Best Practice Recommendations for Replicating Experiments in Public Administration

Richard M. Walker  
*City University of Hong Kong, China*

Gene A. Brewer  
*University of Georgia*

M. Jin Lee  
*City University of Hong Kong, China*

Nicolai Petrovsky  
*University of Kentucky*

Arjen van Witteloostuijn  
*Vrije Universiteit Amsterdam, The Netherlands*

Follow this and additional works at: [https://uknowledge.uky.edu/msppa_facpub](https://uknowledge.uky.edu/msppa_facpub)

Part of the Public Affairs, Public Policy and Public Administration Commons, and the Quantitative, Qualitative, Comparative, and Historical Methodologies Commons

Right click to open a feedback form in a new tab to let us know how this document benefits you.

**Repository Citation**

Walker, Richard M.; Brewer, Gene A.; Lee, M. Jin; Petrovsky, Nicolai; and van Witteloostuijn, Arjen, "Best Practice Recommendations for Replicating Experiments in Public Administration" (2019). *Public Policy and Administration Faculty Publications* 5.  
[https://uknowledge.uky.edu/msppa_facpub/5](https://uknowledge.uky.edu/msppa_facpub/5)

This Article is brought to you for free and open access by the James W. Martin School of Public Policy and Administration at UKnowledge. It has been accepted for inclusion in Public Policy and Administration Faculty Publications by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.
Best Practice Recommendations for Replicating Experiments in Public Administration

Digital Object Identifier (DOI)
https://doi.org/10.1093/jopart/muy047

Notes/Citation Information
Published in Journal of Public Administration Research and Theory, v. 29, issue 4.

© The Author(s) 2018

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

This article is available at UKnowledge: https://uknowledge.uky.edu/msppa_facpub/5
Best Practice Recommendations for Replicating Experiments in Public Administration

Richard M. Walker*, Gene A. Brewer†, M. Jin Lee*, Nicolai Petrovsky‡, Arjen van Witteloostuijn§

*City University of Hong Kong †University of Georgia ‡University of Kentucky §School of Business and Economics, Vrije Universiteit Amsterdam

Abstract

Replication is an important mechanism through which broad lessons for theory and practice can be drawn in the applied interdisciplinary social science field of public administration. We suggest a common replication framework for public administration that is illustrated by experimental work in the field. Drawing on knowledge from other disciplines, together with our experience in replicating several experiments on topics such as decision making, organizational rules, and government–citizen relationships, we provide an overview of the replication process. We then distill this knowledge into seven decision points that offer a clear set of best practices on how to design and implement replications in public administration. We conclude by arguing that replication should be part of the normal scientific process in public administration to help to build valid middle-range theories and provide valuable lessons to practice.

Replication is an important part of the scientific process because it can help to establish the external validity of knowledge and thus the ability to generalize a study’s findings more broadly. Replication thus sits at the heart of scientific progress (e.g., Francis 2012; Freese 2007; Kuhn 1996; Nosek and Lakens 2014). The goal of replication is to confirm or reject theories under similar and dissimilar conditions to build confidence in (or falsify) them. In this way, theory testing is advanced, and knowledge is accumulated. Many disciplines rely heavily on replications to advance knowledge (Hubbard, Vetter, and Eldon 1998; Open Science Collaboration 2015), but they are rare in public administration and management (Walker, Brewer, and James 2017).

Public administration is an interdisciplinary social science field that emphasizes applied (“theory to practice”) applications. In this article, we argue that these two factors suggest the need to craft a replication framework for public administration to advance knowledge. Over half-a-century ago, Simon (1996) argued that public administration is a “design science” that devises courses of action designed to change existing outcomes into preferred ones. Recommendations from public administration scholarship can have tangible consequences for citizens; thus, prescriptions need to be well-founded. If the discipline of public administration is to successfully advance policy and practice, then research findings must be generalizable. In the applied social sciences, the context within which studies are conducted is considered to be a critical variable for the

1 In this article, the term “public administration” includes public management.
development of theory (Freese 2007; Jilke, Meuleman, and Van de Walle 2014).

Most social and behavioral sciences are single disciplines with interdisciplinary interests, whereas public administration is an interdisciplinary discipline with many single interests. For example, much of psychology is concerned with questions about the human psyche and human behavior at the individual level. The interdisciplinary nature of public administration means that we are concerned with individual values, attitudes, and behaviors in social and political settings. Further, public administration research can include many different units of analysis (individuals, organizations, networks, countries, etc.). Although the scope of public administration is not necessarily broader or more important than other disciplines, the nature of public administration is interdisciplinary and spans the social sciences.

To help develop a clear set of arguments and illustrations, this article focuses primarily on the replication of experimental studies. However, much of what we say is relevant to other types of research, which also require rigorous replication. Hence, the steps we outline are generally applicable to all types of research, including qualitative studies (Mele and Belardinelli Forthcoming; Nowell and Albrecht Forthcoming). Public administration scholars have increasingly adopted experimental research designs, which are characterized by random assignment of treatments or subjects into separate treatment and control groups, and conducted in real-world or artificial laboratory-like settings. These designs have been adopted because of their strong internal validity and ability to identify causal relationships (Blom-Hansen, Morton, and Serritzlew 2015; James, Jilke, and Van Ryzin 2017). Yet even when the internal validity of an experimental study is strong, its external validity is often weak.

Researchers face many challenges in publishing replications. Notably, editors often prefer new findings on groundbreaking topics (van Witteloostuijn 2016) rather than the steady hum of scientific progress inching forward. These expectations are reinforced by the reward and promotion practices of universities, in which faculty are expected to publish in leading or high-impact journals that demand strong findings on innovative topics. Such practices have led some to argue that research articles contain many Type I errors in which false positives abound (Ioannidis 2005; van Witteloostuijn 2016). Replication publications are, nonetheless, on the rise across the social sciences (e.g., Francis 2012; Nosek and Lakens 2014) and in public administration (Walker, Brewer, and James 2017), but much more needs to be done. It is therefore important to develop a replication framework for conducting experimental replications, particularly considering the costliness of experiments, the number of published studies with small-Ns, and the contextual nature of the findings reported therein.

This article contributes to the debate on replication in public administration and beyond by offering best practice recommendations for experimental replications with a goal of improving theory and practice. Outlining clear steps for conducting experimental replications will inform researchers’ decisions on whether a study can be replicated and how the replication should be carried out. This article describes seven decision points. Central to these decision points are questions on criteria for determining the type of replication to be implemented. We rely on Tsang and Kwan’s (1999) classification of different types of replications to guide these decisions. Their framework offers conceptual clarity on the type of replication needed to extend the original study to different populations, designs, and analyses. The decision points are (1) deciding whether replicating the study is feasible, (2) assessing the internal validity of the original study and the replication, (3) making choices about statistical power, (4) choosing a critical test case, (5) establishing boundary conditions, (6) establishing content validity, and (7) deciding how to compare the findings.

The first section of this article reviews several rationales for replicating studies in the social sciences in general and catalogs replication experiences in different disciplines. In the second section, we introduce the seven decision points for designing and implementing replications which provides the underpinnings for our best practice recommendations. Following this, we illustrate our best practice recommendations through a replication conducted in Hong Kong in 2017 of Van Ryzin (2013). The foundations for our best practice recommendations are drawn from the replication experiences on the topics reviewed and our experience in replicating field, laboratory, and survey experiments at the Laboratory for Public Management and Policy at City University of Hong Kong (e.g., Walker, Brewer, and James 2017). We conclude by proposing a “common replication framework” that would integrate replication into the normal scientific practices in public administration to improve the quality of scholarship and help develop mid-range theories and robust lessons for practice.

Why Replication

Scientific knowledge is bolstered when research designs have high validity. Researchers may not get it right the first time, but through repeated efforts, their findings amass and converge on the truth. Campbell and colleagues (e.g., Campbell and Stanley 1963; Cook and Campbell 1979) shaped researchers’ thinking when
they distinguished between four essential forms of validity in social science research: internal validity, external validity, construct validity, and statistical conclusion validity. The authors further described various threats to validity and cautioned researchers to minimize them. For many decades, this has been textbook material, and it informs our decision points.

Researchers should strive for high levels of validity, according to Campbell and Stanley (1963). However, there are tradeoffs that work against achieving a perfect balance. On the one hand, research designs tailored to optimize the testing of causal relationships with high internal validity often create artificial or contrived conditions that make the result very difficult to reproduce in other settings (low external validity). On the other hand, research designs aimed at broad generalizability (high external validity) often sacrifice control and rigor in the research design, making the results less certain (low internal validity). Public administration scholars have adopted the use of experimental research designs in the search for stronger internal validity. Nevertheless, in an applied social science discipline like public administration, external validity is also important. Hence, replication research designs often incur a trade-off between internal and external validity. External validity can then be produced through a string of internally valid studies that test boundary conditions.

A Review of the Major Replication Efforts to Date

Reproducibility

Reproducibility is “the extent to which consistent results are observed when scientific studies are repeated” (Open Science Collaboration 2015, 657). Some disciplines have experienced a “reproducibility crisis” because they could not reproduce high visibility research findings. In economics and finance, Dewald, Thursby, and Anderson (1986) attempted to verify findings reported in articles published in the Journal of Money, Credit, and Banking before and after the journal required authors to provide data and code to others upon request. For articles published before the mandate, 14 out of 62 authors contacted responded that they had lost or discarded their data; for articles published subsequently, the most striking observation was the high prevalence of errors, some of which led to substantially different conclusions. More recently, in psychology, the Open Science Collaboration’s (2015) multisite replication of 100 experiments challenged the hitherto prevalent views of much of the accumulated knowledge in the field when only 39% of the original effects were substantively replicated.

The reproducibility debate has led to calls for openness and transparency in all aspects of the research process (Camerer et al. 2016; Ioannidis, 2005; Munafo et al. 2017; Nosek et al. 2015). In our view, the public administration should institutionalize these norms of openness and transparency before a replication crisis emerges in the field. This is essential if replication is to become a normal part of public administration research. Given the relative youthfulness of experimental research designs in public administration, ensuring the full transparency of all published experimental studies is an essential prerequisite for conducting replications (decision point 1). A leading voice in the transparency and openness debate has been the Center for Open Science, which has published Transparency and Openness Guidelines (Nosek et al. 2015). Level 1 of these standards has been adopted by some public administration journals. Whether journals adopt these standards or not, public administration scholars are urged to be open and transparent in their research designs, data and code to facilitate replication.

Replication

Many existing articles provide guidance on the conduct of replications. We identified a sample of these articles by searching for titles containing the term “replicat*” in the Social Science Citation Index in the disciplines of economics, management, psychology, sociology, and public administration for the period 1970–2018. We turned up 971 articles, of which the vast majority were replication studies. Table 1 identifies three broad themes that are discernable in the literature: types of replication, academic practices, and recommendations and guidance for the conduct of replication. The table lists some useful studies and provides a brief summary of their recommendations for improving replication.

The first theme includes scholars working from an exploratory social science perspective and seeking to understand and classify types of replication—we return to this topic in the next section. A second theme involves studies focusing on recommendations to promote the conduct and publication of replications. Recommendations typically include teaching replications in PhD programs to instill replication as a norm and changing the attitudes and behaviors of journals, editors, and reviewers (Hubbard, Vetter, and Little 1998; Schmidt 2009). The third theme promotes the practice of replication. Recommendations include detailed guidance on implementing a direct replication (Brandt et al. 2014), the development of a
<table>
<thead>
<tr>
<th>Authors</th>
<th>Journal</th>
<th>T</th>
<th>A</th>
<th>R</th>
<th>Key Observations on the Conduct of Replications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bettis, Helfat, and Shaver (2016)</td>
<td>Strategic Management Journal</td>
<td>√</td>
<td></td>
<td>√</td>
<td>Proposes narrow and quasi-replication types. Offers guidelines to achieve high-quality replication: Is the focus of the replication an important finding? Is the goal a narrow or quasi-replication? How closely does the replication match the original study? Is it beneficial to add advances in the literature since the publication of the focal study? Are the results interpreted so as to build the literature?</td>
</tr>
<tr>
<td>Brandt et al. (2014)</td>
<td>Journal of Experimental Social Psychology</td>
<td>√</td>
<td></td>
<td></td>
<td>Provides five replication ingredients for a direct replication: (1) Carefully defining the effects and the methods that the researcher intends to replicate. (2) Following as exactly as possible the methods of the original study. (3) Having high statistical power. (4) Making complete details about the replication available. (5) Evaluating replication results, and comparing them critically with the original study.</td>
</tr>
<tr>
<td>Freese and Peterson (2017)</td>
<td>Annual Review of Sociology</td>
<td>√</td>
<td></td>
<td></td>
<td>Replication types based upon new and old data and similarity and differences in research design and a replication policy for sociology</td>
</tr>
<tr>
<td>Hubbard, Vetter, and Little (1998)</td>
<td>Strategic Management Journal</td>
<td></td>
<td></td>
<td></td>
<td>Recommendations on inculcating replication in graduate school to promote replication as normal practice and for journals to have separate sections dedicated to publishing replications</td>
</tr>
<tr>
<td>Jilke et al. (2017)</td>
<td>Public Management Review</td>
<td></td>
<td></td>
<td></td>
<td>Provides guidance on how to empirically achieve measurement equivalence for cross-replication experiments based on the experimental replication and survey measurement equivalence literatures</td>
</tr>
<tr>
<td>Kerr, Shultz, and Lings (2016)</td>
<td>Journal of Advertising</td>
<td>√</td>
<td></td>
<td></td>
<td>Based on a Delphi study an agenda for collective action to increase replication is presented: review the reviewers, normalize replication, openness, and transparency, grow publication opportunities, replicate it differently</td>
</tr>
<tr>
<td>Makel and Plucker (2014)</td>
<td>Educational Researcher</td>
<td></td>
<td></td>
<td></td>
<td>Recommendations on third-party replication studies to improve education policy and practice</td>
</tr>
<tr>
<td>Pederson and Stritch (2018)</td>
<td>Public Administration Review</td>
<td></td>
<td></td>
<td></td>
<td>Provides guidance on key replication ingredients: relevance, number, internal validity, contextual realism, and external</td>
</tr>
<tr>
<td>Schmidt (2009)</td>
<td>Review of General Psychology</td>
<td>√</td>
<td></td>
<td></td>
<td>Replication types: direct and conceptual. Discusses the need for replication to be a mainstream activity by achieving greater recognition of replication activities by scholars, the teaching of replication and changing the editorial behaviors of journals.</td>
</tr>
</tbody>
</table>
“good-enough” replication standard (Singh, Ang, and Leong 2003), and a protocol based on the researchers involved in the replication, the population/subjects studied, and the methods and analysis (Walker, Brewer, and James 2017). Bettis, Helfat, and Shaver (2016) move beyond questions of research design to consider how a replication advances knowledge and how to interpret its results, as do Fabrigar and Wegener (2016), while Jilke et al. (2017) provide detailed guidance on measurement equivalence. Others have provided replication models such as RNICE (relevance, number, internal validity, contextual realism, and external validity), which was published in a public administration journal (Pedersen and Stritch 2018). This model lists topics that need to be examined to evaluate a replication effort and is an important step forward in the emerging debate in public administration.

If replication agendas are to pay dividends, studies must be carefully chosen (Bettis, Helfat, and Shaver 2016). Pedersen and Stritch (2018, 3) suggested that “as a rule, a ‘valuable’ replication aims to provide supporting (or contradicting) evidence about the existence of a phenomenon that audiences care about and deem important.” Detailed impact criteria have been offered by the Netherlands Organisation for Scientific Research, providing a useful guide for public administration scholars: The original study’s findings must be scientifically impactful (e.g., a breakthrough finding with much subsequent work built on it, or a study that has been highly cited) and/or take an important societal perspective (e.g., a critical policy).

Best Practice Recommendations For Designing And Implementing A Replication

This section introduces our main contribution: best practice recommendations for conducting replications in public administration research broadly and in experimental studies specifically. The best practice recommendations consist of key “decision points” to help researchers navigate the thicket of issues they face when planning a replication, such as assessing feasibility and determining the type of replication needed. Replications connect the existing body of knowledge with new knowledge. Therefore, they should be applied systematically (Schmidt 2009). Smart decisions on these issues can increase a researcher’s impact, produce new knowledge for the field, and make scientific progress more efficient.

The best practice recommendations span several steps in the scientific method, such as testing and re-testing hypotheses to accumulate results that eventually converge and form scientific knowledge. Figure 1 depicts the main stages of replication: planning, implementing, and reporting results. The goals of replication and decisions on the type of replication to use (see below) influence the best practice recommendations. Thus, trade-offs between these choices are also considered.

Different Types of Replication

In selecting studies for replication, many decisions must be made, and a classification scheme of different types of replications is very helpful. Tsang and Kwan (1999) proposed a framework that identified four different forms of replication (table 2). The framework contrasts several study elements to create a typology: studies using the “same measurement and analysis” or

<table>
<thead>
<tr>
<th>Singh, Ang, and Leong (2003)</th>
<th>Journal of Management</th>
<th>√</th>
<th>√</th>
<th>Promotes a good-enough replication standard and the focusing of replication research on degree of theory development (limited/substantial) and degree of methodological development (limited/substantial)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tsang and Kwan (1999)</td>
<td>Academy of Management Review</td>
<td>√</td>
<td></td>
<td>Proposed six types of replication (including reproduction) based upon the population (same and different) and measurement and analysis (same and different)</td>
</tr>
<tr>
<td>Walker, James, and Brewer (2017)</td>
<td>Public Management Review</td>
<td>√</td>
<td>√</td>
<td>Provides a replication protocol based upon six types of replication and the researchers involved, population/subjects, methods, and analysis</td>
</tr>
</tbody>
</table>

Note: T = discusses types of replication; A = makes recommendations to the academic community to increase the prevalence of replications; R = makes recommendations for the conduct of replications.

Table 1. (Continued)

“different measurement and/or analysis,” and studies using the “same population” or a “different population.” Tsang and Kwan’s (1999) terminology suggests that different populations consist of different participants (e.g., diverse groups of citizens in different service areas or jurisdictions). Their ideas also apply to the additional dimension of context (e.g., the same citizens experiencing different policy domains or institutional settings).

Tsang and Kwan (1999) called a replication with the same research procedures, measurements, analysis, population (although perhaps a different sample), and context an “exact replication.” Following Schmidt (2009), we have relabeled this a “direct” replication because an “exact” or “literal” replication is actually not possible. At a minimum, time passes, and the subjects change. However, we have moved beyond Schmidt’s (2009, 91) argument that a direct replication is just a “[r]epetition of an experimental procedure.” A study is repeated at a different point in time to test whether the findings hold and whether the concepts are still applicable. A direct replication could also
involve a different sample of public managers, such as frontline workers, from a larger population of public managers. An example of a direct replication was implemented by Grimmelikhuijsen and Porumbescu (2017) on Van Ryzin (2013). The replication yielded similar results and reinforced the findings reported in the original study.

The second type of replication also uses a sample from the same population, but it is labeled “conceptual.” A conceptual replication is an extension of the original study because it involves different measurements and analyses. A conceptual extension introduces new procedures and explores additional ramifications of the original finding(s). For example, using the same population, Van Ryzin, Riccucci, and Li (2017) extended their earlier study of representative bureaucracy (Riccucci, Van Ryzin, and Li 2016) into a new policy domain: emergency preparedness. They were unable to verify their prior findings and concluded that “the symbolic effects of gender representation may be policy-specific” (Van Ryzin, Riccucci, and Li 2017, 1365). Scholars conducting replication studies should be aware that conceptual replication may require many repetitions using similar and dissimilar measurements and analyses to provide clear-cut answers.

The third and fourth types of replication involve varying degrees of extension because they are performed on populations that differ from those in the original study. These replications may also differ by using either the same or new measurements and/or analyses. For example, “empirical generalization” uses the same research design, measures, and analyses, but assesses whether the original findings hold up in different populations. Lee, Moon, and Kim (2017) extended Knott, Millar, and Verkuilen’s (2003) study of American university students’ decision making to Hong Kong and Korea, broadly confirming the original finding that individuals with limited information make incremental decisions. “Generalization and extension” replications use a different population while seeking to extend the original findings by adopting additional measurements and analyses. Under this type of replication, if the results are different from those reported in the original study, the discrepancies may be attributed to an altered research design and/or changes in the population.

The approaches to generalization and extension have varied in the public administration literature. For example, Olsen (2017) examined the hypotheses and findings from psychology to study action bias. George et al. (2017) replicated a Danish experiment on performance information and politicians’ preferences for spending and reform, adding an extension onto the role of strategic goals. As discussed in decision point 7, interpreting the replication results requires finesse. Nevertheless, generalizations and extensions represent an important step in testing boundary conditions (decision point 5) to refine theory and strengthen external validity.

Moving from direct replication to generalization and extension increases the confirmatory power of the replication (Schmidt 2009). A direct replication tests whether the hypotheses and causal relationships in the original study hold up with minimal change. Adding to population, measurement, and analysis through “empirical generalization,” “conceptual,” and “generalization and extension” further tests the concepts and causal relationships in each replication. In an ideal world, with ample time and resources available, every replication would travel through the four stages shown in figure 2 to determine whether, or the extent to which, the findings can be confirmed at each stage—reflecting Singh, Ang, and Leong’s (2003) call for “good enough” replications. As Schmidt (2009, 97) noted, systematically diverse populations, contexts, measures, and analyses are likely to result in variations from the original study that are suited to the multitrait-multimethod research design or a “systematic replication matrix.” To overcome the obstacles of resources and time, a “many-labs” research design has been employed, a topic we return to below.

Planning a Replication

Decision Point 1: Deciding If It Is Feasible to Replicate the Study

The goal of the first decision point is to determine whether it is feasible to replicate a study. Although the original study is likely to have been published in a scholarly journal, as we noted above, most public
administration journals do not require all aspects of a research article’s design, procedures, and analysis to be fully disclosed. Given this, attempting a replication can be challenging.

If all information on the research design, analysis, and results are accessible, the researcher can move directly to decision points 2 and 3 to check for internal and conclusion (statistical) validity and determine which type of replication should be undertaken. However, it may not be possible to replicate a study if the original materials are not available and sufficient. This includes information on participant recruitment, instructions, measures, procedures, and analysis of the final data set (Brandt et al. 2014). In some cases, materials may be available, but the original study was conducted in a different language. For example, the research materials for Weibel, Rost, and Osterloh’s (2010) study were in German. To conduct a replication, these materials had to be requested from the authors and translated meticulously. Translation is costly and translating somewhat technical research materials can result in misinterpretation, raising questions of construct validity (decision point 6).

If the full details are unavailable or the costs of replication are prohibitively high, there are three options. First, the replication can be abandoned. Second, if adequate study data and research materials are available, the researcher can undertake a “checking of analysis” (Tsang and Kwan 1999). If the original findings can be reproduced based on the replicator’s ability to recreate the original study materials, the researcher can move to decision point 2 and onward to consider which type of replication to undertake. Third, if partial details are available, a final option is to use a different type of replication (e.g., a conceptual replication). This would also lead the researcher to decision point 2.

Information on the original materials may be available from the author(s). If the original authors are being consulted on how to access the materials, it may be necessary to consider a role for one or more of them in the replication. Benefits (including access to detailed knowledge on how the original study was conducted and how it can be replicated) can accrue by including the original team members. However, benefits also arise from maintaining independence from the original research team. Autonomy ensures that researchers have less emotional attachment to the findings and can be impartial: analysis of replications in education and psychology indicate that involving the original authors increases the likelihood of producing similar findings to the original article (Makel and Plucker 2014; Makel, Plucker, and Hegarty 2012). If the replicating researcher desires independence from the original authors, but the full materials are unavailable, replication may not be possible, or a different type of replication may need to be undertaken.

Decision Point 2: Assessing the Internal Validity of the Original Study and the Replication
Some public administration experiments are well-designed. They are guided by a clear hypothesis (or a set of hypotheses) derived from a well-specified theoretical perspective. The measures and treatments fit with the theoretical concepts. Alternative explanations are carefully considered and to the greatest extent possible eliminated by the design, which results in a high degree of internal validity. In principle, such cases are suitable for replication. However, replications may still encounter serious threats to internal validity if such threats were present in the original study.

Internal validity is defined by the following questions: did the experimental stimulus, in fact, make some significant difference in the result, and did this covariation result from a causal relationship (Cook and Campbell 1979; Shadish, Cook, and Campbell 2002, 37)? Validity is a property of inferences related to the design that relate to the sureness or truth of the results. Many well-known threats can weaken internal validity (e.g., ambiguous temporal precedence, selection bias, history, maturation, regression, attrition of respondents, testing, and instrumentation). Space does not permit a full review of all threats to internal validity, but when replicating a study, full consideration should be given.

Decision making on internal validity is a two-stage process. First, the research design of the original study must be carefully examined to ensure that internal validity is high and that fatal problems, such as ambiguous temporal precedence, are removed. Many threats to internal validity, such as attrition and history, occur during the implementation of a study. Original studies should clearly document these threats, but this does not always happen. Thus, it is important to systematically access the original study’s design, procedures, and analysis. Second, when replicating, improvements to internal validity should be implemented to enable a “convincing replication” (Brandt et al. 2014; Pedersen and Stritch 2018). Any remaining limitations should be documented when the results are reported.

Decision Point 3: Making Choices about Statistical Power
In public administration, one endemic threat to internal and conclusion validity is low statistical power. The definition of statistical power is the ability of a statistical test (at a level of statistical significance, \( \alpha \), specified by the researcher) to detect an effect size (\( \delta \), also specified by the researcher) that exists within the population. It follows that statistical power is always a contingent quantity—it is not defined except for specified values...
of statistical significance, $\alpha$, and importantly, the effect size $\delta$. Statistical power is a major consideration when deciding what to infer from an experimental finding. Low power can be a parsimonious explanation of nonsignificant findings (Shadish, Cook, and Campbell 2002). Although the concept of statistical power is widely used, practice could be improved, as Lindsay (2015) notes for the field of experimental psychology.

In a power calculation, the level of statistical significance ($\alpha$) is typically specified by convention, often at 0.05 [but see Meyer, van Witteloostuijn, and Beugelsdijk (2017) and Cumming (2014) on how to avoid the pitfalls of such a criterion]. For the effect size $\delta$, researchers tend to follow prior research. In a replication, $\delta$ can be based on the estimate(s) of the prior study. If researchers want to allow for the possibility of publication bias (which would mean the estimates of $\delta$ are greater in magnitude than the population value), they can choose a smaller value of the effect size, $\delta$, for the power calculation (Vazire 2016, 4). Once the level of statistical significance and effect size have been specified, statistical power increases in the number of subjects and decreases in the variance of the outcome. We discuss these two levels in turn.

The clearest solution to low statistical power is to increase the sample size. Often, for practical reasons, the number of subjects available for the replication is relatively limited. Having a limited number of subjects means that a replication can improve on the first experiment by taking steps to enhance statistical power. However, there may be a range of constraints that restrict the researcher from doing this. These include the costs of treatments plus compensation and difficult access to subjects. Those who are interested in replication with a focus on maximizing power can take a number of steps to do so. Shadish, Cook, and Campbell (2002, 46–7) provided a useful catalog of ways to increase power. For one, if the treatments are costly relative to the control conditions, or if each treatment is primarily compared with the control condition, the number of subjects in the control condition can be increased. Indeed, Orr (1999) shows that for $T$ treatments and a control condition, power is maximized if $1/(T \times 2)$ subjects are allocated to each treatment condition, and one-half of the subjects are allocated to the control condition.

Beyond increasing the number of subjects and/or the optimal allocation of subjects to experimental conditions, other strategies to increase statistical power focus on reducing the variance of the outcome. For replications, the variance in outcome can be estimated from the summary statistics of the study being replicated, if they are broken down and reported by experimental conditions. The following three strategies to decrease the outcome variance and thus increase statistical power are all compatible with both direct replication and empirical generalization. These strategies enhance power by reducing heterogeneity first so that likes are compared with likes. They increase power if—and only if—a variable has been measured before the experiment that is expected to correlate with the outcome (cf. Shadish, Cook, and Campbell 2002). How does this work?

The three strategies are matching, blocking, and stratifying. We define and illustrate them using the running example of a “nudge” experiment that seeks to identify what type of electric bill information leads households to reduce their consumption. In this example, there are $N$ households as subjects and $T$ different types of information on the electric bill, and the outcome is electricity consumption. One variable that could be used for matching, blocking, or stratifying is the physical size of each dwelling.

Matching means that there are $N/T$ matches with the same scores on the matching variable. There would be $N/T$ households matched on their square footage, and then the $T$ different treatments would be randomly null of no effect will be rejected in 54% of trials. As noted before, for the specified target effect size, power increases with the sample size and decreases with the level of statistical significance the researcher chooses. Instead of setting the number of subjects in each condition ($N$) and solving for power (which is useful when assessing existing studies to be replicated), it can also be useful to set power at a given value and solve for the number of subjects required. This is useful for designing a new study or replication.

While power calculations are best done in software, we provide a worked, simplified example. For this, we use a between-subjects design, which is representative of the majority of experiments in public administration. Suppose there is a treatment and a control condition (more conditions mean this calculation has to be repeated). The outcome (dependent variable) is a seven-point scale ranging from 1 to 7, with a mean of 3 and standard deviation of 1 in each condition. Suppose we set the level of statistical significance at $\alpha = 0.05$ in a one-tailed test, and the effect size $\delta$ as 0.2 standard deviations, which is 0.2 $\times$ 1 = 0.2 in our example. To calculate power, we use the critical values of the relevant statistical distribution, typically the $t$-distribution. It can be approximated by the standard normal distribution for larger sample sizes, which we use in our example. If the mean of the treated group is 3.2, the power for correctly rejecting the null hypothesis of no positive effect is approximately: $1 - p \{ Z > 1.64 - (0.2 / \sqrt{(1/75)}) \}.$ For a typical $N$ of 75 subjects for a comparison between two conditions, this yields a power of: $1 - p \{ Z > 1.64 - (0.2 / \sqrt{(1/75)}) \} = 1 - p \{ Z > 1.73 \} = 1 - p \{ Z > -0.08 \} = 1 - 0.4641 = 0.5359.$ That is, at the specified level of significance and the specified effect size, the sample size is also subject to diminishing returns, and excessively large samples can raise an ethical problem because they are wasteful and needlessly expose subjects to risk (Bacchetti et al. 2005).

5 Some researchers report a calculation of power using the estimate of the effect size they obtained in the same study. Cumming (2014) shows that such post hoc power can take almost any value and is thus not informative. Statistical power is most useful when calculated in reference to an effect size $\delta$ specified ex ante.

6 While power calculations are best done in software, we provide a worked, simplified example. For this, we use a between-subjects design, which is representative of the majority of experiments in public administration. Suppose there is a treatment and a control condition (more conditions mean this calculation has to be repeated). The outcome (dependent variable) is a seven-point scale ranging from 1 to 7, with a mean of 3 and standard deviation of 1 in each condition. Suppose we set the level of statistical significance at $\alpha = 0.05$ in a one-tailed test, and the effect size $\delta$ as 0.2 standard deviations, which is 0.2 $\times$ 1 = 0.2 in our example. To calculate power, we use the critical values of the relevant statistical distribution, typically the $t$-distribution. It can be approximated by the standard normal distribution for larger sample sizes, which we use in our example. If the mean of the treated group is 3.2, the power for correctly rejecting the null hypothesis of no positive effect is approximately: $1 - p \{ Z > 1.64 - (0.2 / \sqrt{(1/75)}) \}.$ For a typical $N$ of 75 subjects for a comparison between two conditions, this yields a power of: $1 - p \{ Z > 1.64 - (0.2 / \sqrt{(1/75)}) \} = 1 - p \{ Z > 1.73 \} = 1 - p \{ Z > -0.08 \} = 1 - 0.4641 = 0.5359.$ That is, at the specified level of significance and the specified effect size, the sample size is also subject to diminishing returns, and excessively large samples can raise an ethical problem because they are wasteful and needlessly expose subjects to risk (Bacchetti et al. 2005).
assigned within each match. Blocking is conceptually similar to matching but does not necessarily occur for the exact same scores, only for similar scores. In contrast, stratifying means there are fewer than \(N/T\) strata. For instance, strata may be the type of dwelling: apartment/condominium, town home, or single-family home. The \(N\) units would first be grouped into these three strata. Then, within each stratum, the \(T\) different treatments would be randomly assigned.

In addition to using household size, we could use the pretest electricity consumption as the matching, blocking, or stratifying variable. In either case, because the matching, blocking, or stratifying variables are correlated with the outcome, the power of the test to detect the effects of the \(T\)-1 treatments (each compared against the control condition) is increased relative to simply randomly assigning the \(T\) conditions across the \(N\) households. For each of the three strategies, the analysis must consider how the treatment was assigned, especially when the percentage assigned to a condition varies (Gerber and Green 2012).

Apart from the three strategies just discussed, the introduction of covariates is standard practice in public administration. The logic behind controlling for covariates is fundamentally the same as that behind matching, blocking, and stratifying: reducing error variance by comparing treatment and control within more homogeneous subgroups of the subject pool. The difference is that this is done mathematically in the analysis after the experiment has been conducted. Instead of conducting a \(t\)-test or ANOVA of differences between groups, the covariates (e.g., students’ gender, major and location of subjects’ residence) are included as additional variables. For this purpose, it can be helpful to express the analysis as a regression with dummy variables for the treatments. This is mathematically identical to an ANOVA, but some will find it easier to interpret the findings (Angrist and Pischke 2009).

Given the limited resources and growing expertise in replication, together with threats to internal validity, many of the original studies that have been replicated were underpowered. Therefore, the comparison between original and replicated results has been inconclusive. Paying close attention to statistical power when designing a replication study is an essential ingredient for maximizing the insights to be gained from the replication.

Implementing the Replication
Decision Point 4: Choosing the Critical Test Case
The next decision point is where to replicate, or to choose between replication sites. The choice of replication site requires a detailed consideration of context. Recently, several scholars have sought to develop a theory of context in public administration (e.g., Meier, Rutherford, and Avellaneda 2017; O’Toole and Meier 2015). Their motivation stems from the equivocal findings of observational studies in different contexts, which suggests that context matters and is an important aspect of theory in public administration. The multidimensional nature of context across time and space has led to calls for a more intensive study of its key dimensions, such as the political, environmental, and internal factors affecting public organizations (O’Toole and Meier 2015). Scholars further assert that context affects the relationships between variables in different settings by interacting with public administration practice, altering the relationship between the independent and dependent variables and thus reshaping theory. These factors influence site selection decisions.

The first step in choosing the critical test case is to distinguish between two ideal types of studies (with much gray middle ground): confined versus universal. Confined studies produce findings that are purported to be specific to the local context, whereas universal studies claim global generalizability. In public administration, confined studies are the norm: studies in specific institutional contexts are often framed as tests of general theory. Generally, the argument is that public administration phenomena are context-specific and boundary conditions apply (see below).

One example is the role of national culture. What holds in China may not work in Germany, and vice versa. New public management (NPM) practices may be effective in similar contexts (i.e., Anglo-Saxon countries), but much less so or not at all elsewhere. In contrast, universal claims are said to hold everywhere and every time, irrespective of context. For example, public service motivation and red tape have been presented as universal constructs with local variations, whereas agencification and autonomization are viewed as Scandinavian concepts with universal implications. This is not the place to discuss the type of claims researchers should make, but a study’s location on the confined-universal spectrum implies an understanding of context and a choice of replication sites.

In public administration, confined studies are the rule rather than the exception, so we take this as our stepping stone. Our advice for the second step is to apply both Mill’s (1843) method of difference or dissimilarity and agreement or similarity.\(^9\) The initial

\(^8\) The analysis should be presented with and without covariates to aid interpretation.

\(^9\) It should be noted that cloned dependent variables may differ across settings because of unique attributes of the settings.

\(^{10}\) This is closely related to the issue of boundary conditions, which is elaborated in decision point 5.
stage is to select sites where the study is likely to be replicated (the method of agreement) or not (the method of difference). Applying this logic to public administration, we should include theoretically selected sites in the replication design. Take the above example again. Suppose that theory predicts that NPM practices (e.g., competitive tendering) are only effective in Anglo-Saxon countries. A powerful multisite replication would redo the original study in countries such as Australia, Ireland, the United Kingdom, and the United States (where, following the logic of the method of agreement, the original finding is expected to replicate well) and in nation-states like Brazil, China, France, and Germany (where, through the lens of the method of difference argument, the original result is not expected to replicate). This example illustrates the difference in replicating a universal study: site selection is irrelevant because the original result is expected to replicate everywhere.

Many replications fall into the gray zone with neither fully confined nor fully universal findings. In the context of the above NPM example, an argument could be made that the effectiveness of performance pay is positively associated with a country’s degree of individualism. In this case, our advice is to sample replication sites across the individualism spectrum, taking care to select diverse sites. The findings from the replication can then be used in a moderated meta-analysis with individualism as the critical moderator. The prediction is that the effect sizes will be larger with higher individualism (and perhaps small or nonsignificant at the collectivist end of the scale).

Decision Point 5: Establishing Boundary Conditions
A primary motive for replication in public administration is to establish whether research findings work in different contexts and to understand what factors might affect the degree of generalization. Public administration replication often aims to make more explicit the moderators that explain why replication results vary, but boundary conditions may apply to many other aspects of a study, such as substituting a similar stimulus to achieve a desired effect, testing the cause–effect relationship on a different population, moving the experiment to a different setting, or assessing the consistency of results over time. An understanding of boundary conditions is important because, as discussed in decision point 4, the original study may have been confined—designed for a specific and limited purpose rather than for generalization. In any case, the goal is to make all relevant boundary conditions explicit because this permits an understanding of “the accuracy of theoretical predictors for any context” (Busse, Kach, and Wagner 2017, 578). Put simply, it translates theory to practice and helps implement the design science agenda of public administration. The fifth decision point, thus, requires researchers to clearly state boundary conditions and hypothesize their expected findings when replicating an original study.

Choice of replication site will influence boundary conditions. If the strategy is to replicate a study in a very similar setting, the purpose is probably to increase the likelihood of confirming the results or further refine the theory in the original study. The selection of a dissimilar setting suggests that the replication includes extensions that will push the boundaries of the original study to develop or adjust theory. Busse, Kach, and Wagner (2017) propose “inside-out” and “outside-in” approaches to exploring boundary conditions (figure 3). In an inside-out exploration (point A in figure 3), uncertainty about boundary conditions is low, and the accuracy of theoretical predications is high. Replications testing inside-out are likely to implement replications that incrementally change populations or measures and analyses [see Grimmelikhuijsen and Porumbescu’s (2017) replication of Van Ryzin (2013)].

With the outside-in approach (point B in figure 3), boundary conditions are highly uncertain, and the accuracy of theoretical predications is low. A replication based on dissimilarity and uncertain boundary conditions seeks to test and develop a theory for a new context. In this case, the researcher would hypothesize that the results of the original study would not be present in the replication. Confirmation of this hypothesis does not equate to “replication failure” because the focus is on breaking new ground. This means that “the problem statement is less clear and the research project does not follow a fully straightforward plan but more likely involves some feedback loops and iterations” (Busse, Kach, and Wagner 2017, 584).
An inside-out approach to the question of boundary conditions has been applied to the replications conducted in Hong Kong on organizational design. The Hong Kong government was designed by the British colonial powers and thus reflects the combined structures, cultures, and practices of East and West. Externally, public bureaucracy in Hong Kong looks much like the West. However, Hong Kong’s government and public services are much more centralized and hierarchical than those in most Western countries because of its duty-based and leadership-centric culture, which is typical of East countries. Replications of Kaufmann and Feeney’s (2014) study of red tape (Walker, Lee, and James 2017, 457) suggested the following proposition: “The results of replications on organizational design will be in the same direction as the original studies, but weakened by Hong Kong’s diverse culture.” Notable in this view is the likely greater tolerance of rules and decisions made by more senior actors in the hierarchy. This assertion was upheld in the replication, which resulted in nonsignificant findings.

Clearer understanding of boundary conditions in replications can inform public administration on how we use knowledge to solve problems and devise solutions in different contexts. Indeed, exploration of boundary condition fosters theory development, strengthens research validity, and mitigates the research–practice gap (Busse, Kach, and Wagner 2017).

Decision Point 6: Establishing Construct Validity

The sixth decision point concerns questions of construct validity, a topic that has recently been discussed in studies of comparative public administration (Jilke, Meuleman, and Van de Walle 2014; Jilke et al. 2017). In cross-cultural studies, bias has resulted from systemic differences in the measurement instruments (Poortinga 1989). More specifically, the measures used “do not correspond to cross-cultural differences in the construct purportedly measured by the instrument” (Matsumoto and Van de Vijver 2011, 18). At a basic level, translation from one language into the other already comes with subtle differences in meaning (Harzing, Reiche, and Pudelko 2013), implying that replication work that crosses countries and languages requires careful assessment of cross-cultural and cross-language differences, in combination with equally careful translation and interpretation procedures.

Construct validity is important in any experimental replication, and particularly so when replications extend the original study to different populations. Researchers must ensure that the conceptual construct being measured is consistent in meaning across different settings and that it can be “mapped onto a measurement scale in the same way” (Jilke et al. 2017, 1295). This means that a concept and its interpretation by respondents on a scale are equivalent when the respondents hold the same views and variation is only related to context. In generalization and extension replications, researchers can make decisions about the suitability of the measures and explicitly build into their research design measures that are appropriate for the context of the new study. If decisions are not made about construct validity in the replicated study, it may suffer from construct, method, or item bias (Jilke, Meuleman, and Van de Walle 2014). If the measures are not equivalent, the findings can suffer from construct bias (dissimilarity of latent concepts across countries), method bias (all types of biases that come from the methodological and procedural aspects of a survey), and item bias (different people understand or interpret the same survey item differently).

Construct validity can be challenging. The replication of Kaufmann and Feeney’s (2014) study (in Walker, Lee, and James 2017) illustrates some aspects of these challenges. Kaufmann and Feeney (2014) examined red tape and outcome favorability. The term “red tape” does not readily convey the same meaning in English as it does in Chinese. The Chinese translation of “red tape” is “繁文縟禮 (Fan-Wen-Ru-Jie).” The meaning is similar in both languages, but the origins differ substantially. The roots of Fan-Wen-Ru-Jie come from the Chinese history of Confucianism, and different schools of Confucian thought favor different interpretations. Furthermore, the direct back translation of Fan-Wen-Ru-Jie is “complicated wordings (writings or documents) and burdensome etiquette,” even though its definition is accepted as meaning “red tape.” To partially offset this potential bias, the replication of Kaufmann and Feeney (2014) was conducted in the English language. In short, when replicating an original study, the replication cannot simply copy or directly translate the original questionnaire because there may be cultural effects such as item inappropriateness or differential response styles (Matsumoto and Van de Vijver 2011).

Taking questions of construct validity one step further, Jilke et al. (2017) suggested that measurement equivalence should be checked during replication, particularly when populations and measurement or analysis change (i.e., conceptual generalization, empirical generalization, and generalization and extension). Jilke et al. (2017) offered detailed guidance to test various levels of measurement equivalence using multiple group confirmatory factor analysis. Those attempting replications are encouraged to consult this source. Finally, questions of construct validity are bound up with decision point 1 regarding whether the original study can be replicated, and decision point 7 concerning the interpretation of the replication’s results, which we now turn to.
Reporting Results
Decision Point 7: Deciding How to Compare the Findings
Determining how the findings should be compared depends on the number of replications, their power, and the nature and extent of their heterogeneity (see the above discussions of boundary conditions and confined studies). Working with sharp rule-of-thumb threshold differences (e.g., if the $p$ value of the replication is above .05 or the new effect size is half the original, the replication has failed) is unwise (Cumming 2014). Actually, we advise against referring to replications as “successes” or “failures” because these judgments can be highly subjective. As a collateral benefit, removing the success–failure classification from our replication terminology makes replications less threatening for the original authors and hence increases the likelihood that replication work will be undertaken.

Describing the results of replications in more objective terms is also more consistent with the ideals of social science. In any replication, the degree of replicability should be carefully discussed using a series of benchmarks (e.g., power-corrected $p$ values, full confidence intervals, and effect sizes). For an example of such careful reflection, we refer to Bouwmeester et al. (2017), while recalling that care and attention are needed to interpret the statistical evidence (Bordacconi and Larsen 2014). This is an appropriate methodology when only a limited number of replications are being analyzed, perhaps one or two.

When the number of replications that have been conducted increase scholars can compare findings using meta-analytic techniques. Fabrigar and Wegener (2016) illustrate how meta-analytic calculations can be used on a small set of experiments to show how additional replications of the original study change meta-analytic indices. In the example, Fabrigar and Wegener (2016, 75) walk through they show that when meta-analytic calculations are applied to “fragile” replication results—i.e., ones that fail by threshold differences—effect sizes can be detected across the population of studies. As the scale of the results being compared increases, for example, in a many-labs study that involves scholars from different research labs, more traditional meta-analytic techniques can implement focuses on the overall effect size and variations therein.

A Worked Example: Van Ryzin (2013)
In this section, we walk through the decision points to replicate Van Ryzin’s (2013) experimental test of the expectation disconfirmation theory (EDT) of citizen satisfaction. In Van Ryzin’s (2013) online experimental vignette methodology, subjects received either low or high expectations delivered by the mayor of an imaginary US local government. The subjects were asked to view either low- or high-performance street cleanliness photographs. The study confirmed the central hypothesis in EDT that “citizens judge public services not only on experienced service quality but also on an implicit comparison of service quality with prior expectations” (597). EDT was selected for replication because it could better explain citizen satisfaction in a variety of locations and thus had the potential to become a valuable practical framework for governments. Four replications were conducted. The first was an empirical generalization that sought to use, to the extent possible, the same measures and analysis as the original study among a population of Hong Kong citizens. Further replications extended the original study to explore the boundary conditions.

Planning the Replication
The design of the original experiment and information on the treatments and processes were clearly reported in the article (Van Ryzin 2013, 602–6). In the replication, the decision was made to include the original study’s author, who became a co-investigator for the grant funding. This provided access to all of the research materials needed for design and analysis and tacit knowledge on the implementation of the original study. For example, the pretest questions discussed in the article were not published (Van Ryzin 2013, 603). Decision point 1 was passed.

Decision point 2 requires an examination of the original study’s internal validity. EDT is typically studied using observational research methods such as cross-sectional surveys that raise questions of causality, and subjective measures of expectations and performance, and endogeneity. Van Ryzin (2013) applied an experimental research design to address these validity concerns of ambiguous temporal precedence. Careful reading of the article and discussion with the original author suggested that the threats to internal validity were low: subjects were drawn from a nationally representative sample and randomly assigned to relatively simple treatment conditions. Issues of maturation and attrition did not arise. Decision point 2 was thus satisfied.

Decision point 3 focuses on power and sample size. The sample size in the original study was 964 respondents, randomized across four arms of the study (arm 1 = 251, arm 2 = 257, arm 3 = 226, and arm 4 = 230). The original study did not discuss power analysis in relation to sample size. Analysis of the sample sizes reported (Van Ryzin 2013, 602) showed power at 0.95 [for the effect sizes $\delta$ set equal to Van Ryzin’s (2013) findings and the critical value $\alpha = 0.05$]. The original study used path analyses, so the minimum sample size required for power 0.95 was 76, based on MANOVA for repeated measures between factors. Power analysis
Implementing the Replication

To date, all experimental tests of EDT and citizen satisfaction have been based in Western liberal democracies (e.g., Grimmelikhuijsen and Porumbescu 2017). Among these tests, a number have been replications [see Filtenborg, Gaardboe, and Sigsgaard-Rasmussen (2017) on Danish citizens; Grimmelikhuijsen and Porumbescu (2017) on US MTurk subjects]. Hong Kong was selected as a critical test case (decision point 4) because it is an Asian municipality that relies on observational research methods to study citizen satisfaction. The replications of EDT and citizen satisfaction supported the theorized causal relationships (Filtenborg, Gaardboe, and Sigsgaard-Rasmussen 2017; Grimmelikhuijsen and Porumbescu 2017). These confirmatory replications of EDT on citizen satisfaction, together with widespread evidence supporting the theory in studies of goods and products (Oliver 2010), suggested a method of agreement (that the original study would be replicated). We were further motivated to conduct the replication because, if claims of universality could be upheld, EDT offered the prospect of a new and more robust approach to measure citizen satisfaction in the city.

The contexts in Hong Kong and the United States are very different, with large political, environmental, and internal variations. Hong Kong is governed by the Chinese principle of “one-country two systems” and reports directly to Beijing. It is, therefore, a more unitary system than the United States environmentally, social capital is lower in Hong Kong, and the government is very centralized and hierarchical [see Scott (2010) for a thorough review of the Hong Kong context]. However, EDT has been successfully applied to information technology products in Hong Kong (Thong, Hong, and Tam 2006). Given this, we adopted an inside-out approach to the question of boundary conditions raised in decision point 5. We expected that the theory would hold in Hong Kong, despite its very different economic system. To test this expectation and confirm decisions on method of agreement/inside-out replication, the first stage was to conduct an empirical generalization. For this, we sought to retain the original measurement and analysis as much as possible. However, the original vignettes were developed in the American context and were based around a mayor’s announcement to a medium-sized city. In 2017, Hong Kong was a city of 7.36 million people. Therefore, to make the replication as realistic as possible for Hong Kong respondents (see discussion of construct validity below), we changed the treatment from a citywide level to a Hong Kong district. For the generalization and extension replications, the expectation and performance treatments were further changed to capture the Hong Kong context. These replications examined street cleanliness, air quality, and secondary education because these policy issues were relevant in the local context.

EDT is a well-established framework in the management and marketing literature and, as noted above, has experienced increased popularity in public administration studies. The measures and manipulations exhibit good face validity. The topic in Van Ryzin’s EDT study was a technical and unambiguous public service: street cleanliness. The main concept’s construct validity (decision point 6) was therefore not overly problematic. For measurement purposes, we used the same wording for the expectation and performance manipulations and variable measurements as the original study. The performance manipulation in the original study included two pictures of a New York street showing different degrees of cleanliness. In the replication, a picture was taken of a Hong Kong street. One photograph (high performance) showed the original clean street; the other was manipulated using Adobe Photoshop to include litter (low performance). This was done to limit variation in the photographs to only litter (i.e., sharpen the treatment effect), and to enhance validity for the Hong Kong subjects. To avoid construct validity problems associated with translating English into traditional Chinese, all materials were prepared in English, which is widely spoken. Following these decisions and after finalizing the research design, the replication was implemented from February to August 2017.

Reporting the Results of the Replication

As noted above, comparing an original study with a replication can be challenging because of differences in context. However, for Van Ryzin’s (2013) replication, a universal claim was advanced, implying that the results from the original study and the replication would be comparable. We analyzed the replications by implementing the same path analysis as the original study. The results showed the same pattern and direction in the replication as original study, and comparable levels of statistical significance: expectations ⇒ satisfaction (positive), expectation ⇒ disconfirmation (negative), performance ⇒ disconfirmation (positive), performance ⇒ satisfaction (positive), and disconfirmation ⇒ satisfaction (positive). Given that several replications were conducted, it was possible to make further comparisons between the original study, our
study, and other replications (Filtenborg, Gaardboe, and Sigsgaard-Rasmussen 2017; Grimmelikhuijsen and Porumbescu 2017). Table 3 shows the expected relationship between the expectation and performance manipulations, and the satisfaction and mean scores in each study. The mean scores consistently show that performance has a high impact on satisfaction, resulting in high satisfaction when expectations are set low or high, and a low effect on satisfaction when performance is set low, and expectations are set low or high. This brief preview of the results from the several replications of Van Ryzin (2013) suggests that the direction of the findings was comparable and that EDT was validated on three occasions, providing emerging evidence of the theory’s universality.

### Future Research On Replication

Replication needs to become a normal scientific practice in public administration to improve the quality of scholarship, accumulate knowledge, develop mid-range theories, distill lessons for practice, and identify robust size effects and boundary conditions. Replication failures are not necessarily failures of good scientific practice; rather, they should be seen as the scientific process in action.11 The time for action is ripe. If public administration is serious about science, validity, and generalization, scholars should view replications more positively and take up this common framework. Practically, this suggests designing and undertaking replications for knowledge building rather than for instrumental purposes, such as falsifying a study result that, itself, has little impact on theory and practice.

The scholarly community needs to expand the meaning of what we refer to as replication, which could have a more immediate and even greater cumulative effect on our field. Replications are not limited to cloning previous studies or re-running controlled laboratory or field experiments. The findings derived from other research designs, like case studies and survey research, should also be subject to rigorous replication and verification. The same best practices we have suggested, or similar ones, can be used to more efficiently corroborate knowledge derived from non-experimental and quasi-experimental research designs (Fabrigar and Wegener 2016). Researchers should scrutinize these findings and test their boundaries in other samples and settings with different model specifications and research techniques, such as comprehensive literature reviews, meta-analysis, and longitudinal panel studies. Such thorough, persistent efforts to vet important research findings can significantly increase confidence in public administration research and grow the knowledge base in our field. Moreover, rigorous research on boundary conditions will pay big dividends for public administration practice as researchers and practitioners advance the design science ambitions of our field.

Advancing a common replication framework can be achieved through collective action. Replications are an essential part of the scientific method that serves the common interest of the scientific community and beyond. Existing reward structures need to be modified to incentivize replication (Everett and Earp 2015; van Witteloostuijn 2016). Currently, the field does not provide enough incentives for individual researchers to carry out and publish replication studies on a large enough scale. Nevertheless, public administration researchers do know something about solving this “tragedy of the commons.” Collective action problems can be overcome by group interventions (e.g., by altering the incentive structure and spreading the costs) or through individual initiatives (as when some individuals subordinate their self-interest and act for the common good). Both courses of action will help generate more replication studies in public administration.

From the experimental research design perspective taken in this article, many changes have taken

---

**Table 3. Comparison of Expectations and Performance Manipulations on Levels of Satisfaction in the Original Study and Replications**

<table>
<thead>
<tr>
<th>Expected Results</th>
<th>United States</th>
<th>Denmarkb</th>
<th>United States</th>
<th>Hong Kong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low expectations, low performance</td>
<td>Low</td>
<td>3.03</td>
<td>4.1</td>
<td>3.30</td>
</tr>
<tr>
<td>Low expectations, high performance</td>
<td>High</td>
<td>3.97</td>
<td>4.8</td>
<td>4.57</td>
</tr>
<tr>
<td>High expectations, low performance</td>
<td>Low</td>
<td>2.81</td>
<td>4.4</td>
<td>3.75</td>
</tr>
<tr>
<td>High expectations, high performance</td>
<td>High</td>
<td>4.11</td>
<td>5.4</td>
<td>4.92</td>
</tr>
</tbody>
</table>

*Note: Mean scores, 1 = very dissatisfied to 7 = very satisfied, unequal subscripts indicate statistically significant difference within pairs (low expectations, high expectations). Data are available in online supplementary materials.*

aVan Ryzin (2013).
bFiltenborg, Gaardboe, and Sigsgaard-Rasmussen (2017). One decimal place reported.

11 An anonymous reviewer suggested this line of argument.
place in the field since 1992 when Bozeman and Scott argued that experimental methods are not necessarily suited to studying public organizations, and have ethical and logistical barriers (Bozeman and Scott 1992).

The research agenda in public administration now embraces questions well suited for experimental methods, particularly questions of behavior (James, Jilke, and Van Ryzin 2017), institutional design (Meier and Smith 1994), and public service performance (Walker, Boyne, and Brewer 2010). Concerns over ethics and deception have been reduced, and logistical barriers have been lowered as research expertise in the field has grown, and more resources have been dedicated to experimental studies (James, Jilke, and Van Ryzin 2017).

As resources for experiments grow, more experimental laboratories and research centers will become operational, creating opportunities for more extensive replication agendas. When these laboratories are linked and working on common agendas, progress will speed up and become more efficient. Multisite or “many-lab” public administration studies are now feasible (for instance, see Klein et al. 2014; Open Science Collaboration 2015) and we could also amass more replications by making a replication study a mandatory component of PhD programs. Multiple replications can subsequently be used to accumulate estimates that can be used in meta-analyses. Technological advances facilitate this; researchers now have access to larger, more diverse groups of subjects (e.g., through MTurk or eLancing). Survey technologies permit simultaneous implementation of experiments across time and space (e.g., Qualtics). Datasets and protocols can be easily shared and stored. Actually, many-lab studies present an ideal way to explore boundary conditions, and they provide a real opportunity to advance the theory and practice of public administration.

Neither institutions nor individuals can be expected to subordinate their personal interest for the greater good, even though some do. Granting organizations (i.e., Netherlands Organisation for Scientific Research) and journals (i.e., Journal of Behavioral Public Administration) can incentivize replications. Researchers can choose to take up the replication agenda and might earmark a percentage of their publications for replications in their research area. [LeBel (2015) recommends the ratio of 4:1.] Yet over the long term, these contributions may be insufficient to mainstream replication in public administration. This means that public administration replicators must become more strategic and resourceful, as we have alluded to in this article. The field should strive for better replications in the same way that we aim for a better theory, research methods, and utilization. The field should trumpet the need for replications, try to refine and improve the process, and incentivize high-impact work in the same way we pursue other normatively important principles. The implementation of a common replication framework would place public administration at the forefront of the applied social sciences, according to it a leading role in the development of solutions to complex, human-related, real-world problems that involve value judgments. Eventually, replication can become a virtuous cycle of theory testing, refinement, and application.

Conclusions

Replication offers public administration scholars the opportunity to address some substantive questions and advance the theory and practice of public administration, in keeping with Simon’s (1996) notion that public administration is a design science that devises courses of action to change existing situations into preferred ones. This major contribution is centered around testing the boundary conditions of current theory and practice using conceptual, generalization, and extension replications. In this article, we outline a set of best practices for public administration researchers who are planning to conduct replication studies. These practices have been gleaned from our experience in one of the first replication laboratories in public administration peppered with advice from other disciplines. This protocol should help researchers avoid common problems, make wise decisions, and carry out better replications. The important elements include targeting the most impactful research findings for replication, shoring up internal validity before moving on to external validity, and identifying the most important boundary conditions to explore. Many useful practices have been suggested. We hope that this article will trigger debate on replication and that our colleagues will take up this agenda and help advance the scientific ambitions of the field, develop better mid-range theories, and glean useful lessons for practice.

Funding

This work was supported by: University Grants Committee, Research Grants Council (project no. CityU 11611516); Public Policy Research Funding Scheme from the Central Policy Unit of the Government of the Hong Kong Special Administrative Region (project no. 2015.A1.031.16A); City University of Hong Kong Department of Public Policy Conference grant and College of Liberal Arts and Social Sciences Capacity...
Building grant; National Research Foundation of Korea grant funded by the Korean Government (NRF-2017S1A3A2067636).

References


Mill, John S. 1843.