How Dropping Subjects Who Failed Manipulation Checks Can Bias Your Results: An Illustrative Case

Simon Varaine

Double check or don't check.
Why dropping subjects who failed manipulation checks may lead to Type I error

Simon Varaine

+33 6 07 90 02 23
simon.varaine@sciencespo-grenoble.fr

PhD, associate researcher
Univ. Grenoble Alpes, CNRS, Science Po Grenoble*, PACTE,
38000 Grenoble
France
* School of Political Studies Univ. Grenoble Alpes

Abstract

Manipulations checks are post-experimental measures widely used to verify that subjects understood the treatment. Some researchers drop subjects who failed manipulation checks in order to reduce the risk of Type II error. This short report warns that this practice may lead to Type I error. In a survey experiment, subjects were primed with fictional news stories depicting an economic decline versus prosperity. Subjects were then asked whether the news story depicted an economic decline or prosperity. Results indicate that responses to this manipulation check captured subjects' pre-existing beliefs about the economic situation. As a consequence, dropping subjects who failed the manipulation check mixes the effects of pre-existing and induced beliefs, increasing the risk of Type I error. Researchers should either avoid dropping subjects, or rely on highly sensitive manipulation checks or pre-treatment screeners.

Key words

manipulation checks, randomized experiments, survey experiments, causal inference, Type I error

Word count

875

Acknowledgements

I would like to thank Antoine Machut and the team of the Vendredis Quanti network for early discussions about this study. This research is part of the Popeuropa project funded by Sciences Po Grenoble.

How dropping subjects who failed manipulation checks can bias your results. An illustrative case

Abstract

Manipulations checks are post-experimental measures widely used to verify that subjects understood the treatment. Some researchers drop subjects who failed manipulation checks in order to limit the analyses to attentive subjects. This short report offers a novel illustration on how this practice may bias experimental results: in the present case, through confirming a hypothesis that is likely false. In a survey experiment, subjects were primed with fictional news stories depicting an economic decline versus prosperity. Subjects were then asked whether the news story depicted an economic decline or prosperity. Results indicate that responses to this manipulation check captured subjects' pre-existing beliefs about the economic situation. As a consequence, dropping subjects who failed the manipulation check mixes the effects of pre-existing and induced beliefs, increasing the risk of false positive findings. Researchers should avoid dropping subjects based on post-treatment measures and rely on pre-treatment measures of attentiveness.

Keywords: manipulation checks, randomized experiments, survey experiments, causal inference, Type I error

Manipulations checks are post-experimental measures aiming at "ensuring that an experiment actually has been conducted (i.e., that the Independent Variable has been effectively manipulated)" (Sansone et al., 2003, p. 244). They typically take the form of comprehension questions immediately following the experimental treatment to check that subjects paid attention and understood the treatment (Kane & Barabas, 2019). The inclusion of manipulation checks enters standards of best practices in experimental political science (Mutz & Pemantle, 2015). They are particularly important to avoid Type II error – i.e. the false negative conclusion that the research hypothesis is wrong while it is actually true – in case of null results.

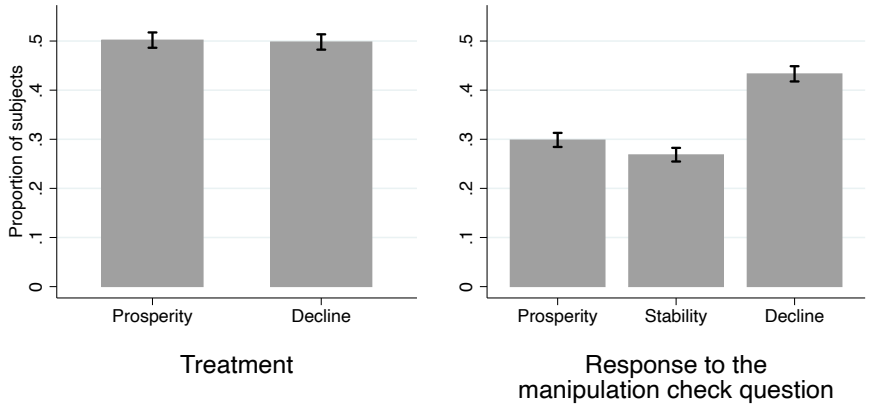
A widespread practice is to exclude participants who failed manipulation checks in order to limit the analyses to subjects who understood the experimental prompt (see political sciences experiments surveyed by Aronow et al., 2019). However, some studies have warned that this may bias the analyses (Aronow et al., 2019; Berinsky et al., 2014; see also Montgomery et al., 2018). Berinsky et al. (2014) showed that individual responses to screeners correlate with a range of personal characteristics. Dropping inattentive subjects may distort the sample to certain "races, ages, and levels of education". More problematically, Aronow et al. (2019) demonstrated that this may bias the estimation of causal effects by creating asymmetry between experimental arm.

This study offers a new illustration on how dropping subjects who failed manipulation checks may bias experimental results. Aronow et al. (2019) presented an illustrative experiment in which dropping subjects lead to under-estimating the effect size of the treatment. The present study presents another experimental case in which dropping subjects increases the risk of Type I error when testing a hypothesis of interest – i.e. drawing a false positive conclusion that confirms the research hypothesis while it is actually wrong.

The experiment was conducted in an online survey filled during April 2019 by nationally representative samples from Denmark, France, Germany, Italy, Spain and the Netherlands. A total of 3949 subjects participated in the experiment, based on the economic threat manipulation from Stenner (2005). Subjects were

randomly assigned to one of two short fictional news stories about the national economic context respectively depicting an improving situation (*prosperity*) or a worsening situation (*decline*)[1]. The initial purpose of the experiment was to test whether subjects' express more nostalgia after the *decline* treatment compared to the *prosperity* treatment.

Figure 1: Share of subjects by experimental treatment and by response to the manipulation check question (with 95% Confidence Interval)



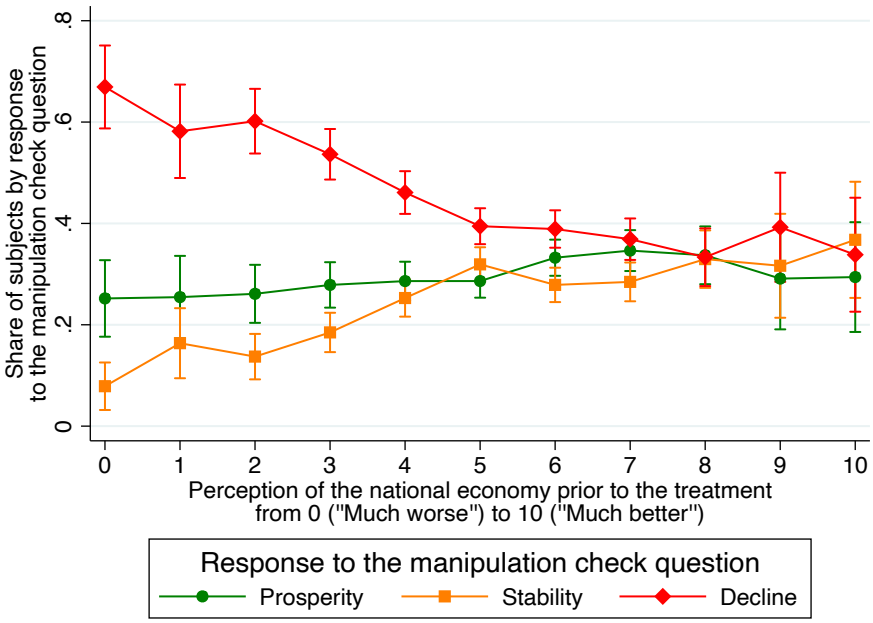
Following the treatment, subjects answered a manipulation check question: "According to the news story, the national economic situation is: Worsening / Stable / Improving". Only 63% of subjects provided the correct answer regarding their experimental treatment – i.e. "Improving" in the *prosperity* treatment and "Worsening" in the *decline* treatment. More problematically, subjects who failed did apparently not respond at random: as shown by Figure 1, 30% of subjects declared that, according to the news story, the economic situation was improving, 27% that it was stable and 43% that it was worsening.

Prior to the experiment, subjects answered a question about their own perception of the economic situation: "Would you say that the economic situation

[1] See the contents of treatments in the online appendix.

now is better or worse to how it was 5 years ago?". Subjects responded with a 11-point scale from 0 ("Much worse") to 10 ("Much better"). Results from a two-way Anova indicate that responses to this question are significantly related with responses to the manipulation check, $F(2, 3731) = 55.79, p < .0001$. As shown by Figure 2, the more favorable subjects' perceptions of the economy, the less they responded that the news story depicted an economic decline and the more they responded that the story depicted economic stability[2]. This means that the manipulation check actually captured some subjects' pre-existing beliefs about the economic situation.

Figure 2: Share of responses to the manipulation check question depending on the perception of the national economy prior to the experiment (with 95% Confidence Interval)

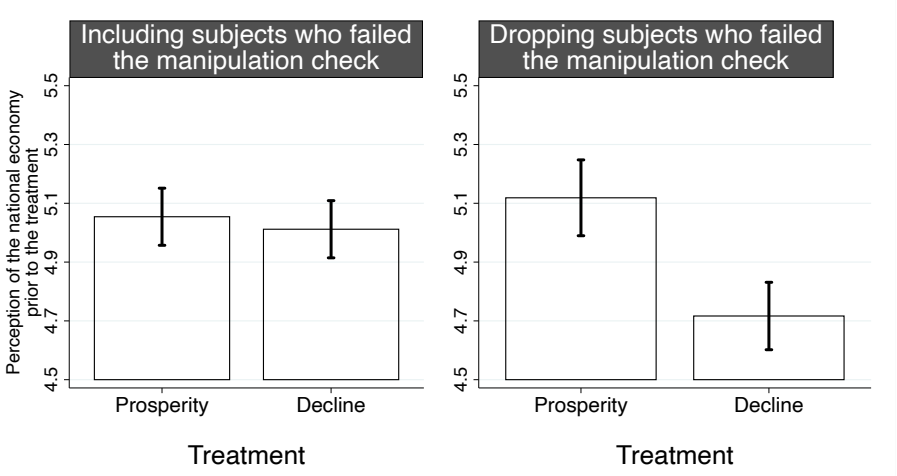


Now, what happens if we drop subjects who failed the manipulation check?

[2]In contrast, there is no clear effect on the probability that subjects responded that the treatment described a economic prosperity.

Figure 3 presents the average perception of the economy prior to the experiment for subjects of the *prosperity* versus *decline* treatments. When including all subjects, there is no significant difference across treatments in the perception of the national economy prior to the survey experiment, $t(3752) = 0.6062, p = .5444$. This is what we expect from randomization: the treatment is independent from the subjects characteristics prior to the experiment. In contrast, when excluding subjects who failed the manipulation check, there is a significant difference across treatments in the perception of the national economy prior to the survey experiment, $t(2366) = 4.5688, p < .0001$. It is impossible that the experimental treatment had a causal effect on responses to a question asked earlier in the survey. Thus, this reflects a selection effect emerging from dropping subjects who failed the manipulation check.

Figure 3: Average perception of the national economy prior to the experiment depending on the experimental treatment (with 95% Confidence Interval)



Suppose that we want to test the effect of the treatment on a dependent variable. After the treatment, the subjects' level of nostalgia was assessed. Subjects indicated on a 5-point scale from "strongly disagree" to "strongly agree" to what extent they agreed that "the society used to be a much better place". As

Table 1: Results from linear regression models of the level of nostalgia

	(1)	(2)	(3)
	Dropping subjects who failed the manipulation check		All subjects
Decline treatment	0.164***	0.114**	0.0603
	(0.0433)	(0.0425)	(0.0337)
Perception of the economy prior to the experiment		-0.138***	
		(0.00989)	
Constant	3.344***	4.054***	3.426***
	(0.0323)	(0.0597)	(0.0237)
Observations	2395	2311	3797

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

shown by regression models in Table 1, when dropping subjects who failed the manipulation check (model 1), one would conclude that the *decline* treatment had a significant positive effect on nostalgia. Yet, this effect is reduced when controlling for the pre-existing perception of the economy (model 2). Since it is impossible to measure all potential pre-existing characteristics of subjects selected through the manipulation check, the best option is to avoid dropping subjects (model 3). This decision is conservative: it is likely to greatly increase noise in the data and reduce effect sizes – with increased risk of Type II error. In model 3, the effect of the treatment is no longer significant[3]. Nonetheless, this is the only way to ensure that, if some treatment effect is observed, it is of genuinely causal nature.

The present study does not advocate against the inclusion of post-treatment manipulation checks. These can be informative tools – especially in development phase of experiments – to assess the degree of attention and comprehension of the treatment in given type of sample. In the present study, the manipulation check reveals that a very large fraction of subjects were inattentive. One possibility is that the pre-treatment question about the subjects' perception of the national economy induced subjects to disregard the content of the experimental vignette. This would explain the high rate of failure in the responses to the manipulation check, and their close correlation with subjects' pre-existing beliefs about the economy. To test fort this, it would be necessary to have a control group for whom the initial question was not included. However, this limitation does not affect the overall conclusion that exclusion based on post-treatment manipulation checks must be avoided.

What are then the options available to researchers? A first option highlighted by the literature is to include pre-treatment questions to gauge subjects' attentiveness. A common tool is instructional manipulation checks – or "screeners". Screeners are similar to classic survey questions but ask participants to ignore

[3]Note that results are essentially unchanged when including country fixed effects (see the online appendix).

the standard response format and instead provide a confirmation that they have read the instruction Berinsky et al. (2014); Oppenheimer et al. (2009). One disadvantage is that screeners may induce subjects to think that researchers want to trap them, which alters their responses to subsequent questions (Hauser & Schwarz, 2015). Alternatively, Kane et al. (2020) recently proposed ready-to use "mock vignettes". A mock vignette mimics common king of descriptive content in political science experiments but appear before the researcher's treatment. All subjects read the same vignette and must then answer factual questions about it, allowing to check for subjects' attentiveness.

The latter tools come with the cost of sacrificing survey space. Another alternative is to rely on timers as a proxy to identify inattentive subjects. Various studies highlight that subjects with short response times are generally less attentive (see for instance Börger, 2016; Wood et al., 2017)[4]. Read et al. (2021) designed a method to identify inattentive subjects based on multiple question timers. Their method does not induce post-treatment selection bias when computed on question timers before the treatment. Besides, it allows to identify slow but nonetheless inattentive subjects.

Depending on the space available in survey, researchers may use these methods to perform analyses on sub-sample(s) of attentive subjects, in order to limit the risk of Type II error without inducing post-treatment bias. However, these measures of attentiveness may correlate with politically relevant variable, such as age, race and education (see Berinsky et al., 2014, Kane et al., 2020). Thus, restricting analyses to attentive subjects comes with the risk of drawing conclusions that are not representative of the population. To mitigate this risk, the best practice should be to report estimates of treatment effects based on both the overall sample, and sub-sample(s) of attentive subjects.

[4]Our study includes a measure of the overall duration of the survey, which confirms this. The share of subjects who failed the manipulation check question is significantly higher among subjects who spent relatively less time in the survey (see the online appendix).

Data Availability

This research is part of the Popeuropa project supported by IDEX Université Grenoble Alpes (IRS 2017-2018), Sciences Po Grenoble and Pacte laboratory. The data, code, and any additional materials required to replicate all analyses in this article are available at the Journal of Experimental Political Science Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/7DXBGG.

Conflicts of Interest

I acknowledge that I have no conflicts of interest or potential conflicts of interest with regard to the submitted work.

Ethics Statement

No IRB approval was sought for this project. The study complies with the European General Data Protection Regulation relative to the protection of personal data, and received the approval of the scientific committee of Sciences Po Grenoble. We obtained informed consent from all participants, who could choose not to answer any questions or withdraw from the study at any time. Compensation was delivered by the survey vendor. This research adheres to APSA's Principles and Guidance for Human Subjects Research. The experiment included short written abstracts of fictional news media prospects about the economy. Given the variety of economic prospects commonly available in public news media, it was considered that no significant deception was induced. Section 1 in the online appendix details the experimental procedure employed.

References

Aronow, P. M., Baron, J., & Pinson, L. (2019). A note on dropping experimental subjects who fail a manipulation check. *Political Analysis*, *27*, 572–589.

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, *58*, 739–753.

Börger, T. (2016). Are fast responses more random? Testing the effect of response time on scale in an online choice experiment. *Environmental and Resource Economics*, *65*, 389–413.

Hauser, D. J., & Schwarz, N. (2015). It'sa trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *Sage Open*, *5*, 2158244015584617.

Kane, J. V., & Barabas, J. (2019). No harm in checking: Using factual manipulation checks to assess attentiveness in experiments. *American Journal of Political Science*, *63*, 234–249.

Kane, J. V., Velez, Y. R., & Barabas, J. (2020). Analyze the Attentive & Bypass Bias: Mock Vignette Checks in Survey Experiments, .

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, *62*, 760–775.

Mutz, D. C., & Pemantle, R. (2015). Standards for experimental research: Encouraging a better understanding of experimental methods. *Journal of Experimental Political Science*, *2*, 192–215.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, *45*, 867–872.

Read, B., Wolters, L., & Berinsky, A. J. (2021). Racing the Clock: Using Response Time as a Proxy for Attentiveness on Self-Administered Surveys. *Political Analysis*, (pp. 1–20). doi:10.1017/pan.2021.32.

Sansone, C., Morf, C. C., & Panter, A. T. (2003). *The Sage Handbook of Methods in Social Psychology*. Thousand Oaks: Sage Publications.

Stenner, K. (2005). *The Authoritarian Dynamic*. Cambridge: Cambridge University Press.

Varaine, S., (2022). Replication Data for: How dropping subjects who failed manipulation checks can bias your experimental results. An illustrative case. *Harvard Dataverse*. doi:10.7910/DVN/7DXBGG.

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, *8*, 454–464.

How dropping subjects who failed manipulation checks can bias your results. An illustrative case

Online Appendix

1. Experimental treatments

During the survey, subjects received the following instruction:

"We are interested in what people can recall about major news stories; We are going to present you a summary of a major news story and then we will ask you how you feel about it."

Prosperity treatment. "The news story was that the [Country] economy might improve dramatically over the next year. The article suggested that the [Country] may enjoy a period of rapid economic growth. According to some of the indicators, the national economy might show considerable gains over the next year or so, with a big drop in inflation and unemployment. The conclusion was that the [Country] may look forward to strong economic growth in the year to come."

Deprivation treatment. "The news story was that the [Country] economy might worsen dramatically over the next year. The article suggested that the [Country] may suffer a period of rapid economic decline. According to some of the indicators, the national economy might show considerable deterioration over the next year or so, with a sharp rise in inflation and unemployment. The conclusion was that the [Country] may be facing a severe economic recession in the year to come."

2. Models with country fixed effects

Table 1: Results from linear regression models of the level of nostalgia

	(1)	(2)	(3)
	Dropping subjects who failed the manipulation check		All subjects
Decline treatment	0.131**	0.0990*	0.0541
	(0.0429)	(0.0423)	(0.0333)
Perception of the economy prior to the experiment		-0.132***	
		(0.0106)	
Country fixed effects (Denmark as reference)			
France	0.171*	-0.0851	0.129*
	(0.0773)	(0.0800)	(0.0572)
Germany	0.124	0.0788	0.0752
	(0.0770)	(0.0767)	(0.0570)
Italy	0.422***	0.219**	0.389***
	(0.0797)	(0.0814)	(0.0565)
Netherlands	-0.176*	-0.180*	-0.152**
	(0.0770)	(0.0766)	(0.0569)
Spain	0.165*	0.0291	0.114
	(0.0779)	(0.0778)	(0.0586)
Constant	3.247***	4.028***	3.336***
	(0.0631)	(0.0892)	(0.0437)
Observations	2395	2311	3797

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3. Time spent on survey and failure to the manipulation check question

Figure 1: Share of subjects who failed the manipulation check depending on the time they spent on the survey (with 95% Confidence Interval)