.

In sum, in this first example we demonstrate that collecting participant explanations could have sped up the process of discovering the scale confound by 20 years, and that it helped us find a new confound present in a different condition, a confound that Leong et al. (2019) had missed.

Example 2. Unkilling a Psychological Mechanism: Gambler's Fallacy

The gambler's fallacy describes the mistaken belief that random streaks are disproportionately likely to end (Marquis de Laplace, 1902). It has traditionally been explained through biased probability beliefs (Kahneman & Tversky, 1972). Xiang et al. (2025) recently challenged this view. They argue that the gambler's fallacy "does not rely on probabilistic reasoning" (p. 12). In their studies, they included a condition where participants were directly asked for the probability that a streak would end. Xiang et al. report that the median response to that question was not systematically higher than the true probability (e.g., 50% for a 50:50 event).

Banki and Simonsohn (2026) reanalyzed their data relying on a statistical strategy with higher power and did find evidence of the gambler's fallacy in elicited beliefs. They nevertheless caution against using participant behavior in one paradigm to infer the rationale for participant behavior in another paradigm, especially when the two paradigms can differ—as they do here—both in complexity and the set of answers available (e.g., in the new paradigm participants can answer with "50:50"; in the traditional one, participants face a forced-choice: e.g., either heads or tails).[6]

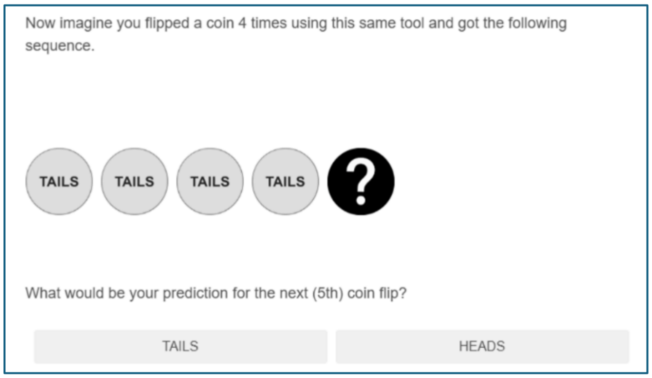
An alternative to using a new paradigm to identify possible mechanisms that could explain findings in an existing paradigm is to run the existing paradigm, eliciting participant explanations. That is what we did. As we detail next, we ran an experiment relying on the traditional gambler's fallacy paradigm, where participants predicted "heads" or "tails" after seeing a streak of length four, and we asked them to explain their prediction. We found ample support for the role of probability beliefs in the gambler's fallacy.

[6] In terms of complexity, the traditional paradigm has participants predict the next outcome (e.g., "heads" or "tails"), which requires simply understanding that a coin has two sides, something presumably everyone understands. In the new paradigm, participants have to provide a probability estimate, which requires understanding probabilities, something many people do not understand.

Method

Our sample, collected on 2025/09/13, consists of 380 US CloudResearch participants who were allowed to begin the study after passing an attention check (45.5% women, $M_{age} = 41.38$). All participants were asked to visit an online coin flip simulator (justflipacoin.com) and familiarize themselves with it by flipping a virtual coin a few times. To verify they had visited the site, we asked them to indicate the total flip count displayed on the page (about 340 million at the time). As pre-registered, we excluded 25 participants who answered incorrectly.

Then participants read "Now imagine you flipped a coin 4 times using this same tool and got the following sequence" and they had to press the "see sequence" button to proceed. When they did, an animation displayed four coin flips, one at a time, counterbalanced to be all heads or all tails. Then participants predicted the 5th flip by choosing "heads" or "tails" (presented in a counterbalanced order). See Figure 4. Then participants were asked to "Please briefly explain why you predicted <their prediction>".



Now imagine you flipped a coin 4 times using this same tool and got the following sequence.

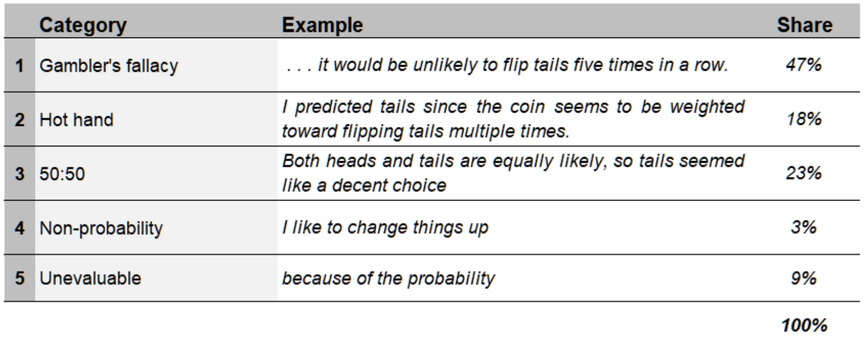
What would be your prediction for the next (5th) coin flip?

Figure 4. *Screenshot of key portion of materials for the gambler's fallacy study*

Results

Replicating the classic gambler's fallacy result, we find that 71.5% of participants predicted an end of the streak. This is significantly different from 50% ($\chi^2(1) = 65.1$, $p < .001$), and similar in magnitude to past estimates (see Rabin, 2002, p. 781). To understand whether this pattern is driven by participants' beliefs about the probability of the next outcome ending the streak, we turn to participant explanations.

We pre-registered a prompt to use with ChatGPT's API to classify participant explanations into 5 pre-determined categories (Gambler's fallacy, hot hand, 50:50 beliefs, a strategy unrelated to probability beliefs, and unevaluable). Figure 5 shows examples of participant explanations the LLM assigned to each category.



	Category	Example	Share
1	Gambler's fallacy	*. . . it would be unlikely to flip tails five times in a row.*	*47%*
2	Hot hand	*I predicted tails since the coin seems to be weighted toward flipping tails multiple times.*	*18%*
3	50:50	*Both heads and tails are equally likely, so tails seemed like a decent choice*	*23%*
4	Non-probability	*I like to change things up*	*3%*
5	Unevaluable	*because of the probability*	*9%*
			100%

Figure 5. *Explanation categories for predicting heads vs tails in Example 2***.**
Note: ChatGPT classified explanations by 355 participants into these five pre-registered categories relying on a pre-registered prompt. Explanation examples are real responses from participants in the study.

In Example 1, our goal was to identify a confound that affected the responses of at least some participants. Knowing that a non-trivial share of observations are confounded is enough to diagnose the problem, and a precise estimate is not particularly useful. Here, in contrast, we are interested in assessing the relative prevalence of explicit gambler's fallacy beliefs, and a precise estimate is particularly useful. So, we decided to supplement the pre-registered LLM classification with a classification done manually by the first author, relying on the same pre-registered categories. Reassuringly, this led to qualitatively similar results. Most notably, the LLM classified 47% of responses as involving gambler's fallacy beliefs, and the first author classified 43.9% of them as such (see Supplement 1). Thus, while some answers are ambiguous and difficult to classify, the overall conclusion that the plurality of respondents provide explicit gambler's fallacy justifications for their gambler's fallacy predictions seems robust to the classification approach.

Having explained why we consider the results to be face-valid, we highlight three key findings. First, as mentioned, about half of the participants explain their predictions using a belief-based gambler's fallacy logic. Among the subset of participants predicting an end of the streak, the LLM assigns 64% of explanations to the gambler's fallacy category (the first author, a similar 61%). This supports the interpretation that the gambler's fallacy findings have received for generations: people's beliefs about probability play a first-order role in the gambler’s fallacy.

Second, there were 81 participants providing explanations classified as "50:50" by the LLM (and 86 by the first author). One may expect that participants providing 50:50 explanations split their responses 50:50 between the streak ending vs continuing, but significantly more than half of those participants predicted an end of the streak, P = 70.4%, $\chi^2(1)=13.4$, $p < .001$. If we assume all of them are correctly classified, so that they all believe the probability is 50% (and, thus, they do not believe in the gambler's fallacy), about 5% of the sample exhibits the gambler's fallacy for non-gambler's fallacy reasons; we interpret this as directionally supporting Xiang et al.’s hypothesis, in that *some* of the evidence for the gambler's fallacy would not arise from biased beliefs. [7]

[7] 9% of responses were classified as unevaluable. They also are disproportionately likely to predict an end of the streak (with 67% chance), but because these explanations are unevaluable, we cannot confidently classify them as showing, or not showing, gambler's fallacy beliefs.

Third, unexpectedly, we find a substantive minority of participants who endorse a belief in the hot hand for the coin, despite it being a transparently random process. This is at odds with the common understanding that people exhibit the gambler's fallacy for random processes and that a belief in the hot hand only arises for human performance (see, e.g., Ayton & Fischer, 2004). Perhaps the key ingredient for hot hand beliefs is not human performance, but a plausible mechanism for the random process to be non-independent. In the spirit of Massey and Wu (2005), some participants are too ready to conclude that the process in front of them is different from the process they were expecting. Note that this bias would also be the direct consequence of the representativeness heuristic (“Fair coins should be 50:50. After four heads in a row, I conclude this is not a fair coin”).

In sum, in this second example we demonstrate that collecting participant explanations sheds substantial new light on an old literature and its recently proposed reexamination.

Example 3. Detecting an Unsuspected Confound: Choosing vs ~~Rejecting~~ Choosing

For this third example, our goal was to illustrate that when participant explanations do not support the mechanism that has been hypothesized for a finding, it should not necessarily be taken as evidence against that mechanism. For instance, when decisions follow subconscious, socially undesirable, or difficult to articulate mental processes, we should not expect participant explanations to reveal those processes.

We searched for studies that we expected would replicate but where we did not anticipate that participants would be able to identify the underlying mechanism. We again prioritized studies that had generated debate, and landed on a choosing vs rejecting study by Shafir and Cheek (2024).

We did not choose this study because we thought it was confounded (we did not think it was), but after collecting new data and reading participant explanations, it became clear that it was.

In the choosing vs rejecting paradigm, participants are offered (typically two) options and are asked to either choose the preferred option or reject the less preferred one(s). The motivating hypothesis is that "positive and negative dimensions of options . . . loom larger when . . . choosing and . . . rejecting, respectively . . " and the key prediction is that options with stronger pros and cons "tend to be chosen and rejected more often" (Shafir, 1993, p. 546).

Following independent replication failures by the Many Labs project (Klein et al., 2018) and by Chandrashekar et al. (2021), Shafir and Cheek (2024) proposed that the reason for the failures was that the "meaning and valence" (p. 1) of the original materials used by Shafir (1993) had changed in the intervening 30 years. Shafir and Cheek (2024) relied on pilot studies to design new stimuli, with which they conducted a pre-registered successful conceptual replication of Shafir (1993).

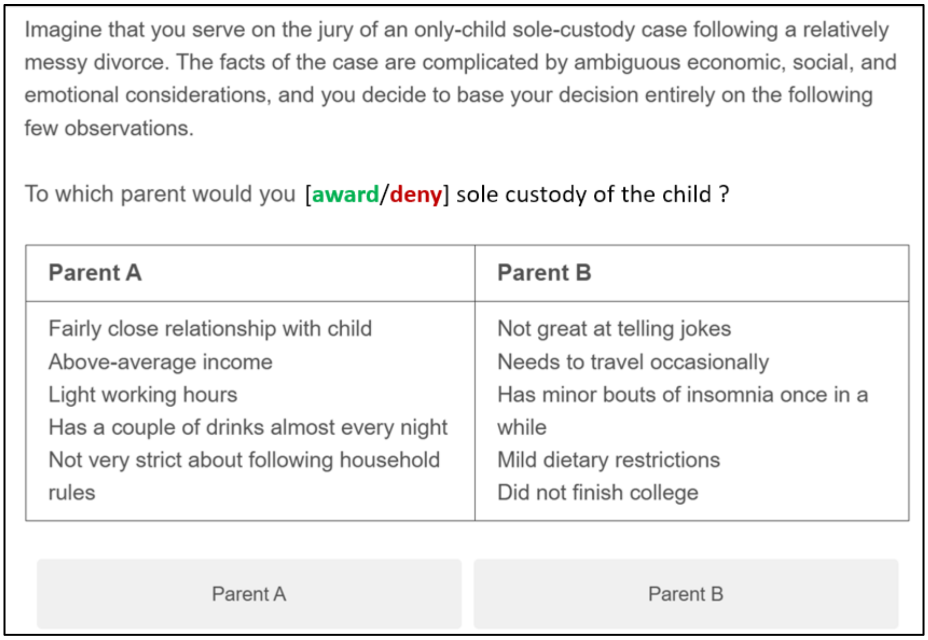
Geiser and Nelson (2026) conducted a replication study that randomly assigned participants either to the original stimuli by Shafir (1993) or the revised ones by Shafir and Cheek (2024). They found a small but significant effect with the original materials (7 pp difference between choosing and not-rejecting) and a significantly larger effect (25 pp) with the new materials.

In our study, we only included the new stimuli, and we asked participants to explain their responses. We obtained three key findings. First, we also replicated the choosing vs rejecting effect with the revised stimuli. Second, as we expected, participant responses did not mention the choosing vs rejecting frame; this was our original motivation for the study, illustrating that absence of process evidence in participant explanations should not be taken as evidence against that process. Third, unexpectedly, but very much in line with the motivation for this example, participant explanations revealed a confound: a substantial share of participants in the reject condition are confused and believe incorrectly that they are choosing. Importantly, this type of confusion biases the results towards the choosing vs rejecting prediction. We followed up this unexpected finding with a replication with three times the sample size; we estimate that about half of the choosing vs rejecting effect is artifactual. These two experiments are presented as Studies 3A and 3B below.

Study 3A. Exploration reveals an unexpected confound

Method

Our sample, collected on 2025/12/08, consists of 200 US CloudResearch participants who were allowed to begin the study after passing an attention check (50.0% women, M_{age} = 39.4). As pre-registered, we excluded 1 participant who copy-pasted text into the survey question. Participants were presented with the scenario from Study 1 by Shafir & Cheek (2024) depicted here in Figure 6.



Imagine that you serve on the jury of an only-child sole-custody case following a relatively messy divorce. The facts of the case are complicated by ambiguous economic, social, and emotional considerations, and you decide to base your decision entirely on the following few observations.

To which parent would you [award/deny] sole custody of the child ?

Parent A	Parent B
Fairly close relationship with child Above-average income Light working hours Has a couple of drinks almost every night Not very strict about following household rules	Not great at telling jokes Needs to travel occasionally Has minor bouts of insomnia once in a while Mild dietary restrictions Did not finish college

Figure 6. *Stimulus for Study 3A and 3B*
Original materials contain either the word "award" or "deny", not both, printed in black.

After participants chose which parent to assign custody to, they were asked to "Please explain how you approached answering the question about the custody case and why you chose to [award/deny] sole custody to <selected parent>" (participants saw either "award" or "deny").

Results

Replicating the choosing vs rejecting finding, the proportion of participants awarding custody to the extreme parent was higher in the choosing condition (*P*=60.6%) than in the rejection condition (*P*=46.3%), $\chi^2(1) = 4.06$, $p = .044$. [8]

As pre-registered, we asked ChatGPT to generate 5 categories into which to classify all participant explanations. In line with our expectations, none of the categories were about the elicitation mode (choosing or rejecting). They focused instead on the parental attributes and the

[8] Shafir (1993) develops an ad-hoc test that purportedly tests whether the same option is chosen and rejected by a plurality of people. Shafir's test, however, does not actually test that null. Moreover, it is nearly equivalent to a difference of proportions test when the sample size is the same across conditions, and it tests an irrelevant null when the sample size is not the same across conditions. See Supplement 2.

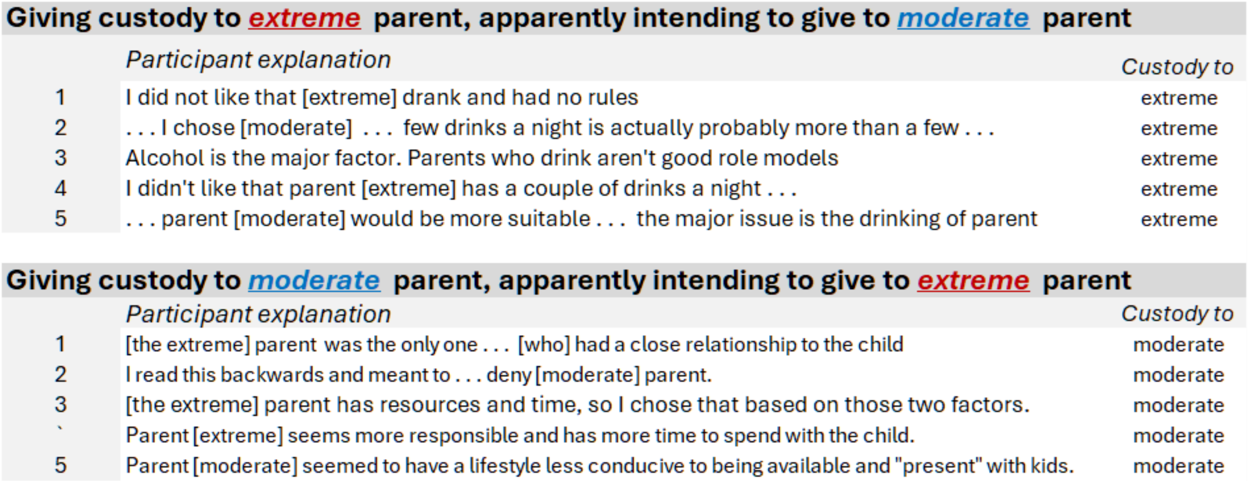
necessary tradeoffs between them (the categories were "making tradeoffs", "alcohol", "stability", "resources", and "holistic evaluation").[9]

Reading participant explanations, something caught our attention: some participants made arguments for one parent, but then assigned custody to the other. We only saw instances of this apparent mismatch in the rejection condition.

To illustrate, let's take a closer look at participants whose explanation was put in the "alcohol" category. As was shown in Figure 6, the extreme parent "has a couple of drinks almost every night". Presumably, this is seen as neutral or negative—but not positive—for parenting purposes. In line with this presumption, in the choosing condition, every participant with an "alcohol explanation" avoids the drinking parent (22 out of 22 participants). In the rejection condition, in contrast, 10 out of 31 participants with an "alcohol explanation" assign the kid to the drinking parent.

The top panel in Figure 7 shows some examples: abridged explanations by participants criticizing alcohol use, yet assigning custody to the drinking parent. To us, they all suggest participant confusion. The bottom panel has analogous examples from participants in another (non-alcohol) explanation category; here, participants verbally endorse the extreme parent but give custody to the moderate one.

[9] The category "holistic" was actually given the name "meta" by the LLM.



Giving custody to *extreme* parent, apparently intending to give to *moderate* parent

	Participant explanation	*Custody to*
1	I did not like that [extreme] drank and had no rules	extreme
2	. . . I chose [moderate] . . . few drinks a night is actually probably more than a few . . .	extreme
3	Alcohol is the major factor. Parents who drink aren't good role models	extreme
4	I didn't like that parent [extreme] has a couple of drinks a night . . .	extreme
5	. . . parent [moderate] would be more suitable . . . the major issue is the drinking of parent	extreme

Giving custody to *moderate* parent, apparently intending to give to *extreme* parent

	Participant explanation	*Custody to*
1	[the extreme] parent was the only one . . . [who] had a close relationship to the child	moderate
2	I read this backwards and meant to . . . deny [moderate] parent.	moderate
3	[the extreme] parent has resources and time, so I chose that based on those two factors.	moderate
`	Parent [extreme] seems more responsible and has more time to spend with the child.	moderate
5	Parent [moderate] seemed to have a lifestyle less conducive to being available and "present" with kids.	moderate

Figure 7. *Participant explanations in Study 3A that endorse one parent and grant custody to the other.*

Notes: We use the labels [extreme] and [moderate] because we counterbalanced across participants whether Parent A or Parent B was the extreme parent, so reading the explanations with the original labels would not be informative.

Our sample size was adequate to detect the choosing vs rejecting effect, but not to explore a subset of participants getting confused in only one of the conditions.[10] We thus conducted a replication with three times the sample size (n=300 per cell), pre-registering that we would use participant explanations as an alternative measure of the intended decision. Specifically, that we would analyze the data in two ways: 1) based on the participant’s decision and 2) based on the participant’s explanation (unless the explanation didn’t allow a clear inference about the intended parent, in which case we went with the participant’s original decision).

[10] The effect size in Study 1 by Shafir & Cheek (2024), comparing proportions of 71% vs 43%, requires n=50 per cell to obtain 80% power. We ran twice that (n=100 per cell), which would give us 98% power to detect an effect of that size.

Study 3B. Confirming Participant Confusion as Choosing vs Rejecting Confound

Method

Our sample, collected on 2026/01/30, consists of 600 US CloudResearch participants who were allowed to begin the study after passing an attention check (48.5% women, M_{age} = 41.6). As pre-registered, we excluded 13 participants who copy-pasted text into the survey question. We used the same materials from Study 3A (see Figure 6), collecting both 1) decisions (choosing/rejecting) and 2) participant explanations for those decisions.

We used a pre-registered prompt to instruct an LLM (Claude Sonnet 4.5) to read participant explanations and infer the parent that the participant wanted to give custody to. When making this inference, the LLM only had access to the explanations, so it was blind to the condition and the participant’s decision. The first author also performed, as pre-registered, an independent evaluation of every explanation that the model had coded as discordant with the participant’s decision.[11] The author coding was blind to the condition and to the participant decision. It was also blind to the LLM’s classification.

Results

Beginning with participant decisions, we replicated the choosing vs rejecting finding again. The proportion of participants awarding custody to the extreme parent was higher in the choosing condition (P=68.0%) than in the rejection condition (P=45.2%), $\chi^2(1) = 31.02$, $p < .0001$.

Turning to participant explanations. The LLM provided an inferred parental preference for 94.5% of explanations, and those matched the participant’s decision in 77.7% of cases. The first

[11] We had pre-registered that the authors would infer parental choice only for the subset of explanations in which the LLM and participant decisions were discordant. However, the task proved less taxing than anticipated, and before any comparison between author and LLM evaluations, the first author inferred parental choice from all explanations.

author's classification produced similar results. A parental preference was inferred for 86% of explanations, and of those, 82.4% matched the participant’s decision. Figure 8 below shows where the 17.6% of non-matching responses (N=89) originate. We refer to them as mistakes, because we believe that a decision that contradicts its justification is a mistaken decision. The figure shows two things. The first one is that mistakes originate almost exclusively in the rejection condition (presumably because it's unusual to be asked which parent should be rejected, so people default to thinking about which one should be chosen). The second one is that they disproportionately involve rejecting the extreme parent (presumably because that's the preferred parent in choice, and these participants mistakenly believe they are being asked to choose a parent).

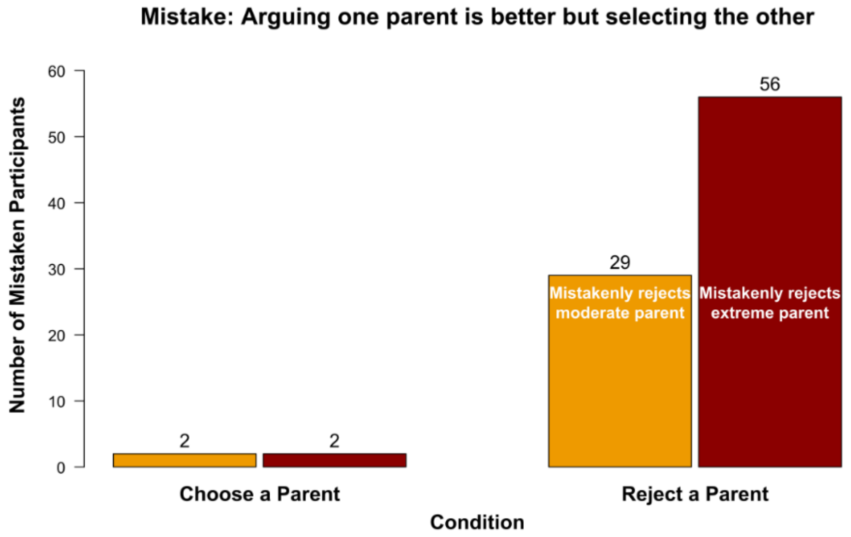


Figure 8. *Mistakes were made.*

Turning to our second pre-registered analysis, we created a new variable with corrected participant decisions, switching the decision to align with the participant explanation when there was a mismatch. We then repeated the difference of proportions test on this new variable. The difference between choosing and rejection remained significant, though it was significantly attenuated, dropping from 23 pp to 11 pp (see Figure 9). The figure suggests that about half of the choosing vs rejecting effect is artifactual.

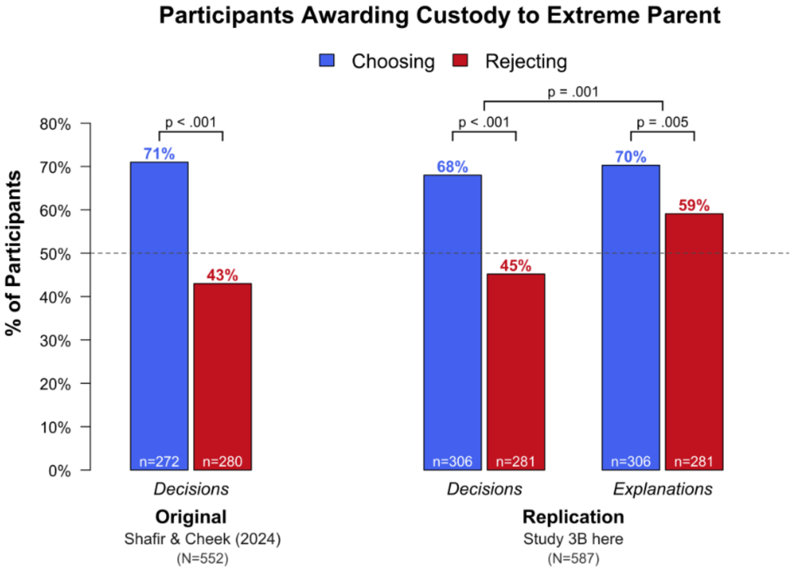


Figure 9 *Parental decisions and explanation-implied decisions in Study 3B*
Note: The N=587 participants in the replication are the same across the two sets of bars. The first set shows participants’ original decisions. The second set shows decisions inferred from the participants’ explanations. The *p*-values for the blue-red bar pairs are obtained with proportions test; the *p*-values for the interaction are obtained through a regression with two observations per participant (one decision-based, one explanation-based), clustering errors by participant. The interaction test was not pre-registered.

An alternative to correcting mistaken responses is to exclude them. This alternative approach is problematic but leads to qualitatively similar results (about half of the effect is artifactual). See Supplement 3.

In sum, in this third example we demonstrate two key points. First, we should not expect participant explanations to necessarily reveal the mechanism behind an effect. For choosing vs rejecting, one may not expect participants to be able to detect that elicitation mode impacted their decisions. In fact, they didn’t detect it. Second, even in studies where one does not expect mechanisms to be revealed by participant explanations, collecting explanations is worthwhile because they can identify unexpected confounds. In this particular case, a 30-year-old influential finding had a first-order confound that explains about half of the effect. This confound went unrecognized by multiple teams of authors who thought very carefully about the design and conducted numerous replications with thousands of participants in total. Despite all the engagement with the study design, the confound was only identified after collecting participant explanations.

How to collect and analyze participant explanations

Psychological research is broad enough, the technology available to process open-ended text is developing fast enough, and our experience collecting and analyzing participant explanations is limited enough that we are reluctant to provide strict recommendations for best practices. Instead, we share some lessons we have learned, hoping that readers will find them useful as they consider what's the best path forward given their research program and currently available technology.

Collecting participant explanations

In terms of eliciting participant explanations, we found that it was useful to explicitly ask participants to explain their decision/action/rating. We always reminded participants of what they had done and used the following template question: "Please explain how you approached <describe the task they performed> and why you <describe the action they took> ". For our three examples, the questions were:

Example 1. 9>221: "Please explain how you approached answering the question about how large the number [9/221] is and why you said <participant’s response>."

Example 2. Gambler's fallacy: "Please briefly explain why you predicted <heads/tails >."

Example 3. Choosing vs Rejecting: "Please explain how you approached answering the question about the custody case and why you chose to [award/deny] sole custody to [Parent A/Parent B]."

Browsing participant explanations

In terms of analyzing the obtained textual data, LLMs make the task obviously easier, but whether authors rely on LLMs, older machine learning tools, or human raters, we recommend that authors themselves always read many of their participants' explanations and that they do this with an open and exploratory mindset, seeking to catch aspects of the experiment they may be missing. Because the amount of text generated can be overwhelming for such a casual inspection, we recommend reading a stratified sample of responses. Specifically, consider the following: sort the data first by condition and then by values of the dependent variable, then browse a few explanations for low, medium, and high values of the dependent variable in each condition.

Systematic processing of textual data

Our systematic analysis of text data took two main approaches, which we refer to as exploratory and confirmatory.

In exploratory analyses, which we always relied on an LLM to perform, we asked the LLM to create five explanation categories and assign each explanation to one of these categories. This is a complement to, but perhaps can often be substituted by, the stratified sampling approach we just mentioned for browsing the data casually. The goal is to identify clusters of answers that may help make sense of the data, flag unexpected patterns, or verify expected patterns. This is what we relied on in the choosing vs rejecting study; there, we found that all clusters were about parental attributes rather than about the elicitation mode. This is also how we found the surprising confound in the "221" condition of the 9 vs 221 study.

For confirmatory analyses, in contrast, one specifies the categories of explanations ahead of time and then instructs an LLM, human raters, or the authors themselves, to classify responses into pre-selected categories. This is what we did in the gambler's fallacy experiment when we classified responses as 50:50, gambler's fallacy, hot hand, idiosyncratic strategy, or unevaluable.

For confirmatory analyses, we recommend that authors pre-register their classification scheme, and if they will rely on LLMs, also the prompts they will use. Participant explanations are notably richer and more diverse than traditional dependent variables. Therefore, we should expect, at least at the beginning, more frequent deviations from pre-registrations. But precisely because participant explanations can identify unexpected patterns, it seems useful to document what patterns are expected. As the pre-registered prompts were often too long for human readers (and for AsPredicted), we pre-registered them as stand-alone text files in ResearchBox. Like all files, these files have a permanent timestamp, and we linked to them from our AsPredicted pre-registration.

Conclusion

For roughly 50 years, psychology has largely avoided asking participants why they did what they did in experiments. We believe this has left enormous amounts of information on the table and has almost surely slowed the pace of discovery, while prolonging pursuits down wrong paths in psychological research.

In this paper, we presented three examples where we simply added a question asking for participant explanations. In all three, we discovered new, first-order facts, despite working within heavily studied and mature research programs. A large amount of information is waiting to be collected in novel research programs. All you need to do is ask.

References

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler’s fallacy: Two faces of subjective randomness? *Memory & Cognition*, *32*(8), 1369-1378.

Banki, D., & Simonsohn, U. (2026). Commentary on Xiang et al. (2025): The Gambler's Fallacy Fallacy Fallacy. In.

Birnbaum, M. H. (1999). How to show that 9> 221: Collect judgments in a between-subjects design. *Psychological Methods*, *4*(3), 243.

Chandrashekar, S. P., Weber, J., Chan, S. Y., Cho, W. Y., Chu, T. C. C., Cheng, B. L., & Feldman, G. (2021). Accentuation and compatibility: Replication and extensions of Shafir (1993) to rethink choosing versus rejecting paradigms. *Judgment and Decision Making*, *16*(1), 36-56.

Geiser, A. E., & Nelson, L. D. (2026). ResearchBox #5801 "Choosing vs Rejecting Replications". In.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, *3*(3), 430-454.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., & Bahník, Š. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443-490.

Leong, L. M., McKenzie, C. R., Sher, S., & Müller-Trede, J. (2019). Illusory inconsistencies in judgment: Stimulus-evoked reference sets and between-subjects designs. *Psychonomic Bulletin & Review*, *26*, 647-653.

Marquis de Laplace, P. S. (1902). *A philosophical essay on probabilities*. Wiley.

Massey, C., & Wu, G. (2005). Detecting regime shifts: The causes of under-and overreaction. *Management Science*, 932-947.

Nisbett, R. E., & Wilson, T. D. (1977). Telling More Than we Can Know - Verbal Reports on Mental Processes. *Psychological Review*, *84*(3), 231-259.

Rabin, M. (2002). Inference by believers in the law of small numbers. *The Quarterly Journal of Economics*, *117*(3), 775.

Shafir, E. (1993). Choosing Versus Rejecting - Why Some Options Are Both Better and Worse Than Others. *Memory & Cognition*, *21*(4), 546-556.

Shafir, E., & Cheek, N. N. (2024). Choosing, rejecting, and closely replicating, 30 years later: A commentary on Chandrashekar et al. *Judgment and Decision Making*, *19*, e5.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly journal of experimental psychology*, *20*(3), 273-281.

Xiang, Y., Dorst, K., & Gershman, S. J. (2025). On the Robustness and Provenance of the Gambler’s Fallacy. *Psychological Science*, *36*(6), 451-464.