International Journal of Aviation Research

Volume 12 | Issue 01 Peer-Reviewed Article

Why Aviation Researchers Often Should Eschew Manipulation Checks

David Trafimow
New Mexico State University

In many sciences, including aviation science, researchers often perform manipulation checks to demonstrate that their experimental manipulations work as hypothesized. And yet, manipulation checks can be problematic in that they can decrease generalizability; increase costs; and incur the risk that an experiment, that otherwise would result in an empirical victory for the researcher, will instead result in an empirical defeat. In addition, researchers overestimate the extent to which manipulation checks facilitate the elimination of alternative hypotheses. Consequently, the decision to use manipulation checks, or for journal editors to require them for publication, should be taken with more care than aviation scientists currently realize. Although there are exceptions, aviation researchers often should eschew manipulation checks.

Recommended Citation:
Trafimow, D (2020). Why Aviation Researchers Often Should Eschew Manipulation Checks. *International Journal of Aviation Research, 12*(01), 43-53.

Aviation research is difficult to perform. One challenge is that many aviation studies are expensive and time-consuming. The equipment is expensive, as is pilot time. Consequently, any measures aviation researchers can take to reduce expenses are worth considering. As I will argue presently, one expense aviation researchers face is due to unnecessary—and sometimes deleterious—manipulation checks. The present goal is to demonstrate that, except on rare occasions, researchers can dispense with manipulation checks, thereby saving time and money. Furthermore, dispensing with manipulation checks can result in an added benefit to aviation researchers in the form of increased generalizability.

What is a manipulation check? A manipulation check is a measure of a construct allegedly influenced by the experimental manipulation, that is hypothesized to intervene between the manipulation and the dependent variable of primary interest (e.g., Sansone, Morf, & Panter, 2008). For example, suppose a marketer wishes to test whether a new advertising campaign influences people's attitudes towards buying the product which, in turn, influences the actual behavior of buying the product. Product sales might be the main dependent variable of interest; but marketing researchers routinely measure attitudes too, as a manipulation check. The idea is to show that the marketing campaign really does influence the hypothesized intervening construct; namely, attitudes towards buying the product. Without this manipulation check, the common criticism is that there is no way to know that the attitude manipulation really "worked" in the sense that it influenced attitudes towards buying the product. Consequently, there is considerable support for researchers performing manipulation checks (see discussions by Berinsky, Margolis, & Sances, 2014; Cozby, 2009; Foschi, 2007; Perdue & Summers, 1986; Sansone et al., 2008).[1]

Why Aviation Researchers Perform Manipulation Checks

Two questions suggest that manipulation checks are indispensable.

- Without a manipulation check, how can a researcher demonstrate that the manipulation worked?
- And if the manipulation might not have worked, isn't the researcher's argument, that the obtained effect is due to the experimental manipulation, thereby compromised?

Based on the two bullet-listed questions, it seems sensible that researchers should perform manipulation checks to demonstrate that their manipulations work as they are supposed to work, and that they have a good case that their effects are due to their experimental manipulations. It also seems sensible that journal reviewers and editors would insist on manipulation checks before being willing to publish research, and that granting agencies would insist on manipulation checks as a condition for funding. After all, manipulation checks are an important component of what might be termed,

[1] Sawyer, Lynch, and Brinberg (1995) provided a Bayesian cost-benefit analysis and Trafimow and Rice (2009) suggested potential disadvantages of manipulation checks.

"responsible science." Thus, it is unsurprising that researchers often perform manipulation checks.

Considering a Successful Experiment Without a Manipulation Check (and Counterfactual)

Imagine a successful experiment as follows. A researcher believes that increased pilot alertness should increase pilot performance on a flight simulator. To test the idea, the researcher randomly assigns pilots to receive a cognitive prime designed to increase alertness (experimental condition) or not (control condition), with performance on a flight simulator as the dependent variable. Suppose that the experiment is successful and that pilots in the experimental condition outperform pilots in the control condition. The researcher submits the manuscript to a journal, and the submission is rejected because of a lack of a manipulation check. In the decision letter, the editor states the following:

> I was fortunate to have obtained the services of two reviewers, both top researchers in the area. Unfortunately, both reviewers recommended rejection, and for the same reason. You had no manipulation check and so there is no way to know that the difference in pilot performance is because pilots in the experimental condition were more alert than pilots in the control condition. It could be that the difference in pilot performance occurred for some other reason. For example, perhaps the cognitive prime caused pilots to want to perform better, and it was the increase in motivation, rather than anything having to do with alertness, that caused the difference in pilot performance. Had you performed a manipulation check, with a separate alertness measure prior to assessing pilot performance, you could have shown that your manipulation really did increase alertness, and then the reviewers and I would have reacted more favorably to your manuscript. As matters now stand, I feel forced to reject your manuscript. I know this is not what you wanted to hear, and I hope that the negative decision will not prevent you from submitting your work to us in the future.

Did our hypothetical editor make the right decision?

Let us consider the counterfactual case where a researcher performed the same experiment, with the same result; but included a successful manipulation check (pilots in the experimental condition scored better on the manipulation check than did pilots in the control condition), as the reviewers and editor recommended. Thus, we have two experimental cases that can be compared easily. Both experiments have an independent variable (IV) and a dependent variable (DV), but only one of the experiments has a manipulation check (MC).

- Original experiment: IV→DV
- Counterfactual experiment: IV→MC→DV

Is it really true that the counterfactual experiment is superior to the original experiment?

Let us consider again the complaint of the editor and reviewers that the cognitive prime in the original experiment could have worked for a reason other than by influencing the alertness of the pilots. Under the assumption that the MC in the counterfactual experiment validly measures alertness—and it is worth stressing that this is an assumption that might not be true—the MC tells us that the manipulation influences alertness. However, the MC does not tell us that the increase in alertness is the factor that causes increased pilot performance on the simulator. Referring back to the decision letter, let us suppose that the cognitive prime influences both alertness and motivation. Well then, despite the manipulation check demonstrating that the manipulation influences alertness, it could nonetheless be possible that the independent variable influences the dependent variable through motivation and not through alertness. Put another way, the competing explanation featuring motivation, as opposed to alertness, is every bit as plausible in the counterfactual experiment as in the original experiment, despite the inclusion of the MC in the counterfactual experiment. Thus, the researcher who performs the counterfactual experiment gains surprisingly little by including the MC.

Worse yet, the counterfactual experiment introduces a problem absent in the original experiment. To see this, consider again that both the original and alternative experiments are successful in the sense that the cognitive prime causes pilot performance to increase, though the reason for the increase is not clear. In the counterfactual experiment, however, it might be that the MC is a necessary component for the increased pilot performance to occur. Perhaps the cognitive prime only works when it is followed by the MC! This is an important, though underestimated problem, with including an MC. The inclusion of an MC reduces generalizability.

And on top of reduced generalizability, including the MC may be costly. Taking the alertness measure as an example, how should the researcher measure alertness? One way is with a questionnaire asking pilots to self-rate themselves on alertness. But this would increase the length of time pilots would spend on the experimental session, and because pilot time is costly, the total cost of the experiment would increase. In addition, it is far from clear that self-evaluations of alertness have sufficient validity. And if the researcher wishes to include various physiological measures of alertness, the costs would be compounded by the necessity of obtaining yet more equipment and assistants who know how to use that equipment.

Finally, adding the MC might cause the experiment not to work. There is no way for the researcher to know what concepts are primed by the mere act of completing the MC, or even what the effects of the passage of the time taken to complete the MC might be. For example, perhaps alertness decreases during the time taken to complete the MC. Unintentional priming effects or effects pertaining to the passage of time might interfere with an experiment that otherwise would work.

In summary, in the case where a researcher performs what otherwise would be a successful experiment, including an MC can decrease generalizability, increase cost, or decrease the probability that the experiment works. Because these disadvantages come with little gain, as the MC is insufficient to eliminate alternative explanations, there is

little point in performing the MC. A consequence is that our hypothetical reviewers and editors made the wrong decision in rejecting the manuscript based on a failure to include an MC.

Exceptions

Despite MCs having important disadvantages, there are exceptions. One of these concerns the rare case where the manipulation is of importance, independently of the researcher's goal. Referring to the example, suppose that it were important to demonstrate that the cognitive prime influences alertness, even in the absence of an effect on simulator performance. It might be that pilot alertness matters with respect to dependent variables not included in the experiment but that nevertheless could become important in future research. To render this possibility salient, let us imagine that the simulator performance is based on a type of aircraft where pilot alertness does not matter; but that there are other types of aircraft, not tested in the experiment, where pilot alertness does matter. In that case, even though manipulating pilot alertness is ineffective in the context of the present experiment, showing that the manipulation nevertheless influences pilot alertness provides a potential step forward for future researchers who test simulated performances with respect to other aircraft.

Another exception is the case where the experiment is not successful. In a failure scenario, it might be important to know the reason for the failure. Is the hypothesis wrong? For example, is it untrue that alertness influences pilot performance? Or is the empirical failure because the manipulation did not work (the cognitive prime did not affect alertness)? If the researcher believes that the failure is due to the manipulation not influencing alertness, that would provide a good reason to attempt a different manipulation. In contrast, if the researcher believes that the manipulation did influence alertness, then the empirical failure could be more strongly attributed to the hypothesis being wrong. Thus, in the case of an empirical failure, an MC could be valuable in helping the researcher make an attribution to the manipulation or the hypothesis. A second attempt makes more sense if the empirical failure is attributed to the manipulation than to the hypothesis.

Contrasting the cases of empirical victory versus empirical defeat indicates a fascinating asymmetry. Although an MC is pointless when the researcher obtains an empirical victory, an MC is perhaps sensible when the researcher suffers an empirical defeat. Thus, in deciding whether to include an MC, one issue the researcher should consider is how confident she is of an empirical victory. More confidence implies not using an MC whereas less confidence implies that an MC might be worth the problems it causes with respect to generalizability, cost, and the increased risk that including the MC might produce with respect to interference with an experiment that otherwise would work.

Discussion

The assertion that an MC is problematic in the context of an empirical victory is obviously controversial. It is possible to imagine arguments against the assertion and this is a convenient place to address them.

1. One argument might be based on the desire to address alternative explanations. In the scenario where a researcher successfully obtains an effect of cognitive priming on pilot performance on a simulator, we have seen that adding an MC fails to eliminate the alternative explanation that perhaps the cognitive prime affects motivation which, in turn, increases pilot performance. But we have not yet considered how to eliminate the alternative interpretation based on motivation. One possibility is to measure motivation and show a lack of an effect on the motivation measure coupled with the positive effect of the cognitive prime on pilot performance. Thus, a good argument can be made that the researcher should include a motivation measure in the experiment.

Several considerations may apply here. One of them is the bias researchers have that effects can be interpreted but that null effects cannot. According to conventional thinking, a null effect can be attributed to insufficient sample size, measure invalidity, and so on.[2] Thus, it is not clear that a lack of an effect on a motivation measure would provide compelling disconfirmation of the alternative explanation. A way around this problem, of course, is to show that some other manipulation of motivation does have an effect on the motivation measure. But such a demonstration incurs additional costs upon the researcher.

Moreover, even if we were to agree to accept the motivation measure as valid, without any additional work, this would not be an MC but rather an explicit test of a clear alternative explanation. Some clarification may help here. In the scenario where a researcher designs a manipulation to influence alertness, the reason for including an alertness measure would be to demonstrate that, indeed, alertness is affected. Thus, the alertness measure is an MC. In contrast, in the scenario where the researcher includes a motivation measure, the reason for including the motivation measure is to eliminate the alternative explanation that the effect of the manipulation on the dependent variable is because of an intervening effect on motivation. Still in contrast to an alertness measure, the hope is for a lack of an effect on the motivation measure. Therefore, the motivation measure is not an MC; but rather it constitutes a test of an alternative, and competing, explanation. It is important not to confuse an MC, designed to show that the manipulation works as hypothesized; with a measure designed to test, and hopefully eliminate, an alternative explanation.

Should the researcher include both an alertness MC and a motivation measure to test the alternative explanation of an effect of the independent variable upon the dependent variable, through motivation? There would be important disadvantages. First, still assuming an empirical victory, there would be an even greater loss of generalizability

[2] Trafimow (2014) suggested that, contrary to conventional thinking, null effects should be taken more seriously.

with two intervening measures than with one or zero intervening measures. The presence of both measures might be a necessary condition for the cognitive prime to influence pilot performance. Second, including both measures implies higher costs than if one or neither measures are used. Third, including the measures might interfere with an experiment that otherwise would result in an empirical victory. Certainly, given these considerations, the alertness MC should not be used. And it is arguable whether the motivation measure should be used. Although I have argued that it can be advantageous to include a measure that might eliminate an alternative explanation, the potential disadvantages of loss of generalizability, cost, and risk of interference with the experiment also should be considered; thereby rendering the decision to include a motivation measure nonobvious.

2. Another argument might be that if one uses an MC, at least the researcher knows that the manipulation did something. But as we have seen, this depends on whether the experiment ends with an empirical victory or empirical defeat. In the latter case, as I stated earlier, an MC can help the researcher distinguish whether the problem is in the hypothesis or in the manipulation. In contrast, however, in the case of an empirical victory, it is already clear that the manipulation works; or else pilot performance would not have been influenced. The mere fact of an empirical victory forces that the manipulation "works" in the sense of an effect on the dependent variable. Thus, what is at issue is not whether the manipulation works; but rather whether it works according to the hypothesis (e.g., through alertness) or for some other reason (e.g., through motivation). But as we already have seen, the MC does not distinguish between these possibilities, and so little is gained to counterbalance the disadvantages in decreased generalizability, increased cost, and the risk of interfering with an experiment that otherwise would be successful.

3. A third argument pertains to the possibility of mediation analyses. If the researcher includes an MC, it opens the door for a mediation analysis testing whether the manipulation works through the alleged mediator to influence the dependent variable. Without an MC, there are only two variables—the independent variable and the dependent variable—and so there is no way to perform a mediation analysis. By facilitating the performance of a mediation analysis, the MC can be argued to be desirable after all, despite the other issues.

This argument has a surface plausibility but depends upon the assumption that mediation analysis is more definitive than it really is. Recently, several researchers have shown it to be invalid (Fiedler, Schott, & Meiser, 2011; Grice et al., 2015; Kline, 2015; Tate, 2015; Thoemmes, 2015; Trafimow, 2015; 2017). For example, Trafimow (2015) used mediation analysis to test two obviously wrong hypotheses against each other. One hypothesis was that planetary mass causes planetary velocity, which causes kinetic energy and momentum. The other hypothesis was that planetary velocity causes planetary mass, which causes kinetic energy and momentum. Although the mediation analysis strongly disconfirmed the former hypothesis, it strongly supported the latter one; despite both of them being wrong. Mediation analysis is subject to a crucial statistical indistinguishability problem whereby many statistical models are consistent with the results of a mediation analysis, thereby rendering inference to the best model extremely

problematic (Spirtes, Glymour, & Scheines, 2000). In fact, Kline (2015) illustrated an impressive number of models that are consistent with the findings resulting from even the simplest of mediation analyses. It is possible to make more sophisticated arguments (e.g., Trafimow, 2017); but that level of sophistication is unnecessary at present. The bottom line is that mediation analysis is a very poor procedure for drawing strong conclusions about mediating variables. A better way, if the mediator is sufficiently important to justify the expense, is to perform a manipulation to drive the alleged mediator towards a floor or ceiling, so it cannot be decreased or increased, respectively, by the independent variable of interest. In other words, by "fixing" the alleged mediator, the effect of the independent variable on the dependent variable should decrease or disappear, thereby providing a strong case for mediation.[3] Thus, that including an MC provides the researcher with the ability to perform a mediation analysis fails to compensate for the disadvantages of loss of generalizability and costs, not to mention risking the possibility that including an MC may cause an experiment that otherwise would work to fail.

Conclusion

To reiterate, we have seen that the researcher who chooses to include an MC risks loss of generalizability, increased costs, and the possibility that an otherwise successful experiment will fail because of the unpredictable effects of participants completing the manipulation check. Are there sufficient advantages to compensate for the disadvantages? Usually, the answer is in the negative. Including an MC fails to satisfactorily address alternative explanations; the explicit consideration of a specific alternative explanation is required for that. An exception might be if the researcher lacks confidence in the hypothesis or in the manipulation, in which case including a manipulation check can be useful in helping the researcher distinguish whether the hypothesis or manipulation is responsible for an empirical defeat.

In those cases where researchers are confident in their hypothesis and manipulation, they should eschew an MC and demonstrate the effect of the independent variable on the dependent variable. Especially for applied purposes, the mere fact that an application is effective may be more important than the reason it is effective. Furthermore, if the reason does matter, then this concern calls for explicit tests of competing hypotheses; not for MCs. Of course, explicit tests of competing hypotheses likely will consume much in the way of resources, but that is an obstacle that a researcher interested in theory must accept. Including a MC fails to relieve the researcher who cares about theory from the necessity to perform the hard work of testing competing possibilities.

[3] Trafimow et al. (2005) provided an example of this. These researchers hypothesized that negative affect mediated between manipulating violations of duties and negative trait attributions. When they independently "fixed" negative affect, the influence of the manipulation on negative trait attributions decreased markedly.

References

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*(3), 739-53. doi: 10.1111/ajps.12081

Cozby, P. C. (2009). *Methods of Behavioral Research*: Tenth Edition. New York, NY: McGraw-Hill.

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology, 47*(6), 1231–1236. doi: 10.1016/j.jesp.2011.05.007

Foschi, M. (2007). Hypotheses, Operationalizations, and Manipulation Checks. In M. Webster, Jr. and J. Sell (Eds.), *Laboratory Experiment in the Social Sciences*, New York: Elsevier (pp. 247-268).

Grice, J. W., Cohn, A., Ramsey, R. R., & Chaney, J. M. (2015). On muddled reasoning and mediation modeling. *Basic and Applied Social Psychology, 37*(4), 214–225. doi: 10.1080/01973533.2015.1049350

Kline, R. B. (2015). The mediation myth. *Basic and Applied Social Psychology, 37*(4), 202–213. doi: 10.1080/01973533.2015.1049349

Perdue, B. C., & Summer, J. O. (1986). Checking the success of manipulations in marketing experiments. *Journal of Marketing Research*, 23(4), 317-26. https://doi.org/10.1177/002224378602300401

Sansone, C., Morf, C. C. & Panter, A. T. (2008). *The Sage Handbook of Methods in Social Psychology*. Thousand Oaks, CA: Sage Publications.

Sawyer, A. G., Lynch, J. G., & Brinberg, D. L. (1995). A Bayesian analysis of the information value of manipulation and confounding checks in theory tests. *Journal of Consumer Research, 21*(4), 581–595. doi: 10.1086/209420

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge, Massachusetts: The MIT Press.

Tate, C. U. (2015). On the overuse and misuse of mediation analysis: It may be a matter of timing. *Basic and Applied Social Psychology, 37*(4), 235–246. doi: 10.1080/01973533.2015.1062380

Thoemmes, F. (2015). Reversing arrows in mediation models does not distinguish plausible models. *Basic and Applied Social Psychology, 37*(4), 226–234. doi: 10.1080/01973533.2015.1049351

Trafimow, D. (2015). Introduction to special issue: What if planetary scientists used mediation analysis to infer causation? *Basic and Applied Social Psychology, 37*(4), 197-201. doi:10.1080/01973533.2015.1064290

Trafimow, D. (2017). The probability of simple versus complex causal models in causal analyses. *Behavior Research Methods, 49*(2), 739-746. doi: 10.3758/s13428-016-0731-3

Trafimow, D., Bromgard, I.K., Finlay, K. A., Ketelaar, T. (2005). The role of affect in determining the attributional weight of immoral behaviors. *Personality and Social Psychology Bulletin, 31*(7), 935-948. doi: 10.1177/0146167204272179

Trafimow, D., & Rice, S. (2009). What if social scientists had reviewed great scientific

works of the past? *Perspectives in Psychological Science*, *4*(1), 65-78. doi: 10.1111/j.1745-6924.2009.01107.x