

# The experimental method in Public Administration: lessons from replication in Psychology

Christina W. Andrews <sup>1</sup>

Michiel S. de Vries <sup>2</sup>

<sup>1</sup> Universidade Federal de São Paulo / Department of Social Sciences, São Paulo / SP – Brazil

<sup>2</sup> Radboud University / Institute for Management Research, Nijmegen – The Netherlands

In recent years, an increasing number of studies adopting the experimental method have appeared in Public Administration journals. It has been argued that the advantage of experiments in behavioral sciences is that researchers can control contextual factors while investigating the effect of manipulations on a variable of interest. Another point is that experiments can be replicated and, thus, increase confidence in research results. However, replications are rarely undertaken, especially in the behavioral sciences. This article examines the results of the “Open Science Reproducibility Project: Psychology,” which replicated 100 experiments previously published in leading Psychology journals. Based on the findings of this project, we present seven recommendations to Public Administration scholars that can improve the quality of their experiments.

**Keywords:** experimental method in the behavioral sciences; research methods in public administration; reproducibility.

## O método experimental na Administração Pública: algumas lições das replicações na Psicologia

Nos últimos anos, um número crescente de estudos adotando o método experimental tem surgido em periódicos de Administração Pública. Tem-se argumentado que a vantagem dos experimentos nas ciências comportamentais é que este permite aos pesquisadores controlar os fatores contextuais enquanto investigam o efeito das manipulações em uma variável de interesse. Outro argumento é que os experimentos podem ser replicados e, assim, aumentar a confiança nos resultados da pesquisa. No entanto, replicações raramente são realizadas, especialmente nas ciências comportamentais. Neste artigo, examinamos os resultados do *Open Science Reproducibility Project: Psychology*, que replicou 100 experimentos publicados anteriormente nas principais revistas de Psicologia. Com base nas conclusões deste projeto, apresentamos sete recomendações a acadêmicos da área de Administração Pública que podem melhorar a qualidade de seus experimentos.

**Palavras-chave:** método experimental nas ciências comportamentais; métodos de pesquisa em administração pública; reprodutibilidade.

## El método experimental en la Administración Pública: algunas lecciones de las repeticiones en Psicología


En los últimos años, han aparecido en revistas de Administración Pública un número creciente de estudios que adoptan el método experimental. Se ha argumentado que la ventaja de los experimentos en las ciencias del comportamiento es que permiten a los investigadores controlar los factores contextuales mientras investigan el efecto de las manipulaciones en una variable de interés. Otro argumento es que los experimentos se pueden reproducir y, por tanto, aumentar la confianza en los resultados de la investigación. Sin embargo, las repeticiones rara vez se realizan, especialmente en las ciencias del comportamiento. En este artículo examinamos los resultados del *Open Science Reproducibility Project: Psychology*, que repitió 100 experimentos publicados anteriormente en las principales revistas de Psicología. Con base en los hallazgos de este proyecto, presentamos siete recomendaciones a los académicos del área de Administración Pública que pueden mejorar la calidad de sus experimentos.

**Palabras clave:** método experimental en ciencias del comportamiento; métodos de investigación en administración pública; reproducibilidad.

DOI: <http://dx.doi.org/10.1590/0034-761220200746>

Article received on September 28, 2020 and accepted on March 08, 2021.

[Original version]

ISSN: 1982-3134 

## 1. INTRODUCTION

Grimmelikhuijsen, Jilke, Olsen and Tummers (2017) noted that eminent Public Administration scholars such as Herbert Simon and Dwight Waldo have stressed the importance Psychology in Public Administration research, but only recently this has been acknowledged more frequently. The authors also noticed that, in recent years, “between 3 percent and 11 percent of all published articles [in Public Administration Journals] are informed by Psychology, a share that has been increasing in recent years” (Grimmelikhuijsen et al., 2017, p. 46). Social Psychology has an especially close affinity to Public Administration whereas both fields seek to understand how the social context influences individual behavior (Sobis & De Vries, 2014).

More recently, Psychology has inspired Public Administration research not only regarding theoretical perspectives but also regarding method. For example, Tepe and Prokop (2017, p. 159) argue that “[...] the setup, conventions, and measurement techniques in psychological experiments provide optimal conditions to explore cognitive evaluation and decision processes” in public management. It should be noted that Social Psychology is split between two competing epistemologies: “social constructionism” and “experimental social psychology” (Jost & Kruglansk, 2002). The latter has been the dominant methodological approach and the one that have influenced Public Administration research; some scholars even speak of an “experimental turn” in the field (Jilke, Van de Walle & Kim, 2016). According to Bouwman and Grimmelikhuijsen (2016), the reason why Public Administration research could benefit from experimental designs is the possibility of controlling for endogeneity and simultaneity. The first problem emerges in the analysis of observational data due to intervening variables that usually remain unaccounted for. The simultaneity problem takes place when the researcher cannot determine the direction of the cause-effect relationship.

Experiments have long been regarded to be advantageous for allowing the researcher to control the variables of interest.<sup>1</sup> This is an important point but one has to acknowledge that the experimental method in the behavioral sciences presents some challenges. As we will discuss in detail below, recent attempts to replicate experiments in the field of Psychology have raised a number of questions about the external validity of research outcomes, i.e., whether results can hold when measurements are made in similar but not identical situations, and whether variables have been sufficiently kept under control.

Recent recommendations for conducting experiments in Public Administration have focused mainly on the advantages of the method (see Grimmelikhuijsen et al., 2017; James, Jilke & Van Ryzin, 2017). Although some limitations of the experimental method have been discussed (e.g., Van de Walle, 2016), as more Public Administration scholars are considering the experimental method to advance the field’s knowledge, it is important to discuss this methodological approach in light of the recent

---

<sup>1</sup> In this article we consider only the experimental methods that involve a control group and the manipulation of one or more variables, including laboratory and survey experiments. Although field experiments also involve the manipulation of one or more variables, these are not discussed here due to the specific features of the method and its relative limited use *vis-à-vis* other experimental methods. For a discussion on field experiments, see Baldassarri and Abascal (2017).

findings in the field of Psychology. Thus, our main reference for this discussion will be the “Open Science Reproducibility Project: Psychology” (Open Science Collaboration, 2015), a collaborative project that replicated 100 Cognitive and Social Psychology studies.

This article aims to answer to the following questions: What lessons Public Administration scholars can learn from the successes and failures of experiments in Social Psychology? Which precautions should these researchers take when opting for the experimental method to investigate problems in Public Administration?

Before we begin our discussion, it is important to clarify the meanings of key terms used throughout this article. The term “replication” or “replicated” refers to experimental studies that follow the same methodological procedures as the original study regardless of the outcome. The term “reproduction” or “reproduced” means that study was replicated and yielded the same results as the original, i.e., that the replication was successful (see Open Science Collaboration, 2015).

In the next section we examine the methodological characteristics of the experimental method in the context of the behavioral sciences. In section 3, we discuss the results of the “Open Science Reproducibility Project: Psychology” (Open Science Collaboration, 2015) and outline the lessons they bring to Public Administration research. Finally, in section 4, we present a synthesis of our recommendations and final remarks.

## 2. THE PRESUMED MERITS OF THE EXPERIMENTAL METHOD

The experimental method is now dominant in Social Psychology research and it is gaining traction in Public Administration. It is a classic and well-developed method that has amassed uncountable successes in several scientific fields. Experiments enable the researcher to hold under control many contextual and interfering factors, enabling the investigator to assess the impact of one or more factors on the variable of interest (see Shadish, Cook & Campwell, 2002). The classic experimental design — the “posttest-only” design — has at least one experimental group and one control group. Of utmost importance in this method is the control group, which should be identical to the experimental group on all relevant aspects in order to assure that changes in the outcome variable are only due to the experimental manipulation of the factor. Researchers try to achieve this by selecting participants sharing the same demographic characteristics (age, gender, education, race) and by randomly assigning subjects to either the experimental or control group.

A second merit of the experimental method is that it requires an operational definition, that is, the researcher needs to use well-defined variables. The underlying research problem might appear abstract — e.g., whether training increases public servants’ empathy towards service users. That is why the researcher must first specify what is to be regarded as “training” and “empathy” and how these factors are to be measured before testing her hypothesis. This operational definition is what allows the replication of experiments. It is expected that the replication will provide the same outcome as the original experiment when the variables are operationally defined and the experimental design is the same.

The characteristics of the experimental method — controlling for confounding variables, operational definition of variables, and enabling replication — can only work appropriately on the

assurance that research findings are objective and unrelated to personal opinion, bias, or prejudice of the researcher.

Notwithstanding its advantages, experiments are liable to some pitfalls. Firstly, although experiments presuppose that results can be reproduced, replication is rarely pursued. We will discuss why this is the case further below.

Brandt et al. (2014, p. 218) have presented the following steps for replicating experiments in the field of Social Psychology:

- “1. Carefully defining the effects and methods that the researcher intends to replicate;
2. Following as exactly as possible the methods of the original study (including participant recruitment, instructions, stimuli, measures, procedures, and analyses);
3. Having high statistical power;
4. Making complete details about the replication available, so that interested experts can fully evaluate the replication attempt (or attempt another replication themselves);
5. Evaluating replication results, and comparing them critically to the results of the original study.”

Even when such procedures are followed, the outcomes of replications can be disappointing. The academic literature offers a few explanations on why the reproduction of experiments is often frustrated. One explanation is known as the “Experimenter Bias Effect” (EBE). Rosenthal and Fode (1963) noted that experiments in which the researcher was convinced of the correctness of the underlying hypothesis corroborated it more often as compared to researchers who were doubtful about the correctness of the hypothesis. However, failures to replicate the EBE bias have also been reported (Barber et al., 1969; Jacob, 1968).

Another and more important explanation for failure to reproduce experimental results is “publication bias”. Because academic journals understandably seek to publish innovative findings, studies that show the correctness of a given hypothesis find their way to academic journals more easily than null results (Ioannidis, Munafo, Fusar-Poli, Nosek & David, 2014; Nosek, Spies & Motyl, 2012). However, novel and unexpected results are also more likely to be statistical flukes (see Backer, 2016). A more general problem is that replications of experiments — either successful or unsuccessful — are unlikely to appear on the pages of well-regarded research journals.

Replications may show that results do not hold in a context other than of the original experiment, exposing the lack of external validity of the findings. Experiments involving subjects with specific features (e.g., undergraduate students) may yield results that are not valid for other types of subjects (such as public servants). Moreover, in the behavioral sciences, experimental findings may be valid within a specific cultural setting but not in other contexts.

Notwithstanding the benefits that the experimental method can bring to Public Administration research, it is worth looking at some of its issues that the “Open Science Reproducibility Project: Psychology” has exposed (Open Science Collaboration, 2015).

### 3. LESSONS FROM THE “OPEN SCIENCE REPRODUCIBILITY PROJECT: PSYCHOLOGY”

In recent years, the field of Social Psychology was witnessed several reports of research misconduct that resulted in a number of publication retractions and the destruction of academic careers (see Van Kolfshoeten, 2014). Since the scandals broke-out, researchers began to wonder whether “a scientific culture that too heavily favors new and counterintuitive ideas over the confirmation of existing results” was to blame (Carpenter, 2012, p. 1558). This context paved the way for a large-scale replication project known as the “Open Science Reproducibility Project: Psychology” (Open Science Collaboration, 2015). Launched in 2012, this project involved more than 270 researchers from several institutions around the world and sought to replicate 100 Psychology studies published in 2008 in three of the field’s most respected journals.<sup>2</sup> In order to assess the success or failure to replicate the original experiments, the replication teams adopted several criteria, including “significance and  $P$  values, effect sizes, subjective assessments of replication teams, and meta-analyses of effect sizes” (Open Science Collaboration, 2015; p. aac4716-2).<sup>3</sup>

The results of the Reproducibility Project sparked a heated debate among Psychology scholars and beyond. While 97% of the original studies had significant results ( $P < 0.05$ ), only 36.1% of the replications reached this standard. The investigation teams also found that in replications the mean for the effect sizes were about half of that found in the original studies ( $M = 0.197$ ,  $SD = 0.257$  *vis-à-vis*  $M = 0.403$ ,  $SD = 0.188$ ). Replication results showed that Cognitive Psychology experiments were far more “reproducible” than Social Psychology studies. While 50% of the Cognitive Psychology experiments were reproduced at the  $P < 0.05$  criterion, only 25% of the Social Psychology experiments filled this criterion. The results of the replications of the original studies published in the *Journal of Personality and Social Psychology (JPSP)* were even more disappointing: in the original experiments the mean for effect sizes was 0.29 ( $SD = 0.10$ ), while for the corresponding replications it was only 0.07 ( $SD = 0.11$ ), i.e., more than four fold smaller than in the original studies.<sup>4</sup>

It is important to stress that the Reproducibility Project: Psychology was the first replication effort to be conducted at this scale. Thus, there is no previous parameter to which the outcomes of these replications could be compared. The jolt that followed the publication of the project’s results is likely to have more to do with the unrealistic expectations of researchers than to a sober-minded assessment of the matter. The authors of the Reproducibility Project were prompt to elucidate this point:

Because reproducibility is a hallmark of credible scientific evidence, it is tempting to think that maximum reproducibility of original results is important from the onset of a line of inquiry

<sup>2</sup> The reports for all replications included in the Reproducibility Project are available at the Open Science Framework website (Retrieved from <https://osf.io/ezcuj/>).

<sup>3</sup> More details on the statistical methods used to evaluate the results of the replication effort can be found in the “statistical analysis” section (Open Science Collaboration, 2015, pp. aac4716-2–aac4716-4)

<sup>4</sup> A more recent study that replicated 21 social science experimental studies that were previously published in *Nature* and *Science* found that 62% of the replications were in the same direction as the original studies and that the average effect sizes were 50% of the original studies (Camerer et al., 2018).

through its maturation. This is a mistake. If initial ideas were always correct, then there would hardly be a reason to conduct research in the first place. A healthy discipline will have many false starts as it confronts the limits of present understanding (Open Science Collaboration, 2015, p. aac4716-7).

As the authors of the replication effort pointed out, neither a successful nor an unsuccessful replication could provide definitive answers regarding the original experimental results (Open Science Collaboration, 2015). Reproduction does not imply that the theoretical interpretation is correct, but only that the results appear to be reliable. On the other hand, failure to reproduce does not mean that the original finding is a false positive. “Replications can fail if the replication methodology differs from the original in ways that interfere with the observing effect”; in addition, “unanticipated factors in the sample, setting, or procedure could still have altered the observed effect magnitudes” (Open Science Collaboration, 2015, p. aac4716-6). At the end, the authors of the replication effort concluded that the project could not establish whether any of the studies’ effects were true or false, adding that only the cumulative results from multiple replications could validate the effects of the original studies.

This brings us to the first recommendation for Public Administration scholars emerging from the outcomes of the Reproducibility Project: considering that reproducibility is an essential component of experimental research, it should be a component of the research design from the start. This means that scholars should engage in collaborative research projects where different teams of researchers would conduct the same experiment using the same methodological procedures. The findings emerging from collaborative investigations would be more likely to find their way to respected academic journals. In addition, because researchers in collaborative projects need to agree on what to investigate, these studies would be more likely to focus on relevant issues for the field of Public Administration.

As mentioned above, Social Psychology studies were less likely to be reproduced than Cognitive Psychology studies (Open Science Collaboration, 2015). One possible reason for this outcome is that the latter tended to use within-subjects research designs and repeated measurements more often than the former. It may be too early to conclude that within-subjects is the best approach for experimental studies in Public Administration, but the matter should not be ignored altogether. Therefore, a second recommendation is that Public Administration scholars should investigate the effect of different research designs —between-subjects *versus* within-subjects, posttest-only *versus* pretest/posttest<sup>5</sup> — on research outcomes.

We will now examine the outcomes of a selection of studies included in the “Open Science Reproducibility Project: Psychology”. We selected 35 replication reports using the following criteria: (a) all the replications of studies originally published in the *Journal of Personality and Social Psychology*,

---

<sup>5</sup> In the posttest only experimental designs the variable of interest is measured in the control group and in the experimental group after the experimental manipulation takes place. In the pretest-posttest experimental design the variable of interest is measured in the control group and in the experimental group before and after the manipulation in the experimental group takes place. This allows the experimenter to assess the baseline measurement of the experimental group, as well as to identify any influence of the experiment on the data, increasing the reliability of the results. See American Psychological Association (APA, n.d.)



due to the connections between Social Psychology and Public Administration; and (b) replications of studies investigating topics that have relevance for the field of Public Administration (values, optimism, communal responsiveness, and conflict). The Excel file listing the 35 original studies of our sample, the replications' outcomes, and excerpts from the replication reports is available at the Open Science Foundation storage website: <https://osf.io/ta746/>.

One aspect worth noticing in this sample is that, in many cases, the replication experimenters did not have a straightforward answer on whether the study was indeed reproduced or not. Although the replication teams applied the evaluation criteria established by the Reproducibility Project, the comments included in the reports show that the picture is less “black-and-white” than one would initially expect. Sometimes the team was able to replicate the main effect, but not some of the additional effects (most studies included multiple experiments). In other replications the effect was in the same direction as seen in the original study but results did not pass the significance criterion ( $P < 0.05$ ). The replication teams usually used samples with a larger number of subjects than the original studies, assuring enough power to detect the alleged effect; in two cases, however, experimenters admitted that their replications lacked sufficient power and, for this reason, considered the replications results inconclusive.

This brings us to the debate about  $P$  values. This debate is not new, going back to Rozeboom's criticism of the null-hypothesis significance test (NHST) (Rozeboom, 1960); the controversy seems far from being settled (see Harlow, Mulaik & Steiger, 2016). The concern over the  $P$  values escalated to the point of moving the American Statistical Association to issue guidelines for its use; it was the first time the association, founded in 1839, has issued such guidelines (Wasserstein & Lazar, 2016). The editors of the *Journal of Basic and Applied Social Psychology* went as far as to banning the use of  $P$  values from the articles published in the journal (Trafimow, 2014; Trafimow & Marks, 2015).

Many statisticians have argued that the  $P$  value is unable to tell anything about the veracity of a given hypothesis. According to Goodman (2008, p. 136), this is due to the very definition of the  $P$  value: “The probability of the observed result, plus more extreme results, if the null hypothesis were true”. The  $P$  value can only make a statement regarding whether the null hypothesis is to be rejected or not, but not about the actual veracity of the alternative hypothesis. According to Goodman, this is just one of the many misinterpretations involving the  $P$  value. He argues that Fisher — the mathematician who introduced NHST — used the term “significance” to mean “worthy of attention in the form of meriting more experimentation, but not proof in itself” (Goodman, 2008, p. 135). Thus, a  $P < 0.05$  does not warrant that  $H_1$  is true. Goodman sustains that the “marriage” between  $P$  value and hypothesis testing was an “unnatural union”. Benjamin et al. (2018) argue that the  $P < 0.05$  threshold yields too many false positives and that the standard for claims of new discoveries should be tightened to  $P < 0.005$ . Armhein and Greenland (2018, p. 4) replied saying that this would only aggravate current problems and proposed instead that “[...] statistics reform should involve completely discarding ‘significance’ and the oversimplified reasoning it encourages”. If one wants to demonstrate that a reliable effect exists, then one should show that there is a robust effect size. Therefore, our third recommendation for conducting experiments in the field of Public Administration is to focus on the effect size and not on the  $P$  value (relevance *versus* significance). This recommendation may not be

easy to follow in the current academic context due to the widespread use of NHST. Experiments cost time and money and scholars hold the reasonable expectation that if the  $P < 0.05$  criterion is satisfied, then there is sufficient reason to publish the results. Nevertheless, this expectation is also what nudges researchers to engage in  $P$ -hacking (see Lindsay, 2020). Nevertheless, at the end, only a strong effect size — along with replications — can support a given hypothesis, unless the main hypothesis is that there is no effect, which rises yet another problem.

Howell (2012, p. 230) noted that experimenters “[...] have only a small chance of finding the effect they are looking for, even if such an effect does exist in the population”. This is the reason why it is important that experiments have enough power. Sufficiently powered experiments are more likely to correctly reject  $H_0$ , reducing the occurrence of type II error. However, if it is not easy to detect an effect, it might be tempting to engage in the “proof of the null hypothesis”, which is another issue that springs from NHST. It is worth to illustrate this point with an example from the field of Public Administration.

Moynihan (2013) designed an experiment to investigate whether higher levels of Public Service Motivation (PSM) (see Perry & Wise, 1990), were associated to budget maximization. According to the author, if this association could be demonstrated, Niskanen’s budget maximization theory would be vindicated (see Niskanen, 1968). It should be noted, however, that Niskanen did not assume that PSM was behind budget maximization. He argued that bureaucrats maximize their budgets because this would bestow prestige on them, which is another way of saying that bureaucrats are as self-interested as everyone else. Moynihan explained the twist in his assumption in the following terms: “Why is a public-spirit budget maximizer more plausible? Bureaucrats might seek to maximize budgets because they sincerely believe in the benefits of their programs” (Moynihan, 2013, p. 182). Although Moynihan’s assumption is disputable, we will focus only on the terms of his experiment. He recruited undergraduate students as subjects and applied the required experimental manipulations. The results did not show a significant linear correlation between budget allocation and levels of PSM, even when outliers were removed from the regression model; thus,  $H_0$  could not be rejected at  $P < 0.05$ . Moynihan affirmed to have conducted a power analysis and that the sample size was enough to detect an effect, although he did not inform the actual power of the experiment. Nevertheless, his conclusion went a long way: he argued that the experiment’s results were “[...] a significant non-significant finding” (Moynihan, 2013, p. 190) that dismantled “[...] another pillar for the budget maximization model” (Moynihan, 2013, p. 190). This corresponds to a classical case of the “proof of the null hypothesis”.

Fisher, in his classical 1935 book, *The Design of Experimentation*, argued that “[...] the null hypothesis is never proved or established, but it is possibly disproved, in the course of experimentation.” (Fisher as cited in Lehman, 2011, p. 64). Since then, statisticians have been alerting about the “proof of the null hypothesis” misinterpretation (see M. P. Lecoutre, Poitevineau & B. Lecoutre, 2003). Failing to detect an effect when there is one is the definition of type II error, but this is not a proof of the null hypothesis. An underpowered experiment will increase the chances for type II error, but this — as expected — cannot prove that the null hypothesis is true. If the experimenter, however, is committed to conducting a powered experiment, still it is not possible to prove that the null hypothesis is true due to the very definition of *power*, which is: “[...] the probability of correctly rejecting a false  $H_0$  when a particular alternative hypothesis is true” (Howell, 2012, p. 230). Thus, a powered experiment is one



that may reject the null hypothesis, but it cannot prove that there is no effect. It is not even possible to calculate a powered sample size because this requires estimating an effect size that is larger than zero (see Howell, 2012). The best alternative to the “proof of the null hypothesis” is to show that the effect size is too small to have any relevant implication to the research problem at hand. Equivalence and noninferiority testing may also be an option (see Streiner, 2003). Thus, our fourth recommendation is to avoid the “prove the null hypothesis” as an experimental design.

What are the characteristics that make experiments in the behavioral sciences more likely to be reproduced? All the replications included in our purposive sample adopted a posttest design; thus it is not possible to infer if this design had any influence in either a favorable or unfavorable outcome of the replication. However, as will be discussed below, there are reasons to recommend the use of pretest-posttest designs whenever possible.

Among the 35 replications we examined, 12 experiments were successfully replicated (25%), 21 were not reproduced, and two were considered inconclusive due to the small sample of the replication. Three of the successful replications used survey designs, while only one among those not replicated adopted this design. This may suggest that survey designs are more reliable than other experimental designs, but survey experiments can also be disappointing. Nosek et al. (2012) describe a survey experiment that included 1,979 participants. As the authors report, the initial results were the dream of any researcher; the hypothesis was supported and the results appeared to be robust and reliable. But, as a matter of caution, the authors decided to replicate the experiment, collecting data from another 1,300 participants. This time around the results were disheartening: “[t]he effect-size had vanished ( $P = 0.59$ )” (p. 616). Large samples can detect weak effect sizes (Streiner, 2003) but because effect sizes can be very weak, large samples are not necessarily a guarantee for successful replications.

Overall, successfully replicated experiments in our sample tended to present simpler research designs, with fewer experimental conditions being tested. Nonetheless, there were cases of more complex designs that were reproduced and of simpler designs that were not.

Social Psychology experiments were more likely to be reproduced when the original effect sizes were robust and the experimental design used high-powered within-subjects manipulations and repeated measurements. Nevertheless, the fact that the mean for effect sizes in the replication experiments was much smaller than the original studies—especially regarding the Social Psychology studies—casts a shadow over large effect-sizes as well. This is why one experimental study alone cannot yield a definitive answer; neither one replication alone can.

Overall, the Reproducibility Project showed that counter-intuitive results and studies that required several experiments and complicated manipulations were correlated with a smaller likelihood of a successful replication (Open Science Collaboration, 2015). This leads to our fifth recommendation: whenever possible, Public Administration researchers should design experiments that are easy to replicate and avoid experimental studies that require complicated steps or too stringent assumptions. This recommendation can also lead to more straightforward results. Reality is complex and investigators strive for realistic results, which may require complex research designs. However, given that reproducibility is an important component of the experimental method, researchers need to balance complexity *versus* reproducibility.

The failure to reproduce the original results in the replications included in our sample was attributed to the following reasons: difficulty in interpreting the original methods (Lewis & Pitt, 2015); differences in sample size (Reinhard, 2014); the effect of unknown moderators (Kelson, 2015; Johnson, Hayes & Graham, 2015); effects of small differences in the experimental procedures (Baranski, 2015; Holubar, 2015; Lane & Gazarian, 2015; Lin, 2013); and, differences in participant demographics/profile/context (Brown et al., 2013; Embley, Johnson & Giner-Sorolla, 2015; Lemn, 2013; Lin, 2013; Marigold, Forest & Anderson, 2015; Mechin & Gable, 2015; Talhelm, Lee & Eggleston, 2015). Therefore, the two main reasons offered to explain the failures to reproduce the original results were different characteristics of participants and differences in the conditions of the experiments. The first issue is not necessarily a problem for Public Administration research; it is even important to know that subjects with different characteristics react differently to experimental manipulations. The second issue, however, is worrisome. Replication of Social Psychology experiments looks like a frail cairn: one has to carefully balance all the components in order to avoid the whole structure to tip over. Because Public Administration has a practical outlook, this can be a major problem. Experiments that cannot withstand the smallest variation in experimental procedures have little if any relevance for the field.

It should be noted that Social Psychology experiments rarely adopt pretest-posttest research designs; posttest-only designs predominate even in experiments that involve several manipulations. However, pretest measurements provide baseline information such as, for example, the initial level of Public Service Motivation in both the control and experimental groups (see Bellé, 2013). The importance of the pretest-posttest design can be illustrated by the replication of Vohs and Schooler's (2008) original study. The original study, which argued that belief in freewill enhance moral behavior, received great attention within the Social Psychology community for its far-reaching consequences. The news that it could not be reproduced was a disappointment among researchers. Neither the original study nor its replication included a pretest measurement. The two questionnaires relevant in one of the study's two experiments — the Freewill and Determinism scale (FWD) and the Positive and Negative Affectivity Schedule (PANAS) — were applied only after the participants had read the texts corresponding to the control and treatment manipulations (treatment = text affirming freewill is an illusion; control = neutral text about consciousness). However, if the control and treatment groups differed in regard to the baseline condition (belief in freewill), then this should have been accounted for in the analysis of the results (see Bonate, 2000). This brings a dilemma: if participants had responded to the FWD questionnaire before the experimental manipulation was applied, this would likely affect the outcome of the experiment. This problem takes place when the baseline variable is inside the head of the subjects participating in an experiment. In clinical trials, baseline levels correspond to physiological factors that can be precisely measured; in such cases, the pretest does not influence the outcome of the posttest. However, in the behavioral sciences, the elaboration of experimental designs often involves difficult choices. All things considered, Public Administration scholars should be aware that, depending on the investigation at hand, the pretest-posttest experimental design may yield more reliable results than the posttest-only design. This brings us to the sixth recommendation: whenever possible, adopt a pretest-posttest design.

Given that Social Psychology experiments are difficult to reproduce, it is reasonable to assume that experimental research should not stand as the sole source of knowledge in the field of Public

Administration. Fortunately, the field does not have a predominant research method — at least not yet — and has relied on many methodological procedures to construct the field’s scholarship. A thoroughly “experimental turn” may be a bad idea for individual researchers and especially for the field as a whole.

Let’s look at two studies investigating a recurrent theme in public management: the issue of whether management tolls yield the same results in the public and private sectors. Robertson and Seneviratne (1995) found that, overall, management change interventions were as effective in the public sector as in the private sector, but this result depended on the outcome variables considered. Banerjee and Solomon (2003, p. 119), in a study about the effectiveness of ecological labeling, argued that “[g]overnment run programs have been far more successful than the private ones. Government support to a labeling program not only increases its credibility and recognition, but also improves financial stability, legal protection and long-term viability”. The findings from these two studies would have been unlikely to emerge if the method of choice was an experiment. To begin with, the concepts of private and public sectors imply so many different aspects that it would be almost impossible to select independent variables to test the effect on the dependent variable. Regime types; intergovernmental relations; central, regional or local government; organizational structure and culture; management and leadership characteristics; motivation and behavior of administrators and staff — these are only a handful of variables that may affect how well public servants respond to incentives, job pressures, tasks, and responsibilities. One may argue that the methods used in these studies are less precise than experiments. However, as argued above, using more than one method and different sources of data can render more reliable results and have more practical relevance. Thus, our seventh recommendation is that Public Administration researchers should avoid relying exclusively on the experimental method.

Finally, it is necessary to mention that the Reproducibility Project: Psychology results have generated some controversy. Gilbert, King, Pettigrew and Wilson (2016) argued that the project failed to account for random error, sustaining that, in addition to the 5% error expected due the 95% confidence interval, the replications used samples from different populations and, in some cases, the replication procedures did not strictly follow the original studies in other aspects as well. On the other hand, Gelman (as cited in Baker, 2016) argued that replications are more reliable guides than original studies because the latter are more likely to be the statistical flukes as compared to replications, particularly due to publication bias. Replications are more thoughtful and planned endeavors, argues Gelman, while original studies showing strong effect sizes tend to find their way to publication too quickly.

#### 4. CONCLUSIONS

This paper addressed the promises of what has been called the “experimental turn” in Public Administration. The argument for the experimental method is that, just as Public Administration and Psychology hold several theoretical affinities, research in the former would benefit from making more use of the experimental method. Experiments have the advantage of controlling and isolating variables as well as allowing within-subjects and between-subjects analyses, effectively controlling for confounding variables, and assuring a high degree of internal validity. Nevertheless, only several experiments yielding the same outcome can provide any confidence on the external validity.

As the “Open Science Reproducibility Project: Psychology” revealed, while experimental studies in Cognitive Psychology are not easily reproduced, Social Psychology studies are even less so (Open Science Collaboration, 2015). As more experimental studies in Public Administration are performed and failure to reproduce their results emerge, the initial enthusiasm for the experimental method among scholars may begin to wither away. This does not imply that scholars should discharge the experimental method altogether but that they should take the necessary precautions when adopting it.

Our examination of the outcomes of the Reproducibility Project led to seven recommendations to Public Administration researchers, which are summarized in Box 1.

### BOX 1 RECOMMENDATIONS DRAWN FROM PSYCHOLOGY STUDIES TO PUBLIC ADMINISTRATION SCHOLARS

- #1 Engage in collaborative research, developing the same experiments simultaneously, making replication a component of the research design.
- #2 Investigate the effect of different research designs on research outcomes.
- #3 Focus on the effect size and not on the *P* value (relevance over significance).
- #4 Avoid the “test of the null hypothesis” approach.
- #5 Whenever possible, choose design experiments that are easy to replicate.
- #6 Consider using pretest-posttest experimental designs when this is feasible.
- #7 Avoid relying exclusively in experimental research to construct the knowledge in the field.

**Source:** Elaborated by the authors.

The single most important recommendation to Public Administration scholars is that replications should become an inherent component of the experimental design. The experimental method should not be an isolated endeavor but a collaborative research project in which replications of experiments in different cultures, regimes, and organizational settings is part and parcel of the research design. The advantage is that collaborative projects — regardless of the results of the experiments — can avoid publication bias, since leading academic journals are certainly interested in the results of research with this characteristic. As more experiments in the field of Public Administration become available, meta-analytical studies also become more feasible. This is of paramount importance, given that even underpowered experiments, when combined, can increase the power of the analysis (Cooper, 2017).

Public Administration scholars should learn from the pitfalls of experiments in Psychology in regard to some specific technical issues as well, such as focusing on the effect size instead of *P* value. There is a reason why the NHST has become so controversial over the years, as we have discussed above. Moreover, while detecting an effect may not be easy, this should not be an incentive to engage in the “proof the null hypothesis”. As discussed above, it is not possible to prove that the null hypothesis

is true even when there is enough power in an experiment to detect an effect. This is because the statistical technique used to calculate powered sample sizes needs to assume an effect size larger than zero (see Howell, 2012). Equivalence and noninferiority testing may be an option when a researcher wants to demonstrate, for example, that one training program is not better than other (Streiner, 2003). Pretest-posttest experimental designs should be used more often, although measuring variables that are “inside the heads” of subjects represent a challenge for this research design.

Not all research topics are suitable to be investigated through experiments. The experimental method requires variables that are already firmly established and measurement instruments that have been properly validated. Public Administration is a discipline that emerged from a specific context: governmental organizations and its interactions with the public. Thus, it is also recommended that experiments testing hypothesis related to the field’s problems use public servants as subjects whenever possible. The outcomes of experiments using undergraduate students enrolled in Western universities as subjects are unlikely to serve as a solid base for practical interventions.

Public Administration scholars not only ask about the “what” and “why” of matters but also about the “how”. The knowledge generated through research in our field needs to be applied to practical problems. It may be reasonable to argue that replications fail because experiments have not followed exactly the same procedures as the original studies. However, in Public Administration, experiments have to be robust enough to withstand some degree of contextual differences. Public Administration investigators deal with a large array of factors and the interrelatedness between them. Therefore, focusing on a hand full of variables may not be sufficient. Experiments may be useful to address issues that are relevant for practitioners and decision-makers, but other methods would still be necessary to investigate knotty social issues. As to experimental designs, complex factorial designs may be necessary, but this does not mean that a straightforward approach can be overlooked.

Finally, our implicit argument throughout this article is that Public Administration scholars need to engage in meta-science, i.e., use scientific tools to reflect about research itself. Understandably, meta-science has gained renewed relevance in the behavioral sciences after the reproducibility crisis in Psychology (see Passmore & Chae, 2019).

All in all, the field of Public Administration can benefit from the experimental method by being aware of its difficulties. If the field is moving towards an “experimental turn”, may this turn be performed with caution and a sense of direction.



## REFERENCES

- American Psychological Association. (s.d.). *APA Dictionary of Psychology*. Retrieved from <https://dictionary.apa.org/pretest-posttest-design>
- Armhein, V., & Greenland, S. (2018). Remove, rather than redefine, statistical significance (correspondence). *Nature Human Behaviour*, 4(1), 4.
- Baker, M. (2016, March 03). Psychology's reproducibility problem is exaggerated — say psychologists. *Nature News*. Retrieved from <http://www.nature.com/news/psychology-s-reproducibility-problem-is-exaggerated-say-psychologists-1.19498#/b4>
- Baldassarri, D., & Abascal, M. (2017). Field experiments across the social sciences. *Annual Review of Sociology*, 43, 41-73.
- Banerjee, A., & Solomon, B. D. (2003). Eco-labeling for energy efficiency and sustainability: a meta-evaluation of US programs. *Energy Policy*, 31, 109-123.
- Baranski, E. (2015). *Replication of "On the relative independence of thinking biases and cognitive ability" by KE Stanovich, RF West (2008, Journal of Personality and Social Psychology)*. Retrieved from <https://osf.io/p3gz2/>
- Barber, T. X., Calverley, D. S., Forgiione, A., McPeake, J. D., Chaves, J. F., ... Bowen, B. (1969). Five attempts to replicate the experimenter bias effect. *Journal of Consulting and Clinical Psychology*, 33(1), 1-6.
- Bellé, N. (2013). Experimental evidence on the relationship between public service motivation and job performance. *Public Administration Review*, 73(1), 143-153.
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... Cesarini, D. (2018). Redefine statistical significance. *Nature - Human Behaviour*, 2(1), 6-10.
- Bonate, P. L. (2000). *Analysis of pretest-posttest designs*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Bouwman, R., & Grimmelikhuijsen, S. (2016). Experimental public administration from 1992 to 2014: a systematic literature review and ways forward. *International Journal of Public Sector Management*, 29(2), 110-131.
- Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, 50, 217-224.
- Brown, B., Brown, K., Attridge, P., DeGaetano, M., Hicks, G., Humphries, D. ... Mainard, H. (2013). *Replication of Study 5 by Centerbar, Schnall, Clore, & Gavin, (2008, JPSP)*. Retrieved from <https://osf.io/wcgx5/>
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... Altmejd, A. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637-644.
- Carpenter, S. (2012). Psychology's bold initiative. *Science*, 335(6076), 1558-1561.
- Cooper, H. M. (2017). *Research synthesis and meta-analysis: A step-by-step approach* (5a ed.). London, UK: Sage.
- Embley, J., Johnson, L. G., & Giner-Sorolla, R. (2015). *Replication of Study 1 by Vohs & Schooler (2008, Psychological Science)*. Retrieved from <https://osf.io/2nf3u/>
- Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science" (Technical Comments). *Science*, 351(6277): 1037b.
- Goodman, S. (2008, July). A dirty dozen: twelve p-value misconceptions. *Seminars in Hematology*, 45(3), 135-140.
- Grimmelikhuijsen, S., Jilke, S., Olsen, A. L., & Tummers, L. (2017). Behavioral public administration: Combining insights from public administration and psychology. *Public Administration Review*, 77(1), 45-56.
- Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (2016). What if there were no significance tests? (Original work published 1997). London, UK: Routledge.
- Holubar, T. (2015). *Replication of "The rejection of moral rebels," Study 4, by Monin, Sawyer, & Marquez (2008, JPSP)*. Retrieved from <https://osf.io/a4fmg/>
- Howell, D. C. (2012). *Statistical methods for psychology*. Belmont, UK: Wadsworth/Cengage Learning.

- Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences*, 18(5), 235-241.
- Jacob, T. (1968). The experimenter bias effect: a failure to replicate. *Psychonomic Science*, 13(4), 239-240.
- James, O., Jilke, S. R., & Van Ryzin, G. G. (2017). Behavioural and experimental public administration: Emerging contributions and new directions. *Public Administration*, 95(4), 865-873.
- Jilke, S., Van de Walle, S., & Kim, S. (2016). Generating Usable Knowledge through an Experimental Approach to Public Administration. *Public Administration Review*, 76(1), 69-72.
- Johnson, K. M., Hayes T., & Graham, J. (2015). *Replication of Study 2 by Amodio, Devine, & Harmon-Jones (2008, Journal of Personality and Social Psychology)*. Retrieved from <https://osf.io/ysxmf/>
- Jost, J. T., & Kruglanski, A. W. (2002). The estrangement of social constructionism and experimental social psychology: History of the rift and prospects for reconciliation. *Personality and Social Psychology Review*, 6(3), 168-187.
- Kelson, K. (2015). *Replication of "The space between us: Stereotype threat and distance in interracial contexts" by P.A. Goff, C.M. Steele, and P.G. Davies (Journal of Personality and Social Psychology, 2008)*. Retrieved from <https://osf.io/7q5us/>
- Lane, K., & Gazarian, D. (2015). *Replication of "The effects of an Implemental mind-set on attitude strength" by Henderson, de Liver, & Gollwitzer (2008, Journal of Personality and Social Psychology)*. Retrieved from <https://osf.io/xqjf4/>
- Lecoutre, M. P., Poitevineau, J., & Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. *International Journal of Psychology*, 38(1), 37-45.
- Lehmann, E. L. (2011). *Fisher, Neyman, and the creation of classical Statistics*. New York, NY: Springer.
- Lemn, K. (2013). *Replication of Blankenship and Wegener (2008, JPSP, Study 5A)*. Retrieved from <https://osf.io/v3e2z/>
- Lewis, M., & Pitts, M. (2015). *Replication of "Errors are Aversive" by Greg Hajcak & Dan Foti (2008, Psychological Science)*. Retrieved from <https://osf.io/tkq9n/>
- Lin, S. (2013). *Replication of Study 7 by Exline, Baumeister, Zell, Kraft & Witvliet (2008, Journal of Personality and Social Psychology)*. Retrieved from <https://osf.io/svz7w/>
- Lindsay, D. S. (2020). Seven steps toward transparency and replicability in psychological science. *Canadian Psychology/Psychologie canadienne*, 61(4), 310-317. Retrieved from <https://doi.org/10.1037/cap0000222>
- Marigold, D. C., Forest, A. L., & Anderson, J. E. (2015). *Replication of "How the head liberates the heart: Projection of communal responsiveness guides relationship promotion" by EP Lemay Jr and MS Clark (2008, JPSP)*. Retrieved from <https://osf.io/mv3i7/>
- Mechin, N., & Gable, P. (2015). *Replication of "Left frontal cortical activation and spreading of alternatives: Test of the action-based model of dissonance" by E Harmon-Jones, C Harmon-Jones, M Fearn, JD Sigelman, P Johnson (2008, Journal of Personality and Social Psychology)*. Retrieved from <https://osf.io/zpwne/>
- Moynihan, D. P. (2013). Does public service motivation lead to budget maximization? Evidence from an experiment. *International Public Management Journal*, 16(2), 179-196.
- Niskanen, W. A. (1968). The peculiar economics of bureaucracy. *The American Economic Review*, 58(2), 293-305.
- Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6), 615-631.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716-1-aac4716-8.
- Passmore, D. L., & Chae, C. (2019). Potential for meta-scientific inquiry to improve the usefulness of HRD research outcomes for practice. *Advances in Developing Human Resources*, 21(4), 409-420.
- Perry, J. L., & Wise, L. R. 1990. The Motivational Bases of Public Service. *Public Administration Review*, 50(3), 367-73

- Reinhard, D. (2014). *Replication of Förster, J., Liberman, N., & Kuschel, S. (2008). The effect of global versus local processing styles on assimilation versus contrast in social judgment, Journal of Personality and Social Psychology, 94, 579-599.* Retrieved from <https://osf.io/mxryb/>
- Robertson, P. J., & S. J. Seneviratne (1995). Outcomes of planned organizational change in the public sector: a meta-analytic comparison to the private sector. *Public Administration Review, 55*(6), 547-558.
- Rosenthal, R., & Fode, K. L. (1963). The effect of experimenter bias on the performance of the albino rat. *Systems Research and Behavioral Science, 8*(3), 183-189.
- Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57*(5), 416-428.
- Shadish, W., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston, MA: Houghton Mifflin Company.
- Sobis, I., & De Vries, M. S. (2014). The social psychology perspective on values and virtue. In M. S. De Vries, & P. S. Kim (Eds.), *Value and virtue in Public Administration: a comparative perspective* (IIAS Series: Governance and Public Management). London, UK: McMillan-Palgrave.
- Streiner, D. L. (2003). Unicorns do exist: A tutorial on “proving” the null hypothesis. *The Canadian Journal of Psychiatry, 48*(11), 756-761.
- Talhelm, T., Lee, M., & Eggleston, C. (2015). *Replication of Poignancy: Mixed Emotional Experience in the Face of Meaningful Endings by Ersner-Hershfield, Mikels, Sullivan, & Carstensen (2008, Journal of Personality and Social Psychology)*. Retrieved from <https://osf.io/fw6hv/>
- Tepe, M., & Prokop, C. (2017). Laboratory experiments: their potential for public management Research. In O. James, S. R. Jilke, & G. G. Van Ryzin (Eds.), *Experiments in public management research: challenges and contributions* (1st ed., pp. 139-164) Cambridge, UK: Cambridge University Press.
- Trafimow, D. (2014). Editorial. *Basic and Applied Social Psychology, 36*(1), 1-2. Retrieved from <https://doi.org/10.1080/01973533.2014.865505>
- Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*(1), 1-2. Retrieved from <http://dx.doi.org/10.1080/01973533.2015.1012991>
- Van de Walle, S. (2016). The experimental turn in public management: How methodological preferences drive substantive choices. In O. James, S. Jilke, & O. Van Ryzin (Eds.), *Experiments in public management research*. Cambridge, UK: Cambridge University Press.
- Van Kolfshoeten, F. (2014). Fresh misconduct charges hit Dutch social psychology. *Science Magazine, 344*(6184), 566-567.
- Vohs, K. D., & Schoolre, J. W. (2008). The value of believing in free will: encouraging a belief in determinism increases cheating. *Psychological Science, 19*(1), 49-54.
- Wasserstein, R., & Lazar, N. (2016). The ASA’s statement on P values: context, process, and purpose. *The American Statistician, 70*(2), 129-133.

### Christina W. Andrews



<https://orcid.org/0000-0003-4461-8892>

Ph.D. in Political Science; Associate Professor in the Department of Social Sciences at the Federal University of São Paulo. E-mail: [christina.andrews@unifesp.br](mailto:christina.andrews@unifesp.br)

### Michiel S. de Vries



<https://orcid.org/0000-0001-6827-2852>

Studied Sociology at the University of Groningen; Worked previously at the University of Amsterdam, Thorbecke Academy in Leeuwarden, and the Free University Amsterdam; Ph.D. in Law at Utrecht University; In 2016 he received a Doctor Honoris Causa in Public Economics from the Masaryk University of Brno; Full professor in Public Administration at the Radboud University of Nijmegen; Distinguished professor in Public Administration at Kaunas University of Technology, and an extraordinary professor at the Free State University in Bloemfontein; He is past president of IASIA and member of the editorial board of numerous journals on Public Administration. E-mail: [m.devries@fm.ru.nl](mailto:m.devries@fm.ru.nl)