FGV EBAPE

RAP
REVISTA DE ADMINISTRAÇÃO PÚBLICA

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

Christina W. Andrews [1]
Michiel S. de Vries [2]

[1] Universidade Federal de São Paulo / Departamento de Ciências Sociais, São Paulo / SP – Brasil
[2] Radboud University / Institute for Management Research, Nijmegen – Países Baixos

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.

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

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.

DOI: http://dx.doi.org/10.1590/0034-761220200746
Artigo recebido em 28 set. 2020 e aceito em 08 mar. 2021.
[Versão traduzida]
ISSN: 1982-3134

1. INTRODUÇÃO

Grimmelikhuijsen, Jilke, Olsen e Tummers (2017) observaram que eminentes acadêmicos como Herbert Simon e Dwight Waldo destacaram a importância da Psicologia na pesquisa em Administração Pública, mas só recentemente essa importância foi reconhecida com maior frequência. Os autores também notaram que, nos últimos anos, "entre 3% e 11% de todos os artigos publicados [em revistas de Administração Pública] são orientados pela Psicologia, uma parcela que vem aumentando nos últimos anos" (Grimmelikhuijsen et al., 2017, p. 46). A Psicologia Social tem uma afinidade especialmente próxima à Administração Pública, pois ambas as áreas procuram compreender como o contexto social influencia o comportamento individual (Sobis & De Vries, 2014).

Mais recentemente, a Psicologia tem inspirado a pesquisa em Administração Pública não só no que se refere às perspectivas teóricas, mas também quanto ao método. Por exemplo, Tepe e Prokop (2017, p. 159) argumentam que "[...] a configuração, as convenções e as técnicas de medição de experimentos psicológicos fornecem condições ideais para explorar a avaliação cognitiva e os processos de decisão" na gestão pública. Deve-se apontar que a Psicologia Social está dividida entre duas epistemologias concorrentes: o "construtivismo social" e a "psicologia social experimental" (Jost & Kruglansk, 2002). A última tem sido a abordagem metodológica dominante e a que mais influenciou a pesquisa em Administração Pública; alguns acadêmicos inclusive falam de uma "virada experimental" na área (Jilke, Van de Walle & Kim, 2016). Segundo Bouwman e Grimmelikhuijsen (2016), a razão pela qual a pesquisa em Administração Pública poderia se beneficiar de projetos experimentais é a possibilidade de controlar a endogeneidade e simultaneidade. O primeiro problema surge na análise de dados observacionais devido a variáveis intervenientes que geralmente não são identificadas. O problema da simultaneidade ocorre quando o pesquisador não pode determinar a direção da relação causa-efeito.

Os experimentos há muito são considerados vantajosos por permitirem ao pesquisador controlar as variáveis de interesse.[1] Esta é uma questão importante, mas é preciso reconhecer que o método experimental nas ciências comportamentais apresenta alguns desafios. Como discutiremos em detalhes abaixo, as tentativas recentes de replicar experimentos na área da Psicologia levantaram uma série de questões sobre a validade externa dos resultados das pesquisas, ou seja, se os resultados podem ser mantidos quando as medições são feitas em situações semelhantes, mas não idênticas, e se as variáveis foram mantidas suficientemente sob controle.

Recomendações recentes para a realização de experimentos em Administração Pública têm focado principalmente nas vantagens do método (ver Grimmelikhuijsen et al., 2017; James, Jilke & Van Ryzin, 2017). Embora algumas limitações do método experimental tenham sido discutidas (por exemplo, Van de Walle, 2016), à medida que mais estudiosos da Administração Pública venham a considerar o uso do método experimental para desenvolver o conhecimento da área, é importante que esta abordagem metodológica possa ser discutida à luz das recentes descobertas no campo da Psicologia. Assim, nossa principal referência para esta discussão será o *Open Science Reproducibility*

[1] Neste artigo, consideramos apenas os métodos experimentais que envolvem um grupo de controle e a manipulação de uma ou mais variáveis, incluindo experimentos de laboratório e que usam *surveys*. Embora os experimentos de campo também envolvam a manipulação de uma ou mais variáveis, esses não são discutidas aqui devido às características específicas do método e seu uso relativamente limitado *vis-à-vis* outros métodos experimentais. Para uma discussão sobre experimentos de campo, ver Baldassarri e Abascal (2017).

Project: Psychology (Open Science Collaboration, 2015), um projeto colaborativo que replicou 100 estudos nas áreas de Psicologia Cognitiva e Psicologia Social.

O presente artigo tem como objetivo responder às seguintes perguntas: Quais são as lições que os acadêmicos da área de Administração Pública podem aprender dos sucessos e fracassos dos experimentos em Psicologia Social? Quais são os cuidados que esses pesquisadores devem tomar ao optar pelo método experimental para investigar problemas de Administração Pública?

Antes de começarmos a discussão, é importante esclarecer os significados de termos-chave que serão utilizados ao longo deste artigo. O termo "replicação" ou "replicado" refere-se a estudos experimentais que seguem os mesmos procedimentos metodológicos do estudo original, independentemente do resultado. O termo "reprodução" ou "reproduzido" significa que o estudo foi replicado e produziu os mesmos resultados que o original, ou seja, que a replicação foi bem-sucedida (ver Open Science Collaboration, 2015).

Na próxima seção examinaremos as características metodológicas do método experimental no contexto das ciências comportamentais. Na seção 3, discutiremos os resultados do projeto *Open Science Reproducibility Project: Psychology* (Open Science Collaboration, 2015) e descreveremos as lições que esses resultados trazem para a pesquisa em Administração Pública. Finalmente, na seção 4, apresentaremos uma síntese de nossas recomendações e as considerações finais.

2. OS MÉRITOS PRESUMIDOS DO MÉTODO EXPERIMENTAL

O método experimental é agora dominante na pesquisa em Psicologia Social e está ganhando força na Administração Pública. É um método clássico e bem desenvolvido que acumulou sucessos incontáveis em diversos campos científicos. Os experimentos permitem que o pesquisador mantenha sob controle diversos fatores contextuais e intervenientes, permitindo que o pesquisador avalie o impacto de um ou mais fatores na variável de interesse (ver Shadish, Cook & Campwell, 2002). O delineamento experimental clássico — o delineamento pós-teste — tem pelo menos um grupo experimental e um grupo de controle. Nesse método, o grupo de controle é de extrema importância e deve ser idêntico ao grupo experimental em todos os aspectos relevantes de forma a garantir que as mudanças na variável de interesse resultem apenas da manipulação experimental desse fator. Os pesquisadores tentam alcançar isso selecionando participantes que compartilham as mesmas características demográficas (idade, gênero, educação, raça) e designando aleatoriamente os participantes para o grupo experimental ou o grupo de controle.

Um segundo mérito do método experimental é que ele requer uma definição operacional, ou seja, o pesquisador precisa utilizar variáveis bem definidas. O problema central da pesquisa pode parecer abstrato — por exemplo, se o treinamento aumenta a empatia dos funcionários públicos em relação aos usuários do serviço. Por essa razão o pesquisador deve primeiro especificar o que deve ser considerado como "treinamento" e "empatia" e como esses fatores devem ser medidos antes de testar sua hipótese. Essa definição operacional é o que permite a replicação dos experimentos. Espera-se que a replicação forneça o mesmo resultado do experimento original quando as variáveis são operacionalmente definidas e o delineamento experimental for o mesmo.

As características do método experimental — controle de variáveis intervenientes, definição operacional de variáveis e possibilidade de replicação — só podem operar apropriadamente sob a

condição de que os resultados da pesquisa são objetivos e não estão relacionados à opinião pessoal, parcialidade ou preconceito do pesquisador.

Apesar de suas vantagens, os experimentos estão sujeitos a algumas armadilhas. Em primeiro lugar, embora os experimentos pressuponham que os resultados possam ser reproduzidos, a replicação raramente é empreendida. Discutiremos por que isso acontece mais adiante.

Brandt et al. (2014, p. 218) apresentaram as seguintes etapas para a replicação de experimentos na área da Psicologia Social:

> "1. Definir cuidadosamente os efeitos e métodos que o pesquisador pretende replicar;
> 2. Seguir da maneira mais exata possível os métodos do estudo original (incluindo recrutamento de participantes, instruções, estímulos, medidas, procedimentos e análises);
> 3. Garantir alto poder estatístico;
> 4. Disponibilizar detalhes completos sobre a replicação, para que os especialistas interessados possam avaliar em profundidade a tentativa de replicação (ou tentar outra replicação eles mesmos);
> 5. Avaliar os resultados da replicação e compará-los criticamente com os resultados do estudo original."

Mesmo quando esses procedimentos são seguidos, os resultados das replicações podem ser decepcionantes. A literatura acadêmica oferece algumas explicações sobre por que a reprodução dos experimentos muitas vezes não acontece. Uma explicação é conhecida como *Experimenter Bias Effect* (EBE). Rosenthal e Fode (1963) observaram que experimentos nos quais o pesquisador estava convencido da correção da hipótese subjacente a corroboravam com mais frequência em comparação com pesquisadores que duvidavam da correção da hipótese. No entanto, fracassos em replicar o EBE também foram relatados (Barber et al., 1969; Jacob, 1968).

Outra explicação mais importante para o fracasso na reprodução dos resultados experimentais é o "viés de publicação". Uma vez que os periódicos acadêmicos, compreensivelmente, buscam publicar descobertas inovadoras, estudos mostrando a correção de uma dada hipótese encontram mais facilmente seu caminho para os periódicos acadêmicos do que os resultados nulos (Ioannidis, Munafo, Fusar-Poli, Nosek & David, 2014; Nosek, Spies & Motyl, 2012). No entanto, resultados novos e inesperados também têm maior chance de serem acasos estatísticos (ver Backer, 2016). Um problema mais geral é que as replicações de experimentos — bem ou malsucedidas — provavelmente não aparecerão nas páginas de periódicos de pesquisa bem-conceituados.

Replicações podem mostrar que os resultados não se sustentam em um contexto diferente do experimento original, expondo a falta de validade externa dos resultados. Experimentos envolvendo indivíduos com características específicas (ex. estudantes de graduação) podem produzir resultados que não são válidos para outros tipos de participantes (como servidores públicos). Além disso, nas ciências do comportamento, as descobertas experimentais podem ser válidas em um ambiente cultural específico, mas não em outros contextos.

Não obstante os benefícios que o método experimental pode trazer para a investigação na Administração Pública, vale a pena olharmos para alguns dos problemas que o *Open Science Reprodutibility Project: Psychology* expôs (Open Science Collaboration, 2015).

3. LIÇÕES DO *OPEN SCIENCE REPRODUTIBILITY PROJECT: PSYCHOLOGY*

Nos últimos anos, a área da Psicologia Social testemunhou vários relatos de má conduta científica que resultaram em uma série de retratações de publicações e na destruição de carreiras acadêmicas (ver Van Kolfschooten, 2014). Desde que os escândalos estouraram, pesquisadores começaram a se perguntar se "uma cultura científica que favorece demasiadamente ideias novas e contraintuitivas em lugar da confirmação de resultados existentes" não seria a culpada (Carpenter, 2012, p. 1558). Este contexto abriu o caminho para um projeto de replicação em larga escala conhecido como *Open Science Reprodutibility Project: Psychology* (Colaboração Open Science de 2015). Lançado em 2012, este projeto envolveu mais de 270 pesquisadores de diversas instituições em todo o mundo e procurou replicar 100 estudos de Psicologia publicados em 2008 em três das revistas mais respeitadas da área.[2] Para avaliar o sucesso ou fracasso na replicação dos experimentos originais, a equipe responsável adotou diversos critérios, incluindo "significância e valores *P*, tamanho de efeito, avaliações subjetivas das equipes de replicação e meta-análises dos tamanhos do efeito" (Open Science Collaboration, 2015; p. Aac4716-2).[3]

Os resultados do *Reprodutibility Project* geraram um acalorado debate entre os estudiosos da Psicologia que se estendeu para além desta área do conhecimento. Enquanto 97% dos estudos originais mostraram resultados significativos ($P < 0,05$), apenas 36,1% das replicações alcançaram esse padrão. As equipes de investigação também descobriram que a média para os tamanhos do efeito nas replicações eram cerca da metade do que a constatada nos estudos originais (M = 0,197, DP = 0,257 *vis à vis* M = 0,403, DP = 0,188). Os resultados das replicações mostraram que os experimentos de Psicologia Cognitiva eram muito mais "reproduzíveis" do que os estudos de Psicologia Social. Enquanto 50% dos experimentos em Psicologia Cognitiva foram reproduzidos segundo o critério $P < 0,05$, apenas 25% dos experimentos em Psicologia Social atenderam esse critério. Os resultados das replicações dos estudos originais publicados no *Journal of Personality and Social Psychology* (*JPSP*) foram ainda mais decepcionantes: nos experimentos originais a média do tamanho de efeito foi de 0,29 (SD = 0,10), enquanto que para as replicações correspondentes a média foi de apenas 0,07 (DP = 0,11), ou seja, mais de quatro vezes menor do que nos estudos originais.[4]

É importante ressaltar que o *Reprodutibility Project*: *Psychology* foi o primeiro estudo desse tipo a ser realizado nessa escala. Assim, não há parâmetro anterior que possa ser comparado aos resultados dessas replicações. É provável que o choque que se seguiu à publicação dos resultados seja devido mais às expectativas irreais dos pesquisadores do que a uma avaliação equilibrada da questão. Os autores do *Reprodutibility Project* foram claros ao elucidarem este ponto:

> Como a reprodutibilidade é uma característica de qualidade das evidências científicas confiáveis, é tentador pensar que a reprodutibilidade máxima dos resultados originais é fundamental do

[2] Os relatórios de todas as replicações incluídas no Projeto de Reprodutibilidade estão disponíveis no site Open Science Framework (Recuperado de https://osf.io/ezcuj/).

[3] Mais detalhes sobre os métodos estatísticos usados para avaliar os resultados do esforço de replicação podem ser encontrados na seção "análise estatística" (Open Science Collaboration, 2015, pp. aac4716-2–aac4716-4).

[4] Um estudo mais recente que replicou 21 estudos experimentais na área de Ciências Sociais que foram anteriormente publicados nas revistas *Nature* e *Science* revelou que 62% das replicações estavam na mesma direção que os estudos originais e que os tamanhos de efeito médios eram 50% daqueles dos estudos originais (Camerer et al., 2018).

> início de uma linha de investigação até o seu amadurecimento. Isto é um erro. Se as ideias iniciais estivessem sempre corretas, dificilmente haveria uma razão para realizar investigações. Uma disciplina científica saudável terá muitos começos falsos ao se confrontar com os limites do entendimento atual (Open Science Collaboration, 2015, p. Aac4716-7).

Como os autores do esforço de replicação apontaram, nem uma replicação bem-sucedida nem uma malsucedida poderia fornecer respostas definitivas sobre os resultados experimentais originais (Open Science Collaboration, 2015). A reprodução não implica que a interpretação teórica esteja correta, mas apenas que os resultados parecem ser confiáveis. Por outro lado, a falha na replicação não significa que o achado original seja um falso positivo. "As replicações podem falhar se a metodologia de replicação difere da original de maneiras que interferem com efeito observado"; além disso, "fatores imprevistos na amostra, no ambiente ou no procedimento podem ainda ter alterado as magnitudes dos efeitos observados" (Open Science Collaboration, 2015, p. aac4716-6). No final, os autores do esforço de replicação concluíram que o projeto não conseguiu estabelecer se qualquer um dos efeitos dos estudos eram verdadeiros ou falsos, acrescentando que apenas os resultados cumulativos de várias replicações poderiam validar os efeitos dos estudos originais.

Isso nos leva à primeira recomendação para os pesquisadores de Administração Pública que emergem dos resultados do *Reprodutibility Project*: considerando que a reprodutibilidade é um componente essencial da pesquisa experimental, a replicação deve integrar o delineamento experimental desde o início. Isso significa que os pesquisadores devem se envolver em projetos de pesquisa colaborativos, onde diferentes equipes de pesquisadores conduziriam o mesmo experimento usando os mesmos procedimentos metodológicos. As descobertas que surgirem de investigações colaborativas teriam maior chance de chegar às revistas acadêmicas respeitadas. Além disso, uma vez que os pesquisadores de projetos colaborativos precisam chegar a um consenso sobre o que investigar, esses estudos provavelmente abordariam questões relevantes para a área de Administração Pública.

Conforme mencionado acima, os estudos de Psicologia Social foram menos propensos a serem reproduzidos do que os estudos de Psicologia Cognitiva (Open Science Collaboration, 2015). Uma possível razão para esse resultado é que esta última área de investigação tende a usar delineamentos experimentais intrassujeito e medições repetidas mais frequentemente do que o primeiro. Pode ser muito cedo para se concluir que o delineamento intrassujeito é a melhor abordagem para estudos experimentais na Administração Pública, mas a questão não deve ser totalmente ignorada. Portanto, uma segunda recomendação é que os pesquisadores em Administração Pública devem investigar o efeito de diferentes delineamentos de pesquisa — entressujeitos *versus* intrassujeito, pós-teste *versus* pré-teste/pós-teste[5] — sobre os resultados da investigação.

Vamos agora examinar agora os resultados de uma seleção de estudos incluídos no *Open Science Reproducibility Project: Psychology*. Selecionamos 35 relatórios de replicação usando os seguintes critérios: (a) todas as replicações de estudos originalmente publicados no *Journal of Personality and*

[5] No delineamento de pesquisa pós-teste, a variável de interesse é medida no grupo de controle e no grupo experimental após a manipulação. No delineamento experimental pré-teste/pós-teste, a variável de interesse é medida no grupo de controle e no grupo experimental antes e depois da manipulação no grupo experimental. Isso permite ao experimentador avaliar a magnitude do valor de base do grupo experimental, bem como identificar qualquer influência do experimento nos dados, aumentando a confiabilidade dos resultados. Ver American Psychological Association (APA, s.d.).

Social Psychology, devido às conexões entre Psicologia Social e Administração Pública; e (b) replicação de estudos que investiguem temas relevantes para o campo da Administração Pública (valores, otimismo, capacidade de resposta comunal e conflito). O arquivo Excel listando os 35 estudos originais de nossa amostra, os resultados das replicações e trechos dos relatórios de replicação está disponível no site de armazenamento da *Open Science Foundation*: https://osf.io/ta746/.

Um aspecto que vale a pena destacar nessa amostra é que, em muitos casos, os pesquisadores que realizaram as replicações não tinham uma resposta clara sobre se o estudo foi de fato reproduzido ou não. Embora as equipes de replicação tenham aplicado os critérios de avaliação do *Reproducibility Project*, os comentários incluídos nos relatórios mostram que o retrato é menos "preto e branco" do que se esperaria inicialmente. Às vezes, a equipe foi capaz de replicar o efeito principal, mas não alguns dos efeitos adicionais (a maioria dos estudos incluiu vários experimentos). Em outras replicações, o efeito foi na mesma direção observada no estudo original, mas os resultados não atenderam o critério de significância ($P < 0,05$). As equipes de replicação geralmente usaram amostras com um número maior de participantes do que os estudos originais, garantindo potência estatística suficiente para detectar o efeito da hipótese testada; em dois casos, entretanto, os experimentadores admitiram que suas replicações careciam de potência suficiente e, por isso, consideraram os resultados das replicações inconclusivos.

Isso nos leva ao debate sobre os valores *P.* Este debate não é novo, remontando às críticas de Rozemboom ao teste de hipótese nula (Rozemboom, 1960); a controvérsia sobre a significância estatística parece longe de ser resolvida (ver Harlow, Mulaik & Steiger, 2016). A preocupação sobre os valores *P* se acirrou a ponto de levar a Associação Americana de Estatística a emitir diretrizes para seu uso; foi a primeira vez que a associação, fundada em 1839, elaborou uma diretriz desse tipo (Wasserstein & Lazar, 2016). Os editores do *Journal of Basic and Applied Social Psychology* chegaram a banir o uso dos valores *P* dos artigos publicados no periódico (Trafimow, 2014; Trafimow & Marks, 2015).

Muitos estatísticos argumentaram que o valor *P* é incapaz de dizer algo sobre a veracidade de uma dada hipótese. Segundo Goodman (2008, p. 136), isso se deve à própria definição do valor *P*: "A probabilidade do resultado observado, acrescida aos resultados mais extremos, se a hipótese nula fosse verdadeira". O valor *P* só pode fazer uma afirmação sobre se a hipótese nula deve ser rejeitada ou não, mas não sobre a veracidade real da hipótese alternativa. De acordo com Goodman, esta é apenas uma das muitas interpretações errôneas envolvendo o valor *P.* Ele argumenta que Fisher — o matemático que desenvolveu o teste de hipótese — usou o termo "significância" no sentido de "digno de atenção na forma de merecer mais experimentação, mas não prova em si" (Goodman, 2008, p. 135). Assim, um $P < 0,05$ não garante que a H_1 seja verdadeira. Goodman sustenta que o "casamento" entre o valor *P* e o teste de hipótese foi uma "união não natural". Benjamin et al. (2018) argumentam que o limite $P < 0,05$ produz muitos falsos positivos e que o padrão para reivindicações de novas descobertas deveria ser restrito a $P < 0,005$. Armhein e Greenland (2018, p. 4) responderam que isso só agravaria os atuais problemas e propuseram, em vez disso, que "[...] a reforma estatística deve descartar completamente a 'significância' e o raciocínio simplista que ela encoraja". Se alguém deseja demonstrar que um efeito confiável existe, então é necessário mostrar que existe um tamanho de efeito robusto. Portanto, a nossa terceira recomendação para a realização de experimentos na área da Administração Pública é focar no tamanho de efeito e não no valor *P* (relevância *versus* significância). Essa recomendação pode não ser fácil de ser seguida no atual contexto acadêmico devido ao uso

generalizado do teste de hipótese. Os experimentos custam tempo e dinheiro e os pesquisadores mantêm a expectativa razoável de que, se o critério $P < 0,05$ for satisfeito, então há razão suficiente para publicar os resultados. No entanto, essa expectativa também é o que estimula os pesquisadores a se envolverem em *P-hacking* (ver Lindsay, 2020). No entanto, ao final, apenas um tamanho de efeito forte — associado a replicações — pode apoiar uma dada hipótese, a menos que a hipótese principal seja a de que não há efeito algum, o que levanta outro problema.

Howell (2012, p. 230) observou que os pesquisadores "[...] têm apenas uma pequena chance de encontrar o efeito que procuram, mesmo que tal efeito exista na população". Esta é a razão pela qual é importante que experimentos tenham poder estatístico suficiente. Experimentos com poder suficiente têm maior probabilidade de rejeitar corretamente a H_0, reduzindo a ocorrência de erro do tipo II. No entanto, se não for fácil detectar um efeito, o pesquisador pode se sentir tentado a adotar a "prova da hipótese nula", que é outro problema que surge do teste de hipótese. Vale a pena ilustrar este ponto com um exemplo da área da Administração Pública.

Moynihan (2013) delineou um experimento para investigar se níveis mais elevados de Motivação do Serviço Público (MSP) (ver Perry & Wise, 1990), estavam associados à maximização do orçamento. Segundo o autor, se essa associação pudesse ser demonstrada, a teoria maximização do orçamento de Niskanen seria vindicada (ver Niskanen, 1968). Deve-se notar, entretanto, que Niskanen não presumiu que a MSP estava por trás da maximização do orçamento. Ele argumentou que os burocratas maximizam seus orçamentos porque isso lhes confere prestígio, o que é outra maneira de dizer que os burocratas são tão autointeressados quanto qualquer outra pessoa. Moynihan explicou a distorção em sua suposição nos seguintes termos: "Por que um maximizador de orçamento de espírito público é mais plausível? Os burocratas podem tentar maximizar os orçamentos porque acreditam sinceramente nos benefícios de seus programas" (Moynihan, 2013, p. 182). Embora a suposição de Moynihan seja discutível, vamos nos concentrar apenas nos termos de seu experimento. Ele recrutou estudantes de graduação como participantes e aplicou as manipulações experimentais necessárias. Os resultados não mostraram uma correlação linear significativa entre maximização do orçamento e níveis de MSP, mesmo quando os *outliers* foram removidos do modelo de regressão; assim, H_0 não poderia ser rejeitada a $P < 0,05$. Moynihan afirmou ter realizado uma análise de poder estatístico, embora ele não tenha informado o poder estatístico do experimento. No entanto, sua conclusão foi abrangente: ele argumentou que os resultados do experimento foram "[...] um achado significativo não significativo" (Moynihan, 2013, p. 190) que desmontou "[...] mais um pilar do modelo de maximização do orçamento" (Moynihan, 2013, p. 190). Isto corresponde a um caso clássico de "prova da hipótese nula".

Fisher, em seu livro clássico de 1935, *The Design of Experimentation*, argumentou que "[...] a hipótese nula nunca é comprovada ou estabelecida, mas é possivelmente refutada, no curso da experimentação" (Fisher como citado em Lehman, 2011, p. 64). Desde então, os estatísticos têm alertado sobre o erro de interpretação da "prova da hipótese nula" (ver M. P. Lecoutre, Poitevineau & B. Lecoutre, 2003). Deixar de detectar um efeito quando ele existe é a definição de erro do tipo II, mas isso não é uma prova da hipótese nula. Um experimento usando uma amostra com poder insuficiente aumentará as chances de erro do tipo II, mas isso — como esperado — não pode provar que a hipótese nula é verdadeira. Se o pesquisador, no entanto, está empenhado em conduzir um experimento com potência suficiente, ainda assim não é possível provar que a hipótese nula é verdadeira devido à própria definição de *poder estatístico*, a saber: "[…] a probabilidade de rejeitar corretamente um falso H_0 quando uma determinada hipótese alternativa é verdadeira" (Howell, 2012, p. 230). Assim,

um experimento com suficiente poder estatístico é aquele que pode rejeitar a hipótese nula, mas não pode provar que não há efeito. Nem sequer é possível calcular uma amostra com potência suficiente, pois isso requer estimar um tamanho de efeito maior do que zero (ver Howell, 2012). A melhor alternativa para a "prova da hipótese nula" é mostrar que o tamanho de efeito é muito pequeno para ter qualquer implicação relevante para o problema de pesquisa em mãos. Os testes de equivalência e de não-inferioridade também podem ser uma opção (ver Streiner, 2003). Assim, nossa quarta recomendação é evitar a "prova da hipótese nula" como delineamento experimental.

Quais são as características que tornam os experimentos nas ciências do comportamento mais prováveis de serem reproduzidos? Todas as replicações incluídas em nossa amostra intencional adotaram um delineamento que usa pós-teste apenas; portanto, não é possível inferir se esse delineamento teve alguma influência em um resultado favorável ou desfavorável da replicação. No entanto, como será discutido abaixo, há razões para recomendar o uso de delineamentos pré-teste/pós-teste sempre que possível.

Entre as 35 repetições examinadas, 12 experimentos foram replicados com sucesso (25%), 21 não foram reproduzidos, e dois foram considerados inconclusivos devido ao tamanho pequeno da amostra na replicação. Três das replicações bem-sucedidas usaram *surveys*, enquanto apenas uma entre as não-replicadas adotou esse delineamento experimental. Isso pode sugerir que os delineamentos experimentais usando *surveys* são mais confiáveis do que outros delineamentos experimentais, mas experimentos que usam *surveys* também podem ser decepcionantes. Nosek et al. (2012) descrevem um experimento que usou um *survey* com 1.979 participantes. Conforme relatam os autores, os resultados iniciais eram o sonho de qualquer pesquisador; a hipótese se sustentou e os resultados pareciam robustos e confiáveis. Mas, por uma questão de cautela, os autores decidiram replicar o experimento, coletando dados de outros 1.300 participantes. Desta vez, os resultados foram desanimadores: "[o] tamanho de efeito havia desaparecido (P = 0,59)" (Nosek et al., 2012, p. 616). Amostras grandes podem detectar tamanhos de efeito fracos (Streiner, 2003), mas porque os tamanhos de efeito podem ser muito fracos, amostras grandes não são necessariamente uma garantia de sucesso nas replicações.

No geral, experimentos replicados com sucesso em nossa amostra apresentaram uma tendência a ter delineamentos experimentais mais simples, com menos condições experimentais sendo testadas. No entanto, houve casos de delineamentos mais complexos que foram reproduzidos e de delineamentos mais simples que não foram.

Os experimentos de Psicologia Social eram mais propensos a serem reproduzidos quando os tamanhos de efeito originais eram robustos, tinham alto poder estatístico, adotavam delineamentos intrassujeito e incluíam medições repetidas. No entanto, o fato de que a média de tamanhos de efeito nos experimentos replicados foi muito menor do que nos estudos originais — especialmente no que diz respeito aos estudos de Psicologia Social — também lança uma sombra sobre os tamanhos de efeito. É por isso que um único estudo experimental não pode produzir uma resposta definitiva; uma única replicação também não.

Em geral, o *Reproducibility Project* mostrou que resultados contraintuitivos e estudos que demandavam vários experimentos e manipulações complicadas estavam correlacionados a uma probabilidade menor de replicação bem-sucedida (Open Science Collaboration, 2015). Isto leva a nossa quinta recomendação: sempre que possível, os pesquisadores em Administração Pública devem planejar experimentos que são fáceis de replicar e evitar estudos experimentais que exigem etapas complicadas ou pressupostos muito exigentes. Essa recomendação também pode levar a

resultados mais diretos. A realidade é complexa e os investigadores se esforçam para obter resultados realistas, o que pode exigir delineamentos de pesquisa complexos. No entanto, considerando-se que a reprodutibilidade é um componente importante do método experimental, os investigadores precisam equilibrar a complexidade *versus* a reprodutibilidade.

A não-reprodução dos resultados originais nas replicações incluídas em nossa amostra foi atribuída aos seguintes motivos: dificuldade de interpretação dos métodos originais (Lewis & Pitt, 2015); diferenças no tamanho da amostra (Reinhard, 2014); o efeito de moderadores desconhecidos (Kelson, 2015; Johnson, Hayes & Graham, 2015); efeitos de pequenas diferenças nos procedimentos experimentais (Baranski, 2015; Holubar, 2015; Lane & Gazarian, 2015; Lin, 2013); e diferenças na demografia/perfil/contexto dos participantes (Brown et al., 2013; Embley, Johnson & Giner-Sorolla, 2015; Lemn, 2013; Lin, 2013; Marigold, Forest & Anderson, 2015; Mechin & Gable, 2015; Talhelm, Lee & Eggleston, 2015). Portanto, as duas principais razões oferecidas para explicar as falhas na reprodução dos resultados originais foram as diferentes características dos participantes e as diferenças nas condições dos experimentos. A primeira questão não é necessariamente um problema para a pesquisa em Administração Pública; é até importante saber que sujeitos com características diferentes reagem diferentemente às manipulações experimentais. A segunda questão, entretanto, é preocupante. A replicação dos experimentos de Psicologia Social parece uma frágil pilha de pedras: é preciso equilibrar cuidadosamente todos os componentes para evitar que toda a estrutura desabe. Como a Administração Pública tem uma orientação prática, isso pode ser um grande problema. Os experimentos que não podem suportar a menor variação nos procedimentos experimentais têm pouca ou nenhuma relevância para a área.

Deve-se observar que os experimentos de Psicologia Social raramente adotam delineamentos de pesquisa pré-teste/pós-teste; os delineamentos que adotam apenas o pós-teste predominam mesmo em experimentos que envolvem várias manipulações. No entanto, as medições do pré-teste fornecem informações sobre a linha de base, como, por exemplo, o nível inicial de Motivação do Serviço Público tanto no grupo de controle como no experimental (ver Bellé, 2013). A importância do delineamento pré-teste/pós-teste pode ser ilustrado pela replicação do estudo de Vohs e Schoolre (2008). O estudo original, que argumentou que a crença no livre-arbítrio melhoraria o comportamento moral, recebeu grande atenção dentro da área de Psicologia Social devido as suas consequências de longo alcance. A notícia de que o experimento não pôde ser reproduzido foi uma decepção entre os pesquisadores. Nem o estudo original nem a sua replicação incluíram uma medida pré-teste. Os dois questionários relevantes para um dos dois experimentos incluídos no estudo — a escala *Freewill and Determinism* (FWD) e a escala *Positive and Negative Affectivity Schedule* (PANAS) — foram aplicados apenas depois que os participantes haviam lido os textos correspondentes ao controle e à manipulação de tratamento (tratamento = texto que afirma o livre arbítrio é uma ilusão; controle = texto neutro sobre a consciência). No entanto, se os grupos de controle e tratamento diferiam em relação à condição basal (crença no livre arbítrio), isso deveria ter sido levado em consideração na análise dos resultados (ver Bonate, 2000). Isso traz um dilema: se os participantes tivessem respondido ao questionário FWD antes que a manipulação experimental foi aplicada, isso provavelmente afetaria o resultado do experimento. Esse problema ocorre quando a variável do nível de base está dentro da cabeça dos indivíduos que participam de um experimento. Em estudos clínicos, os níveis basais correspondem a fatores fisiológicos que podem ser medidos com precisão; nesses casos, o pré-teste não influencia o resultado do pós-teste. No entanto, nas ciências comportamentais, a elaboração de delineamentos experimentais frequentemente envolve escolhas difíceis. Em todo o caso, os

pesquisadores em Administração Pública devem estar cientes de que, dependendo da investigação em mãos, o delineamento experimental pré-teste/pós-teste pode produzir resultados mais confiáveis do que o delineamento que usa apenas o pós-teste. Isso nos leva à sexta recomendação: sempre que possível, deve-se adotar um delineamento pré-teste/pós-teste.

Tendo em vista que os experimentos de Psicologia Social são de difícil reprodução, é razoável supor que a pesquisa experimental não deva ser a única fonte de conhecimento na área da Administração Pública. Felizmente, a área não possui um método de pesquisa predominante — pelo menos não ainda — e tem se baseado em diversos procedimentos metodológicos para construir a base de seus conhecimentos. Uma completa "virada experimental" pode ser uma má ideia para pesquisadores individualmente e especialmente para a área como um todo.

Vamos agora comentar rapidamente dois estudos que investigam um tema recorrente na gestão pública: a questão se práticas gestão oferecem os mesmos resultados nos setores público e privado. Robertson e Seneviratne (1995) descobriram que, em geral, as intervenções de mudança gerencial foram tão eficazes no setor público quanto no setor privado, mas esse resultado dependia de quais variáveis de interesse eram consideradas. Banerjeea e Solomon (2003, p. 119), em um estudo sobre a eficácia da rotulagem ecológica, argumentaram que "programas governamentais têm tido muito mais sucesso do que os programas privados. O apoio do governo a um programa de rotulagem não apenas aumenta sua credibilidade e reconhecimento, mas também melhora a estabilidade financeira, a proteção legal e a viabilidade a longo prazo". Os resultados desses dois estudos teriam pouca probabilidade de serem revelados se o método de pesquisa escolhido fosse um experimento. Para começar, os conceitos de setor privado e público implicam tantos aspectos diferentes que seria quase impossível selecionar variáveis independentes para testar o efeito sobre a variável dependente. Tipos de regime; relações intergovernamentais; governo central, regional ou local; estrutura e cultura organizacional; características de gestão e liderança; motivação e comportamento de administradores e funcionários — essas são apenas algumas variáveis que podem afetar quão bem os servidores públicos respondem a incentivos, pressões de trabalho, tarefas e responsabilidades. Pode-se argumentar que os métodos usados nesses estudos são menos precisos do que os experimentos. No entanto, como argumentado acima, o uso de mais de um método e diferentes fontes de dados podem oferecer resultados mais confiáveis e com maior relevância prática. Assim, nossa sétima recomendação é que o pesquisador na área de Administração Pública evite confiar exclusivamente no método experimental.

Por fim, é necessário mencionar que os resultados do *Reproducibility Project: Psychology* geraram controvérsias. Gilbert, King, Pettigrew e Wilson (2016) argumentaram que o projeto falhou em contabilizar o erro aleatório, argumentando que, além do erro de 5% esperado devido ao intervalo de confiança de 95%, as replicações usaram amostras de diferentes populações e, em alguns casos, os procedimentos de replicação não seguiram estritamente os estudos originais em outros aspectos também. Por outro lado, Gelman (como citado em Baker, 2016) argumenta que as replicações oferecem orientações mais confiáveis do que os estudos originais porque esses últimos têm maior probabilidade de serem acasos estatísticos em comparação com as replicações, particularmente devido ao viés de publicação. As replicações são empreendimentos mais cuidadosos e planejados, argumenta Gelman, enquanto os estudos originais que mostram grandes tamanhos de efeito tendem a encontrar seu caminho para a publicação muito rapidamente.

4. CONCLUSÕES

Este artigo abordou as promessas do que se convencionou chamar de "virada experimental" na Administração Pública. O argumento a favor do método experimental é que, assim como a Administração Pública e a Psicologia têm afinidades teóricas, a pesquisa na primeira área se beneficiaria com um maior uso do método experimental. Os experimentos têm a vantagem de controlar e isolar variáveis, bem como de permitir análises intrassujeitos e entressujeitos, controlando de maneira eficaz variáveis intervenientes, e garantindo um alto grau de validade interna. No entanto, apenas vários experimentos produzindo o mesmo resultado podem fornecer qualquer confiança sobre a validade externa.

Como o *Open Science Reproducibility Project: Psychology* revelou, embora os estudos experimentais de Psicologia Cognitiva não foram facilmente reproduzidos, os estudos de Psicologia Social foram ainda menos (Open Science Collaboration, 2015). À medida em que mais estudos experimentais na área de Administração Pública forem sendo realizados e os fracassos em reproduzir seus resultados começarem a aparecer, o entusiasmo inicial pelo método experimental entre os pesquisadores da área pode começar a se dissipar. Isso não significa que pesquisadores devam descartar totalmente o método experimental, mas que devem tomar os devidos cuidados ao adotá-lo.

Nossa análise dos resultados do *Reproducibility Project* levou a sete recomendações para pesquisadores da Administração Pública, que estão resumidas no Quadro 1.

QUADRO 1 RECOMENDAÇÕES EXTRAÍDAS DE ESTUDOS DE PSICOLOGIA PARA PESQUISADORES EM ADMINISTRAÇÃO PÚBLICA

# 1	Envolver-se em pesquisas colaborativas, desenvolvendo os mesmos experimentos simultaneamente, tornando a replicação um componente do delineamento da pesquisa.
# 2	Investigar o efeito de diferentes delineamentos de pesquisa nos resultados.
# 3	Concentrar-se no tamanho de efeito e não no valor *P* (relevância *predomina sobre* significância).
# 4	Evitar a abordagem do "teste da hipótese nula".
# 5	Sempre que possível, escolher delineamentos experimentais que sejam fáceis de replicar.
# 6	Considerar o uso delineamentos experimentais pré-teste/pós-teste quando isso for viável.
# 7	Evitar depender exclusivamente da pesquisa experimental para construir o conhecimento na área.

Fonte: Elaborado pelos autores.

A principal recomendação aos pesquisadores da área de Administração Pública é para que a replicação se torne um componente inerente do delineamento experimental. O método experimental não deve ser um empreendimento isolado, mas um projeto de pesquisa colaborativa em que a replicação de experimentos em diferentes culturas, regimes e contextos organizacionais faça parte do plano da pesquisa. A vantagem é que projetos colaborativos —independentemente dos resultados dos experimentos — podem evitar o viés de publicação, pois as revistas acadêmicas de ponta certamente têm interesse nos resultados de pesquisas com essa caraterística. À medida que mais experimentos

na área da Administração Pública se tornarem disponíveis, estudos meta-analíticos também se tornarão mais viáveis. Isso é de suma importância, uma vez que mesmo experimentos com baixo poder estatístico, quando combinados, podem aumentar o poder da análise (Cooper, 2017).

Pesquisadores na área de Administração Pública deveriam aprender com as dificuldades dos experimentos em Psicologia também em relação a algumas questões técnicas específicas, como focar no tamanho de efeito em lugar do valor *P*. Há uma razão por que o teste de hipótese se tornou tão polêmico ao longo dos anos, como foi discutido acima. Além disso, embora detectar um efeito possa não ser fácil, isso não deve ser um incentivo para alguém se engajar na "prova da hipótese nula". Conforme discutido acima, não é possível provar que a hipótese nula é verdadeira mesmo quando o experimento tem poder estatístico suficiente para detectar um efeito. Isso é consequência do fato de que a técnica estatística usada para calcular os tamanhos das amostras tem que assumir um tamanho de efeito maior do que zero (ver Howell, 2012). Os testes de equivalência e não-inferioridade podem ser uma opção quando o pesquisador quiser demonstrar, por exemplo, que um programa de treinamento não é melhor do que outro (Streiner, 2003). Os delineamentos experimentais pré-teste/pós-teste devem ser usados com maior frequência, embora medir variáveis que estão "dentro das cabeças" dos indivíduos represente um desafio para este delineamento experimental.

Nem todos os tópicos de pesquisa são adequados para serem investigados por meio de experimentos. O método experimental demanda variáveis que já se encontram firmemente estabelecidas e instrumentos de medição que tenham sido adequadamente validados. A Administração Pública é uma disciplina que surgiu de um contexto específico: as organizações governamentais e suas interações com o público. Assim, recomenda-se também que os experimentos testando hipóteses relacionadas aos problemas da área use servidores públicos como participantes experimentais sempre que possível. Os resultados de experimentos que usam como sujeitos experimentais alunos de graduação matriculados em universidades ocidentais provavelmente não servirão como uma base sólida para intervenções práticas.

Os pesquisadores em Administração Pública não perguntam apenas sobre "o quê" e o "porquê" das questões, mas também sobre o "como". O conhecimento gerado por meio de pesquisas em nossa área precisa ser aplicado a problemas práticos. Pode ser razoável argumentar que as replicações falham porque os experimentos não seguiram exatamente os mesmos procedimentos dos estudos originais. No entanto, na Administração Pública, os experimentos devem ser robustos o suficiente para suportar algum grau de diferenças contextuais. Os investigadores em Administração Pública lidam com uma grande variedade de fatores e a interrelação entre eles. Portanto, focar em algumas poucas variáveis pode não ser suficiente. Experimentos podem ser úteis para abordar questões que são relevantes para profissionais e tomadores de decisão, mas outros métodos ainda seriam necessários para investigar questões sociais complexas. Quanto aos delineamentos experimentais, análises fatoriais complexas podem ser necessárias, mas isso não significa que uma abordagem direta possa ser negligenciada.

Finalmente, o nosso argumento implícito ao longo deste artigo é que os pesquisadores na área de Administração Pública precisam se envolver em metaciência, ou seja, usar ferramentas científicas para refletir sobre a prática científica. Compreensivelmente, a metaciência ganhou relevância renovada nas ciências comportamentais após a crise de reprodutibilidade na Psicologia (ver Passmore & Chae, 2019).

Ao fim e ao cabo, a área da Administração Pública pode se beneficiar do método experimental ao estar ciente de suas dificuldades. Se a área está se movendo para uma "virada experimental", que esta virada possa ser realizada com cautela e senso de direção.

REFERÊNCIAS

American Psychological Association. (s.d.). *APA Dictionary of Psychology*. Recuperado de 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, março 03). Psychology's reproducibility problem is exaggerated — say psychologists. *Nature News*. Recuperado de 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.

Banerjeea, 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)*. Recuperado de https://osf.io/p3gz2/

Barber, T. X., Calverley, D. S., Forgione, 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)*. Recuperado de 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)*. Recuperado de 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, julho). 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)*. Recuperado de https://osf.io/a4fmg/

Howell, D. C. (2012). *Statistical methods for psychology*. Belmond, 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)*. Recuperado de 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)*. Recuperado de 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)*. Recuperado de 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 slassical Statistics*. New York, NY: Springer.

Lemn, K. (2013). *Replication of Blankenship and Wegener (2008, JPSP, Study 5A)*. Recuperado de https://osf.io/v3e2z/

Lewis, M., & Pitts, M. (2015). *Replication of "Errors are Aversive" by Greg Hajcak & Dan Foti (2008, Psychological Science)*. Recuperado de https://osf.io/tkq9n/

Lin, S. (2013). *Replication of Study 7 by Exline, Baumeister, Zell, Kraft & Witvliet (2008, Journal of Personality and Social Psychology)*. Recuperado de 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. Recuperado de 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)*. Recuperado de 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)*. Recuperado de 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*. Recuperado de 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., & Campwell, D. T. (2002). *Experimental and quasi-exprimental 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)*. Recuperado de 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. Recuperado de https://doi.org/10.1080/01973533.2014.865505

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1-2. Recuperado de 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 Kolfschooten, 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

Doutora em Ciência Política; Professora Associada no Departamento de Ciências Sociais da Universidade Federal de São Paulo. E-mail: christina.andrews@unifesp.br

Michiel S. de Vries

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

Estudou Sociologia na Universidade de Groningen; Trabalhou anteriormente na University of Amsterdam, Thorbecke Academy em Leeuwarden e na Free University Amsterdam; Doutor em Direito pela Universidade de Utrecht; Em 2016 recebeu o título de Doutor Honoris Causa em Economia Pública pela Universidade Masaryk de Brno; Professor Titular no Instituto de Pesquisa em Administração na Radboud University of Nijmegen; Distinto professor de Administração Pública da Kaunas University of Technology e extraordinário professor da Free State University de Bloemfontein; Ele foi presidente da IASIA e membro do conselho editorial de várias revistas de Administração Pública. E-mail: m.devries@fm.ru.nl