Methods in Psychology 3 (2020) 100032

Contents lists available at ScienceDirect

Methods in Psychology

journal homepage: www.journals.elsevier.com/methods-in-psychology

Reducing the number of non-naïve participants in Mechanical Turk samples

Ethan A. Meyers [*], Alexander C. Walker, Jonathan A. Fugelsang, Derek J. Koehler

Department of Psychology, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, N2L 3G1, Canada

ARTICLE INFO

Keywords:
Mechanical Turk
Nonnaivete
HIT restriction
Data quality

ABSTRACT

Using participants who have been previously exposed to experimental stimuli (referred to as non-naïveté) can reduce effect sizes. The workforce of Amazon's Mechanical Turk is particularly vulnerable to this problem and solutions are usually cost and time inefficient and of mixed effectiveness. In response to this problem and its currently underwhelming solutions, we tested various participant recruitment strategies designed to recruit participants naïve to frequently used experimental stimuli. We collected samples using maximum HIT restrictions (50 for Experiment 1 and 2, 500 for Experiment 2) and TurkPrime's (now CloudResearch) naiveté feature and compared them to samples recruited with standard restrictions (95% HIT approval rating). In these comparisons, we replicated past findings where using nonnaïve (vs. naïve) participants has been shown to reduce effect sizes and affect performance on a variety of tasks (e.g., the Cognitive Reflection Test, a Public Goods Game). We demonstrate that restricting by the maximum number of HITs heavily reduces the number of "experienced" research subjects in samples but necessitates some sacrifice in data quality and collection speed. We discuss the pragmatics of our method, its limitations, and future directions for solving the problem of non-naïveté on Mechanical Turk. For those looking to avoid this issue, we recommend setting a maximum HIT restriction of 50 when recruiting participants.

1. Introduction

Over the last decade, online crowdsourcing platforms, primarily Amazon's Mechanical Turk (MTurk), have transformed how scholars conduct scientific research. The appeal of quick and convenient data collection at a minimal cost has lured ever more researchers to use the internet to recruit participants and run their experiments. This is not to say the transition from slow in-lab studies to rapid online research is without problems. One such problem is the over exposure of participants to frequently used experimental tasks and methodologies, particularly on MTurk.

1.1. Consequences of prior exposure

Prior exposure to tasks, typically referred to as "non-naïveté", has non-trivial consequences for social science research. When completing an experimental task for the second time participants' responses may be wildly discrepant from their initial responses. Importantly, these changes are often beyond random within-subjects error and can lead to significantly different estimates of effect sizes. For example, Chandler et al. (2015), had participants complete a set of classic cognitive and behavioral tasks (e.g., anchoring problems, gain/loss framing) on two separate occasions. In doing so they were able to demonstrate how the influence of various experimental stimuli (e.g., a high or low anchor) may be curtailed by prior exposure. That is, effect sizes at time two (when participants were nonnaïve) were shown to be substantially smaller compared to those at time one, in some instances being reduced to one-half of the original effect size. Similarly, an individual's tendency to cooperate in public goods games reduces over repeated exposure to the game, leading previously found effects to fail to replicate in research conducted later in time (Rand et al., 2012; Rand et al., 2014; Rand, 2018). However, having been previously exposed to experimental tasks does not necessarily dictate that behavior in them will not be consistent across exposures. For example, basic tasks from the areas of perception, memory, and language show no reduction in effect sizes as a result of prior exposure (Zwaan et al., 2018). This is important as it implies that the effect of nonnaïvete varies as a function of what is being measured (consistent with, Chandler et al., 2019; Robinson et al., 2019). For the purposes of this paper we will focus on cognitive and behavioral tasks for which participant nonnaïvete has demonstrated (and in one case, inferred) consequences.

Not surprisingly, various reviews of MTurk as a recruiting platform have listed prior exposure to experimental stimuli as a cause of concern (Chandler and Shapiro, 2016; Hauser et al., 2019; Keith et al., 2017;

[*] Corresponding author.
E-mail address: emeyers@uwaterloo.ca (E.A. Meyers).

https://doi.org/10.1016/j.metip.2020.100032
Received 12 August 2019; Received in revised form 25 June 2020; Accepted 19 August 2020
Available online 21 August 2020

Stewart et al., 2017, but see Robinson et al., 2019 for an opposing view). Furthermore, some researchers have highlighted this oversaturation of frequently-used measures and tasks as reason for social scientists to actively pursue alternatives to MTurk (see Peer et al., 2017). Before discussing current solutions to the problem of non-naïveté on MTurk and similar online crowdsourcing platforms, we will first outline its origins.

1.2. How has prior exposure become a problem?

Subject pools can become exposed to experimental stimuli in two broad ways: Externally and Internally. External exposure includes the discovery of stimuli from sources outside the context of an experiment. This includes learning about stimuli on the internet (e.g., through YouTube videos), discovering stimuli in an educational setting (e.g., undergraduate psychology classes) or even being exposed to stimuli while watching television game shows (e.g., “Split or Steal”). Widely used behavioral tasks typically find themselves being presented in these mediums because of the natural interest humans have in them. These external sources of exposure will be very difficult to eliminate. However, as external exposure typically requires participants to actively remember the “correct” answer to a problem, this form of exposure is unlikely to have much of an effect on task responding. For example, consider the Cognitive Reflection Test (CRT; Frederick, 2005). Since its explosion in popularity within social science research (e.g., Bialek and Pennycook, 2018; Haigh, 2016; Toplak et al., 2011), it has subsequently been widely distributed and discussed online.[1] As this test features problems that cue an incorrect intuitive response, concerns within the scientific community grew that once participants had completed the task previously or had learned the answer elsewhere, they were likely not to get fooled again. Indeed, an initial experimental test of this concern demonstrated that naïve participants (those who had completed very few Human Intelligence Tasks [HITs] on MTurk) scored far worse than experienced participants (those who had completed at least a moderate amount of HITs on MTurk) on the CRT (Chandler et al., 2014). However, recent research contradicts this claim, strongly challenging the notion that prior exposure increases performance on the CRT (Meyer et al., 2018; Stagnaro et al., 2018). Importantly, the public dissemination of the CRT represents one way potential subjects can be exposed to stimuli prior to participating in research and the potential consequences (or lack thereof) exposure may have.

Internal sources of exposure refer to exposure to stimuli within the context of a study. To be specific, someone who has participated in numerous studies has likely been exposed to a host of experimental paradigms, several of which they may have completed multiple times. While the MTurk worker population may appear to be large enough to be able to conduct multiple experiments on, while only sampling any given individual once, this is simply not the case. Specifically, the average laboratory conducting research on MTurk samples only approximately 7,300 unique participants (Stewart et al., 2015; although the data of Difallah et al., 2018, suggest the number is higher). As such, a researcher using the same measures in multiple MTurk studies is likely to be repeatedly exposing some of their participants to these measures. In addition, making this even more likely, is the fact that most of the HITs on MTurk are completed by a very small subset of workers known as “Super Turkers” (Bohannon, 2011; Ford, 2017). This pattern is even more concerning as Super Turkers are likely to have completed a vast number of studies for other labs as well, further increasing the likelihood that they have been exposed to a specific paradigm. This is especially worrisome when using popular paradigms that are widely shared (e.g., a prisoner's dilemma, the CRT). As it has been clearly demonstrated that participants do not necessarily respond in the same manner when given the same task or measure twice (Chandler et al., 2015), internal sources of exposure are very likely to contribute to prior exposure effects. Nevertheless, as the issues of prior exposure and oversaturation of materials within a population pool have amplified in severity, solutions have been increasingly presented.

1.3. Solutions to prior exposure

Recent detailed reviews of MTurk as a sampling population have singled out prior exposure as an issue and forwardly discussed existing and potential solutions to this problem (see Keith et al., 2017). Unfortunately, most of the current solutions offered take place during the stage of data cleaning and not data collection (Deetlefs et al., 2015). These methods, primarily statistical, typically require meticulous data trimming while only offering mixed effectiveness. When using such trimming methods, one may commit far more resources (e.g., time) than intended to solve a problem without being sure that one has actually solved it. Furthermore, remuneration given to nonnaïve participants cannot be refunded, so researchers using these methods end up paying for a significant number of data points that they never intend to analyze. Therefore, a seemingly far more efficient and cost-effective solution would be to have a method for reducing the proportion of nonnaïve responders during the stage of data collection.

One method implemented by Chandler et al. (2014), was to restrict participant recruitment on MTurk by the number of HITs a potential participant has completed. In order to test this method, they launched two identical studies simultaneously, varying only the qualifications required for participation. In the “inexperienced participants” study, they recruited workers who had completed at least one HIT but no more than three. In the “experienced participants” study, workers were required to have completed at least four HITs with no maximum cap. Importantly, they found that the quality of worker data did not vary as a function of prior experience with MTurk. Thus, they provided initial support that restricting by the number of HITs completed may be a way to reduce non-naïveté in an MTurk sample while simultaneously preserving data quality. Their method however, suffered from a few flaws. First, their technique of restricting worker recruitment to those who completed one to three HITs is not possible on Amazon's requester website (the primary website where MTurk studies are set up and launched) or through TurkPrime (now CloudResearch), a popular 3rd party extension for running studies on MTurk (Litman et al., 2017). Therefore, even if a researcher wanted to implement this method, there is currently no easy way to do so. Second, this technique, due to the restrictiveness of its qualification, leads to significantly slower data collection mitigating one of the largest advantages to using the MTurk platform (i.e., rapid data collection). Third and finally, the population who have completed at most three HITs and at least one is likely too small to be able to sample from in the long-term as it would require consistent refreshing of the earliest stages of this participant pool. Thus, even if one was able to implement this method, it is likely too restrictive for its own good.

Discussion of restricting by the maximum number of HITs as a potential solution to non-naïveté has since continued in the literature but its effects remain uncertain.[2] For instance, Hauser et al. (2019), suggested that while researchers may be tempted to restrict by the maximum number of HITs to reduce non-naïveté, the effects on data quality and data collection are relatively unknown. So, while Chandler et al. (2014) highlighted the potential value of a maximum HITs restriction, they were not able to provide the research community with a viable strategy to reduce non-naïveté. Notably this was not the purpose of their work, but it was of ours.

[1] One of the authors of this paper found a YouTube video discussing the test with over 4.5 million views at the time of writing.

[2] During the publication process we became aware of a manuscript by Robinson et al. (2019) wherein they report to have explored the issue of nonnaïveté on MTurk and have tested the data quality and effectiveness of recruiting inexperienced participants by way of a maximum HIT restriction.

1.4. The present work

The aim of this paper was to empirically test the effectiveness of various recruitment criterion at reducing the number of nonnaïve participants in online samples and provide an effective solution to non-naïveté for social scientists using MTurk. Across two experiments we recruited multiple samples all via different recruitment strategies. In Experiment 1 we tested a criterion of limiting recruitment by the maximum number of HITs an individual worker could have completed in order to participate (50) and compared it to standard recruitment practices (i.e., 95% HIT approval rating). We elected to use 50 HITs as the maximum cap as it addresses the three shortcomings of the Chandler et al. (2014) method discussed previously. First, it is both possible and easy to implement on Amazon's interface and TurkPrime. Second, it is unlikely to drastically slow the speed of data collection. Third, it would represent a population that could potentially be sampled from long-term (we address these points further in the general discussion). In Experiment 2, in addition to the recruitment settings above, we also tested a larger maximum cap of 500 HITs and Turk Prime's new naiveté feature. In both of these experiments we implemented a variety of tasks where it has been previously shown that prior exposure to them significantly effects performance on them. When comparing across our recruited naïve and experienced samples, we expected to observe a pattern of results consistent with prior research, such as larger effect sizes in the naïve group (in tasks where effect sizes are compared) and predictable differences in group means (in tasks where the means of the samples are compared).

2. Experiment 1

2.1. Methods

Data and materials for both experiments are available online on the Open Science Framework here: https://osf.io/q94n8/. This experiment was not preregistered.

2.1.1. Participants and design

A sample of 398 participants was recruited from MTurk and received $1.00 upon completion of an 8-minute online questionnaire. In order to compare recruitment strategies, we simultaneously launched two separate HITs ($n = 199$) on MTurk which differed only in exclusion criteria. To create our "experienced" sample participants were recruited under the condition that they be U.S. residents and possess an MTurk HIT approval rating greater than or equal to 95% (henceforth referred to as the standard sample). To create our "naïve" sample, participants were recruited under the condition that they be U.S. residents who had completed up to a maximum of 50 HITs on MTurk. Participants were prevented from participating in both HITs. Using these two recruitment strategies we were able to compare a non-naïveté reduction method (i.e., a maximum HIT restriction) against a widely used method of recruitment (i.e., a 95% HIT approval rating providing us with our standard sample[3]).

2.1.2. Measures/tasks

We selected five measures to assess the differences between our naïve and standard samples. Four of these measures have been used previously to demonstrate the effects of non-naïveté (these include the CRT, the Asian Disease Problem, Anchoring Problems, and a Public Goods Game). We also chose to assess the effects of non-naïveté in the Prisoner's Dilemma, as choices within this paradigm tend to change as participants become more experienced with the game (Grujić et al., 2012). Lastly, we selected an attention check to assess whether data quality differed between our samples. Specifically, we wanted to assess whether the failure rate for our attention check item would be greater in our naïve, as opposed to our standard sample.

2.1.2.1. Cognitive Reflection Test. The CRT (Frederick, 2005) is a test that is used to measure an individual's tendency to reject an intuitive but incorrect response in favor of a correct but deliberative answer. Participants were presented with the original three CRT items created by Frederick (2005). The number of correct responses for each participant was summed, resulting in each participant possessing a CRT score that ranged from zero to three.

2.1.2.2. Asian Disease Problem. The Asian Disease Problem is a popular problem used to test how different representations of the same problem result in different responses (Tversky and Kahneman, 1981). For this task, participants were randomly assigned to receive either a gain or loss framing of a vignette describing a hypothetical disease. They were presented with two possible treatment programs that were framed as either gains (e.g., "If Program A is adopted, 200 people will be saved") or losses (e.g., "If Program A is adopted, 200 people will die") and asked to choose between a certain (e.g., "200 people will be saved") versus a risky option (e.g., "2/3rd chance 600 people will die") with an equivalent expected value of life-loss.

2.1.2.3. Anchoring. We used an anchoring task previously implemented by Chandler and colleagues (2015; originally from Jacowitz and Kahneman, 1995). In this task participants are first asked to guess whether a specific value (e.g., the height of Mount Everest; the average number of babies born per day in the U.S.) is above or below a randomly assigned anchor (i.e., either a high [45,500 feet; 50,000 babies born per day] or low [2,000 feet; 100 babies born per day] anchor). Next, participants are asked to state their exact estimate for a presented question (e.g., estimate the height of Mount Everest). Anchoring refers to peoples' tendency to fail to adjust appropriately from a provided anchor, and, as a consequence, have their estimates biased towards the randomly assigned anchor. Two different anchoring vignettes were used in the current study ("How tall is Mount Everest?" and "On average, how many babies do you think are born per day in the United States?").

2.1.2.4. Public Good's Game. We implemented a one-shot Public Good's Game (PGG; see Rand et al., 2014) where participants were asked to decide how much of their received allotment they would contribute to the public good and how much they would keep for themselves. Participants were instructed that they were playing with three other participants, all of whom would be presented with an identical decision. All of the money that was contributed to the public project was doubled and then evenly distributed to all players, regardless of their contribution to the group. The instructed goal of each player was to maximize their own personal profit which could be maximally achieved by not contributing anything to the public project.

2.1.2.5. Prisoner's dilemma. We used a Prisoner's Dilemma (adapted from Poundstone, 1992) where our participants were instructed that they had committed a crime with an undisclosed partner. Caught by the police, participants were provided with two possible options of which they must choose one: betray or stay silent. They were informed that their partner was also presented an identical decision, and that they would not know their partner's decision at the time of making their own. The normative choice in this dilemma is to betray your partner.

2.1.2.6. Attention check. We used an attention check (taken from Grujic et al., 2012) that provided a short descriptive passage to participants explaining that half of online research participants do not carefully read questions. Participants were asked what the study had them do and were provided with four options "Make various judgments", "Identify Lions", "Look at Tigers" and "Other _". Included in the passage were instructions

[3] Note that this restriction requires all participants in this group to have completed at least 100 HITs.

Table 1
Mean values of all variables of interest and comparison between our findings and past findings across all of our measures.

Tasks	Naïve (<50 HITs)	Standard (95% HIT Approval)	Our Finding	Past Finding	Replicated?
Anchoring: Births	77,133	18,227	Greater anchoring in naïve sample	Greater anchoring in naïve sample	Yes
Anchoring: Mount Everest	21,909	21,224	No difference in anchoring	Greater anchoring in naïve sample	No
Cognitive Reflection Test	1.10 (*1.13*)	1.96 (*1.18*)	More correct answers for standard sample	More correct answers for standard sample	Yes
Gain/Loss Framing	70% vs. 30%	60% vs. 41%	Larger framing effect in naïve sample	Larger framing effect in naïve sample	Yes
Public Goods Game	8.12 (*4.27*)	6.58 (*4.76*)	Naïve participants donated more	Naïve participants donated more	Yes
Prisoner's Dilemma	19.8%	24.9%	Equal proportion of betrayals	No past finding	–
Attention Check - Pass	89%	95%	–	–	–
Completion Time	577 (*309*)	399 (*306*)	–	–	–
Prior Exposure	1.89 (*2.32*)	3.36 (*1.59*)	–	–	–

to select the final option and enter “Decision Making” into the blank space.

2.1.3. Procedure

Participants in both the naïve and standard samples completed an identical online questionnaire involving the tasks described above in a random order. Prior to exiting the study, participants responded to various demographic questions and indicated which of the presented tasks (if any) they had previously seen or completed.

2.2. Predictions

We predicted that larger effects (when effect sizes were to be compared) and discrepant group means (when group means were to be compared) would be observed in the naïve sample relative to the standard sample. In particular, we predicted that the discrepancy between high- and low-anchored estimates would be larger for the naïve sample compared to the standard sample (larger anchoring effect). Similarly, we predicted that participants in the naïve sample would more often choose the certain option when it is framed as a gain and the risky option when it is framed as a loss (larger framing effect). In addition, we hypothesized that the naïve sample would score lower on the CRT and donate more to the public works project compared to the standard sample (discrepant group means). We also expected proportionately fewer betrayals in the naïve sample. Finally, we predicted that the attention check failure rate would be greater for the naïve sample.

2.3. Results

Prior to analysis, we excluded all participants who failed the attention check measure. We excluded 22 participants in the naïve sample (new $n = 177$) and 10 in the standard sample (new $n = 189$). The naïve sample demonstrated a significantly higher failure rate compared to the standard sample, $\chi^2(1) = 4.89, p = .027$. Importantly, the passing rate of the naïve sample for this attention check was 88%, suggesting that the quality of data for these participants may be comparable to the quality of a standard sample.

Where predicted, one-sided tests were used to maximize power. As a manipulation check, participants in the standard sample reported having seen and/or completed more measures prior to participating in this study compared to participants in the naïve sample, $t(364) = 7.12$, $p < .001$, $d = 0.74$, 95 CI [0.56, 0.92]. Standard sample participants were significantly older ($M_{\text{age}} = 39$) than naïve sample participants ($M_{\text{age}} = 32$) on average, $t(364) = 5.73, p < .001, d = 0.60$ [0.39, 0.81]. No differences in annual household income, ethnicity, education or gender across our samples were found. Participants in the naïve sample took significantly longer to complete the study ($M = 577$ s) relative to participants in the standard sample ($M = 399$ s), $t = 5.52, p < .001, d = 0.58$ [0.37, 0.79]. Finally, while the studies were launched simultaneously, the standard sample took two and a half hours to collect while the naïve sample was collected in just under 10 hours.

Four out of the five behavioral measures revealed results consistent with past research (see Table 1). Experience showed a large effect on the CRT as the naïve sample performed significantly worse ($M = 1.10$) compared to the standard sample ($M = 1.96$), $t(250) = 5.87$, $p < .001$, $d = 0.74$ [0.52, 0.96]. Notably, the difference between these two samples was almost an entire correct question. As the test features only three questions a discrepancy of this amount suggests experience may play a large role in performance on this task. When deciding how much to contribute to a public works project, naïve sample participants donated ($M = \$8.12$) significantly more to the project relative to the standard sample participants ($M = \$6.58$), $t(286) = 4.17$, p < .001, $d = 0.49$ [0.26, 0.76].[4] When choosing a solution to an upcoming Asian disease, both samples experienced a framing effect where they preferred the guaranteed option when it was framed as a gain while tending to choose the risky option when it was framed as a loss. Importantly, the naïve sample demonstrated a larger framing effect (70% choosing sure option when framed as gain vs. 30% when framed as loss) compared to the standard sample (60% vs. 41%), $Z = 2.57$, $p = .010$.

Prior to analyzing anchoring responses, we removed two subjects for giving impossibly large guesses (one said the height of Mount Everest was 25,000,000 feet, while the other estimated 1,000,000 babies were born each day in the US). We then standardized the scores and removed outliers who had guesses beyond three standard deviations of the mean on either guess (14 subjects).[5] When guessing how many babies are born in the US each day on average, both samples anchored their guesses. Importantly, the naïve sample demonstrated a larger anchoring effect ($d = 0.61$ [0.31, 0.94]) compared to the standard sample ($d = 0.21$ [0.09, 0.50]; $F(1, 345) = 6.12$, $p = .011$). However, this was not the case when estimating the height of mount Everest as both samples anchored to a very extreme but not significantly different amount (both d's > 1.5), $F(1, 343) < 1$.

When deciding whether to betray their partner in crime in the Prisoner's Dilemma, the naïve sample nominally betrayed less (35 compared to 47 betrayals), but not proportionally less as predicted compared to the standard sample, $\chi^2(1) = 1.36$, $p = .243$. While we speculated that experienced participants might betray more given that decisions in these dilemmas can change over repeated trials, we are not aware of any research that demonstrates experienced participants betraying more than naïve participants in a one-shot version of the game (as we have implemented here).

[4] These results are reported further excluding participants who wanted to donate more than \$20, leaving the naïve sample with $n = 138$, and the experienced sample with $n = 150$.

[5] These results are nearly identical if outliers are removed from their specific analyses.

2.4. Discussion

The results of Experiment 1 show that when comparing our recruitment strategy to standard practice we replicate various prior exposure findings. We found the predictably larger effect sizes (e.g., larger anchoring and framing effects) and discrepant group means (e.g., fewer CRT questions correct, greater PGG donations) that one would expect to find if comparing a naïve sample vs. a nonnaïve sample. We also found that while the naïve sample did fail the attention check more often, the passing rate for that sample was still quite high (88%). If this metric is to be considered a marker for data quality, this result suggests that less experienced workers do provide reputable data (consistent with the findings of Chandler et al., 2014). Finally, our recruitment method undoubtedly slowed data collection, but the delay was a matter of *hours,* not days or weeks.

A criticism of the recruitment strategy used in Experiment 1 is that 50 HITs could be too restrictive of a maximum cap (see General Discussion for the addressing of this point). As such, we looked to assess the viability of other recruitment strategies to reduce non-naïveté and compare them to those tested in Experiment 1. Specifically, we implemented a more inclusive maximum cap of 500 HITs completed in addition to another alternative recruitment strategy: the naiveté feature recently introduced on TurkPrime. For an additional $0.25 per participant this feature can restrict up to the top 10% most active workers on the website from participating in an experiment with the goal of preventing "Super Turkers" from being potential subjects. Here, along with the two conditions used in Experiment 1, we tested the effectiveness of these recruitment strategies at reducing the number of nonnaïve participants in a MTurk sample.

3. Experiment 2

3.1. Methods

Data and materials are available online on the Open Science Framework here: https://osf.io/q94n8/. This experiment was preregistered, and the preregistration is available on the OSF.

3.1.1. Participants and design

A collective sample of 1082 participants was recruited from MTurk and participants received $1.00 upon completion of a 10-minute online questionnaire. In order to compare recruitment strategies, we simultaneously launched four separate HITs on MTurk (each $n \sim 250$) which differed only in exclusion criteria. In our standard sample, participants were recruited under the condition that they be U.S. residents and possess a Mechanical Turk HIT approval rating greater than or equal to 95%. In our "naiveté feature" sample, we implemented Turk Prime's naiveté feature which filters out (up to) the top 10% of workers by HITs completed on MTurk. That is, the feature excludes up to the top 10% most active workers on the website, depending on the value set by the researcher. While the intention of this feature is to reduce non-naïveté, the data reported below suggest it is similar to the standard sample. As such, throughout the results section the naiveté feature sample will be grouped with the standard sample under the label of "experienced" group for purposes of broader comparisons. In our "naïve" samples, participants were recruited under the condition that they be U.S. residents who had completed up to a maximum of 500 HITs on MTurk in one sample (labelled the 500 HITs sample) and 50 HITs in another (labelled the 50 HITs sample).

3.1.2. Measures/tasks

We selected eight measures to assess differences between the "naïve" and "experienced" samples. Four of these measures were kept from the previous study (the CRT, PGG, ADP, and Anchoring). Three new measures, the Retrospective Gambling Fallacy, the Allow/Forbid Task, and a Quote Attribution task were directly imported from Chandler et al. (2015). Finally, a conjunction fallacy problem was included as it has been speculated to be robust to prior exposure effects (Hauser et al., 2019). Additionally, our attention check differed from that featured in Experiment 1. For the exact materials used visit the supplement available on the Open Science Framework.

3.1.2.1. Retrospective Gambling Fallacy. In this task (taken from Oppenheimer and Monin, 2009), participants were asked to imagine a man in a casino rolling dice. In one condition, the man rolls 3 sixes as they are walking by, whereas in the other condition, he rolls 2 sixes and a three. Participants then estimated how many times the man had rolled the dice before they had walked by.

3.1.2.2. Quote attribution. We implemented the quote attribution task of Chandler et al. (2015) which was a conceptual replication of a task designed by Lorge and Curtiss (1936). Here, participants were shown a quote ("I have sworn to only live free, even if I find bitter the taste of death") attributed to either George Washington or Osama bin Laden, depending on condition. Participants indicated the extent to which they agreed with the quote (1 = *strongly agree*, 9 = *strongly disagree*).

3.1.2.3. Allow/forbid. In the allow/forbid task (Rugg, 1941) participants were asked to indicate whether the U.S. should allow speeches against democracy (Allow Condition), or if the U.S. should forbid speeches against democracy (Forbid Condition). Participants responded with either "Yes" or "No". As the statements are logical compliments, the proportion of "Yes" responses in the Allow condition should be approximately equal to the proportion of "No" responses in the Forbid Condition.

3.1.2.4. Conjunction fallacy. Participants completed the classic Linda problem (Tversky and Kahneman, 1983) where they were provided with a personality description of Linda and asked, "which is more probable: that she is a bank teller, or that she is a bank teller and is active in the feminist movement?"

3.1.2.5. Attention check. We implemented a two-question attention check taken from Peer et al. (2014). Here, participants were asked whether they "would prefer to live in a warm city rather than a cold city" and whether they "would prefer to live in a city with many parks, even if the cost of living was higher." Both questions were answered on a 1–7 Likert scale. However, participants were instructed to not provide their actual preferences but to select "2" for the first question and then add 3 to that value and use that result to answer to the second question. Answers that deviated from 2 to 5 on each respective question were deemed to have failed the check.

3.1.3. Procedure

Participants in all samples completed an identical online questionnaire involving various one-shot tasks used widely in behavioral science research. Participants completed the tasks in a randomized order. Lastly, prior to exiting the study, participants responded to various demographic questions and indicated which of the presented tasks (if any) they had previously seen or completed.

3.2. Results

Prior to analysis, we excluded all participants who failed the attention check measure. We excluded 110 participants in the 50 HITs sample (new $n = 146$), 106 participants in the 500 HITs sample (new $n = 145$), 84 participants in the naïveté sample (new $n = 168$), and 47 in the standard practice sample (new $n = 203$). Consistent with our prediction, the more experienced participants were, the more they passed the attention check, $\chi^2(3) = 42.24$, $p < .001$. Unlike in Experiment 1, double the number of participants failed to pass the attention check in the naïve samples compared to the standard sample. However, these passing rates were fairly consistent with the data from Peer et al. (2014). Our standard practice sample had an identical passing rate compared to their "high productivity" group (500+ HITs completed, minimum 95% HIT approval rating), but, our HIT restriction samples were below their "low productivity" group (100 or fewer HITs completed, 95% HIT approval rating).

As a manipulation check, more experienced participants reported having seen and/or completed more of our items prior to participating in this study compared to more naïve participants, $F(1, 660) = 28.13$, $p < .001$. More experienced participants tended to be older than less experienced participants, $F(3, 656) = 6.84$, $p < .001$, and feature proportionally more males, $\chi^2(6) = 12.64, p = .049$. No differences in annual household income, ethnicity, or education across the samples were found. Participants in the 50 HITs sample took significantly longer to complete the study ($M = 679$ s) relative to participants in other samples ($M = 450$ s across the three samples), $t(658) = 4.36$, $p < .001$. Finally, while the studies were launched simultaneously, the experienced sample and the naiveté feature sample took around 1 h and 40 min to collect, the 500 HITs sample took two and a half hours to collect, and the 50 HITs sample was collected in 16 and a half hours.

Table 2 below presents the means on all variables of interest across each sample. Experience had a large effect on the CRT as the "naïve" samples performed significantly worse compared to "experienced" samples, $F(3, 658) = 25.78, p < .001, n^2 = .11$. Similar to Experiment 1, the difference between the two HITs restriction samples and the standard practices sample was almost an entire correct question. When deciding how much to contribute to a public works project, naïve participants contributed significantly more to the project relative to experienced participants, $F(3, 534) = 9.54$, p $< .001$, $n^2 = .05$.[6] When choosing solutions to an Asian Disease problem, all samples showed a framing effect where they preferred the guaranteed option when it was framed as a gain while tending to choose the risky option when it was framed as a loss. However, no sample showed a larger or smaller framing effect compared to the rest (all p's $> .663$).

When guessing how many babies are born in the US each day on average, no sample anchored their guesses more than another, $F(3, 641) < 1$. This was also the case when estimating the height of Mount Everest as all samples anchored to an extreme, but not significantly different, amount (all d's > 1.3), $F(1, 641) < 1$.

When judging the likelihood of the group Linda belongs to, more naïve participants committed the conjunction error than experienced participants, $\chi^2(3) = 9.56$, $p = .023$. Notably, this effect has previously been discussed as one that should be robust to the effect of prior exposure. However, we find evidence that naïve participants do succumb more to the conjunction fallacy. When guessing how many times a man had rolled the set of dice before the final result (two sixes and a three versus three sixes) the HITs restriction samples showed a greater difference between their guesses (26 and 24 times for the 50 HITs and 500 HITs respectively), than the more experienced samples (11 and 5 for the naiveté feature and standard practices respectively) but it was not significant, $F(3, 652) < 1$, $p = .441$.

When deciding the permissibility of the US allowing or forbidding speeches against democracy, all samples had a greater portion of participants saying that they should not forbid it than allow it. However, this proportion did not vary across samples, $B = -0.08$, $SE = 0.11$, $p = .471$. When rating their agreement with a quote attributed to either George Washington or Osama bin Laden, all participants rated the quote as more agreeable when it was attributed to George Washington, but this did not differ across samples, $F(3, 652) = 1.02$, $p = .383$.

Table 2
Mean values of all variables of interest across all recruitment conditions.

Tasks	50 HITs	500 HITs	Naïveté Feature	Standard Practice
Allow/Forbid	86% vs. 93%	79% vs. 94%	82% vs. 94%	79% vs. 93%
Anchoring: Births	4,744	33,143	26,495	50,326
Anchoring: Mount Everest	21,698	20,751	19,967	22,979
Cognitive Reflection Test	1.06 (*1.12*)	1.08 (*1.14*)	1.39 (*1.20*)	2.00 (*1.19*)
Conjunction Fallacy	75%	79%	88%	83%
Gain/Loss Framing	69% vs. 39%	69% vs. 33%	80% vs. 32%	82% vs. 42%
Public Goods Game	9.20 (*4.29*)	8.68 (*4.03*)	8.33 (*4.52*)	6.57 (*4.71*)
Quote Attribution	1.79	1.33	2.05	1.97
Retrospective Gambler's Fallacy	25.70	23.61	10.88	4.44
Completion Time	679 (*553*)	486 (*222*)	458 (*203*)	422 (*238*)
Prior Exposure	2.13 (*2.80*)	3.06 (*3.11*)	2.60 (*1.91*)	3.71 (*1.87*)
Attention Check - Pass	57%	58%	67%	81%

Notes: The Allow/Forbid values are the proportion who say they should allow speeches against democracy versus the proportion who say they should not forbid speeches against democracy. Anchoring values are the mean difference scores taken from subtracting the average guess in the Low anchoring condition from the average guess in the High anchoring condition. Attention check values are the failure rate. Cognitive Reflection Test is scored on correctness out of three items. The Conjunction Fallacy values represent the proportion of the sample selecting that it is more likely that Linda is both a Bank Teller and active in the feminist movement. Gain/Loss framing values are the percentage of each group electing the certain option in the gain frame versus the certain option in the loss frame. The value displayed for the Public Goods Game is the average sum (out of $20) contributed to the public works project. Quote attribution values are the difference between the mean profoundness ratings of the quote said by George Washington minus the mean profoundness ratings of the quote said by Osama bin Laden. The values for the Retrospective Gambler's Fallacy are the differences between the mean guesses for how many times three sixes were rolled versus two sixes and a two. Completion time is reported in seconds. Prior exposure is the number of questions they completed in the study that they remembered having seen or completed prior to this study. Relevant (*standard deviations*) are included.

3.3. Discussion

Experiment 2 tested the effectiveness of three recruitment strategies (setting a maximum HITS restriction of 50, a max of 500, and using TurkPrime's naiveté feature) at reducing non-naïveté from MTurk samples by comparing their data to a sample collected using standard recruitment practices (95% HIT approval rating). The results support four conclusions. First, that the maximum 50 HIT restriction is perhaps only slightly more effective at reducing non-naïveté than the 500 HIT restriction. As is evident in Table 2, the 500 HIT restriction mirrors the 50 HIT restriction in many tasks and the few discernible differences between the two samples (e.g., greater self-reported prior exposure in the 500 HIT restriction sample) slightly favor the 50 HIT restriction. Second, the 500 HIT restriction appears to be a much more efficient recruitment strategy compared to the 50 HIT restriction. This is because the two maximum HIT restrictions have highly similar patterns of data, while the 500 HIT restriction boasts a much faster recruitment time (but again, this is only a matter of *hours*). Third, TurkPrime's naiveté feature only modestly reduced non-naïveté. The sampling method can be most accurately depicted as having recruited participants that were somewhere in-between standard practices and our maximum HIT restriction samples in terms of naiveté. That is, on some measures the data were identical to standard practices, on others the data were similar to the naïve samples.

[6] Participants who wanted to donate more than $20 to the public works project were excluded from these analyses, leaving the 50 HIT sample with $n = 116$, the 500 HIT sample with $n = 123$, the naiveté feature sample with $n = 140$, and the standard practice sample with $n = 159$.

Fourth, MTurk experience can be treated as a viable proxy for participant non-naïveté. Restrictions that recruited more experienced participants led to responses consistent with non-naïveté as defined by prior research.

Relative to Experiment 1, the attention check passing rates of this experiment were comparatively much worse. We take this in part to reflect the increased difficulty of the implemented attention check, which in this experiment spanned across two questions instead of just one. Attention checks can vary in how difficult they are to pass (Oppenheimer et al., 2009), and this difficulty generally reflects discriminating between workers who are of a reasonable quality. Importantly, the failure rates observed in these samples were not abnormal compared to attention- and comprehension-based filtering in experimental work (e.g., Chandler et al., 2014; Meyer et al., 2018). However, we cannot dismiss the difference in failure rates between the standard sample and the HIT restriction samples. Despite the difficulty difference across studies, the results of Experiment 2 would suggest that employing a maximum HIT restriction will lead to a decrease in data quality compared to implementing standard practices.

4. General discussion

Before this research, scientists conducting studies on MTurk who worried about prior exposure to their stimuli had three choices: find another platform to sample from, change their experiment's materials, or attempt to tackle any potentially unwanted effects of prior exposure post-data collection. Here we present evidence for a fourth, and arguably more desirable choice: a maximum HIT restriction. We implemented this recruitment strategy and demonstrated how it can protect against prior-exposure effects across a variety of measures. In cognitive and behavioral tasks where the effect sizes of our samples were compared (i.e., Gain/loss framing, Anchoring effects), our naïve samples (i.e., 50 and 500 HIT restriction samples) tended to show significantly larger effect sizes. In tasks where these groups were directly compared (i.e., the CRT, Public Goods game) the expected trends emerged. That is, more naïve participants answered fewer CRT items correctly and made greater contributions in a Public Goods game. However, our tested solutions are not without cost. While data quality in Experiment 1 appeared comparative across samples, Experiment 2, a likely more accurate test, showed that data quality is reduced under a maximum HIT restriction. Collectively, our results suggest that setting a maximum HIT restriction is an effective solution to the issue of non-naïveté, but researchers should be mindful of the potential for a significant reduction in data quality.

One of the most attractive features of recruiting via MTurk is its cost effectiveness. Its cheap labor force allows for large samples to be collected at minimal expense relative to other sampling platforms (e.g., Survey Monkey, Prolific Academic) and classic survey methods (e.g., paying in-lab participants). When there is a concern of non-naïveté, researchers who want to keep costs low are faced with several potentially less-desirable options. First, they can use a more expensive platform in hopes that participants will have lower rates of prior exposure. Second, they can recruit more participants than necessary and attempt to filter out those who have been previously exposed to the experimental paradigm post data collection. Finally, they can use the naiveté feature on TurkPrime, a previously untested method that according to our data leads to only a modest reduction of nonnaïve sampling at an extra cost. However, setting a maximum HIT restriction allows researchers to continue using MTurk for a low cost as it largely filters at the recruitment stages of data collection, rather than after. Using this method, researchers mostly pay for data that they intend to use, not data they intend to filter out. However, as data quality is likely to be reduced, researchers may still find themselves filtering out low-quality participants post-collection. While the expressed benefit of pre-collecting filtering may not directly manifest itself within the context of one study, multiple experiments or lines of research using our strategy can certainly reap large, long-term financial benefits.

A further advantage of our method is its effectiveness. While other methods (e.g., trimming) may inappropriately remove naïve participants who are thought to be experienced, or fail to remove some portion of experienced participants, our method nearly guarantees that naïve and only naïve participants will be sampled. In order for nonnaïve workers to pass our filter, "experienced" workers would have to create new and/or alternative worker accounts (something that goes against Amazon's Participants agreement and is incredibly difficult to do). Further, they must have only completed a maximum of 50 (or 500) HITs on their alternative account. The hurdle (and violation of policy) for a worker to maintain a naïve account is high, and there is little motivation for an experienced worker with a high-ranking (e.g., a Masters qualification) to do such a thing. Therefore, researchers looking to eliminate experienced workers from their MTurk sample should feel confident when using our recommended filter.

Another seductive feature of MTurk is how quickly data can be collected. To put it into perspective, each "experienced" sample (a minimum of 200 participants in every sample) was recruited in less than 3 h. Any solution designed to work at the recruitment stages of sampling needs to account for this benefit. Launched at the same time, our strictest method recruited the same number of participants in under 10 h for the first study and 16 h for the second. While data collection did take longer (as expected), it is unlikely that this increase in *hours* (as compared to days, weeks or months) is a deterrent to researchers looking to reduce the effects of prior exposure in their MTurk samples. However, researchers using MTurk to recruit unique populations (e.g., individuals with past illicit substance use) may in fact find this method prohibitive. In these cases, participant recruitment is already greatly extended (e.g., a matter of days or weeks rather than hours). Combining our recommended HIT restriction with other restrictions in sampling may increase recruitment time such that it becomes prohibitively long.[7] In sum, the increase in time to recruit appears to be negligible for researchers collecting data from the general population, however, the feasibility of this method for those recruiting hard-to-obtain samples is unknown and may act as a deterrent.

It could be argued that the first cap we set of 50 HITs completed is too restrictive. In particular, one could argue that if enough researchers adopt this recruitment method it may drain the pool of participants faster than it can be replenished. While we do agree this could occur, we think it is unlikely for two reasons. First, it has been shown that an average laboratory conducting research on MTurk receives approximately 1,900 new workers every three months (Stewart et al., 2015) and that this number may be a very conservative estimate (Robinson et al., 2019). Therefore, the likelihood of the average laboratory running out of naïve workers appears to be low. Second, while the current study primarily investigated a recruitment method that utilized a 50 HIT cap, we also tested two other recruitment restrictions. In general, it appeared that the cap of 500 has most of the advantages of restricting by 50 HITs (as most of the data are similar), while avoiding some of the drawbacks (e.g., the time it took to recruit the full sample was within an hour of collecting the standard practices sample). Thus, researchers may be justified in loosening the tight restriction from 50 HITs completed to 500. Further, we also assessed the naiveté feature recently introduced in TurkPrime. We found that this feature failed to produce data that consistently discriminated from the standard practices sample. So, while the intention may be that it reduces the most active workers, it appears that based on our metrics, it only modestly reduced non-naïveté.

As social science research continues its increased use of online-sampling methods, concerns of over-exposing subjects to frequently used paradigms will be ever-present. To counteract this problem, we recommend setting a maximum HIT restriction when recruiting participants. We suggest that researchers could set this limit to 500, but those who are willing to make maximal trade-offs to ensure they are collecting a naïve sample should set this limit to 50.

[7] We thank an anonymous reviewer for this suggestion.

CRediT author statement

Ethan A Meyers: Conceptualization, Methodology, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Visualization, Project administration. **Alexander C. Walker:** Conceptualization, Methodology, Writing – Original Draft, Writing – Review & Editing. **Jonathan A. Fugelsang:** Methodology, Writing – Review & Editing, Supervision, Funding acquisition. **Derek Koehler:** Methodology, Writing – Review & Editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank the members of the Reasoning and Decision-Making Lab at the University of Waterloo for their helpful discussions pertaining to this project. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

References

Bialek, M., Pennycook, G., 2018. The cognitive reflection test is robust to multiple exposures. Behav. Res. Methods 50 (5), 1953–1959. https://doi.org/10.3758/s13428-017-0963-x.

Bohannon, J., 2011. Social science for pennies. Science 334. https://doi.org/10.1126/science.334.6054.307.

Chandler, J., Mueller, P., Paolacci, G., 2014. Non-naïveté among amazon mechanical turk workers: consequences and solutions for behavioral researchers. Behav. Res. Methods 46 (1), 112–130. https://doi.org/10.3758/s13428-013-0365-7.

Chandler, J., Paolacci, G., Peer, E., Mueller, P., Ratliff, K.A., 2015. Using nonnaive participants can reduce effect sizes. Psychol. Sci. 26 (7), 1131–1139. https://doi.org/10.1177/0956797615585115.

Chandler, J., Rosenzweig, C., Moss, A.J., Robinson, J., Litman, L., 2019. Online panels in social science research: expanding sampling methods beyond mechanical turk. Behav. Res. Methods 51, 2022–2038.

Chandler, J., Shapiro, D., 2016. Conducting clinical research using crowdsourced convenience samples. Annu. Rev. Clin. Psychol. 12, 53–81. https://doi.org/10.1146/annurev-clinpsy-021815-093623.

Difallah, D., Filatova, E., Ipeirotis, P., 2018. Demographics and dynamics of mechanical turk workers. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 135–143.

Deetlefs, A.M.J., Chylinski, M., Ortmann, A., 2015. MTurk 'Unscrubbed': Exploring the Good, the 'super', and the Unreliable on Amazon's Mechanical Turk, pp. 1–44. https://doi.org/10.2139/ssrn.2654056.

Frederick, S., 2005. Cognitive reflection and decision making. J. Econ. Perspect. 19 (4), 25–42. https://doi.org/10.1257/089533005775196732.

Ford, J.B., 2017. Amazon's mechanical turk: a comment. J. Advert. 3367 https://doi.org/10.1080/00913367.2016.1277380.

Grujić, J., Eke, B., Cabrales, A., Cuesta, J.A., Sánchez, A., 2012. Three is a crowd in iterated prisoner's dilemmas: experimental evidence on reciprocal behavior. Sci. Rep. 2, 638.

Haigh, M., 2016. Has the standard cognitive reflection test become a victim of its own success? Adv. Cognit. Psychol. 12 (3), 145–149.

Hauser, D., Paolacci, G., Chandler, J., 2019. Common concerns with mturk as a participant pool: evidence and solutions. In: Kardes, F.R., Herr, P.M., Schwarz, N. (Eds.), Handbook of Research Methods in Consumer Psychology. https://doi.org/10.4324/9781351137713-17.

Jacowitz, K.E., Kahneman, D., 1995. Measures of anchoring in estimation tasks. Pers. Soc. Psychol. Bull. 21 (11), 1161–1166. https://doi.org/10.1177/01461672952111004.

Keith, M.G., Tay, L., Harms, P.D., 2017. Systems perspective of amazon mechanical turk for organizational research: review and recommendations. Front. Psychol. 8 (AUG) https://doi.org/10.3389/fpsyg.2017.01359.

Litman, L., Robinson, J., Abberbock, T., 2017. TurkPrime.com: a versatile crowdsourcing data acquisition platform for the behavioral sciences. Behav. Res. Methods 433–442. https://doi.org/10.3758/s13428-016-0727-z.

Lorge, I., Curtiss, C.C., 1936. Prestige, suggestion, and attitudes. J. Soc. Psychol. 7 (4), 386–402. https://doi.org/10.1080/00224545.1936.9919891.

Meyer, A., Zhou, E., Frederick, S., 2018. The non-effects of repeated exposure to the cognitive reflection test. Judgm. Decis. Mak. 13 (3), 246–259.

Oppenheimer, D.M., Monin, B., 2009. The retrospective gambler's fallacy: unlikely events, constructing the past, and multiple universes. Judgm. Decis. Mak. 4 (5), 326–334.

Oppenheimer, D.M., Meyvis, T., Davidenko, N., 2009. Instructional manipulation checks: detecting satisficing to increase statistical power. J. Exp. Soc. Psychol. 45 (4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009.

Peer, E., Brandimarte, L., Samat, S., Acquisti, A., 2017. Beyond the turk: alternative platforms for crowdsourcing behavioral research. J. Exp. Soc. Psychol. 70, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006.

Peer, E., Vosgerau, J., Acquisti, A., 2014. Reputation as a sufficient condition for data quality on amazon mechanical turk. Behav. Res. Methods 1023–1031. https://doi.org/10.3758/s13428-013-0434-y.

Poundstone, W., 1992. Prisoner's Dilemma. Doubleday, New York, NY, USA.

Rand, D.G., 2018. Non-naïvety may reduce the effect of intuition manipulations. Nature Human Behaviour 2 (9), 602. https://doi.org/10.1038/s41562-018-0404-6.

Rand, D.G., Greene, J.D., Nowak, M.A., 2012. Spontaneous giving and calculated greed. Nature 489 (7416), 427–430. https://doi.org/10.1038/nature11467.

Rand, D.G., Peysakhovich, A., Kraft-Todd, G.T., Newman, G.E., Wurzbacher, O., Nowak, M.A., Greene, J.D., 2014. Social heuristics shape intuitive cooperation. Nat. Commun. 5, 1–12. https://doi.org/10.1038/ncomms4677.

Robinson, J., Rosenzweig, C., Moss, A.J., Litman, L., 2019. Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. PloS One 14 (12), e0226394. https://doi.org/10.1371/journal.pone.0226394.

Rugg, D., 1941. Experiments in wording questions: II. Publ. Opin. Q. 5, 91–92. https://doi.org/10.1086/265467.

Stagnaro, M.N., Pennycook, G., Rand, D.G., 2018. Performance on the cognitive reflection test is stable across time. Judgm. Decis. Mak. 13 (3), 260–267.

Stewart, N., Chandler, J., Paolacci, G., 2017. Crowdsourcing samples in cognitive science. Trends Cognit. Sci. 21 (10), 736–748. https://doi.org/10.1016/j.tics.2017.06.007.

Stewart, N., Ungemach, C., Harris, A.J.L., Bartels, D.M., Paolacci, G., Chandler, J., 2015. The average laboratory samples a population of 7, 300 Amazon Mechanical Turk workers. Judgm. Decis. Mak. 10 (5), 479–491.

Toplak, M.E., West, R.F., Stanovich, K.E., 2011. The Cognitive Reflection Test as a Predictor of Performance on Heuristics-And-Biases Tasks, pp. 1275–1289. https://doi.org/10.3758/s13421-011-0104-1.

Tversky, A., Kahneman, D., 1981. The framing of decisions and the psychology of choice. Science 211 (4481), 453–458. https://doi.org/10.1126/science.7455683.

Tversky, A., Kahneman, D., 1983. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. Psychol. Rev. 90 (4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293.

Zwaan, R.A., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., Zeelenberg, R., 2018. Participant nonnaiveté and the reproducibility of cognitive psychology. Psychon. Bull. Rev. 25, 1968–1972. https://doi.org/10.3758/s13423-017-1348-y.