AARHUS UNIVERSITY

Coversheet

This is the accepted manuscript (post-print version) of the article.

Contentwise, the post-print version is identical to the final published version, but there may be differences in typography and layout.

How to cite this publication

Please cite the final published version:

Pedersen, M. J., & Stritch, J. M. (2018). RNICE Model: Evaluating the Contribution of Replication Studies in Public Administration and Management Research. *Public Administration Review*, 78(4), 606-612. <https://doi.org/10.1111/puar.12910>

Publication metadata

Title:	RNICE Model: Evaluating the Contribution of Replication Studies in Public Administration and Management Research
Author(s):	Mogens Jin Pedersen, Justin M. Stritch
Journal:	<i>Public Administration Review</i> , 78(4), 606-612
DOI/Link:	https://doi.org/10.1111/puar.12910
Document version:	Accepted manuscript (post-print)

General Rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

RNICE Model: Evaluating the Contribution of Replication Studies in Public Administration and Management Research

Replication studies relate to the scientific principle of replicability and serve the significant purpose of providing supporting (or contradicting) evidence regarding the existence of a phenomenon. However, replication has never been an integral part of public administration and management research. Recently, scholars have issued calls for more replication, but academic reflections on when replication adds substantive value to public administration and management research are needed. This concise article presents a conceptual model, RNICE, for assessing when and how a replication study contributes knowledge about a social phenomenon and advances knowledge in the public administration and management literatures. The RNICE model provides a vehicle for researchers who seek to evaluate or demonstrate the value of a replication study systematically. We illustrate the practical application of the model using two previously published replication studies as examples.

Replicability is a key tenet of the scientific method and positivist epistemological approaches to research. As Epstein (1980: 796) notes, “[t]here is no more fundamental requirement in science than the replicability of findings be established.” The validity of scientific claims depends on the extent to which analyses are reproducible and analytical results are reliable and generalizable to other situations and subjects (Campbell and Stanley 1963).

Replication—the process of repeating previous research efforts with the aim of confirming or extending previous findings—speaks to the principle of replicability.

Replication serves the important purpose of providing supporting (or contradicting) evidence regarding the existence of a phenomenon (Collins 1992; Mackey 2012). In statistical terms, replication lowers the probability of Type I or Type II error in the testing of any null hypothesis (Cohen 1994; Phye, Robinson, and Levin 2005; Robinson and Levin 1997) and provides a control for systematic extraneous variables possibly having confounded previous findings (Krauth 2000; Schmidt 2009). Moreover, meta-analysis is widely thought to be the platinum standard of evidence (Stegenga 2011)—and replications of previous research findings are a requirement for conducting them (Makel and Plucker 2014).¹ While replication is not a panacea (Makel, Plucker, and Hegarty 2012; Pashler and Wagenmakers 2012), dismissing replication implies a value of novelty over truth (Nosek, Spies, and Motyl 2012) and a misconception of science. As Tukey (1969: 84) notes, “[c]onfirmation comes from repetition. Any attempt to avoid this statement leads to failure and more probably to destruction.”

Replication studies play an integral role in the natural sciences for testing the reliability and demonstrating the generalizability of scientific findings (Madden, Easley, and Dunn 1995). However, replication has never been an integral part of public administration and management research. As Walker, James, and Brewer (2017, 11) note: “*Replication is not very widespread in public management, partly because replications are difficult to publish*

and faculty tenure and promotion practices often reflect a ‘publish or perish’ mentality.”

Recently, however, replication has become a more salient subject of scholarly attention and debate in the field. Researchers have issued calls to undertake replication studies, emphasizing how replication is an essential part of the scientific process that can help produce a sounder base of knowledge for our field (Walker, James, and Brewer 2017; Walker, Lee, and James 2017).

Scholars leading the field in experimental research are spearheading today’s explicit discussion of replication in public administration and management research (see the recent *Public Management Review* symposium, titled “Replication, Experiments and Knowledge in Public Management Research”). However, the discussion of replication should not get compartmentalized among groups of public administration and management scholars. Indeed, the steps *Public Administration Review* is currently taking to promote openness and transparency may enhance scholarly replication and help bring the issue to a wider audience, including scholars using survey and administrative sources of data (see Perry 2016).

While researchers have made much progress over the last 25 years, a major concern with extant public administration and management research is that many empirical tests of theory come from a small number of data sets (see O’Toole and Meier 2015). At the same time, an implicit appreciation for the principle of replication is also emerging. The field is beginning to see the use of meta-analysis (Homberg, McCarthy, and Tabvuma 2015; Gerrish 2016). Scholars are starting to replicate observational studies in different public management contexts (Meier et al. 2015). Similarly, scholars are examining both the stability of findings (Bullock, Stritch, and Rainey 2015) and the equivalence of measures in multiple country contexts (Kim et al. 2013). However, academic reflections on when and how a specific replication study adds substantive value to public administration and management research remain implicit and sparse.

This concise article suggests a simple conceptual model for evaluating how a replication study contributes to knowledge about a social phenomenon. The model is parsimonious and applicable for conceptualizing and evaluating when and how a replication study contributes to public administration and management research. Moreover, the model applies to scholars using a variety of different methodological approaches and data sources. The model defines the five main characteristics of replication studies: Relevance, Number, Internal Validity, Contextual Realism, and External Validity. Analyzing these characteristics can help to identify and clarify the contribution of a replication study to the public management and administration research community. This model is of utility to both authors situating the contributions of a replication study and reviewers assessing the value of the contribution.

We do not suggest that replication in public administration and management should be guided by idiosyncratic practices and standards unique to the discipline. On the contrary, we believe that the social sciences should work towards a shared framework for evaluations of replication studies. Although framed in the context of public administration and management research, the model we present is, therefore, extendable and applicable to other social science disciplines.

By framing the model in the specific context of public administration and management, we hope to help cultivate and promote replication within the discipline. Not least, we seek to draw attention to the issue from public administration and management scholars who are not yet engaged in today's discussions about replication in the field (Walker, James, and Brewer 2017; Walker, Lee, and James 2017). As previously mentioned, replication has never been an integral part of public administration and management research. However, if replication is to become fully integrated into the discipline, the value of replication research needs to be broadly recognized and a wide variety of scholars must be interested and directly involved in

replication. Replication efforts must not be pigeonholed to subgroups of scholars, such as researchers taking behavioral approaches to public administration (see Grimmelikhuijsen et al. 2017) or experimental researchers. Our model provides a conceptual framework for considering replication that is concise, but flexible enough to accommodate the diversity of substantive interests and methodological approaches in public administration and management research.

Current Frameworks for Replication

We are not the first to present a framework for assisting and guiding research efforts to replicate existing scientific findings in the social sciences. For example, Munafò et al. (2017) provide a set of measures that may improve research efficiency and robustness of scientific findings by directly targeting specific threats to reproducible science. Implementable by researchers, institutions, journals, and funders, the measures constitute practical and achievable steps toward improving rigor and reproducibility (e.g., blinding, pre-registration of study protocols, multi-site studies and team science consortia, reporting checklists, open data and materials, transparency and openness guidelines, open science practices in hiring and promotion). Similarly, Nosek et al. (2015) provide a set of standards to be included in journals' procedures and policies for publication promoting transparency, openness, and reproducibility in scientific practice. As another example, Open Science Collaboration (2015) recognizes that there is no single standard for evaluating replication success. They emphasize and use five indicators to evaluate reproducibility (significance and *P* values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes). Focusing specifically on replication in public administration and management research, Walker, James, and Brewer (2017) present a typology of different types of replication studies and outline a protocol to assist researchers who implement replications. Their protocol lists different types

of replications and highlights key characteristics and challenges for each type based on the changes to subjects, methods, and analyses relative to the original study.

The examples above provide valuable insights and guidance. Munafò et al. (2017) and Nosek et al. (2015) provide measures and standards that may promote replication endeavors if incorporated into scientific practices and processes. Open Science Collaboration (2015) presents a set of indicators researchers can use to determine whether a replication study succeeds in reproducing the findings of the original study. Finally, Walker, James, and Brewer (2017) suggest a useful taxonomy for distinguishing among different types of replications and outline the main features of each type in terms of subjects, methods, and analysis.

However, we are not aware of an existing framework for replication that may assist researchers and reviewers in systemic thought and evaluation of when and how a replication study contributes to knowledge about a given social phenomenon. In the following section, we present our model for conceptualizing and evaluating the value of a replication study. We then illustrate the practical application of the model with two examples.

The Conceptual Model

Figure 1 shows our conceptual model, RNICE, and the notation that defines the model:

[Figure 1 here]

The model suggests that the contribution and value of a replication study (V)—i.e., the extent of refinement and improvement to existing knowledge that the replication study provides to public administration and management research—is a function of five criteria: (1) Relevance, (2) Number, (3) Internal Validity, (4) Contextual Realism, and (5) External Validity.²

Relevance (R): *The relevance of the particular phenomenon subject to study.* Replication should address theoretically interesting and currently relevant research questions.

Additionally, replication studies could examine research claims that are widely held or accepted in the field but might have been insufficiently investigated in previous studies. 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.

Number (N): *The quantity (n) of other replication studies examining the phenomenon subject to study.* For any phenomenon, each additional replication has, *ceteris paribus*, less value than the preceding replication. However, many replications are not exact, and the type of replication (see Walker, James, and Brewer 2017) is a moderating factor to this logic. For example, say that several replication studies have examined a specific phenomenon using different subjects while applying the same methods and analytical approaches. In this context, an additional replication using different methods and analytical approaches may be of more value than an additional replication using different subjects but similar methods and analytical approaches as the extant studies.

Internal Validity (I): *The extent to which the internal validity of the replication is higher than in previous studies.* Referring broadly to the approximate truth of an inference (knowledge claim) about whether a given research design successfully identifies a causal relationship (or, in some cases, a correlation) for the population being studied³, this criterion relates to the research design of the replication relative to that of the original study or previous replications. Improvements in internal validity may, for example, result from the use of methods or estimators for examining the phenomenon subject to study that more effectively minimizes threats to internal validity (e.g., ambiguous temporal precedence,

history, maturation, selection, attrition, regression to the mean, testing, and instrumentation; see Gliner, Morgan, and Leech 2009; Posavac 2016; Shadish, Cook, and Cambell 2002). A properly implemented experimental design, or for that matter a high-quality quasi-experimental design, tend to offer strong claims to internal validity (Angrist and Pischke 2009; James, Jilke, and Van Ryzin 2017). Similarly, research design features such as a greater number of data observations, more reliable variable measures, and correcting for multiple testing can increase the value of a replication study by minimizing the reasons why causal (or correlational) inferences about the existence and size of covariation between variables could be wrong (Shadish, Cook, and Cambell 2002). For example, more data observations and more reliable variable measures may address issues relating to low statistical power and measurement error—issues increasing the risk of incorrectly concluding that two variables do not covary when they, in fact, do (Type II error). Similarly, correcting for multiple testing addresses the issue of fishing and the error rate problem—an issue increasing the risk of incorrectly concluding that two variables covary when they do not (Type I error).

Contextual Realism (C): *The extent to which the contextual realism (or ecological validity) of the replication is higher than in previous studies.* This criterion refers to the degree that the methods, materials, and setting of a research study approximates the real-world being examined (Reis and Judd 2014)—and therefore the extent to which the conclusions of a study are generalizable to the settings and situations in which the phenomenon subject to study would naturally occur (Wegener and Blakenship 2007). Improvements in contextual realism may, for example, result from more natural (less artificial and constricted) environments; research settings having more in common with real-life settings and situations. For example, findings from properly implemented field experiments or observations in the field (e.g.,

survey research or administrative data) may be more generalizable to real-world settings than findings from laboratory experiments or simulations.

External Validity (E): *The extent to which the replication improves scholarly knowledge about the scope conditions of the phenomenon subject to study.* This criterion refers to the validity of inferences about whether a causal relationship (or, in some cases, a correlation) holds over variation in persons, settings, treatments, and measurement variables (Shadish, Cook, and Cambell 2002). “Valuable” replications demonstrate the generalizability (or lack thereof) of previous findings to other units of analysis, settings, and/or situations.

Improvements in knowledge may, for example, result from testing whether the findings of a previous study on the effects of management on the performance of U.S. school teachers are replicable when examining data on European school teachers, U.S. unemployment caseworkers, or European unemployment caseworkers. Regardless of whether previously observed relationships are reproduced, replications can improve our knowledge of the boundary conditions of different phenomena, an important and often overlooked part of building knowledge.

RNICE Model Application

We illustrate the practical application of the RNICE model using two cases of replication in public administration and management research: (1) Bellé’s (2015) replication of Ariely, Bracha, and Meier’s (2009) study of motivational crowding effects; and (2) Rho’s (2013) replication of O’Toole and Meier’s (2004) study of the determinants and consequences of public service contracting. Walker, James, and Brewer (2017) specifically highlight Belle’s (2015) and Rho’s (2013) studies as examples of replications in public administration and management research and, thus, provide the basis for our case selection. By applying the

model to both a replication of an experimental study and of an observational study of administrative data, we show how the RNICE model can be applied to different types of research.

Bellé (2015)

Bellé (2015) examines the effects of performance-contingent monetary rewards on public employee performance, and the contingencies that may moderate these effects. As part of his investigation, Bellé (2015) replicates a study by Ariely, Bracha, and Meier (2009) which demonstrates that financial incentives and reward transparency (or observability) may interact negatively with respect to activities with a prosocial impact. In a randomized laboratory experiment, Ariely, Bracha, and Meier (2009) had 161 Princeton University undergraduates repeatedly click two keys on a computer keyboard for up to five minutes. The researchers donated to charity according to the number of clicks performed by a student. A randomly assigned subgroup of students who were offered a performance-contingent monetary reward outperformed those who were not offered a financial incentive. Among the students receiving a financial incentive, a random subsample whose performance and pay were kept secret tended to outperform those whose performance and pay were publicly observable.

Bellé's (2015) replication involves a randomized field experiment with nurses at a local health authority in Italy. As part of a mandatory training program, the nurses were tasked with assembling surgical kits for shipment to health care practitioners in a former war zone facing a humanitarian emergency. By random assignment, a subgroup of participants received fixed pay according to their normal hourly wage, while another subgroup received fixed pay plus a performance-contingent monetary reward. In addition, all nurses were randomly assigned to either a "disclosed" or a "secret" condition: some were told that the

individual performance of each participant would be displayed on a bulletin board whereas others were told that this information would be anonymous and not publicly disclosed.

Using the RNICE model as a vehicle to systematically evaluate the contribution of Bellé's (2015) replication, we conclude that the replication adds significant refinement and improvement to scholarly knowledge about public administration and management phenomena. As we elaborate below, Bellé's (2015) replication appears to add value in terms of four of the RNICE model criteria: Relevance, Number, Contextual Realism, and External Validity.

In terms of the Relevance criterion, public sectors across the globe continue to adopt performance-related pay, although evidence on its effectiveness in the public sector is mixed and inconclusive (Perry, Engbers, and Jun 2009). Bellé's (2015) replication advances our knowledge about the approaches to performance-related pay that have the greatest effects on public employee performance and, therefore, appears timely for a broad audience of scholars and decision-makers. What are the relative effects of different types of performance-contingent monetary rewards schemes in the public sector? In the context of performance-related pay, should an employee's performance and performance-contingent pay be disclosed or kept secret? Bellé's (2015) replication addresses these critical questions.

In terms of the Number criterion, Ariely, Bracha, and Meier (2009) replicate their "click for charity" experiment in a similar laboratory experiment involving 151 Massachusetts Institute of Technology students who were asked to cycle on an exercise bike, with researchers donating \$1 per mile. However, outside of Ariely, Bracha, and Meier's (2009) replication of their own work, Bellé (2015) appears to be the first to replicate Ariely, Bracha, and Meier's (2009) results.

In terms of the Contextual Realism criterion, Bellé's (2015) replication seems to offer stronger claim to contextual realism (or ecological validity) than Ariely, Bracha, and Meier's

(2009) study. Bellé's (2015) field setting and the nurses' task of assembling surgical kits has arguably more in common with real-life workplace settings and situations than Ariely, Bracha, and Meier's (2009) laboratory setting and the students' task of repeatedly clicking two keys on a computer keyboard as fast as possible.

In terms of the External Validity criterion, Bellé's (2015) replication expands our knowledge about the generalizability of Ariely, Bracha, and Meier's (2009) findings to other units of analysis, environmental settings, and situations. Whereas Ariely, Bracha, and Meier (2009) look at students in a laboratory setting in the United States, Bellé (2015) examines public employees in public sector work environments in Italy. Similarly, Bellé's (2015) replication informs us of whether Ariely, Bracha, and Meier's (2009) findings hold over variation in work task, treatment, and measurement variables.

Based on our reading of Bellé (2015) and Ariely, Bracha, and Meier (2009), Bellé's (2015) replication does not satisfy the Internal Validity criterion. Both studies—Bellé (2015) and Ariely, Bracha, and Meier (2009)—appear to have strong claim to internal validity. For example, both studies implement experimental designs that appear well thought out and executed and we are unable to pinpoint substantive sources to statistical bias or causal-reasoning error allowing us to differentiate between the two in terms of internal validity.

Rho (2013)

Rho (2013) examines the determinants and consequences of government contracting by replicating prior research by O'Toole and Meier (2004). Using pooled cross-sectional data from more than 1,000 Texas school districts (years 1997-1999), O'Toole and Meier (2004) show how government contracting is associated with cost savings (spending on the core instructional function) and service quality (educational performance). Furthermore, O'Toole

and Meier (2004) also examine how potential determinants (e.g., student enrollment, teacher turnover, revenue per pupil) correlate with contracting.

Rho's (2013) replication also uses data from Texas school districts, but comprising years 1997-2008, and thus a much longer time frame than O'Toole and Meier's (2004). Except for this distinction, Rho's (2013) replication involves the same methods, estimators, and variable measures as that of O'Toole and Meier (2004).

Using the RNICE model to systematically evaluate the contribution of Rho's (2013) replication, we conclude that the replication adds value to public administration and management literature by property of two of the RNICE model criteria: Relevance and Internal Validity.

In terms of the Relevance criterion, government contracting has been (and is) widely used as a mechanism for public service provision, both in the United States and abroad. However, evidence on the effects of contracting on performance and public service quality are mixed (Boyne 1998a; Domberger and Jensen 1997). Similarly, although research has expanded our knowledge about the determinants of contracting, much remains to be learned and understood (Brown and Potoski 2003; O'Toole and Meier 2004; Romzek and Johnson 2005). Thus, continued research in the area is needed. Examining the determinants and consequences of government contracting is thus a timely and relevant issue.

In terms of the Internal Validity criterion, Rho's (2013) use of data including more time periods reduces the concern that results are confounded by omitted variables affecting both contracting decision as well as outcomes and/or potential determinants of contracting (e.g., an exogenous event such as a change in government policy at a given point in time). Rho's (2013) replication thus appears to have stronger claim to internal validity than O'Toole and Meier's (2004) original study.

The value added of Rho's (2013) replication in terms of the Number criterion appears to be minimal. Rho (2013) is the first to do an exact replication of O'Toole and Meier's (2004) study—which adds value in terms testing the robustness of a previously reported results. At the same time, numerous studies have explored the relationship between contracting and performance (see Hodge 2000). While most of these studies focus on economic performance (costs and efficiency), others have targeted service quality. Research on contracting determinants is sparser, but does exist (see Boyne 1998b).

As Rho (2013) employs the same units of analyses, settings, treatments, and measurement variables as O'Toole and Meier (2004), Rho's (2013) replication does not contribute to knowledge in terms of the Contextual Realism criterion or the External Validity criterion.

Case Analysis Summary

By assessing and pinpointing the contribution of the two replication case studies using the RNICE model, we have presented an outline of how researchers and reviewers can operationalize the model to systematically evaluate and consider the value of a replication study. Using the model, we find that the replications by Bellé (2015) and by Rho (2013) both provide substantive refinement and improvement to public administration and management research—but in different respects. Both replications address timely research issues that a broad audience of scholars and decision makers find important, thus fulfilling the Relevance criterion. However, whereas Bellé's (2015) replication yields added value in terms of number (quantity of other replications), contextual realism, and external validity relative to the initial study, Rho's (2013) contribution refers to stronger claims to internal validity.

A replication study will satisfy between zero and all of the RNICE model's five criteria. As a baseline, replications satisfying none of the five criteria do not hold significant value.

Similarly, the value of replications satisfying only the Relevance criterion is also negligible. Discounting these exceptions, the number of model criteria a study satisfies should not be interpreted as automatically prescribing the magnitude of a replication's contribution. For example, Bellé's (2015) replication satisfies four criteria, whereas Rho's (2013) replication satisfies two criteria. Yet Bellé's (2015) replication does not necessarily offer twice the value as Rho's (2013) as a consequence thereof. The RNICE model can be used to systematically evaluate or situate when and how a replication improves knowledge about a social phenomenon, but should not be used as a formal value or metric of one replication study's value relative to another.

Having said that, the RNICE model is useful for structuring efforts to specify the magnitude of a replication's contribution. We can illustrate how this is true using Rho's (2013) replication. The RNICE model allows us to systematically identify the primary parameters of Rho's (2013) contribution: relevance and internal validity. Focusing on the internal validity component, we recognize that the use of data to include more time periods increases the internal validity of Rho's (2013) replication relative to O'Toole and Meier's (2004) original study. At the same time, Rho's (2013) replication could be built upon to exhaust potential gains in internal validity. For example, given the longitudinal nature of the data, panel data methods capturing unit specific fixed effects might arguably strengthen Rho's (2013) claim to internal validity. The use of experimental or quasi-experimental approaches would increase the strength of Rho's (2013) claim to internal validity even more.⁴ Importantly, our initial application of the RNICE model is what guides and structures our journey towards this conclusion.

Moreover, the RNICE model does not imply that the five model criteria are necessarily equally important. Indeed, the satisfaction of some of the criteria may yield greater contribution and value than the satisfaction of others. The relative importance of the criteria

depends on the phenomenon subject to study and the previous research. For example, say that two replication studies examine the same phenomenon. One study yields value added in terms of the Internal Validity criterion, while the other study contributes in terms of the External Validity criterion. Whether one study contributes more than the other depends on the previous research. Improvement in claim to internal validity may be more valuable than improvement in claim to external validity if the previous studies examine the phenomenon among a wide variety of people and settings using research designs with weak claim to internal validity. Conversely, the satisfaction of the External Validity criterion may yield greater value than the satisfaction of the Internal Validity criterion if the previous studies examine the phenomenon among the same people and setting using research designs with strong claim to internal validity.

Path Forward

By which criteria can we conceptualize and assess the value of a replication study to public administration and management research? Addressing this question is important for two reasons. First, replication speaks to the scientific principle of replicability and provides supporting (or contradicting) evidence regarding our current understanding of public administration and management phenomena. Replication studies increase certainty when findings are reproduced and can lead to theory development, innovation, and refinement when they are not. Second, the development of a common model for assessing replication studies can assist and improve future research efforts heeding recent calls for more replication in public administration and management research (Walker, James, and Brewer 2017; Walker, Lee, and James 2017).

Emphasizing five characteristics of replication studies that add substantive refinement and improvement to scholarly knowledge about public administration and management

phenomena, the RNICE model provides a conceptual reference point that can be used for evaluating how a replication study—extant or prospective—further knowledge within the field. Recognizing that our model is only a first step towards establishing a common conceptual standard for assessing and evaluating replication, we strongly encourage future dialogue and efforts to refine and improve our model. What are the main characteristics of replication that contribute knowledge about a social phenomenon and advance knowledge in the public administration and management literatures? We call on fellow scholars to join the debate. Ultimately, the social sciences should implement a shared framework for evaluating the value of a replication study in their *modus operandi*. Establishing a conceptual model on what constitutes valuable replication in public administration and management could be a starting point.

Notes

1. The notion of meta-analysis as the platinum standard of evidence is conditioned by the quality of the studies being analyzed—garbage in, garbage out.
2. As we elaborate later, the RNICE model can be used to systematically evaluate or situate when and how a replication improves knowledge about a social phenomenon, but the RNICE model should not be used as a formal value or metric of one replication study's value relative to another. Similarly, we return to discuss the relative importance of the five model criteria.
3. Note that our criteria definition is based on established validity typology (see Angrist and Pischke 2009; Shadish, Cook, and Cambell 2002) but incorporates the concept of 'statistical conclusion validity'—the validity of inference about the covariation among variables based on the data. We do so for the purpose of parsimony and because internal validity and statistical conclusion validity are closely related. Both are concerned with study operations and the relationship between variables. Statistical conclusion validity relates to errors in

assessing statistical covariation, whereas internal validity relates with causal-reasoning errors. As Shadish, Cook, and Campbell (2002, 63) note, internal validity thus depends substantially on statistical conclusion validity in quantitative research. Moreover, we merge the concepts because we seek to develop a conceptual model in which all model criteria are relevant for evaluating the value of a replication—regardless of the particular social phenomenon subject to study. Some theoretically interesting and timely research questions concern description and correlation rather than causal identification. In such cases, evaluation of causal-reasoning errors appears irrelevant.

4. We acknowledge that when working administrative or observational data, experimental variation (e.g., a natural experiment) is often a fortunate happenstance. Furthermore, choices of panel data methods, while allowing for the *possibility* of stronger internal validity (Zhu 2014), are sensitive to specification and assumptions about data structures and the nature of change over time (Stritch 2017).

References

- Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Most Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.
- Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. *American Economic Review* 99(1): 544–55.
- Bellé, Nicola. 2015. Performance-Related Pay and the Crowding Out of Motivation in the Public Sector: A Randomized Field Experiment. *Public Administration Review* 75(2): 230–41.
- Boyne, George A. 1998a. Bureaucratic Theory Meets Reality: Public Choice and Service Contracting in U.S. Local Government. *Public Administration Review* 58(6): 474–84.

- Boyne, George A. 1998b. The Determinants of Variations in Local Service Contracting: Garbage In, Garbage Out? *Urban Affairs Review* 34(1): 150–63.
- Brown, Trevor L., and Matthew Potoski. 2003. Transaction Costs and Institutional Explanations for Government Service Production Decisions. *Journal of Public Administration Research and Theory* 13(4): 441-468.
- Bullock, Justin B., Justin M. Stritch, and Hal G. Rainey. 2015. International Comparison of Public and Private Employees' Work Motives, Attitudes, and Perceived Rewards. *Public Administration Review* 75(3): 479–89.
- Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago, IL: Rand-McNally.
- Cohen, Jacob. 1994. The Earth is round ($p < .05$). *American Psychologist* 49(12): 997–1003.
- Collins, Harry M. 1992. *Changing Order: Replication and Induction in Scientific Practice* (2e). London, UK: Sage.
- Domberger, Simon, and Paul Jensen. 1997. Contracting Out by the Public Sector: Theory, Evidence, Prospects. *Oxford Review of Economic Policy* 13(4): 67–78.
- Epstein, Seymour. 1980. The Stability of Behavior: II. Implications for Psychological Research. *American Psychologist* 35(9):790–806.
- Gerrish, Ed. 2016. The Impact of Performance Management on Performance in Public Organizations: A Meta-Analysis. *Public Administration Review* 76(1): 48-66.
- Gliner, Jeffrey A., George A. Morgan, and Nancy L. Leech. 2009. *Research Methods in Applied Settings: An Integrated Approach to Design and Analysis* (2nd ed.). New York, NY: Routledge.
- Grimmelikhuijsen, Stephan, Sebastian Jilke, Asmus L. Olsen, and Lars Tummers. 2017. Behavioral Public Administration: Combining Insights from Public Administration and Psychology. *Public Administration Review* 77(1): 45-56.

- Hodge, Graeme A. 2000. *Privatization: An International Review of Performance*. Boulder, CO: Westview Press.
- Homberg, Fabian, Dermot McCarthy, and Vurain Tabvuma. 2015. A Meta-Analysis of the Relationship between Public Service Motivation and Job Satisfaction. *Public Administration Review* 75(5): 711-722.
- James, Oliver, Sebastian R. Jilke, and Gregg G. Van Ryzin. 2017. *Experiments in Public Management Research: Challenges and Contributions*. Cambridge: Cambridge University Press.
- Kim, Sangmook, et al. 2013. Investigating the Structure and Meaning of Public Service Motivation across Populations: Developing an International Instrument and Addressing Issues of Measurement Invariance. *Journal of Public Administration Research and Theory* 23(1): 79–102.
- Krauth, Joachim. 2000. *Experimental Design: A Handbook and Dictionary for Medical and Behavioral Research* (Techniques in the Behavioral and Neural Sciences Vol. 14). Amsterdam, The Netherlands: Elsevier Science.
- Mackey, Allison. 2012. Why (or Why Not), When, and How to Replicate Research, In *Replication Research in Applied Linguistics*, edited by Graeme Porte, 21–46. Cambridge, UK: Cambridge University Press.
- Madden, Charles S., Richard W. Easley, and Mark G. Dunn. 1995. How Journal Editors View Replication Research. *Journal of Advertising* 24(4): 77–87.
- Makel, Matthew C., and Jonathan A. Plucker. 2014. Facts Are More Important Than Novelty: Replication in the Education Sciences. *Educational Researcher* 43(6): 304–16.
- Makel Matthew C., Jonathan A. Plucker, and Boyd Hegarty. 2012. Replications in Psychology Research: How Often Do They Really Occur? *Perspectives in Psychological Science* 7(6): 537–42.

- Laurence J. O'Toole, Jr., and Kenneth J. Meier. 2015. Public Management, Context, and Performance: In Quest of a More General Theory. *Journal of Public Administration Research and Theory* 25(1): 237–56.
- Munafò, Marcus R. et al. 2017. A Manifesto for Reproducible Science. *Nature Human Behavior* 1(1): 0021. DOI:10.1038/s41562-016-0021
- Meier, Kenneth J., Simon C. Andersen, Laurence J. O'Toole Jr, Nathan Favero, and Søren C. Winter. 2015. Taking Managerial Context Seriously: Public Management and Performance in US and Denmark Schools. *International Public Management Journal* 18(1): 130–50.
- Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl. 2012. Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth over Publishability. *Perspectives in Psychological Science* 7(6): 615–31.
- Nosek, Brian A. et al. 2015. Promoting an Open Research Culture. *Science* 348(6242): 1422-1425.
- Open Science Collaboration. 2015. Estimating the Reproducibility of Psychological Science. *Science* 349(6251): aac4716. DOI:10.1126/science.aac4716
- O'Toole, Laurence J., Jr., and Kenneth J. Meier. 2004. Parkinson's Law and the New Public Management? Contracting Determinants and Service-Quality Consequences in Public Education. *Public Administration Review* 64(3): 342–52.
- Pashler, Harold, and Eric-Jan Wagenmakers. 2012. Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science* 7(6): 528–530.
- Perry, James L. 2017. Practicing What We Preach! Public Administration Review Promotes Transparency and Openness. *Public Administration Review* 77(1): 5-6.

- Perry, James L., Trent A. Engbers, and So Yun. 2009. Back to the Future? Performance-Related Pay, Empirical Research, and the Perils of Persistence. *Public Administration Review* 69(1): 39–51.
- Phye, Gary, D., Daniel H. Robinson, and Joel R. Levin (eds.). 2005. *Empirical Methods for Evaluating Educational Interventions*. San Diego, CA: Elsevier Academic Press
- Posavac, Emil J. 2011. *Program Evaluation: Methods and Case Studies* (8th ed.). New York, NY: Routledge.
- Reis, Harry T., and Charles M. Judd. 2014. *Handbook of Research Methods in Social and Personality Psychology* (2nd ed.). Cambridge: Cambridge University Press.
- Rho, Eunju. 2013. Contracting Revisited: Determinants and Consequences of Contracting Out for Public Education Services. *Public Administration Review* 73(2): 327–37.
- Robinson, Daniel H., and Joel R. Levin. 1997. Reflections on Statistical and Substantive Significance, with a Slice of Replication. *Educational Researcher* 26(5): 21–6.
- Romzek, Barbara S., and Jocelyn M. Johnston. 2005. State Social Services Contracting: Exploring the Determinants of Effective Contract Accountability. *Public Administration Review* 65(4): 436-449.
- Schmidt, Stefan. 2009. Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences. *Review of General Psychology* 13(2): 90-100.
- Shadish, William R., Jr., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.
- Stegenga, Jacob. 2011. Is Meta-Analysis the Platinum Standard of Evidence? *Studies in History and Philosophy of Biological and Biomedical Sciences* 42(4): 497-507.

- Stritch, Justin M. 2017. Minding the Time: A Critical Look at Longitudinal Design and Data Analysis in Quantitative Public Management Research. *Review of Public Personnel Administration* 37(2): 219-244.
- Tukey, John W. 1969. Analyzing Data: Sanctification or Detective Work. *American Psychologist* 24(2): 83–91.
- Walker, Richard M., Oliver James, and Gene A. Brewer. 2017. Replication, Experiments and Knowledge in Public Management Research. *Public Management Review*. DOI:10.1080/14719037.2017.1282003
- Walker, Richard M., M. Jin Lee, and Oliver James. 2017. Replications of Experimental Research: Implications for the Study of Public Management. In Oliver James, Sebastian R. Jilke, and Gregg G. Van Ryzin, eds., *Experiments in Public Management Research: Challenges and Contributions*, pp. 439-60. Cambridge, UK: Cambridge University Press.
- Wegener, Duane T., and Kevin L. Blakenship. 2007. Ecological Validity. In Roy F. Baumeister and Kathleen D. Vohs, eds., *Encyclopedia of Social Psychology Vol. 1*, pp. 275-76. Thousand Oaks, CA: SAGE Publications Inc.
- Zhu, Ling. 2013. Panel Data Analysis in Public Administration: Substantive And Statistical Considerations. *Journal of Public Administration Research and Theory* 23(2): 395-428.

Figure 1

The RNICE Model