P-Hacking, *P*-Curves, and the PSM–Performance Relationship: Is there evidential value?

Dominik Vogel
University of Hamburg
Department of Socioeconomics
Von-Melle-Park 9, 20146 Hamburg, Germany
dominik.vogel-2@uni-hamburg.de
ORCID: 0000-0002-0145-7956

Fabian Homberg
Luiss – Libera Università Internazionale degli Studi Sociali Guido Carli
Department of Business & Management
Viale Romania, 32, 00197 Roma, Italy
fhomberg@luiss.it
ORCID: 0000-0002-5119-2685

Short biographies

Dominik Vogel is an Assistant Professor of Public Management at the University of Hamburg, Germany. His research focuses on the motivation of public employees, leadership, and human resource management in the public sector, interaction of citizens with public administrations, and performance management.

Fabian Homberg is Associate Professor of Human Resource Management and Organizational Behavior at LUISS Guido Carli University, Rome, Italy. His current research interests include public service motivation and incentives in private- and public-sector organizations.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/puar.13273

Abstract

Recent developments in the social sciences have demonstrated that we cannot uncritically aggregate the published research on a particular effect to conclude about its presence or absence. Instead, questionable research practices such as *p*-hacking (conducting additional analyses or collecting new data to obtain significant results) and selective publication of significant results can produce a body of published research that misleads readers even if it contains many significant results. It is, therefore, necessary to assess the evidential value of the research on a certain effect, i.e., one must rule out that it is the result of questionable research practices. We introduce the *p*-curve method to public administration research and apply it to the research on the relationship between public service motivation (PSM) and individual performance, in order to demonstrate how the evidential value of a body of published research can be assessed. We find that this particular literature contains evidential value.

Evidence for Practice

- Public servants with high public-service motivation show better performance at the workplace.
- However, a set of significant findings on a particular effect is not necessarily an indicator that this effect actually exists, and as evidence-based decision making becomes more popular in public organizations, decision-makers need to rely on trustworthy and unbiased evidence.
- Introducing and applying the *p*-curve method, our findings reveal that the literature on the effect of public service motivation on individual performance is unbiased, i.e., it seems not to be based on *p*-hacking and publication bias and, thus, provides a good foundation for evidence-based decisions in public organizations.

Notwithstanding the efforts to enhance transparency in leading journals, with the aim to safeguard "quality and integrity of the science we use to build knowledge in public administration" (Perry 2017), the recent past saw rising levels of distrust in the results published by scholars (Hubbard 2015). This phenomenon is driven not only by the general public's increased distrust in scientific evidence on a broad range of topics, from climate change to dietary recommendations, but also the heightened skepticism among scientists themselves in multiple disciplines. The most prominent example of such an inner-disciplinary challenge involving published research is the "crisis of confidence" or "replication crisis" in psychology (Pashler and Wagenmakers 2012). In light of the recent developments in public administration and the stronger links forged with psychology (Grimmelikhuijsen et al. 2017; Jilke, Meier, and van Ryzin 2018), the replication crisis has become increasingly relevant for public administration research.

Besides cases of actual fraud, the starting point of the replication crisis in psychology was that only 39 out of 100 studies could be replicated in a large-scale replication attempt (Open Science Collaboration 2015). This study evoked the long-standing criticism (e.g., Ioannidis 2005) of the use of questionable research practices with additional evidence. It seemed to establish that such practices result in highly inflated type I error rates (wrongly concluding that there is an effect when there is actually none). Such practices include hypothesizing after the results are known (HARKing), conducting additional analyses or collecting new data to obtain significant results (*p*-hacking), and selectively reporting significant results. These behaviors naturally lead to the publication of literature that overwhelmingly consists of significant results (publication bias) (Earp and Trafimow 2015). As Starbuck (2016, 172) has stated, "Not only are HARKing and p-Hacking widespread, but sad to say, editors, reviewers, and colleagues often advise researchers to use these practices."

In light of the replication crisis, advocates of rigorous psychological research have not only advertised for a change in research practices (Wagenmakers et al. 2011) and promoted approaches such as pre-registration, open data initiatives, or open peer-review, but also developed and applied methods to assess the credibility of the published literature without the necessity to replicate all of it. One of the most prominent of these is the *p*-curve method (Simonsohn, Nelson, and Simmons 2014a), which uses the *p* values of published research works to assess their "evidential value." A body of published research contains evidential value if it is not solely the result of *p*-hacking and selective reporting of significant effects. "The only objective of testing for evidential value is to rule out selective reporting as a likely explanation for a set of statistically significant findings" (Simonsohn, Nelson, and Simmons 2014a, 535)

The aforementioned dynamics coincide with the recent calls for more rigorous and transparent research practices in public administration research (e.g., Perry 2017; Zhu, Witko, and Meier 2018; Jilke, Meier, and van Ryzin 2018). We, therefore, argue that we—as a scholarly community—should not only learn from other disciplines how research practices can be improved to increase the rigor of our methods but also develop transparency-enhancing practices within our discipline. A starting point would be the assessment of the evidential value of the published research. In other words, we should not only look forward but also backward, on the literature we build on to create further research, in order to avoid developing ideas on potentially shaky grounds.

As we cannot assess all public administration research at once and as it would also not produce meaningful insights to pan over all research areas, a thoughtful decision has to be made about where to start. A reasonable starting point would be to focus on a research question that is of major importance to the field. The few concepts originating in the field of public administration are potential candidates. Among those, a frequently studied

relationship, which is significant to the entire field, is the relation of public service motivation (PSM) (Perry and Wise 1990) to the performance of public employees. Given the prominent position of this relationship in the research conducted in the last two decades (Ritz, Brewer, and Neumann 2016) and its emphasized relevance for practitioners (Christensen, Paarlberg, and Perry 2017), it is a reasonable starting point for assessing the evidential value of public administration research. The critical role public servants and their behavior play in the efficiency and effectiveness of public service further justifies the focus on PSM and individual performance. From a more technical point of view, the PSM and performance literature is mature enough to facilitate the aggregation of studies, as a sufficient number of studies are available.

Hence, in this article, we ask if the published research pertaining to the effect of PSM on individual performance contains evidential value. In other words, we assess the extent to which the findings presented by the many primary studies of this relationship yield trustworthy results. In order to do so, we apply the *p*-curve method to this focal set of public administration research.

Overall, we aim to make three contributions to the literature. First, we seek to promote evidence-based argumentation in the field of public administration, showcasing the *p*-curve method and its potential value to public administration research. Second, we wish to offer an evidence-based assessment of the PSM–performance relationship that aggregates the findings from various studies. This assessment is based on the unique and rarely used *p*-curve method, in order to synthesize the existing literature instead of resorting to more the traditional means of a narrative review or a correlation-driven meta-analysis. Third, the results of the *p*-curve analysis inform debates on the trustworthiness of public administration scholarship. Ultimately, understanding the methods of research synthesis is also valuable for practitioners, as it facilitates a more accurate interpretation of empirical findings.

The Effect of PSM on Individual Performance

Public service motivation (PSM) is defined as "an individual's predisposition to respond to motives grounded primarily or uniquely in public institutions and organizations" (Perry and Wise 1990, 368). It comprises four dimensions: attraction to public policy-making, commitment to public interest, compassion, and self-sacrifice (Perry 1996). PSM is one of the most intensively studied concepts of public administration (Ritz, Brewer, and Neumann 2016), used to understand a variety of phenomena such as people's job choices and sector attraction (Korac, Saliterer, and Weigand 2019; Asseburg and Homberg 2020; Asseburg et al. 2020), the performance of public employees (Bellé 2013; Pedersen, M. 2015; Andersen, Heinesen, and Pedersen, L. 2014), the differences between public and private sector employees (Kroll and Vogel, D. 2018), and many more. Public administration scholars also use it to derive advice regarding the practice of public administration (Christensen, Paarlberg, and Perry 2017; Perry 2014).

One of the key questions of public administration research is whether it influences public servants' performance. Perry and Wise (1990) proposed this relationship as one of the three key effects of PSM and assumed that PSM is positively related to individual performance.

Scholars have used different theories to demonstrate a relationship between PSM and individual performance. Perry and Wise (1990) referred to work on motivational aspects of job characteristics. They argued that public servants with high PSM derive autonomy, task identity, and task significance from their work, which leads to increased perceptions of meaningfulness, which, in turn, drives them to perform well. The latter aligns with well-established models of job design theory (Oldham and Fried 2016; Vogel, D. and Willems 2020) that have received vast empirical support. Additionally, a high PSM is assumed to lead to increased organizational commitment, which also increases motivation and enhances

performance. According to Christensen et al. (2013), managers enjoying higher levels of PSM may value extra-role helping behaviors stronger in performance assessments.

Later work (e.g., Vandenabeele 2007; Andrews 2016) refers to the self-determination theory (Deci and Ryan 2004) to explain the reason employees with a higher PSM perform better. This literature argues that people with high PSM have internalized a specific set of values on which they build a public service identity. As Wright, Hassan, and Christensen (2017, 113) explain, "employees with higher PSM are expected to work harder and longer in their efforts to support the organization's mission because they find both the work to be especially meaningful and congruent with their own values." PSM is, therefore, an internalized autonomous motivation—a very strong type of motivation—and belongs to the core self of a highly self-determined employee. As a result, such an employee is more motivated to exert effort in their job.

Although some studies have separately analyzed the effects of the four PSM dimensions on individual performance, they usually do not theorize about it a priori. Vandenabeele (2009), for example, found no direct effect of compassion on individual performance and argued that this specific dimension does not fit in with public servants' professional context, as it could be seen as favoritism. Cheng (2015) found significant effects only for the commitment to the public interest dimension and argued that it is especially this dimension that drives the PSM–performance relationship.

In their initial article, Perry and Wise (1990) did not further specify what comprises individual performance. Hence, researchers have used the term flexibly, analyzing both measures of in-role and extra-role performance. This includes self-assessed in-role performance, supervisors' performance evaluations of subordinates, outcomes such as students' test scores, and self-assessments of organizational citizenship behavior. Following

Andersen, Boesen, and Pedersen, L. (2016), individual performance constitutes the performance ascribed to an individual employee. This could be in-role performance (the work an individual is hired to do) as well as extra-role performance (activities that go beyond the narrow tasks that employees are obligated to fulfill). Chen et al. (2009, 120) subsumed as extra-role performance all activities that are beneficial for the organization but "not explicitly required of employees." Homberg, Vogel, R., and Weiherl (2019) identified organizational citizenship behavior (OCB) as one particular form of extra-role behavior aimed at maintaining the status-quo of the organization. In this article, we follow their categorization.

Besides the direct effects, many studies have also considered the indirect effects of PSM on performance (e.g., Vandenabeele 2009; Gould-Williams, Mostafa, and Bottomley 2015). Several authors have studied the mediating or moderating effect of person–organizations fit or person-job fit, arguing that a high PSM can only result in better performance if employees have the opportunity to fulfill their need for contributing to society and helping others. Ritz, Vandenabeele, and Vogel, D. (forthcoming) give an overview of additional mediation and moderations studies, including, for example, transformational leadership, mission match, or public-service orientation of the organization.

Both observational (e.g., Vandenabeele 2009) and experimental (e.g., Bellé, 2013) empirical evidence supports such theoretical reasoning. The *p*-curve method allows the determination of whether the literature on the PSM–performance effect contains evidential value, i.e., if it reflects a true effect and, therefore, "[…] we can rule out selective reporting as the sole explanation of those findings" (Simonsohn, Nelson, and Simmons 2014a, 535).

What Is Evidential Value and Why Should We Care About It?

We first need to clarify that empirical research is mainly interested in detecting true effects. A true effect is a "genuine association [between two variables of interest] that is not entirely due to chance or biases (confounding, misclassification, selection biases, selective reporting, or other)" (Ioannidis 2008, 640). A body of studies on the same effect, which overwhelmingly reports insignificant results for a tested relationship, is usually indicative of the absence of a true effect. However, there are two separate but interrelated phenomena explaining the way a set of published studies on the same effect, with predominantly significant results, could also reflect a null effect: publication bias and *p*-hacking. Publication bias describes the practice of mostly publishing significant or "positive" results while insignificant or contradictory results are "stored in file-drawer" (Simonsohn, Nelson, and Simmons 2014a). It has been discussed for decades (Sterling 1959) and is well documented in various fields (Doucouliagos and Stanley 2013; Rost and Ehrmann 2017).

As significant results are much more likely to get published, there is an incentive to "produce" significant results by engaging in various practices that are commonly termed as *p*-hacking. The basis of *p*-hacking are the various decisions scholars need to make during the course of a research project that can also be used to influence the outcome of a study (Simmons, Nelson, and Simonsohn 2011). These decisions encompass, among others, the question of whether additional data should be gathered, which participants should be excluded, how measured variables should be aggregated to latent constructs, what statistical methods should be used to analyze the data, what control variables should be included, among others (Simmons, Nelson, and Simonsohn 2011). As a result of *p*-hacking, the probability of reporting an effect when there is actually no true effect (type I error) is no longer 5% (for a significance threshold of $p = 0.05$); it can be much higher. Hence, even a

body of research with a lot of significant results can be based on a nonexistent effect (Simonsohn, Nelson, and Simmons 2014a).

The *p*-curve method (explained in detail in the next section), therefore, aims to distinguish a set of significant results that is likely to be the result of selective reporting (i.e., publication bias and *p*-hacking) from a set of significant results that is not (Simonsohn, Nelson, and Simmons 2014a, 535). If the *p*-curve method rules out the possibility that a set of significant results is solely the result of selective reporting, then the set is deemed to contain "evidential value."

Taking into account the fact that the literature on a certain effect could reflect selective reporting on a nonexistent effect rather than being an honest reflection of a true effect is of vital importance for the field of public administration. It is essential since scholars build additional research on published research, which is only fruitful if the published research reflects a true effect. Moreover, policy-advice and managerial consulting are derived from the research published by public administration scholars (Newman, Cherney, and Head 2016).

The *p*-curve Method

The *p*-curve method (Simonsohn, Nelson, and Simmons 2014a) is a meta-analytical method, which means that it is designed to draw inference from the results of other studies, in contrast to methods that draw inference from primary data. Contrary to well-known meta-analytical methods such as fixed-, random-, and mixed-effects models (Hedges and Olkin 1985), the *p*-curve method does not use the reported effect sizes and their variances as the data source but rather the *p* values of all kinds of significance tests (e.g., *t*-test, correlation, *ANOVA*, regression). It, therefore, also does not infer on some kind of average effect size but considers the evidential value of the analyzed studies.[1]

The *p*-curve method is based on the distribution of significant *p* values of a set of findings (i.e., the *p*-curve). It compares the shape of this distribution with the shape that is to be theoretically expected in case of a true effect. This is possible due to the fact that the nature of *p* values determines a predictable distribution of these values. If there is a true effect, the *p* values always follow a right-skewed distribution with more small (e.g., $p = 0.01$) than large values (e.g., $p = 0.04$) (Wallis 1942; Hung et al. 1997; Simonsohn, Nelson, and Simmons 2014a).

In order to understand the underlying logic of the distribution of *p* values, it helps to imagine the distribution of *p* values of many studies on a nonexistent effect (i.e., if the null hypothesis is true). A *p* value indicates the probability of observing a result at least as extreme as the one observed if there is actually no effect. Hence, if there is no true effect, all *p* values are equally likely; the *p* values are uniformly distributed (Hung et al. 1997).

If there is a true effect, the distribution of the *p* values depends on the statistical power of the tests that the *p* values are derived from. The statistical power is the probability of observing a significant result if there is a true effect (Ellis 2010, 52), and it regularly increases with the number of observations a test is based on. The higher the statistical power of a test, the more frequently we will observe small *p* values. Hence, we obtain a more right-skewed distribution. With a statistical power of 80%, for example, 75% of the observed *p* values are expected to be smaller or equal to 0.035, and only 5% would be greater than 0.2 (Hung et al. 1997, 17). Figure A1 in Appendix A depicts the distribution of *p* values with varying power.

Regardless of the statistical power, the distribution of *p* values is right-skewed if they represent a true effect. In every case, smaller *p* values occur more often than bigger ones. This is not only true for the whole distribution but also for the subset of significant *p* values (i.e., $p < 0.05$).

The central problem of drawing inference from the distribution of *p* values is the same it is for all meta-analytic methods: We do not know if the published research we want to analyze is affected by publication bias or *p*-hacking. However, if we cannot take all the conducted studies into account and do not know if they have been *p*-hacked, our inferences will be biased. As a result, we might conclude that there is a true effect when, in fact, the published research consists mainly of false-negative findings (significant results that are based on a null effect).

As it is challenging to detect publication bias and *p*-hacking (Banks, G., Kepes, and Banks, K. 2012), the *p*-curve method assumes, by default, that there is selective reporting. It makes use of the fact that if there is a true effect, small *p* values are more likely than greater *p* values within the subset of significant *p* values as well. Hence, the distribution of significant *p* values has to be right-skewed. The *p*-curve method assumes that all significant effects ultimately get published or that there is at least no biasing selection effect. This allows for an unbiased inference from the subset of significant findings (Simonsohn, Nelson, and Simmons 2014a; Simonsohn, Simmons, and Nelson 2015). Based on this assumption, the *p*-curve method tests if the distribution of significant *p* values is right-skewed—as one would expect in case the *p* values stem from tests of a true effect.

One might argue that focusing only on significant results might bear another risk: biased inference due to *p*-hacking. Unlike publication bias, which only affects the number of published insignificant results, *p*-hacking also affects the number of significant results by adding more significant findings to the body of published research and thereby affecting the distribution of significant *p* values. As the *p*-curve method inferences from the distribution of significant *p* values, it is affected by *p*-hacking. As shown in figure A2 in Appendix A, the *p*-curve nevertheless allows to differentiate between a set of *p*-hacked studies that reflects a null effect and a set of studies that suffers from *p*-hacking but is based on a true effect. In the first

case, the distribution is left-skewed, while it is still right-skewed for a true effect (if there is a minimum of statistical power). A right-skewed distribution, therefore, indicates evidential value (Simonsohn, Nelson, and Simmons 2014a, 536).[2]

The Evidential Value of the Literature on the PSM–Performance Effect

Given the prominence and practical as well as theoretical relevance of PSM (Ritz, Brewer, and Neumann 2016; Christensen, Paarlberg, and Perry 2017) and its effect on performance, we use the *p*-curve method to assess whether the published research on the PSM–performance effect contain evidential value. The *p*-curve method allows the determination of whether the literature on the PSM–performance effect contains evidential value, i.e., if it reflects a true effect and, therefore, "[…] we can rule out selective reporting as the sole explanation of those findings" (Simonsohn, Nelson, and Simmons 2014a, 535). Building on the aforementioned logic, the subsequent sections apply the *p*-curve method to the branch of the literature that studies PSM and performance.

Literature Search

In order to create the database for the *p*-curve analysis, we conducted a systematic search on the Web of Science Core Collection and Social Sciences Citation Index and searched looked for reference lists of retrieved articles. The studies had to have been published in English. The search terms were "Public Service Motivation" (including its variants, e.g., PSM) and "performance" or "organizational citizenship behavior" (including its variants, e.g., OCB). The studies could either be published as a journal article or a book chapter. The time range for the search was 1990 to 2017. The start point marks the publication of Perry & Wise (1990), establishing the PSM research field. We started working on this project in 2017. This

way, 295 studies were initially identified. After screening them, 216 were excluded, as they did not address the PSM–performance relationship.

The remaining 79 studies were assessed in detail. They had to meet two criteria to be included in the *p*-curve analysis: First, the studies had to feature an empirical analysis of PSM in relation to an individual performance measure. As performance is a multi-dimensional concept, we needed to be clear about the type of performance that is relevant to our analysis. Thus, we decided to focus on the core aspects of in-role and extra-role performance, consequently excluding other measures such as job satisfaction and organizational commitment, which are regularly summarized under the label of performance. As our main interest is individual performance, we also excluded studies focusing on organizational performance measures. Second, studies needed to use a research design and analytical method that is based on *p* values or where calculating a *p* value is possible. Figure 1 summarizes the search procedure, which yielded 40 studies on the PSM–performance relationship. Of them, 35 reported significant effects and were eligible to be included in the *p*-curve analysis.

[Figure 1 here]

The Selection of p Values and Effect Sizes

In order to extract the effects from the 40 identified studies with significant or insignificant results on the PSM–performance relationship, we followed the procedure recommended by Simonsohn, Nelson, and Simmons (2014a; 2015). The first important step is to select *p* values that meet three criteria: they have to test the hypothesis of interest, have to have a

uniform distribution under the null, and have to be statistically independent of other *p* values (Simonsohn, Nelson, and Simmons 2014a, 542). To meet these criteria, only one *p* value per study can be selected. If a paper reports multiple statistical models (e.g., multiple regression models), we included the *p* value extracted from the "most complete" model. In the case of cross-sectional designs, it was the model with all the control variables. If an article tested a moderation effect of PSM with another independent variable, we used the model without the interaction term, as the *p* value of the main effect would not be uniformly distributed under the null otherwise (Simonsohn, Nelson, and Simmons 2014a, 543). If multiple measures of performance were tested or the dimensions of PSM were analyzed separately, we selected the effect with the smallest *p* value. Following the advice of Simonsohn, Nelson, and Simmons (2014a), we created a "*p*-curve disclosure table," making our selections transparent. The table is displayed in Supplementary Material A.

In order to overcome the limitation of including only one *p* value per study, we conducted two additional analyses. These analyses ensured that our results are not mainly driven by our selection procedure. For the additional analyses, we replaced the *p* values for studies with more than one *p* value. In this case, we sorted the *p* values by size. For the first alternative, the median value was used (or the one to the left of the median if there was an even number of *p* values). The second alternative replaced the *p* values with the largest one.

As the field of public administration does not follow strict reporting standards for statistical results, we faced an additional challenge in identifying and selecting the *p* values for the *p*-curve analysis. For the majority of the articles selected, it was possible to calculate a partial correlation from the regression or structural equation model results using the formula by Aloe (2014).[3] However, in 15 of the 40 articles, the reported test statistics were not sufficient to calculate an exact *p* value. We, therefore, developed a decision tree (displayed in Supplementary Material B) to be able to extract an exact *p* value nevertheless. The measures

taken range from contacting the authors twice to ultimately using the reported correlation coefficient in order to calculate a *p* value.

The *p* values of the 40 studies listed in Appendix B were ultimately used to carry out the *p*-curve method. Five of them were not incorporated in the calculation of the *p*-curve, as they were insignificant, and the *p*-curve method only uses significant *p* values. As already explained, the reason for excluding insignificant *p* values is that we cannot assume that the published insignificant findings are an unbiased sample of all insignificant studies conducted (publication bias). We nevertheless calculated an exact *p* value for these five studies in order to rule out possible reporting errors in the original studies.

Results

In order to execute the *p*-curve method, Simonsohn, Nelson, and Simmons (2014a) developed an *R*-based (R Core Team 2020) web app (http://www.p-curve.com). We used the *p*-curve app version 4.06 to generate the results described in this section. The data and analysis code is available at the Open Science Framework: https://doi.org/10.17605/OSF.IO/PTFYE.

Figure 2 displays the results of the *p*-curve analysis. The solid line represents the observed distribution of the significant *p* values obtained from the 35 studies on the effect of PSM on individual performance. The dashed line depicts the expected distribution of *p* values if the studies would have been conducted with 33% power. The dotted line represents the expected uniform distribution of *p* values if there was no true effect. In other words, the right-skewed deviation from the dotted line implies that the literature potentially has evidential value, but further checks will be necessary to substantiate this assessment.

[Figure 2 here]

Table 1 displays more detailed results of the *p*-curve analysis. First, we statistically examined whether the *p*-curve is right-skewed, as this is what we would expect from a set of studies that is performed on a true effect. The corresponding results are depicted in the first row. The first column reports the results of a binomial test, which determines if the share of *p* values that are smaller than 0.025 (i.e., the left half of the graph) is significantly greater than the share we can expect if there is no true effect. This is the case, as the share of *p* values smaller than 0.025 constitutes 86%, which is significantly larger than 50% ($p < 0.001$). We take this as a first indication in concluding that the PSM–performance literature contains evidential value.

However, as this binomial test is prone to ambitious *p*-hacking, Simonsohn, Simmons, and Nelson (2015) developed the continuous test, which is reported in the second and third columns of Table 1. The continuous test requires that either the half *p*-curve test be significant with $p < .05$ or both, and the full *p*-curve and the half *p*-curve test be significant with $p < .1$. In our case, both tests were significant with $p < .0001$ ($Z = –20.41$ and $Z = –21.08$). Hence, the results from the continuous test provide further support for the initial assessment that evidential value is present in the PSM–performance literature. Both tests indicate that the distribution of *p* values is significantly right-skewed, which is what one would expect for a set of studies reflecting a genuine effect.

One might argue that the mildly increasing slope of the *p*-curve towards *p* values between 0.04 and 0.05 might be an indication of *p*-hacking. However, this increase could also be a result of the fact that only a significance level was given for two studies and we had to assume a *p* value near the threshold of 0.05. When these two studies were excluded in a robustness analysis, the proportion of *p* values between 0.04 and 0.05 decreased to 6%, which is not distinctive from the *p* values between 0.03 and 0.04 (see alternative A in Supplementary Material D). As already mentioned, we conducted additional robustness

checks in which we replaced *p* values for the studies that reported multiple *p* values, as they tested multiple performance measures or multiple dimensions of PSM. The results are reported in Supplementary Material D. Replacing *p* values with the median ranked *p* value or with the biggest *p* value does not result in substantive differences from the main results shown in Table 1.

[Table 1 here]

Our overall conclusion, therefore, is that the published literature pertaining to the effect of PSM on individual performance contains evidential value and is, therefore, not substantially affected by selective reporting (*p*-hacking and publication bias).

Nonetheless, while evidential value might be present, it may not be adequate. As the *p*-curve method allows us to test whether the studies' evidential value is inadequate, we applied this procedure to further substantiate the results. The evidential value is perceived to be inadequate if it is obtained from a set of studies that has less than 33% power. The *p*-curve method assesses this by testing whether the *p*-curve of a set of studies is flatter than the *p*-curve would be if the same set of studies were to be conducted with 33% power. The *p*-curve analysis suggests that 71% of the studies would have *p* values smaller than 0.025 on conducting them with 33% power. The second row in Table 1 depicts the corresponding results. In the case of the considered PSM–performance studies, the *p* values for the full *p*-curve and the half *p*-curve test are greater than .99. Hence, the observed *p*-curve is not flatter than it would be if the studies were conducted with 33% power. We, therefore, conclude that the evidential value of the 35 assessed studies on the PSM–performance relationship is not inadequate, substantiating the overall assessment.

A final problem emerges, as employees' self-report of performance is known to be a biased measure (Meier and O'Toole 2013). Thus, we conducted two additional *p*-curve analyses, which can be interpreted as robustness checks. The first one excludes the studies that rely on self-reported (n = 29) or hypothetical (n = 1) performance and only considers the ones with supervisor-assessed performance (n = 4) and objective performance measures (n = 6). A second analysis only includes the studies that used objective performance measures. The results, which are reported in Supplementary Material E, again indicate that the studies contain evidential value for the effect of PSM on performance. However, they are less unequivocal, which is partly due to the small number of studies available. Nevertheless, it has to be noted that the six studies using objective performance measures are based on 8,087 observations.

Readers might also be interested in the average effect size of the PSM–performance relationship. We used the *metafor* package (Viechtbauer 2010) for *R* version 3.6.2 (R Core Team 2020) to estimate an average effect size using a multi-level random-effects model with the effects extracted for the *p*-curve analysis (effect with smallest *p* value, median *p* value, and largest *p* value). The meta-analysis revealed an average effect of r = .188, 95% *CI* [.139, .236], d = 0.383, which can be classified as a small effect (Cohen 1988).

Discussion

Synthetizing 40 primary studies focused on the relationship between PSM and individual performance, we found that this branch of the literature displays evidential value. The results of the *p*-curve analysis indicate that the published literature on the PSM–performance effect is not substantially affected by researchers tweaking their analyses to obtain significant results (*p*-hacking) and selective publication of significant results (publication bias).

This finding has several implications for the research community. First, it justifies committing resources to study the PSM–performance link. Researchers are not tilting at windmills here. Instead, they are generating insights of the utmost importance for the conduct of public organizations. Consequently, this field of research helps bridge the research-relevance gap and contributes to the forging of a stronger exchange with public management practitioners. We believe that there is still a necessity for additional research on the PSM–performance link. For example, we found only nine studies that do not use self-reported measures of performance, which was challenging, as we should interpret self-reported measures very carefully (Meier and O'Toole 2013). Additionally, there is more to learn about the exact theoretical mechanisms that link PSM and individual performance, including mediators and moderators, as well as the role of the PSM dimensions (see Ritz, Vandenabeele, and Vogel, D. forthcoming, for a detailed discussion). From an ad hoc look at the reported effect sizes, they appear substantially smaller for objective measures of performance, and a third of the studies using objective measures are even insignificant (see Appendix B).

Second, the mere fact that we need techniques to identify literatures affected by *p*-hacking, HARKing, publication bias, or other distortions highlights the problems embedded in the research practices applied in the field. For example, Starbuck (2016, 171) described HARKing and *p*-hacking as "so common, indeed, that some researchers misperceive them as legitimate." When undesired practices are taken for granted, there is a danger that they will become institutionalized. For example, similar dynamics have been observed in the literature pertaining to corruption, where petty corruption can become systemic through movement on a "slippery slope of corruption." (Anand, Ashforth, and Joshi 2004). Here, the latter describes the process of institutionalization (Anand, Ashforth, and Joshi 2004). While the responsibility to improve research practices lies on all shoulders involved (i.e., the individual researchers,

reviewers, editors, as well as research funding agencies and universities), the *p*-curve analysis is a tool that enables researchers to detect the problems arising in a given piece of literature before it becomes institutionalized.

Third, taking the subject area of PSM and performance as an example, public administration research, in general, may benefit from developing more standardized practices. For example, when building the database for this study (and other works of research synthesis), we noted a large variance in the reporting standards of the primary studies. While some studies clearly reported all data required for the *p*-curve analysis presented here, most were less detailed. For some studies, we even needed to contact the authors and, in cases where we did not get an answer, even had to approximate the *p* value. Moreover, meta-analysts regularly face similar difficulties while trying to find the necessary information to produce effect size. This situation highlights a particularly valuable aspect of using techniques of research synthesis: The process of gathering the underlying data required for applications such as the *p*-curve or meta-analysis reveals shortcomings in the reporting practices of individual researchers. Public administration could use the reporting standards of other disciplines, such as those of the American Psychological Association (American Psychological Association 2020), to overcome these issues. As a rule of thumb, authors should aim to provide enough information such that others can recalculate *p* values or calculate standardized effect sizes such as Cohen's *d*. For t-tests, this requires a *t* value and the test's degrees of freedom. For regression tables, unstandardized coefficients, a measure of uncertainty of the estimated coefficient (*t* value or standard error), and the number of observations used for the model are required. Structural equation models should be accompanied by a table of unstandardized path coefficients and standard errors of the paths. In Supplementary Material F, we have provided a set of four principles that should be followed while reporting statistical results besides

citing examples for reporting the most common tests (also see American Psychological Association 2020).

Ultimately, the *p*-curve analysis technique offers a fruitful soil for developing research projects and validating a stream of research. We looked at the PSM–performance link in this article; similarly, other parts of the literature, such as the one on red tape or publicness, can be subjected to the *p*-curve analysis. Literatures characterized by the presence of a large set of quantitative studies are ideal. There might even be the need for a large-scale collaborative project for assessing the public administration literature more systematically. Nonetheless, one problem of the field is its lack of repeated testing of the same effect, which is driven by the logic of "novelty" that is applied by most journals to qualify papers for publication ("the Neophilia disease;" Antonakis 2017). This condition imposes a severe boundary for research synthesis techniques in general and *p*-curve analysis in particular. Only a minority of relationships have been tested often enough to make such an analysis feasible. However, as the *p*-curve method can be applied to any kind of set of published research, public administration scholars can also use it by defining alternative sets of published research. This could be the research on a specific construct, regardless of the tested effect, the articles published in a certain journal, or even the work of individual researchers.

Furthermore, *p*-curve analyses could complement meta-analyses to deliver a more holistic synthesis of a given field. Meta-analysts are interested in quantifying the average effect in a given literature, whereas the *p*-curve method does not indicate how large an effect is or whether it is theoretically meaningful. "The only objective of testing for evidential value is to rule out selective reporting as a likely explanation for a set of statistically significant findings" (Simonsohn, Nelson, and Simmons 2014a, 535). Hence, it is a perfect complement to traditional meta-analysis.

Limitations

We believe that the *p*-curve method provides valuable insights into the study of public administration. However, the limitations of this method have to be acknowledged (see e.g., Simonsohn, Nelson, and Simmons 2014a; Simonsohn, Simmons, and Nelson 2015; Ulrich and Miller 2015; Bruns and Ioannidis 2016). First of all, we have been able to assess the evidential value of the published research on the PSM–performance effect. Still, we cannot draw any conclusions regarding the underlying theory for this effect, and we cannot assess if the *p*-curve analysis is based on a valid test of this theory. Furthermore, we cannot derive any conclusion regarding the size of the effect of PSM on individual performance. We only infer from the results that the effect is not zero, but it could nevertheless be very small. Second, the *p*-curve method often fails to detect a lack of evidential value when a high amount of *p*-hacking is present (Simonsohn, Nelson, and Simmons 2014a). This is because unhacked results produce a strong right-skewed curve while *p*-hacking shifts the curve only mildly towards a left-skewness. The *p*-curve method, therefore, misses a lack of evidential value if a lot of *p*-hacked studies are combined with a small number of true effects. Finally, one of the underlying assumptions of the *p*-curve technique is that all significant findings will ultimately get published, which is debatable (Simonsohn, Nelson, and Simmons 2014a). However, the current academic publication system gives preference to studies presenting significant findings—meaningful null-results are still the exception. Hence, while the underlying assumption may be perceived as too strong, it contains a kernel of truth that accurately reflects current publishing practices. One must, therefore, interpret the *p*-curve as a conservative test of evidential value (Simonsohn, Nelson, and Simmons 2014a, 546) due to its lower power to detect evidential value on application to correlational studies. Ultimately, the *p*-curve method does what it is supposed to do (Simonsohn, Simmons, and Nelson 2016): It assesses if the tested set of studies is solely based on selective reporting, not whether there

is a causal effect. Hence, the *p*-curve method appears to be a viable tool to synthesize a mature branch of the literature.

Conclusion

Research progresses by building ideas on the works of others. The *p*-curve method is a tool that allows scholars to evaluate the grounds on which new ideas stand. It also helps practitioners determine what results they can rely on. We opted to take a closer look at the relationship between PSM and performance, and our results reveal promising grounds for future research. The results of the *p*-curve analysis support the presence of evidential value in this literature. Hence researchers can trust previous findings, and it seems worthwhile to further study the PSM–performance link. However, the field of public administration is much wider, and other sub-streams in the literature should be subjected to the *p*-curve analysis in order to develop a more comprehensive picture of the current state of the literature. For example, perceptions of red tape in relation to organizational- and individual-level outcomes seem to be good candidates for future *p*-curve applications. However, one needs to keep in mind that a synthesis of research relies on the primary studies serving as its basis. Hence, in order to enhance the accuracy of the *p*-curve method, publication bias methods, and meta-analyses in general, there is a strong need to give more space to meaningful non-findings and replication research in the published literature (Kepes, Banks, G., and Oh 2014).

Notes

[1] There is also an adaptation of the *p*-curve method that estimates a true effect size from the published significance tests (Simonsohn, Nelson, and Simmons 2014b), but it serves a different purpose and will, therefore, not be discussed in this article.

[2] For a more formal discussion on the effects of p-hacking on the distribution of p values, see Supplementary 3 of (Simonsohn, Nelson, and Simmons 2014a). For a discussion on the effects of more sophisticated p-hacking procedures, see (Simonsohn, Simmons, and Nelson 2015)

[3] $r_p = \frac{t_x}{\sqrt{t_x^2+(n-p-1)}}$, where t_x is the *t* value of the regression coefficient of PSM, *n* is the number of observations, and *p* is the number of estimated parameters. If the standard error was given instead of the *t* value, we first transformed the coefficient into a *t* value by dividing it by the standard error ($t = \frac{b}{SE}$)

References

Aloe, Ariel M. 2014. "An Empirical Investigation of Partial Effect Sizes in Meta-Analysis of Correlational Data." *The Journal of General Psychology* 141 (1): 47–64. doi:10.1080/00221309.2013.853021.

American Psychological Association. 2020. *Publication Manual of the American Psychological Association: The Official Guide to APA Style.* 7th ed. Washington, D.C. American Psychological Association.

Anand, Vikas, Blake E. Ashforth, and Mahendra Joshi. 2004. "Business as Usual: The Acceptance and Perpetuation of Corruption in Organizations." *Academy of Management Perspectives* 18 (2): 39–53. doi:10.5465/ame.2004.13837437.

Andersen, Lotte B., Andreas Boesen, and Lene H. Pedersen. 2016. "Performance in Public Organizations: Clarifying the Conceptual Space." *Public Administration Review* 76 (6): 852–62. doi:10.1111/puar.12578.

Andersen, Lotte B., Eskil Heinesen, and Lene H. Pedersen. 2014. "How Does Public Service Motivation Among Teachers Affect Student Performance in Schools?" *Journal of Public Administration Research and Theory* 24 (3): 651–71. doi:10.1093/jopart/mut082.

Andrews, Christina. 2016. "Integrating Public Service Motivation and Self-Determination Theory." *International Journal of Public Sector Management* 29 (3): 238–54. doi:10.1108/IJPSM-10-2015-0176.

Antonakis, John. 2017. "On Doing Better Science: From Thrill of Discovery to Policy Implications." *Leadership Quarterly* 28 (1): 5–21. doi:10.1016/j.leaqua.2017.01.006.

Asseburg, Julia, Judith Hattke, David Hensel, Fabian Homberg, and Rick Vogel. 2020. "The Tacit Dimension of Public Sector Attraction in Multi-Incentive Settings." *Journal of Public Administration Research and Theory* 30 (1): 41–59. doi:10.1093/jopart/muz004.

Asseburg, Julia, and Fabian Homberg. 2020. "Public Service Motivation or Sector Rewards? Two Studies on the Determinants of Sector Attraction." *Review of Public Personnel Administration* 40 (1): 82–111. doi:10.1177/0734371X18778334.

Banks, George C., Sven Kepes, and Karen P. Banks. 2012. "Publication Bias." *Educational Evaluation and Policy Analysis* 34 (3): 259–77. doi:10.3102/0162373712446144.

Bellé, Nicola. 2013. "Experimental Evidence on the Relationship Between Public Service Motivation and Job Performance." *Public Administration Review* 73 (1): 143–53. doi:10.1111/j.1540-6210.2012.02621.x.

Bruns, Stephan B., and John P. A. Ioannidis. 2016. "P-Curve and P-Hacking in Observational Research." *PloS one* 11 (2): e0149144. doi:10.1371/journal.pone.0149144.

Chen, Zhixia, Robert Eisenberger, Kelly M. Johnson, Ivan L. Sucharski, and Justin Aselage. 2009. "Perceived Organizational Support and Extra-Role Performance: Which Leads to Which?" *Journal of Social Psychology* 149 (1): 119–24. doi:10.3200/SOCP.149.1.119-124.

Cheng, Kuo-Tai. 2015. "Public Service Motivation and Job Performance in Public Utilities." *International Journal of Public Sector Management* 28 (4/5): 352–70. doi:10.1108/IJPSM-08-2015-0152.

Christensen, Robert K., Laurie Paarlberg, and James L. Perry. 2017. "Public Service Motivation Research: Lessons for Practice." *Public Administration Review* 77 (4): 529–42. doi:10.1111/puar.12796.

Christensen, Robert K., Steven W. Whiting, Tobin Im, Eunju Rho, Justin M. Stritch, and Jungho Park. 2013. "Public Service Motivation, Task, and Non-Task Behavior: A Performance Appraisal Experiment with Korean MPA and MBA Students." *International Public Management Journal* 16 (1): 28–52. doi:10.1080/10967494.2013.796257.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Hillsdale, NJ: Erlbaum.

Deci, Edward L., and Richard M. Ryan, eds. 2004. *Handbook of Self-Determination Research.* Rochester, NY: University of Rochester Press.

Doucouliagos, Chris, and T. D. Stanley. 2013. "Are All Economic Facts Greatly Exaggerated? Theory Competition and Selectivity." *Journal of Economic Surveys* 27 (2): 316–39. doi:10.1111/j.1467-6419.2011.00706.x.

Earp, Brian D., and David Trafimow. 2015. "Replication, Falsification, and the Crisis of Confidence in Social Psychology." *Frontiers in Psychology* 6:621. doi:10.3389/fpsyg.2015.00621.

Ellis, Paul D. 2010. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results.* Cambridge: Cambridge Univ. Press.

Gould-Williams, Julian S., Ahmed Mohammed Sayed Mostafa, and Paul Bottomley. 2015. "Public Service Motivation and Employee Outcomes in the Egyptian Public Sector: Testing the Mediating Effect of Person-Organization Fit." *Journal of Public Administration Research and Theory* 25 (2): 597–622. doi:10.1093/jopart/mut053.

Grimmelikhuijsen, Stephan, Sebastian Jilke, Asmus L. Olsen, and Lars G. Tummers. 2017. "Behavioral Public Administration: Combining Insights from Public Administration and Psychology." *Public Administration Review* 77 (1): 45–56. doi:10.1111/puar.12609.

Hedges, Larry V., and Ingram Olkin. 1985. *Statistical Method for Meta-Analysis.* New York: Academic Press.

Homberg, Fabian, Rick Vogel, and Julia Weiherl. 2019. "Public Service Motivation and Continuous Organizational Change: Taking Charge Behaviour at Police Services." *Public Administration* 97 (1): 28–47. doi:10.1111/padm.12354.

Hubbard, Raymond T. 2015. *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science.* Los Angeles: Sage.

Hung, H. M. J., Robert T. O'Neill, Peter Bauer, and Karl Kohne. 1997. "The Behavior of the P-Value When the Alternative Hypothesis Is True." *Biometrics* 53 (1): 11. doi:10.2307/2533093.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS medicine* 2 (8): e124. doi:10.1371/journal.pmed.0020124.

———. 2008. "Why Most Discovered True Associations Are Inflated." *Epidemiology* 19 (5): 640–48. doi:10.1097/EDE.0b013e31818131e7.

Jilke, Sebastian, Kenneth J. Meier, and Gregg G. van Ryzin. 2018. "Editorial." *Journal of Behavioral Public Administration* 1 (1). doi:10.30636/jbpa.11.9.

Kepes, Sven, George C. Banks, and In-Sue Oh. 2014. "Avoiding Bias in Publication Bias Research: The Value of "Null" Findings." *Journal of Business and Psychology* 29 (2): 183–203. doi:10.1007/s10869-012-9279-0.

Korac, Sanja, Iris Saliterer, and Benedikt Weigand. 2019. "Factors Affecting the Preference for Public Sector Employment at the Pre-Entry Level: A Systematic Review." *International Public Management Journal* 22 (5): 797–840. doi:10.1080/10967494.2018.1430086.

Kroll, Alexander, and Dominik Vogel. 2018. “Changes in Prosocial Motivation over Time: A Cross-Sector Analysis of Effects on Volunteering and Work Behavior.” *International Journal of Public Administration* 41 (14): 1119–31. doi:10.1080/01900692.2017.1347945.

Liberati, Alessandro, Douglas G. Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C. Gøtzsche, John P. A. Ioannidis, Mike Clarke, P. J. Devereaux, Jos Kleijnen, and David Moher. 2009. “The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration.” *PLoS medicine* 6 (7): e1000100. doi:10.1371/journal.pmed.1000100.

Meier, Kenneth J., and Laurence J. O'Toole. 2013. “I Think (I Am Doing Well), Therefore I Am: Assessing the Validity of Administrators' Self-Assessments of Performance.” *International Public Management Journal* 16 (1): 1–27. doi:10.1080/10967494.2013.796253.

Newman, Joshua, Adrian Cherney, and Brian W. Head. 2016. “Do Policy Makers Use Academic Research? Reexamining the “Two Communities” Theory of Research Utilization.” *Public Administration Review* 76 (1): 24–32. doi:10.1111/puar.12464.

Oldham, Greg R., and Yitzhak Fried. 2016. “Job Design Research and Theory: Past, Present and Future.” *Organizational Behavior and Human Decision Processes* 136:20–35. doi:10.1016/j.obhdp.2016.05.002.

Open Science Collaboration. 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349 (6251): aac4716. doi:10.1126/science.aac4716.

Pashler, Harold, and Eric-Jan Wagenmakers. 2012. “Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?” *Perspectives on Psychological Science* 7 (6): 528–30. doi:10.1177/1745691612465253.

Pedersen, Mogens J. 2015. "Activating the Forces of Public Service Motivation: Evidence from a Low-Intensity Randomized Survey Experiment." *Public Administration Review* 75 (5): 734–46. doi:10.1111/puar.12325.

Perry, James L. 1996. "Measuring Public Service Motivation: An Assessment of Construct Reliability and Validity." *Journal of Public Administration Research and Theory* 6 (1): 5–22. doi:10.1093/oxfordjournals.jpart.a024303.

———. 2014. "The Motivational Bases of Public Service: Foundations for a Third Wave of Research." *Asia Pacific Journal of Public Administration* 36 (1): 34–47. doi:10.1080/23276665.2014.892272.

———. 2017. "Practicing What We Preach! Public Administration Review Promotes Transparency and Openness." *Public Administration Review* 77 (1): 5–6. doi:10.1111/puar.12705.

Perry, James L., and Lois R. Wise. 1990. "The Motivational Bases of Public Service." *Public Administration Review* 50 (3): 367–73. doi:10.2307/976618.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria. https://www.R-project.org/.

Ritz, Adrian, Gene A. Brewer, and Oliver Neumann. 2016. "Public Service Motivation: A Systematic Literature Review and Outlook." *Public Administration Review* 76 (3): 414–26. doi:10.1111/puar.12505.

Ritz, Adrian, Wouter Vandenabeele, and Dominik Vogel. forthcoming. "Public Service Motivation and Individual Job Performance." In *Managing for Public Service Performance: How People and Values Make a Difference*, edited by Peter Leisink, Lotte B. Andersen, Gene A. Brewer, Christian B. Jacobsen, Eva Knies, and Wouter Vandenabeele. Oxford: Oxford University Press.

Rost, Katja, and Thomas Ehrmann. 2017. “Reporting Biases in Empirical Management Research: The Example of Win-Win Corporate Social Responsibility.” *Business & Society* 56 (6): 840–88. doi:10.1177/0007650315572858.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” *Psychological Science* 22 (11): 1359–66. doi:10.1177/0956797611417632.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014a. “P-Curve: A Key to the File-Drawer.” *Journal of Experimental Psychology: General* 143 (2): 534–47. doi:10.1037/a0033242.

———. 2014b. “P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results.” *Perspectives on Psychological Science* 9 (6): 666–81. doi:10.1177/1745691614553988.

Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson. 2015. “Better P-Curves: Making P-Curve Analysis More Robust to Errors, Fraud, and Ambitious P-Hacking, a Reply to Ulrich and Miller (2015).” *Journal of Experimental Psychology: General* 144 (6): 1146–52. doi:10.1037/xge0000104.

———. 2016. “P-Curve Won't Do Your Laundry, but Will Identify Replicable Findings.” Accessed April 14, 2020. http://datacolada.org/49.

Starbuck, William H. 2016. “60th Anniversary Essay.” *Administrative Science Quarterly* 61 (2): 165–83. doi:10.1177/0001839216629644.

Sterling, Theodore D. 1959. “Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa.” *Journal of the American Statistical Association* 54 (285): 30–34. doi:10.1080/01621459.1959.10501497.

Ulrich, Rolf, and Jeff Miller. 2015. "P-Hacking by Post Hoc Selection with Multiple Opportunities: Detectability by Skewness Test? Comment on Simonsohn, Nelson, and Simmons (2014)." *Journal of Experimental Psychology: General* 144 (6): 1137–45. doi:10.1037/xge0000086.

Vandenabeele, Wouter. 2007. "Toward a Public Administration Theory of Public Service Motivation." *Public Management Review* 9 (4): 545–56. doi:10.1080/14719030701726697.

———. 2009. "The Mediating Effect of Job Satisfaction and Organizational Commitment on Self-Reported Performance." *International Review of Administrative Sciences* 75 (1): 11–34. doi:10.1177/0020852308099504.

Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses in R with the Metafor Package." *Journal of Statistical Software* 36 (3): 1–48. doi:10.18637/jss.v036.i03.

Vogel, Dominik, and Jurgen Willems. 2020. "The Effects of Making Public Service Employees Aware of Their Prosocial and Societal Impact: A Microintervention Study." *Journal of Public Administration Research and Theory* 30 (3): 485–503. doi:10.1093/jopart/muz044.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, and Han L. J. van der Maas. 2011. "Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011)." *Journal of Personality and Social Psychology* 100 (3): 426–32. doi:10.1037/a0022790.

Wallis, W. A. 1942. "Compounding Probabilities from Independent Significance Tests." *Econometrica* 10 (3/4): 229. doi:10.2307/1905466.

Wright, Bradley E., Shahidul Hassan, and Robert K. Christensen. 2017. "Job Choice and Performance: Revisiting Core Assumptions About Public Service Motivation."

International Public Management Journal 20 (1): 108–31. doi:10.1080/10967494.2015.1088493.

Zhu, Ling, Christopher Witko, and Kenneth J. Meier. 2018. "The Public Administration Manifesto II: Matching Methods to Theory and Substance." *Journal of Public Administration Research and Theory.* doi:10.1093/jopart/muy079.

Tables and Figures

Table 1: Results of the *p*-curve analysis with 35 significant studies on the PSM–performance relationship

	Binomial Test *(Share of results $p < .025$)*	**Continuous Test** *(Aggregate with the Stouffer Method)*	
		Full *p*-curve (*p*'s < .05)	**Half *p*-curve (*p*'s < .025)**
1) Studies contain evidential value. *(Right skew)*	$p < .0001$	$Z = -20.41$, $p < .0001$	$Z = -21.08$, $p < .0001$
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	$p = .9861$	$Z = 14.21$, $p > .9999$	$Z = 18.59$, $p > .9999$
Power of tests included in the *p*-curve *(correcting for selective reporting)*	*Estimate: 99%* *90% Confidence interval: (99%, 99%)*		

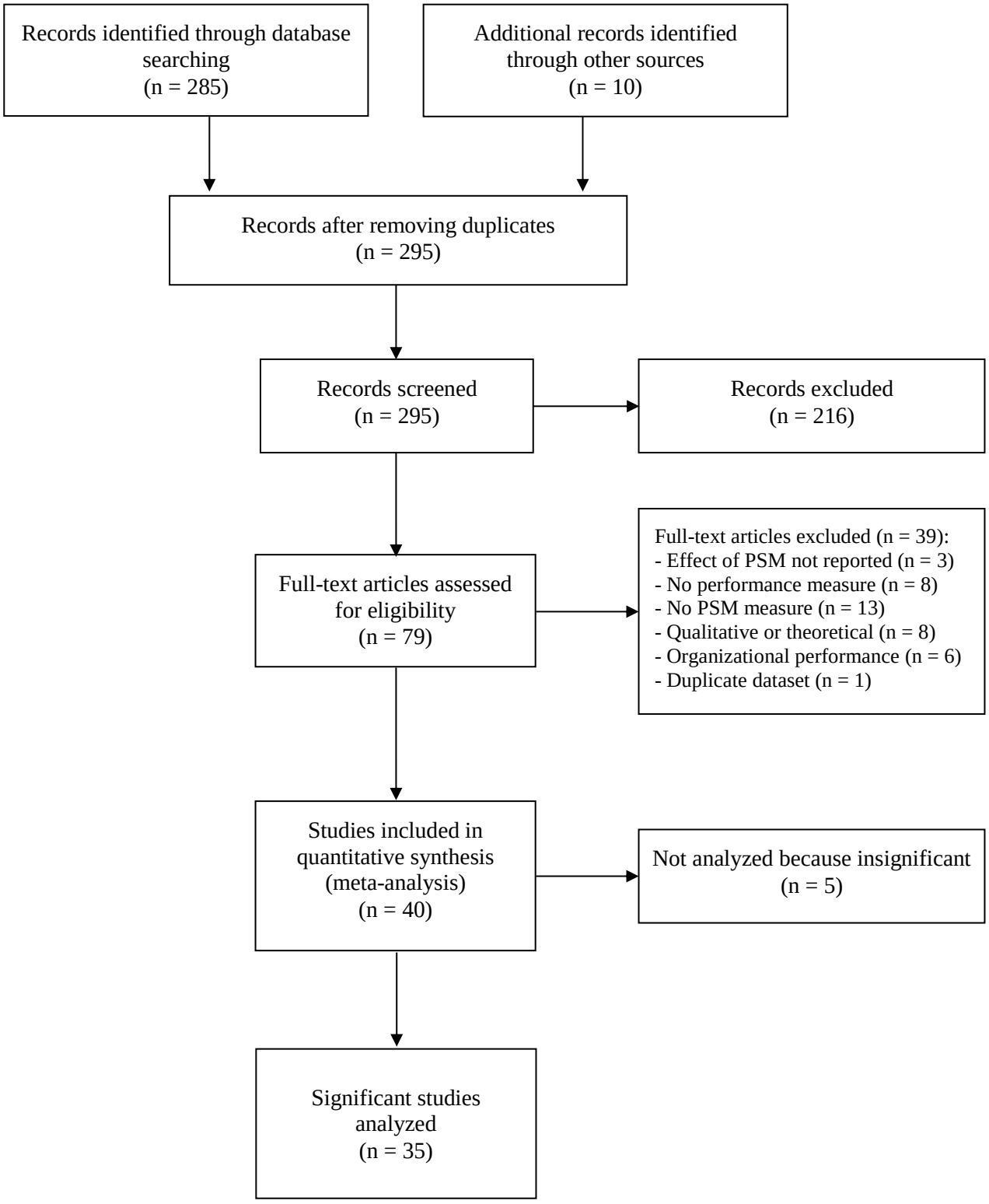

Figure 1: The search flow: Results of the search for articles based on Liberati et al. (2009)

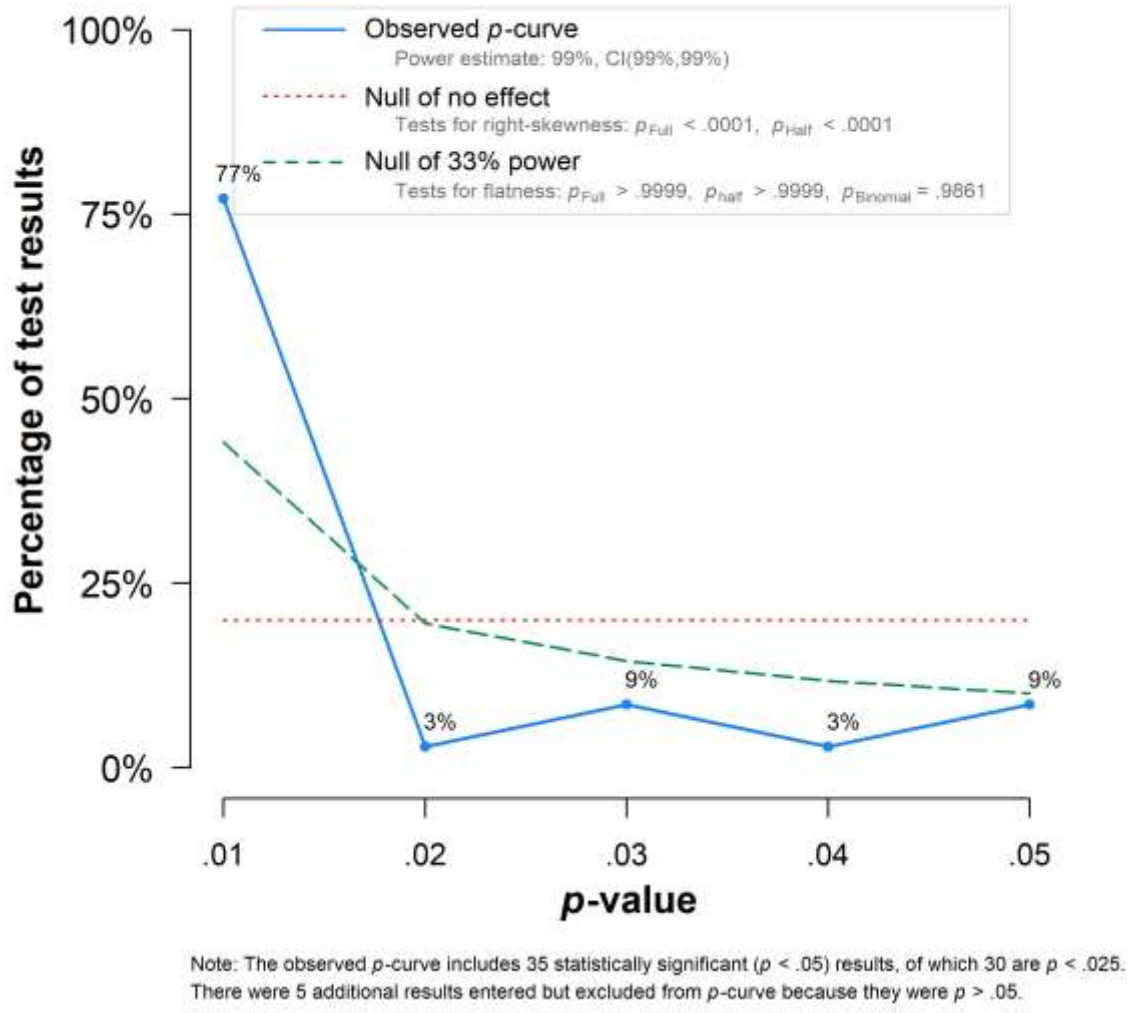

Figure 2: The *p*-curve for 35 statistically significant results on the effect of public service motivation on individual performance. Graph obtained from the *p*-curve app (http://www.p-curve.com).

Appendix A: Distribution of *p* values

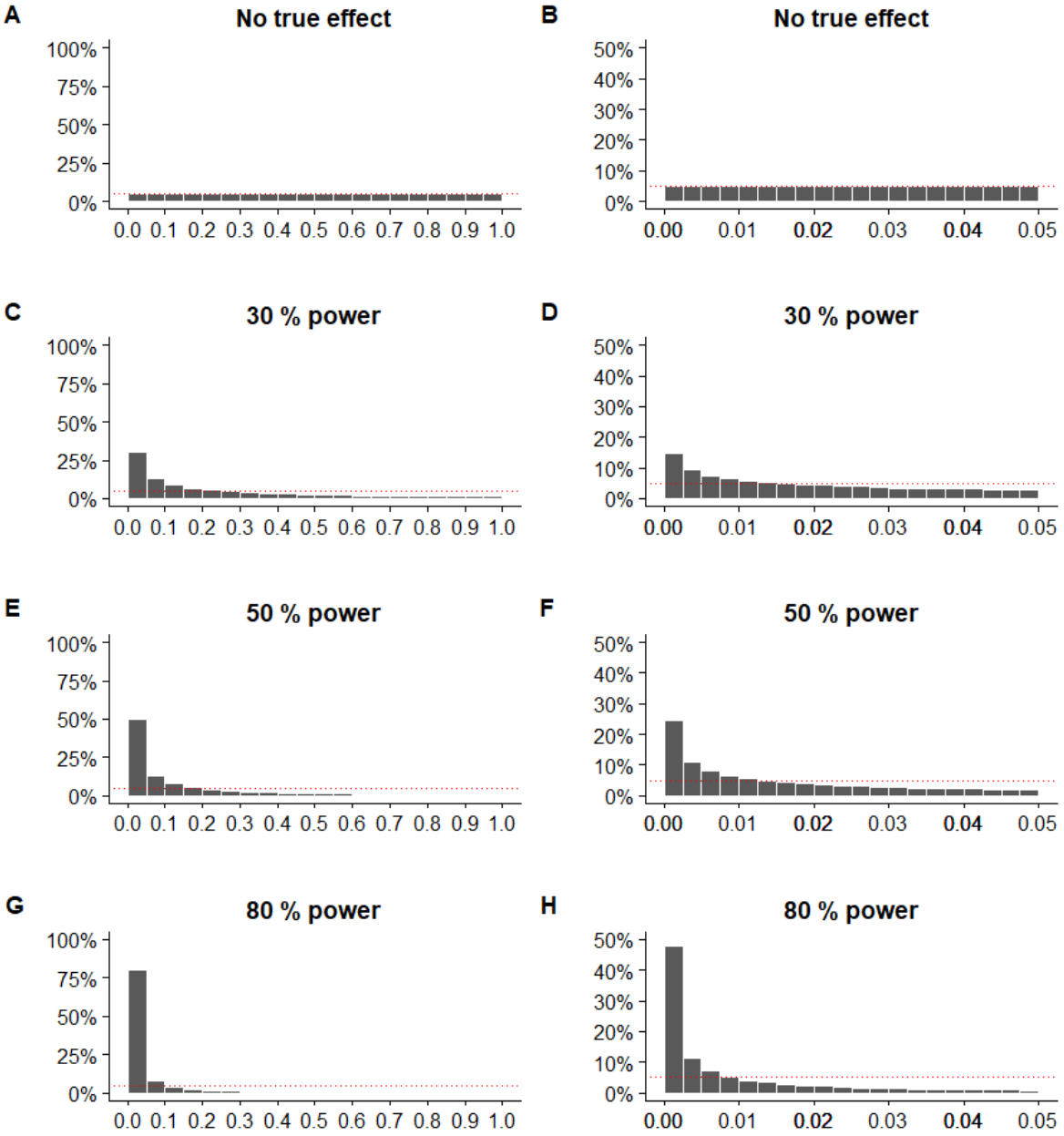

Figure A1: The expected distribution of *p* values for tests with different sample sizes. The first column (Graphs A, C, E, G) depicts the distribution of the full range of *p* values (0.00–1.00). The second column (Graphs B, D, F, H) displays only the significant *p* values ($p < .05$). Graphs A and B show a null effect, while Graphs C–H show a true effect of Cohen's $d = 0.4$. The dotted line represents the proportion of significant *p* values that could be expected if there is no true effect (i.e., type I error rate = 5%). For each graph, one million random samples from a normal distribution with a sample size of *n* and a true effect of Cohen's $d = 0.0$ (graphs A–B) or $d = 0.4$ (graphs C–H) were simulated, and a two-sample *t*-test was performed for each sample. Sample sizes are $n = 15$ (graphs A–D), $n = 26$ (graphs E–F), and $n = 51$ (graphs G–H) per group. Graphs and underlying simulations are created by the authors based on the explanations by Simonsohn, Nelson, and Simmons (2014a).

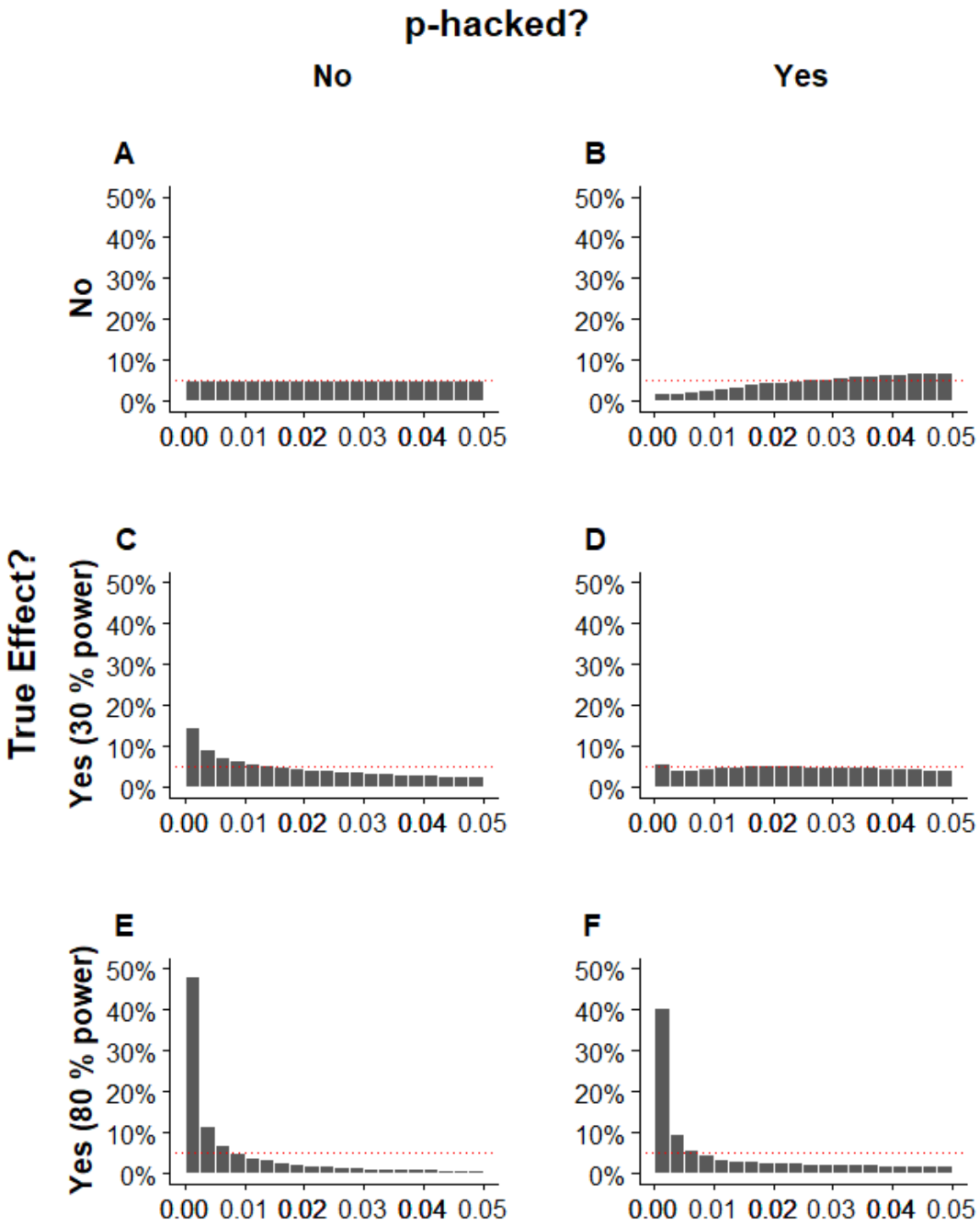

Figure A2: The distribution of significant *p* values with and without *p*-hacking. The graphs show only the significant *p* values ($p < .05$). The first column (Graphs A, C, E) depicts the distribution of significant *p* values if *p*-hacking is absent. The second column (Graphs B, D, F) displays the distribution of significant *p* values if *p*-hacking is present. The area below the dotted line represents the proportion of significant *p* values that could be expected if there is no true effect (i.e., type I error rate). For each graph, one million random samples from a normal distribution with a sample size of *n* and a true effect of Cohen's $d = 0.0$ (graphs A–B) or $d = 0.4$ (graphs C–F) were simulated, and a two-sided two-sample *t*-test was performed for each sample. Sample sizes are $n = 15$ (Graphs A–D) and $n = 51$ (Graphs E–F) per group. For the *p*-hacked graphs (B, D, F) if the difference was not significant, five additional, independent observations were added to each sample, up to a maximum of 25 additional observations. Graphs and underlying simulations are created by the authors based on the explanations by Simonsohn, Nelson, and Simmons (2014a).

Appendix B: Included Studies

Table A1: Included studies with extracted effect sizes and exact *p* values

Study	Type of perf. measure	Effect	Exact *p* value
Alonso and Lewis 2001	Self-assessed	r(7270) = 0.2051	<.00001
Andersen and Serritzlew 2012[c]	Objective	r(506) = 0.138	.00182
Andersen, Heinesen, and Pedersen, L. 2014	Objective	r(453) = 0.154	.00098
Andersen, Heinesen, and Pedersen, L. 2016	Objective	r(704) = 0.184	.00991
Bellé 2013[a]	Objective	r(87) = 0.272	.00992
Bottomley et al. 2016[b]	Self-assessed	r(830) = 0.49	<.00001
Bright 2007[c]	Self-assessed	r(187) = 0.035	.63256
Caillier 2010	Self-assessed	r(305) = 0.035	.54124
Caillier 2014	Self-assessed	r(767) = 0.081	.02469
Caillier 2016	Self-assessed	r(898) = 0.474	<.00001
Camilleri and van der Heijden 2007[a]	Self-assessed	r(1186) = 0.057	.04951
Campbell and Im 2016	Self-assessed	r(444) = 0.337	<.00001
Cheng 2015	Self-assessed	r(2221) = 0.347	<.00001
Cun 2012[a]	Self-assessed	r(492) = 0.148	.00097
Gould-Williams, Mostafa, and Bottomley 2015[b]	Self-assessed	r(306) = 0.52	<.00001
Jin, McDonald, and Park, J. 2018	Self-assessed	r(678) = 0.251	<.00001
Kim 2006[b]	Self-assessed	r(1583) = 0.525	<.00001
Koumenta 2015	Self-assessed	r(489) = 0.156	.00052
Leisink and Steijn 2009	Self-assessed	r(3727) = 0.033	.04390
Levitats and Vigoda-Gadot 2017	Self-assessed	r(188) = 0.146	.04443
Lynggaard, Pedersen, M., and Andersen 2018	Objective	r(5640) = 0.022	.09847
Mostafa, Gould-Williams, and Bottomley 2015[b]	Self-assessed	r(613) = 0.62	<.00001
Mostafa and Leon-Cazares 2016[a]	Self-assessed	r(833) = 0.107	.00196
Naff and Crum 1999	Self-assessed	r(8070) = 0.110	<.00001
Palma and Sepe 2017	Self-assessed	r(586) = 0.091	.02735

Palma, Hinna, and Mangia 2017	Self-assessed	r(264) = 0.182	.00289
Pandey, Wright, and Moynihan 2008[c]	Self-assessed	r(156) = 0.228	.00396
Park, S. and Rainey 2008	Self-assessed	r(6772) = 0.029	.01699
Pedersen, M. 2015	Hypothetical	r(523) = 0.162	.00019
Potipiroon and Faerman 2016	Supervisor-assessed	r(185) = 0.166	.02317
Resh, Marvel, and Wen 2018	Objective	r(575) = 0.110	.00818
Ritz et al. 2014[b]	Self-assessed	r(538) = 0.39	<.00001
Schwarz et al. 2016	Supervisor-assessed	r(243) = 0.136	.03336
Shim and Faerman 2017	Self-assessed	r(385) = 0.306	<.00001
van Loon 2017a[b]	Self-assessed	r(329) = 0.180	.00100
van Loon 2017b	Supervisor-assessed	r(41) = 0.207	.18290
van Loon, Vandenabeele, and Leisink 2017[b]	Self-assessed	r(1030) = 0.213	<.00001
Vandenabeele 2009	Self-assessed	r(3491) = 0.099	<.00001
Wright, Hassan, and Christensen 2017	Supervisor-assessed	r(413) = 0.049	.31935
Xiaohua 2008[a]	Self-assessed	r(315) = 0.183	.00106

Note: [a] Significance threshold was used to calculate an effect size, since the exact *p* value could not been extracted from the article and authors could not be contacted, did not respond, or were unable to provide the necessary information; [b] Correlation between PSM and performance was used to calculate an effect size since the exact *p* value could not been extracted from the article and authors could not be contacted, did not respond, or were unable to provide the necessary information; [c] Information necessary to calculate effects size provided by authors. Full references to the included studies are displayed in the supplementary material.