First version: 2026/02/16
This version: 2026/06/26

Confounds Protocols: Collecting & Analyzing Participant Explanations to Identify Confounds (and Sometimes Mechanisms)

Daniel Banki
Esade Business School
bankidaniel@gmail.com

Uri Simonsohn
Esade Business School
urisohn@gmail.com

Abstract
To protect against unexpected confounds, we propose collecting participant explanations in psychology experiments by default. We introduce the "Confounds Protocols", a five-step procedure for collecting and analyzing such data. We demonstrate their value by replicating three well-known studies that had already sparked published debate. In each example, participant explanations allowed us to discover something new and fundamental, be it confounds or mechanisms, about these mature paradigms.

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, and by default asking participants in our experiments to explain what they did. Asking participants to provide verbal reports of their experiences was common in the early years of experimental psychology; notably Wilhelm Wundt, recognized as the first experimental psychologist, relied heavily on such data in the late 1800s (see Boring, 1929). The popularity of collecting participant explanations, however, plummeted with behaviorism, and it never really recovered. The notable exception is the subfield of "problem solving" within cognitive psychology, where think-aloud protocols are central (Ericsson & Simon, 1980, 1993). As influential and important as that exceptional subfield is, it is that; the exception. Collecting, or at least reporting participant explanations in psychology experiments remains unusual.[1]

A recurring justification to not ask participants for explanations is the work by Nisbett and Wilson (1977). Experimental psychologists today act as if that paper demonstrated participant explanations are never useful. Here we propose, in contrast, that participant explanations are almost always useful. The key point of departure is our focus on confound identification. While participant explanations sometimes can't tell us whether the expected psychological mechanism is at play, they typically can tell us whether an unexpected confound is.

Building on the work of Ericsson and Simon (1980), we begin by noting that participant explanations are expected to be accurate when they explain a recent & deliberate action.[2] This

[1] Historically, this was the case in economics too, but there is a recent literature using “open-ended” questions as data (Haaland et al., 2025). That recent literature includes some papers in experimental economics (Andre, 2025; Braghieri et al., 2024), and we are aware of an earlier experimental paper by Brown (2005) examining the causes of the endowment effect. Our focus in this paper, as we explain later, differs from theirs (and that of the problem-solving subfield) because we propose collecting participant explanations with the specific goal of identifying confounds, and we put forward a specific process, "Confounds Protocols", to do so.

[2] Ericsson and Simon (1980, p. 220) distinguish among three levels of verbalization. Levels 1 & 2 are likely to be accurate, and those rely on working memory, while level 3 is not necessarily expected to be accurate, and it involves elaborating above and beyond working memory. In our demarcation, we spell out what it is required for something

demarcation is simple and useful. Participant explanations are expected to be inaccurate when they involve something 1) that is not recent (e.g., "why did you become a nurse?"), 2) that is not an action, either because it asks about a counterfactual, something the participant did not do (e.g., "What would you have done if the confederate hadn't apologized?") or because it asks the participant to generalize from a given action (e.g., "Why are you risk seeking for losses?"). Participant explanations will also be inaccurate when 3) what the participant did is not deliberate (e.g., "Why do you find Alex attractive?"). In addition, we note that explanation accuracy requires that 4) participants lack a motive to lie, such as when honest responding may be socially undesirable (e.g. "Why did you sit so far away from the minority confederate?") or when it may reveal the participant is not taking the experiment seriously (e.g., "Why did you say that you spent 69 minutes on this task?").

Even if every hypothesis psychologists studied required an experimental design that led to violating at least one of these four conditions for explanation accuracy (an extreme premise), it would still be valuable to collect participant explanations in most experiments, because psychology experiments require operationalizations, those operationalizations can introduce unexpected confounds, and *those confounds* may (and usually will) satisfy these conditions. We shouldn't expect a participant to report that a 150-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 logic is similar to the Wason (1968) task, where one ought to seek information that may disconfirm, not only confirm expectations.

---

that is being explained by a participant to be in their working memory, proposing that it involves asking about something they did recently and deliberately.

When as researchers we choose not to collect participant explanations, not only do we implicitly assume that participants are so unlikely to be right about why or how they did what they did that there is no point in asking them (a questionable assumption), but we also implicitly assume we are so unlikely to have an overlooked confound in our experiment that there is no point in collecting data that may flag it (an indefensible assumption). 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 it is the participants who are wrong. The current approach, to almost never collect or at least report participant explanations, hides situations when researchers misunderstand their experimental results; as we show in the three examples discussed below, sometimes those misunderstandings last decades. The optimal strategy, it seems to us, consists of *always* collecting participant explanations. We can ignore information we deem irrelevant, but we cannot use information we did not collect.

In the next three sections, we illustrate our proposed use of participant explanations by revisiting three well-known psychological findings. 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 the multiple teams engaged in the debate.

When working on each of the three examples, we were surprised by our 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. Working on these examples also helped us develop what we hope is a reusable and generalizable five-step approach for collecting and analyzing participant explanations with the goal of detecting confounds. We refer to it as the "Confounds Protocols". Figure 1 overviews the

five steps; to avoid repetition, and to allow for a concrete exposition, we present the protocols in detail within our discussion of Example 1. Because we developed the protocols informed by our examples, some of the analyses we performed, especially the LLM prompts we rely on to categorize participant explanations, deviated from our pre-registrations. Our qualitative conclusions do not hinge on those deviations, but because we believe the prompts we developed later, as part of the Confounds Protocols, were superior, we report those results here and we relegate the results for the pre-registered prompts to the supplement.

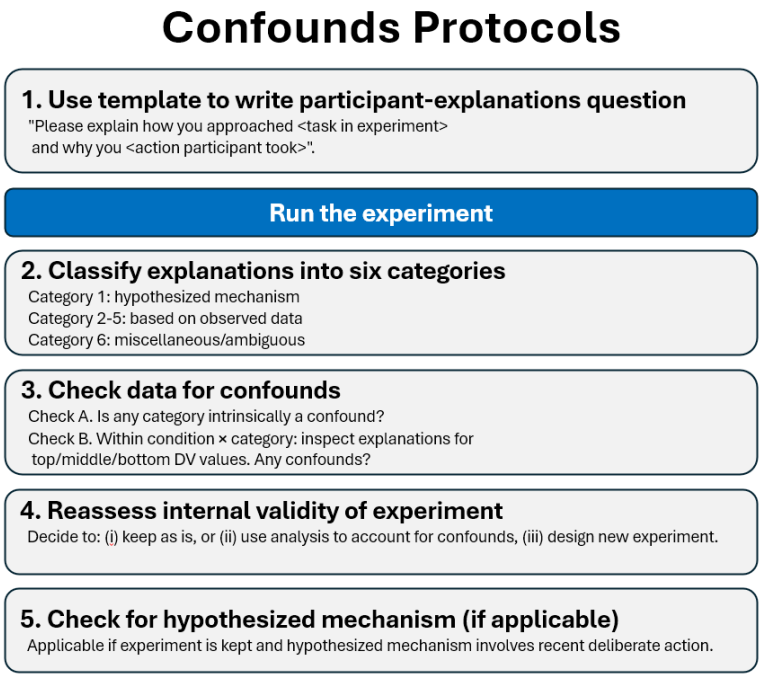

Figure 1. *Five-step process for using participant explanations to find confounds*

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.[3] Leong et al. propose that "some participants [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.[4] We wondered whether asking participants to explain their responses would have flagged the scale confound from the very beginning.

[3] 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 "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.

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

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 2 has a screenshot from our survey.

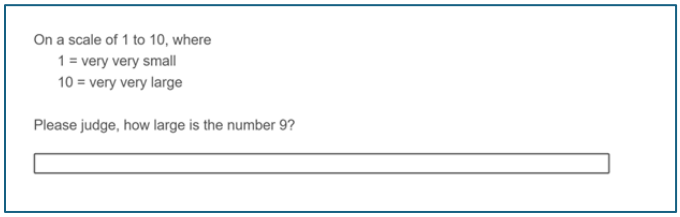

Figure 2. *Scale used to evaluate "9" and "221", same as Birnbaum (1999).*[5]

We then elicited participant explanations. This takes us to the first step in the Confounds Protocols: crafting a question to elicit participant explanations. We propose that the question should (i) ask about a concrete action a participant took, and do so (ii) without building in any assumptions by the experimenter about the meaning or cause of that action. To illustrate, a question that would violate the first principle is "Please share any thoughts you had while answering the question about the number '9'". One that would violate the second is "tell us which sets of numbers you compared '9' to when assessing its magnitude" (this assumes participants used sets of numbers as reference, but they may not have done that).

We proposed the following simple (and not especially original) template question that satisfies both principles: "Please explain how you approached <task participants carried out in the

[5] 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

experiment> and why you <action participant took>". For this first example, it became "Please explain how you approached answering the question about how large the number [9/221] is and why you said <participant's response on the 1-10 scale>".

Before continuing to the results of this first example, a note on *when* to ask participants for explanations. All three examples in this paper involve brief single-action experiments. In such experiments, the only time to ask for a participant explanation is immediately after the single action. But, in multi-action experiments (e.g., where participants rate 10 different morally ambiguous acts) one could ask after the first action, after the last, at a random point, ask for all of them, have a pilot with just one question where participant explanations are collected, etc. The optimal strategy, among these or possibly others, depends on the specifics of the multi-action experiment in question.

Results

Beginning with the participant ratings of number magnitude, we replicated the original finding: "9" (M=7.82) was rated as larger than "221" (M = 5.62), t(194.9) = 5.48, $p < .001$.[6]

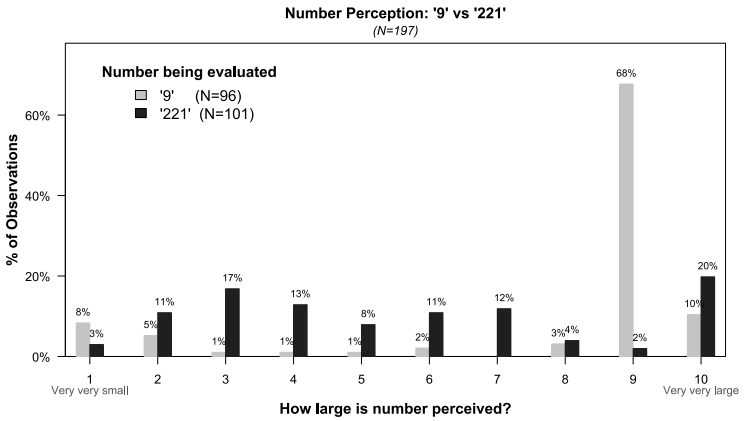

Figure 3. *Distribution of responses to the question "How large is the number [9/221]?"*

[6] The means in the original study, back in 1999, were lower, M=5.13 and M=3.10 for "9" and "221", respectively.

The distribution of responses in Figure 3 (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.

This takes us to Step 2 in the Confounds Protocols. It involves organizing participant explanations into categories to facilitate the exploration of potential confounds. We propose by default creating six different categories. The first category is set ex ante, based on the hypothesis of interest to the researcher. The next four categories are meant to be exploratory, generated ex post based on the obtained data. The idea is to classify explanations into categories that are meaningfully distinct and internally cohesive. The sixth category is an explicit catch-all "ambiguous/unclear" category to avoid distorting the other categories when attempting to fit all of the explanations into just 5 categories. (This is a default approach to categorization, which will often need adjustments. In our second example, for instance, we have four competing hypothesized mechanisms, so we define four categories ex ante, rather than only one.) To carry out this second step via LLMs, we developed a template prompt. It includes instructions to generate a helper text file with descriptions of each category (see Supplement 1).

In Step 3 of the Confounds Protocols we actively look for confounds, first by examining the category descriptions, then by using those categories to systematically examine participant explanations. In Example 1, we begin, then, exploring the categories of explanations created. Starting with the category created ex-ante with Birnbaum's hypothesis of reference sets, a small minority of responses (n=14) were assigned to this category. This is not necessarily problematic, as it is unclear whether the participants would be aware of the mechanism Birnbaum had in mind (conscious deliberation about number size could perhaps begin only after the effect of interest, the

automatic generation of a reference set, had already occurred below awareness). Moving on to the other categories of explanations, however, we do encounter problematic evidence.

The category with the largest share of explanations (n=80) maps onto the artifact pointed out by Leong et al. That category, "scale", is described by the LLM as "The explanation treats the response scale or an explicitly constructed rating range as the comparison frame […]" The remaining three categories simply involved which arbitrary comparison participants relied on. [7]

Following with the Confounds Protocols we now sort the data by condition, explanation category and dependent variable. Within each condition×category bin, we then read participant explanations with high, medium, and low dependent variable values to identify possible confounds. This process leads to very explicit evidence of the confound: participants in the "scale" category are unambiguously using the scale as the point of reference. Figure 4 displays explanations in the "9" condition, within the "scale" category, for participants who assessed the magnitude as a 9.

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 4. *Participant explanations in the "scale" category*

[7] The three other categories were related to what comparison group had been used. The categories were thus "large", "middle", and "small". For example, the explanation "Numbers run to infinity, with that in mind 221 is very small" was assigned to the "small" category.

In light of Leong et al. (2019) we expected this result, but further following Step 3 in the Confounds Protocols, we also came across an unexpected result. Specifically, when we moved on to participant explanations in the "221" condition, specifically those with high values of the dependent variable in the "scale" category, we realized there was a scale confound also in that condition, one we had not expected. Neither had 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 in that bin made it clear that some participants in "221" condition also used 1-10 as a point of reference and judged 221 as very large for that reason. For example, they wrote "[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.". This second confound probably explains an interesting pattern in Figure 3 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 ceiling of the scale in the "221" condition than in the "9" condition.

In Step 4 of the Confounds Protocols, we reassess the internal validity of the experiment in light of the evidence uncovered in Step 3. In this step, when no confounds are uncovered, one keeps the study as is. When confounds are uncovered, one must decide whether the impact of the confounds can be removed via statistical analysis, or whether the problem seems unfixable and a new experimental design is warranted. This decision obviously is subjective and needs to be made on a case-by-case basis. In this particular case, the confound seems overwhelmingly consequential and to us only seems addressable by running new studies with a different scale (like those run by Leong et al., those that failed to replicate Birnbaum's findings).

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 realize the confound impacted both conditions, not only the "9" condition.

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).

Other researchers, however, analyzed the same data relying on alternative statistical strategies, with higher power, to detect a minority of participants exhibiting the gambler's fallacy, and did find evidence of it in elicited beliefs (Banki & Simonsohn, in press; Choi & DeKay, 2026). Those re-analyses showed that at least some of the evidence of the gambler's fallacy in predictions comes from gambler's fallacy beliefs. In this second example, we rely on participant explanations to provide a more direct and quantitative assessment of the relative role of biased beliefs in the gambler's fallacy.

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.[8] We found ample support for the role of probability beliefs in the gambler's fallacy.

[8] Banki and Simonsohn (in press) report a variant of this design inspired by the study presented here.

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 349 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 5. Then participants were asked to "Please briefly explain why you predicted <their prediction>".

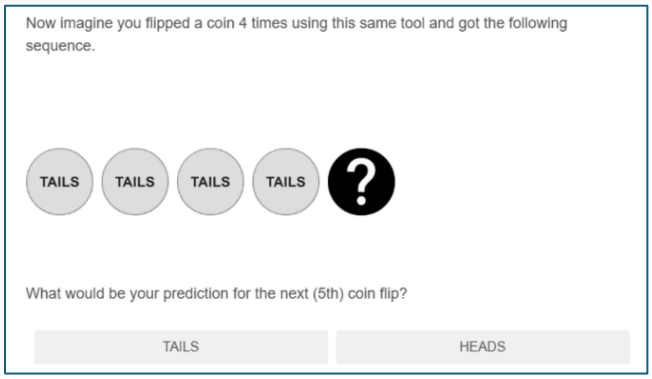

Figure 5. *Screenshot of key portion of materials for the gambler's fallacy study*

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).

Moving now to Step 2 in the Confounds Protocols, we classified responses into categories. This experiment is highly unusual in that we had not just 1 but 4 hypothesized mechanisms for behavior. The 4 mechanisms are: participants 1) believe that streaks are overly likely to end (gambler's fallacy), 2) believe they are overly likely to continue (hot hand), 3) believe they are irrelevant for predicting the next outcome (rationality), or 4) make predictions about heads vs. tails that are unrelated to probabilities. The fourth mechanism is the hypothesis by Xiang et al. (2025) that motivated this example. We thus carried out a modified Step 2 where we pre-specified the four categories of interest and created a fifth catch-all category we could explore potential confounds with (full prompt in Supplement 2, some excerpts in footnote).[9]

Moving on to Step 3, Figure 6 shows the share of explanations assigned to each of the categories. We see that about half of participants, *P*=51%, gave an explanation that involved an explicit gambler's fallacy, while just *P*=6% of them gave a non-probability based one.

	Category	Example	Share
1	Gambler's fallacy	*. . . it would be unlikely to flip tails five times in a row.*	*51%*
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*	*16%*
4	Non-probability	*I like to change things up*	*6%*
5	Unevaluable	*Just the odds I suppose.*	*9%*
			100%

Figure 6. *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.

[9] The full text of the prompt for classifying the gambler's fallacy explanations is in Supplement 2. We include key quotes to give a sense of the instructions here. Prompt:"[. . .] Your job is to classify the explanations into 5 categories.. . . "GF" = The explanation says or clearly implies that a long sequence makes the end of the sequence more likely now [examples redacted] . . . "HH" = The explanation says or clearly implies that a long sequence makes the continuation of the sequence more likely now . . . [examples redacted] . . . "50_50" = The explanation says or clearly implies that each flip is independent or 50/50 and justifies picking randomly . . . ".

We now move on to exploring participant explanations for different values of the dependent variable within each explanation category bin (note that there were no conditions in this study). In the gambler's fallacy and hot-hand categories, participants' explanations matched their prediction (e.g., that the streak ends in the former).

We take a paragraph-long detour on the hot hand. We note that through participant explanations in a pilot study we realized that some participants in gambler's fallacy experiments exhibit the hot-hand fallacy (which is why we included it as an ex-ante category). This challenges what seems to be the consensus view that people exhibit the gambler's fallacy for random processes and a belief in the hot hand 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.

Moving on from the hot hand. In the "50:50 (rational)" bin, one may expect that participants 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, $P = 69.6\%$, $\chi^2(1)=8.6$, $p = .003$. 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), then a small minority of participants exhibiting the gambler's fallacy do so for reasons other than biased beliefs (the mechanism that Xiang et al. hypothesized was entirely behind the gambler's fallacy).

In terms of Step 4, reassessing internal validity. We interpret a belief in the hot-hand as a confound that attenuates the effect of the gambler's fallacy. This confound is difficult to identify with prediction data (when participants predict heads or tails), but easy with participant

explanations. Thus, while the Confounds Protocols have identified a confound, our read is that the study is as high in internal validity as one may hope.

In terms of Step 5, examining the impact of the hypothesized mechanism, by the nature of that example, this was done as part of Step 3. 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 a hypothesized mechanism, it should not necessarily be taken as evidence against that mechanism. We looked for studies 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, the "enriched option", "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) ran a replication with both sets of stimuli, finding a larger effect with the new (25 pp) compared to original stimuli (7 pp).

In our study, we only included the new stimuli, and we asked participants to explain their responses. We obtained three key findings. First, we replicated the choosing vs rejecting effect with the revised stimuli: more people selected the "enriched" option when choosing rather than rejecting. Second, as we expected, participant responses did not mention the choosing vs rejecting frame (this was not surprising, as doing so would have required counterfactual thinking, i.e., imagining a condition they were not assigned to). Third, unexpectedly, participant explanations revealed a confound: a substantial number of participants in the rejection condition evidently misunderstood the instructions and incorrectly believed they were asked to choose rather than to reject. Because of this unexpected result, we decided to run a replication with a larger sample, pre-registering that we would also re-analyze the data using as the dependent variable the decision implied by the participant explanation. We report only this replication in this paper, and the first study, which can be thought of as a pilot, in Supplement 3.

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. Participants were presented with the scenario from Study 1 by Shafir & Cheek (2024) reprinted here in Figure 7; participants were asked to select one parent to receive full custody of their child. One parent had extreme attributes, the "enriched" option, and one had more moderate ones, the "impoverished" option. Which parent appeared on the left (as Parent A) was counterbalanced.

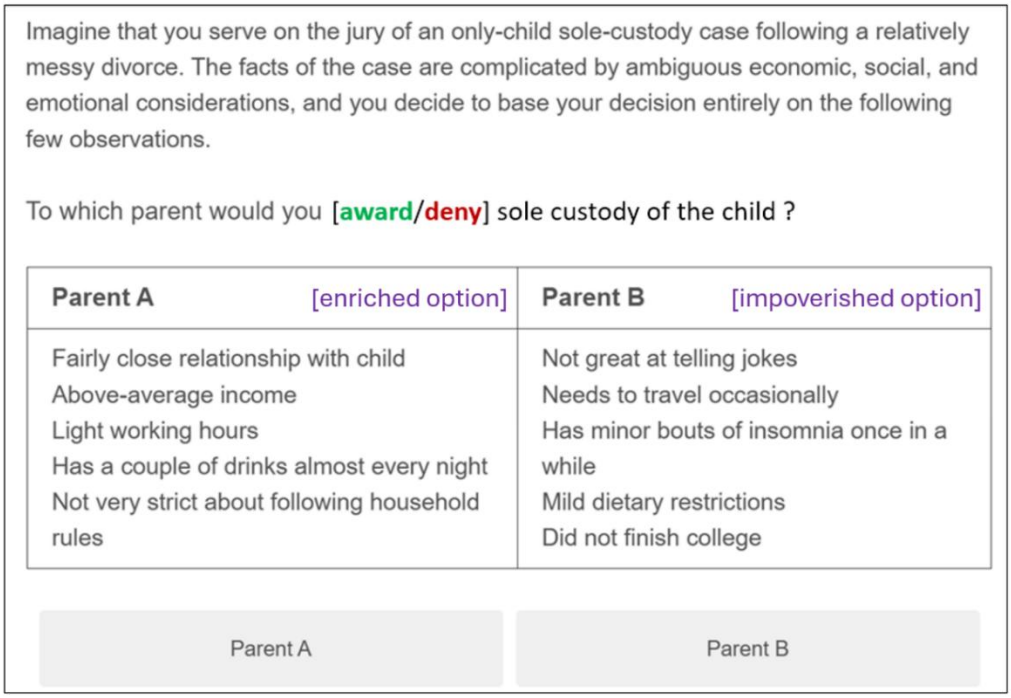

Figure 7. *Stimulus for Example 3*
Original materials contain either the word "award" or "deny", printed in black, and not the text in [brackets].

After participants made their decision, in line with Step 1 in the Confounds Protocols, we asked them 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

Beginning with participant award decisions, we replicated the choosing vs rejecting finding. The proportion of participants awarding custody to the enriched parent was higher in the choosing condition (P=68.0%) than in the rejection condition (P=45.2%), $\chi^2(1) = 31.02$, $p < .0001$.[10]

[10] 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

Moving on to the participant explanations. To carry out Step 2 in the Confounds Protocols, we relied on our template prompt and asked an LLM to classify them into six categories. The first and pre-determined category involved the mechanism implied by the choosing vs rejecting hypothesis, which we operationalized for the LLM as follows *"an explanation that mentions that their choice was affected by the frame (i.e., whether they were asked about choosing or rejecting)".* Four additional categories were created by the LLM based on participant responses, plus the customary 6th unclear/ambiguous category.

After we had the 6 categories, we proceeded to Step 3. We began inspecting the categories created by the LLM. As expected, very few explanations (n=8) were classified into the predetermined category with the hypothesized framing effect mechanism. The other four categories, created by the LLM, reflected which parental attributes participants were explaining their decision with (e.g., one category was 'sobriety' and had explanations that emphasized the drinking of alcohol in their justification). None of the four categories corresponds to a confound.

Continuing with Step 3 in the Confounds Protocols, we then sorted explanations by condition, explanation category, and dependent variable, and systematically explored participant explanations at different dependent variable responses within each condition×category bin. We replicated the unexpected finding from the pilot: for many participants in the rejection condition the explanation unambiguously pointed towards giving the child to one parent, but they *rejected* that parent (awarding custody to the other). For example, one participant wrote "Parent A drinks a lot and are loose on rules. It could put the child in a bad situation. I choose parent B", but they *rejected* Parent B, and put the child in the bad situation they feared with Parent A.

---

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 5).

Moving on to Step 4, reassessing the internal validity of the experiment. The data as is lacks internal validity, many participants did not act in the way they had been interpreted to have acted. This can be addressed by a different statistical analysis of the existing data, or a new experimental design. Were we writing a paper about the choosing vs rejecting effect, we would take both approaches. We would formally take confusion into account when analyzing these data, and we would design a new experiment where participants are less likely to reject thinking they are actually choosing (e.g., we could ask participants to place a big red X on top of the parent they deny custody to). But for the purposes of this paper on participant explanations, we just focus on the reanalysis of these data.

To formally take confusion into account, as pre-registered, we instructed an LLM to read participant explanations and infer the parent the participant wanted to give custody to (or indicate it was unclear). When making this inference, the LLM only had access to the explanations, it was blind to the condition and the participant's decision. The first author did the same exercise, also evaluating only the explanations, blind to any other information.

The LLM provided an inferred parental preference for n=517 or 88.1% of all participants, the first author did for n=505 participants or 86% of them. We report results for the author inferences in the main text and for the LLM in footnotes (both were pre-registered). [11]

The inferred preference matched the participant's decision in 70.9% of cases. Figure 8 tabulates the mismatches, referred to as mistakes, by condition. The figure shows mistakes originate almost exclusively in the rejection condition (presumably because it's unusual to be asked

[11] Our pre-registration indicated the first author would only classify explanations where there was a disagreement between the LLM and the participants, expecting the task would be prohibitively taxing, but it wasn't, so we classified all of them.

which parent should be rejected), and disproportionately involve rejecting the enriched parent (presumably because that's the preferred parent in choice). [12]

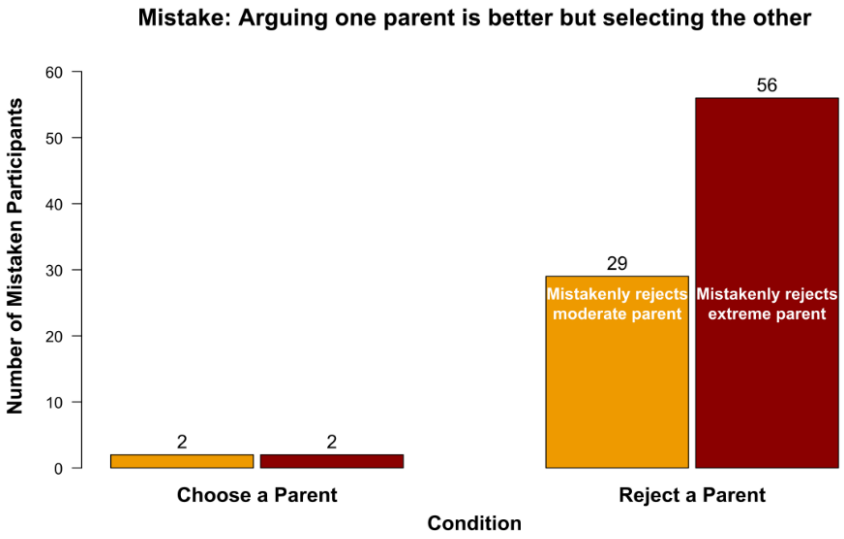

Figure 8. *Mistakes were made (in Example 3).*

As pre-registered, we redid the difference of proportions test for which parent was awarded custody across frames, but setting as the dependent variable the decision implied by the explanation rather than the explicit decision. The difference between choosing and rejection remained significant (p=.001), though it was significantly attenuated (p=.004), dropping from 23 pp to 13 pp (see Figure 9).[13] In our study, then, about half of the choosing vs rejecting effect was artifactual. Lastly, in terms of Step 5, examining via participant explanations the hypothesized mechanism. Examining the n=8 participants the LLM assigned to the "frame" category we found they were all false positives, none actually mentioned the impact of frame on attribute weights.

[12] When relying on the LLM evaluations of the participant explanations we obtained similar numbers: inferred guess matched for 72.4% of participants, in the award condition there are 2 mismatches and in the rejection condition 90.
[13] Our pre-registration did not specify what we would do with participants whose explanations did not clearly identify a preferred parent. In our main analysis we left their decision unchanged (i.e., we only changed participant decisions when it was unambiguous they should be changed). If instead, we drop participants with unclear explanations, the choosing vs rejecting effect attenuates further, to P = 66.5% vs. P = 56.8%, $\chi^2(1)=5.085$, p=.024, using the first author's imputations, and to 67.3% vs. 58.3%, $\chi^2(1)=4.48$, p=.034 using the LLM's.

However, for participants to provide in their explanations the mechanism of interest, they need to engage in counterfactual thinking (how they would have reacted had they been in a different, unknown-to-them condition) and perhaps non-deliberate mechanism (it's ambiguous whether the choosing vs rejecting frame has a conscious or subconscious effect on attribute weight). The hypothesized mechanism, therefore, is expected to produce participant explanations that violate two of the four necessary conditions for participant explanations accuracy. We thus do not think Step 5 is diagnostic/applicable to this study.

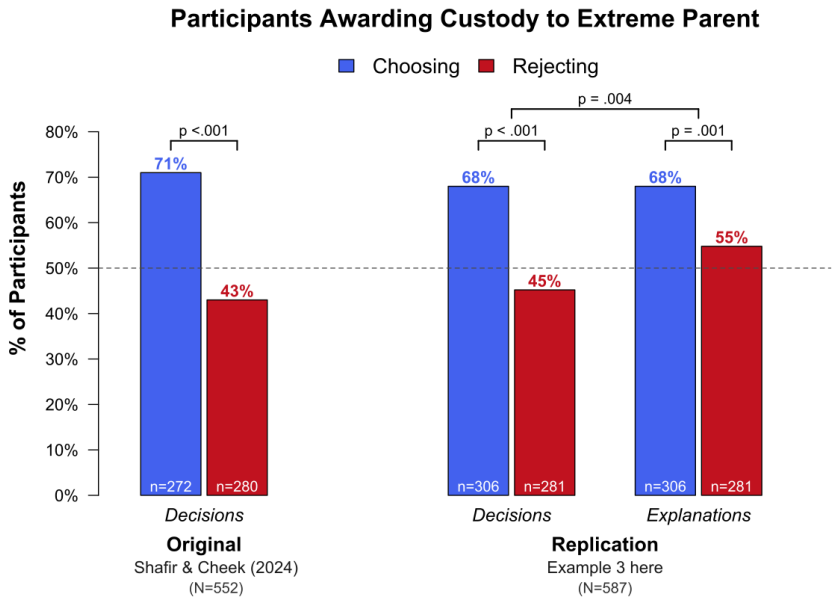

Figure 9. *Parental decisions and explanation-implied decisions in Example 3*
Note: The N=587 participants in the replication are the same across the two sets of bars. The first set shows participants' original decisions. The second set shows decisions inferred from the participants' explanations (left unchanged if no inference was possible). The *p*-values for the blue-red bar pairs are obtained with proportions test; the *p*-values for the interaction are obtained through a linear regression with two observations per participant (one decision-based, one explanation-based), clustering errors by participant. The interaction test was not pre-registered.

In sum, 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 across four published

papers and one unpublished replication; all teams thought very carefully about the design, and yet, the confound was only identified by collecting participant explanations.

Conclusion

For nearly a century, experimental psychologists have largely avoided asking participants why they did what they did in experiments. We believe this has left enormous amounts of information on the table, especially in terms of discovering confounds, and has almost surely slowed the pace of discovery, while prolonging pursuits down wrong paths in psychological research. As promising as participant explanations are for discovering confounds, we cannot tell whether they are capable of detecting all confounds, a large majority of them, or simply very many of them. Within any given experiment, the absence of evidence of confounds in participant explanations should not be interpreted as a guarantee that the study is free of confounds.

In this paper, we presented three examples where, relying on our proposed Confounds Protocols, we were able to discover new first-order facts about the experiments we revisited, despite working within heavily studied and mature research programs. A large amount of information is waiting to be collected in novel research programs, all we have to do is ask.

References

Andre, P. (2025). Shallow meritocracy. *Review of Economic Studies*, *92*(2), 772-807.

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. (in press). Commentary on Xiang et al. (2025): The Gambler's Fallacy Fallacy Fallacy. *Psychological Science*.

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

Boring, E. G. (1929). *A history of experimental psychology*. Century Company.

Braghieri, L., Schwardmann, P., & Tripodi, E. (2024). Talking across the Aisle. *Unpublished manuscript*.

Brown, T. C. (2005). Loss aversion without the endowment effect, and other explanations for the WTA–WTP disparity. *Journal of Economic Behavior & Organization*, *57*(3), 367-379.

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.

Choi, Y., & DeKay, M. L. (2026). A Gambler's Fallacy for Probability Judgments when Event Sequences Are Truly Random: A Reanalysis of Xiang, Dorst, and Gershman's (2025) Data. In *https://osf.io/preprints/psyarxiv/yxgt8_v3*.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological review*, *87*(3), 215.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data (Revised Edition)*. MIT Press. https://doi.org/10.7551/mitpress/5657.001.0001

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

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.