Psychon Bull Rev (2018) 25:1968–1972
DOI 10.3758/s13423-017-1348-y

BRIEF REPORT

Participant Nonnaiveté and the reproducibility of cognitive psychology

Rolf A. Zwaan[1] · **Diane Pecher**[1] · **Gabriele Paolacci**[2] · **Samantha Bouwmeester**[1] · **Peter Verkoeijen**[1,3] · **Katinka Dijkstra**[1] · **René Zeelenberg**[1]

Published online: 25 July 2017

Abstract Many argue that there is a reproducibility crisis in psychology. We investigated nine well-known effects from the cognitive psychology literature—three each from the domains of perception/action, memory, and language, respectively—and found that they are highly reproducible. Not only can they be reproduced in online environments, but they also can be reproduced with nonnaïve participants with no reduction of effect size. Apparently, some cognitive tasks are so constraining that they encapsulate behavior from external influences, such as testing situation and prior recent experience with the experiment to yield highly robust effects.

Keywords Replication · Reproducibility · Perception · Memory · Language

A hallmark of science is reproducibility. A finding is promoted from anecdote to scientific evidence if it can be reproduced (Lykken, 1968; Popper, 1959). There is growing awareness that problems exist with reproducibility in psychology. A recent estimate is that fewer than half of the findings in cognitive and social psychology are reproducible (Open Science Collaboration, 2015). In addition, there have been several been high-profile, preregistered, multi-lab failures to replicate well-known effects psychology (Eerland et al., 2016; Hagger et al., 2016; Wagenmakers et al., 2016). A similar multi-lab replication psychology that was considered successful yielded an effect size that was much smaller than the original (Alogna et al. 2014). These findings have engendered pessimism about reproducibility.

Coincident with the start of the reproducibility debate was the advent of online experimentation. Crowd-sourcing websites, such as Amazon Mechanical Turk, offered the prospect of more efficient, powerful, and generalizable ways of testing psychological theories (Buhrmester, Kwang, & Gosling, 2011). The lower monetary costs and the more time-efficient way of conducting experiments online rather than in a physical lab allowed researchers to recruit larger numbers of participants across broader geographical, age, and educational ranges of participants compared with undergraduates (Paolacci & Chandler, 2014). However, online experimentation presents challenges, typically associated with the loss of control over the testing environment and conditions (Bohannon, 2016). Most relevant to the reproducibility debate, online participant pools are large but not infinite, and hundreds of studies are conducted on the same participant pool every day, familiarizing participants with study materials and procedures (Chandler, Mueller, Paolacci, 2014; Stewart et al., 2015). Of particular concern for reproducibility, participants may participate in studies in which they have participated before. A recent preregistered study found sizable reductions in decision-making effects among participants had previously participated in the same studies, suggesting that nonnaïve participants may pose a threat to reproducibility (Chandler et al., 2015). Indeed, nonnaïve participants have

Electronic supplementary material The online version of this article (doi:10.3758/s13423-017-1348-y) contains supplementary material, which is available to authorized users.

✉ Rolf A. Zwaan
rolfzwaan@gmail.com

[1] Department of Psychology, Educational, and Child Sciences, Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3000 DR Rotterdam, Netherlands

[2] Rotterdam School of Management, Erasmus University Rotterdam, Rotterdam, Netherlands

[3] Learning and Innovation Center, Avans University of Applied Sciences, Breda, The Netherlands

been implicated in failures to replicate and declining effect sizes (DeVoe & House, 2016; Rand et al., 2014).

Although concerns with reproducibility span the entire field of psychology and beyond, results in cognitive psychology are typically conceived as comparatively robust (Open Science Collaboration, 2015). We put a sample of these findings to a particularly stringent test by running them under circumstances that are increasingly representative of current practices of data collection but also are documented as challenging for reproducibility. In particular, we conducted the first preregistered replication of a large set of cognitive psychological effects in the most popular online participant pool (Crump, McDonnell, & Gureckis, 2013 and Zwaan & Pecher, 2012 for non-preregistered replications on MTurk). Most importantly, we examined whether reproducibility depends on participant nonnaïveté by conducting the same experiments twice on the same participants a few days apart.

Research suggests that access to knowledge obtained from previous participation (e.g., from alternative conditions or elaboration) can affect people's responses and may reduce effect sizes when participants accordingly adjust their intuitive responses towards what is perceived as normatively correct (Chandler et al., 2015). However, studies in cognitive psychology typically have nontransparent research goals, making memory of previous experiences irrelevant. Accordingly, a reduction of effect size due to repeated participation should be close to zero.

We tested the hypothesis that cognitive psychology is relatively immune to nonnaïveté effects in a series of nine preregistered experiments (https://osf.io/shej3/wiki/home/; see Table 1 for descriptions of each experiment). We selected these experiments for the following reasons. First, we wanted a broad coverage of cognitive psychology. Therefore, we selected three experiments each from the domains of perception/action, memory, and language, arguably the major areas in the field of cognitive psychology. Second, we selected findings that are both well known and known to be robust. After all, testing immunity to nonnaïveté effects presupposes that one finds effects in the first place. Third, we selected tasks that lend themselves to online testing. And fourth, we selected tasks that our team had experience with.

Table 1 Brief descriptions of and references to all replicated experiments

Number	Task	Description	Reference
1	*Simon task*	*Choice-reaction time task that measures spatial compatibility. Responses are faster when a visual target (a red square is presented on the left of the screen) is spatially compatible with the response (pressing the left button) than when the target is spatially incompatible with the response (presented on the right of the screen).*	Craft and Simon (1970)
2	*Flanker task*	*Response inhibition task in which relevant information is selected and inappropriate responses in a certain context are suppressed. Responses are faster for congruent trials in which compatible distractors flank a central target (AAAAA) than for incongruent trials in which incompatible distractors flank a central target (AAEAA).*	Eriksen and Eriksen (1974)
3	*Motor priming (a = masked, b = unmasked)*	*A task with a priming procedure in which responses to stimuli (arrow probes <<) are required that are primed by presented compatible (<<) or incompatible (>>) items. Responses are slower for compatible items when primes are masked but faster when primes are visible.*	Forster and Davis (1984)
4	*Spacing effect*	*Learning task in which learning (of words) is spaced over time. Recall of words is higher for spaced item repetitions with intervening items than for massed items immediately repeated after their first presentation.*	Greene (1989)
5	*False memories*	*Memory task that assesses false memory of recognition performance of items that have not been presented before in a word list but tend to be recognized as presented before because they are semantically related to the words in the list.*	Roediger and McDermott (1995)
6	*Serial position (a = primacy, b = recency)*	*Memory task that examines recall probability based on a word's position in a list. Recall is higher for the first and last words in the list and lowest for items in the middle of the list.*	Murdock (1962)
7	*Associative priming*	*Implicit memory task which requires a response to a target word that is preceded by prime word. Responses are faster when the prime is related than when the prime is unrelated.*	Meyer and Schvaneveldt (1971)
8	*Repetition priming (a = low frequency, b = high frequency)*	*Implicit memory task in which speed of response depends on previous exposure to an item and the word frequency of that item. Responses are faster for repeated than for new items. This repetition effect is larger for low frequency words than high frequency words.*	Forster and Davis (1984)
9	*Shape simulation*	*Sentence-verification task that requires a response on whether the object in a picture was present in the previous sentence. Yes responses are faster when the picture matches the implied shape mentioned in sentence than when it mismatches.*	Zwaan, Yaxley, and Stanfield (2002)

Although these findings have proven to be highly reproducible in the laboratory, their robustness in an online environment has not yet been established in preregistered experiments. More importantly, it is unknown whether these findings are robust to the presence of nonnaïve participants. We tested this hypothesis by replicating each study in the most conservative case—in which *all* participants encountered the study before.

General method

Detailed method descriptions for each experiment can be found in the *Supplementary Materials*. Participants were tested in two waves using the Mechanical Turk platform. Approval for data collection was obtained from the Institutional Review Board in the Department of Psychology at Erasmus University Rotterdam. All experiments were programmed in Inquisit. The Inquisit scripts used for collecting the data can be found at https://osf.io/ghv6m/. At the end of wave 1 of each experimental task, participants were asked to provide the following information: age, gender, native language, education. At the end of both waves, we asked the following questions, all of which could be responded to by selecting one of the alternatives "not at all," "somewhat," or "very much": "I'm in a noisy environment"; "There are a lot of distractions here"; "I'm in a busy environment"; "All instructions were clear"; "I found the experiment interesting"; "I followed the instructions closely"; "The experiment was difficult"; "I did my best on the task at hand"; "I was distracted during the experiment."

In all experiments, different versions of materials and, in some cases, key assignments were created. Different versions ensured counterbalancing of stimulus materials and key assignments. Participants were randomly assigned to one of the versions when they participated in wave 1. Then, upon return 3 or 4 days later for wave 2, half of the participants were assigned to the exact same version of the experiment and the other half were assigned to a different version such that there was zero overlap between the stimuli in the first and second wave. Participants who had participated in one of the experiments were not prohibited from participating in the other experiments.

Sampling plan

For each experiment, we started with recruiting 200 participants: 100 on Monday and 100 on Thursday. Three or four days after the first participation, each participant was invited to participate again. Our goal was to have a final sample size of 80 participants per condition (same items or different items on the second occasion), taking into account nonresponses and the exclusion criteria below. Whenever we ended up with fewer than 80 participants per condition, we recruited another batch. Because we expected null effect for the crucial interactions, power analyses could not be used to determine our sample sizes, because these analyses require that one predicts an effect and that one has strong arguments for its magnitude. Hence, we decided to obtain more observations than is typically done in previous experiments examining the same effects. By doing so, our parameter estimates are relatively precise.

Exclusion criteria

Data from participants with an accuracy <80% in RT tasks or an accuracy <10% in memory tasks or a mean (reaction time) RT longer than the group $M + 3SD$ were excluded. Data from each participant in the RT tasks were trimmed by excluding trials where the trial RT deviated more than $3SD$ from the subject M. From the remainder, participants were excluded (starting with those who participated last) to create equal numbers of participants per counterbalancing version.

Participants were recruited via Amazon Mechanical Turk. The subjects participated in two waves, held approximately 3 days apart. In the second wave, half of the subjects participated in an exact copy of the experiment they had participated in before; the other half participated in a version that had an identical instruction and procedure but used different stimuli. A recent study demonstrated that certain findings replicated with the same but not with a different set of (similar) stimuli (Bahník & Vranka, 2017). Our manipulation allowed us to examine whether changing the surface features of an experiment (i.e., the stimuli) affects the reproducibility of its effect in the same sample of subjects. Each experiment had a sample size of 80 per between-subjects condition (same stimuli vs. different stimuli).

General results

Detailed results per experiment are described in the *Supplementary Materials*. Data for all experiments can be found here: https://osf.io/b27fd/. The results can be summarized as follows. First, the first wave yielded highly significant effects for all nine experiments, with in each case Bayes factors in excess of 10,000 in support of the prediction. Second, each effect was replicated during the second wave. Third, effect size did not vary as a function of wave; Bayes factors showed moderate to very strong support for the null hypothesis. Fourth, it did not matter whether subjects had previously participated in the exact same experiment or one with different stimuli. The main results are summarized in Fig. 1. The x-axis displays the wave-1 effect sizes and the y-axis the wave-2 effect sizes. The blue dots indicate the same-stimuli condition and the red dots the different-stimuli condition. The numbers indicate the specific experiment (e.g., 5 = false memory).

In the preregistration, we stated that "Bayesian analysis will be used to determine whether the effect size difference

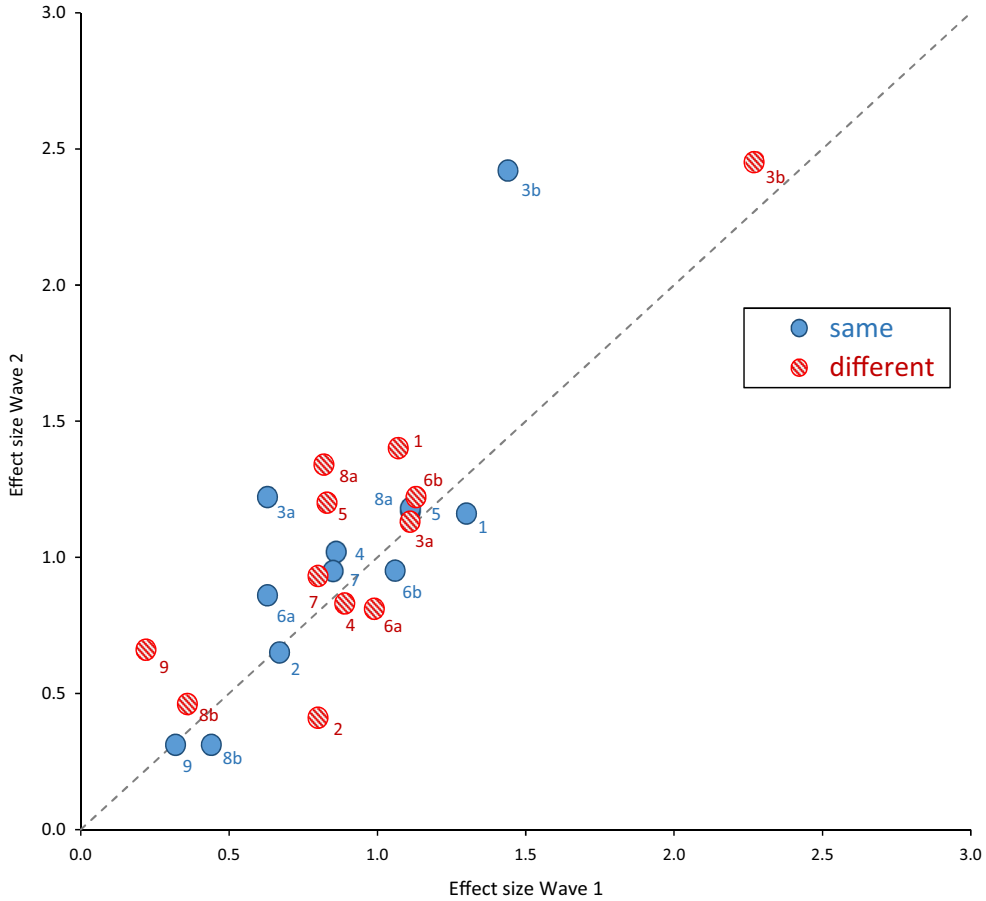

Fig. 1 Wave 1 effect size versus wave 2 effect size (Cohen's *d*). Effect sizes were computed in JASP (JASP Team, 2017). Diagonal line represents equal effect sizes. For each experiment separate effect sizes are plotted for same materials between sessions (blue solid dots) and different materials between sessions (red striped dots). Labels correspond to the different experiments listed in Table 1.

between waves 1 and 2 better fits a 0% reduction model or a 25% reduction model." However, the absence of a reduction in effect sizes from wave 1 to wave 2—the wave 2 effect sizes were, if anything, larger than the wave 1 effect sizes—rendered the planned analysis meaningless. We therefore did not conduct this analysis.

General discussion

Overall, these results present good news for the field of psychology. In contrast to findings in other parts of the field (Chandler et al., 2015), the effects we studied were reproducible in samples of nonnaïve participants, which are increasingly becoming the staple of psychological research. What the tasks used in this research have in common is that they (1) use within-subjects designs and (2) have opaque goals. Although it is clear that participants may learn something from their previous experience with the experiments (e.g., response times were often faster in wave 2 than in wave 1), this learning did not extent to the nature of the manipulation. We should note that it is not impossible that some of our participants had previously participated in similar experiments. For these participants, wave 1 would actually be wave N+1 and wave 2 would be wave N+2. Nevertheless, it appears that the tasks used in this study are so constraining that they encapsulate behavior from contextual variation and even from recent relevant experiences to yield highly reproducible effects. We should add a note of caution. What we have examined are the basic effects with each of these paradigms. In the literature, one often finds variations that are designed to examine how the basic effect varies as a function of some other factor, such as manipulations of instructions, stimulus materials (e.g., emotional vs. neutral stimuli), subject population (patients vs. controls) of the addition of a secondary task. The jury is still out on whether such secondary findings are as robust as the more basic findings we have presented here.

Author contributions R.A. Zwaan developed the study concept. All authors contributed to the study design. Testing and data collection were performed by D. Pecher and S. Bouwmeester. D. Pecher and R.A. Zwaan performed the data analysis. R.A. Zwaan, D. Pecher, and G. Paolacci drafted the manuscript, and all other authors provided critical revisions.

All authors approved the final version of the manuscript for submission. We thank Frederick Verbruggen and Hal Pashler for helpful feedback on a previous version of this paper.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S., Birt, A. R., & Zwaan, R. A. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science, 9,* 556–578.

Bahník, S., & Vranka, M. A. (2017). If it's difficult to pronounce, it might not be risky. The effect of fluency on judgment of risk does not generalize to new stimuli. *Psychological Science, 28*, 427–436.

Bohannon, J. (2016). Mechanical Turk upends social sciences. *Science, 352,* 1263–1264.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3–5.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46,* 112–130.

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science, 26,* 1131–1139.

Craft, J. L., & Simon, J. R. (1970). Processing symbolic information from a visual display: Interference from an irrelevant directional cue. *Journal of Experimental Psychology, 83,* 415–420.

Crump, M. J. C., McDonnell, J. V., & Gureckis, V. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One, 8*, e57410.

DeVoe, S. E., & House, J. (2016). Replications with MTurkers who are naïve versus experienced with academic studies: A comment on Connors, Khamitov, Moroz, Campbell, and Henderson (2015). *Journal of Experimental Social Psychology, 67,* 65–67.

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., & Prenoveau, J. M. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science, 11,* 158–171.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a non search task. *Perception and Psychophysics, 16,* 143–149.

Forster, K. I., & Davis, C. (1984). Repetition Priming and Frequency Attenuation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 4.

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 371–377.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., ... Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*, 546–573.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90,* 227–234.

Murdock, B. B., Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64,* 482–488.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. doi:10.1126/science.aac4716

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23,* 184–188.

Popper, K. R. (1959). *The Logic of Scientific Discovery, translation of Logik der Forschung*. Oxford: Routledge.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, A. W., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications, 5,* 4677.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 803–814.

Stewart, N., Ungemach, C., Harris, A. J. L., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,30 Amazon Mechanical Turk workers. *Judgment and Decision Making, 10,* 479–491.

JASP Team (2017). JASP (Version 0.8.1.2)[Computer software].

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Jr., & Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science, 11,* 917–928.

Zwaan, R. A., & Pecher, D. (2012). Revisiting mental simulation in language comprehension: Six replication attempts. *PLoS One, 7*, e51382.

Zwaan, R. A., Yaxley, R., & Stanfield, R. (2002). Language comprehenders mentally represent the shape of objects. *Psychological Science, 13,* 168–171. **Experiment 1**.