

# Exploring New Avenues for the Meta-Analysis Method in Personality and Social Psychology Research

---

Dissertation

zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

der Fakultät HW

Bereich Empirische Humanwissenschaften

der Universität des Saarlandes

---

vorgelegt von

**Julius Frankenbach**

aus Wiesbaden

Saarbrücken, 2023

Dekan: Univ.-Prof. Dr. Peter Loos, Universität des Saarlandes

Berichterstatter: Univ.-Prof. Dr. Malte Friese, Universität des Saarlandes

Univ.-Prof. Dr. Cornelius J. König, Universität des Saarlandes

PD Dr. Dorota Reis, Universität des Saarlandes

Tag der Disputation: 01.03.2023

## Summary

This dissertation addresses theoretical validity and bias in meta-analytic research in personality and social psychology research. The conceptual starting point of the dissertation is research on ego depletion (Baumeister et al., 1998). In this line of research, hundreds of studies documented an experimental effect that probably does not exist, as was later revealed by extensive replication work (Hagger et al., 2010, 2016). This debacle has presumably been caused by dysfunctional structures and procedures in psychological science, such as widespread publication bias (Carter & McCullough, 2014). Unfortunately, these dysfunctions were (and in some cases still are) also prevalent in other areas of psychological research beside ego depletion (Ferguson & Brannick, 2012; Open Science Collaboration, 2015). Because extensive replication research is too costly to be repeated for all past work, it has been a contentious question what to do with research data that has been generated during an era of questionable research practices: should this research be abandoned or can some of it be salvaged? In four research papers, this dissertation project attempts to address these questions. In part I of the dissertation project, two papers highlight and analyze challenges when summarizing past research in social psychology and personality research. Paper 1 (Friese et al., 2017) attempted to find summary evidence for the effectiveness of self-control training, a research field related to ego depletion, but came to a sobering conclusion: The summary effect was small, likely inflated by publication bias, and could not be attributed beyond doubt to a theoretical mechanism. Paper 2 (Friese & Frankenbach, 2020) reported on a simulation study that showed how multiple sources of bias (publication bias, *p*-hacking) can interact with contextual factors and each other to create significant meta-analytic evidence from very small or even zero true effects. Part II of the dissertation project is an attempt to advance social-psychological and personality theory with

meta-scientific work despite an unknowable risk of bias in the literature. In part II, two papers (Frankenbach et al., 2020, 2022) make use of one key idea: Re-using existing raw research data to test novel theoretical ideas in secondary (meta-)analyses. Results revealed that this idea helps towards both goals of the dissertation project, that is, advancing theory while reducing risk-of-bias. The general discussion analyses promises and limitations of such secondary data analyses in more detail and attempts to situate the idea more broadly in the psychological research toolkit by contrasting integrative versus innovative research. Further discussion covers how conceptual and technological innovations may facilitate more secondary data analyses in the future, and how such advances may pave the way for a slower, more incremental, but truly valid and cumulative psychological science.

## Zusammenfassung

Die vorliegende Dissertation behandelt theoretischen Validität und Verzerrung (Bias) von meta-analytischer Forschung in der Persönlichkeits- und Sozialpsychologie. Der konzeptuelle Ausgangspunkt der Dissertation ist die Forschung zu „Ego Depletion“ (Baumeister et al., 1998). In dieser Forschungslinie haben hunderte von Studien einen Effekt belegt, der, wie sich später durch umfangreiche Replikationsarbeiten (Hagger et al., 2010, 2016) herausstellte, vermutlich nicht existiert. Dieses Debakel wurde mutmaßlich mitverursacht durch dysfunktionale Strukturen und Prozesse in der psychologischen Forschung, insbesondere Publikationsbias („publication bias“). Unglücklicherweise lagen (und liegen) diese Dysfunktionalitäten neben Ego Depletion auch in anderen psychologischen Forschungsbereichen vor (Ferguson & Brannick, 2012; Open Science Collaboration, 2015). Da aus Kostengründen nicht alle Forschungsarbeiten der Vergangenheit repliziert werden können, ergibt sich eine kritische Frage: Wie soll mit psychologischer Forschung umgegangen werden, die unter mutmaßlich verzerrenden Bedingungen generiert wurde? Sollte diese Forschung ad acta gelegt werden oder können Teile davon weiterverwendet werden? Das vorliegende Dissertationsprojekt versucht im Rahmen von vier Forschungsbeiträgen sich diesen Fragen anzunähern. Im ersten Teil der Dissertation beleuchten und analysieren zwei Forschungsbeiträge Probleme und Herausforderungen, die sich bei der Zusammenfassung von bestehender Forschung der Sozial- und Persönlichkeitspsychologie ergeben. Der erste Beitrag (Friese et al., 2017) versucht in einer Meta-Analyse Evidenz für die Wirksamkeit von Selbstkontrolltrainings zu finden, aber kommt zu einem ernüchternden Ergebnis: Die Gesamteffekte sind klein, mutmaßlich durch Publikationsbias fälschlich überhöht und können überdies nicht zweifelsfrei einem theoretischen Kausalmechanismus zugeordnet werden. Der zweite Beitrag (Friese & Frankenbach, 2020)

umfasst eine Simulationsstudie, die aufzeigt, wie verschiedene Formen von Bias (Publikationsbias und sog. „*p*-hacking“) miteinander und mit Kontextfaktoren interagieren können, wodurch signifikante, meta-analytische Effekte aus sehr kleinen wahren Effekten oder sogar Nulleffekten entstehen können. Der zweite Teil der Dissertation versucht, trotz eines unbestimmbaren Bias-Risikos, Fortschritte in der sozial- und persönlichkeitspsychologischen Theorie zu erzielen. Zu diesem Zweck wird in zwei Forschungsbeiträgen (Frankenbach et al., 2020, 2022) auf eine Schlüssel-Idee zurückgegriffen: Die Testung von neuen theoretischen Hypothesen unter Wiederverwendung von existierenden Forschungsdaten in Sekundärdatenanalysen. Die Ergebnisse zeigen, dass dieser Ansatz tatsächlich dazu beitragen kann, theoretische Fortschritte mit vermindertem Verzerrungsrisiko zu machen. Die anschließende, übergreifende Diskussion behandelt Möglichkeiten und Limitationen solcher Sekundärdatenanalysen und versucht, den Ansatz in einer Gegenüberstellung von integrativer und innovativer Forschung übergreifender in die psychologische Forschungsmethodik einzuordnen. Im Weiteren wird diskutiert, wie konzeptuelle und technologische Entwicklungen in der Zukunft Sekundärdatenanalysen erleichtern könnten und wie solche Fortschritte den Weg ebnen könnten für eine langsamere, inkrementelle, aber wahrhaft valide und kumulative psychologische Wissenschaft.

## **Acknowledgements**

This work would not have been possible without the unwavering support of my supervisor Malte Friese. Thank you for many hours of inspiring discussion, for offering insight and direction, and for an open door. I also thank my colleagues Veronika Job, Jacob Juhl, Helena Kilger, David Loschelder, Constantine Sedikides, Marcel Weber, and Tim Wildschut for the fruitful collaboration. I am grateful to Cornelius König for agreeing to review my thesis. Special thanks to all research assistants that supported the projects, and to the many authors of primary studies who volunteered much time and effort to provide data for the meta-analyses.

## Index of Publications

This publication-oriented dissertation (German: publikations-orientierte Dissertation) is based on four manuscripts, three of them published and one in press. The author of this dissertation is the sole first author of two of these manuscripts and shared first author for the other two. Two articles are published with *SAGE journals*, who permit reproducing articles in theses as the published, typeset version. The other two are or will be published with the *American Psychological Association* (APA). APA does not permit reproducing articles in theses. For these two articles, links to publicly available preprint are included in the dissertation. The manuscript that is currently in press has been accepted for publication at *Psychological Bulletin*. This article is currently publicly available as a preprint and also referenced as such. For paper 1 of part II, a corrigendum has been published and can be retrieved under <https://doi.org/10.1177/08902070211026136>.

Part I, Paper 1: Friese, M., Frankenbach, J., Job, V., & Loschelder, D. D. (2017). Does self-control training improve self-control? A meta-analysis. *Perspectives on Psychological Science*, 12(6), 1077–1099. <https://doi.org/10.1177/1745691617697076>

Part I, Paper 2: Friese, M., & Frankenbach, J. (2020). P-Hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods*, 25(4), 456–471. <https://doi.org/10.1037/met0000246>

Part II, Paper 1: Frankenbach, J., Wildschut, T., Juhl, J., & Sedikides, C. (2021). Does neuroticism disrupt the psychological benefits of nostalgia? A meta-analytic test. *European Journal of Personality*, 35(2), 249–266. <https://doi.org/10.1002/per.2276>



Part II, Paper 2: Frankenbach, J., Weber, M., Loschelder, D. D., Kilger, H., & Friese, M. (2022). *Sex drive: Theoretical conceptualization and meta-analytic review of gender differences* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/9yk8e>

## Content

Summary .....	I
Zusammenfassung.....	III
Acknowledgements.....	V
Index of Publications .....	VI
General Introduction .....	1
The Present Research.....	5
Introduction to Part I: Challenges and Limitations of the Meta-Analysis Method.....	6
Introduction to Part II: Exploring Solutions to Conceptual Heterogeneity and Risk-of-Bias in Meta-Analyses.....	10
Part I, Paper 1: “Does self-control training improve self-control? A meta-analysis.”.....	14
Part I, Paper 2: “P-Hacking and publication bias interact to distort meta-analytic effect size estimates.” .....	15
Part II, Paper 1: “Does neuroticism disrupt the psychological benefits of nostalgia? A meta- analytic test.”.....	16
Part II, Paper 2: “Sex drive: Theoretical conceptualization and meta-analytic review of gender differences.” .....	17
General Discussion .....	18
Key Idea: Secondary Data Analyses to Test Novel Research Questions.....	18
The Role of Theory in Secondary Data Analyses .....	19
How Secondary Data Analyses Can Reduce Risk of Bias.....	20
Integration versus Innovation in Psychological Research.....	22

To Reduce Risk-of-Bias, What About Open Science? .....	24
How Technological Advancements May Enable More Secondary Data Analyses .....	25
Conclusion.....	30
References.....	31

## General Introduction

In 2015, I handed in my thesis manuscript, entitled “Versatile means of overcoming ego depletion: A narrative review”, in pursuit of a bachelor’s degree in psychology. As the title says, the work was concerned with ego depletion, a simple, social-psychological theory of self-control (Baumeister et al., 1998). The theory posits that self-control relies on a limited, domain-general, psychological resource that gets depleted when self-control is exerted. This assumption was at the time supported by a myriad of empirical experiments in which participants completed two consecutive tasks that required self-control (sometimes referred to as the *sequential task paradigm*). The ubiquitous finding was that performance was poorer on the second task, supporting the assumption that some form of finite self-control resource or energy was consumed in the first task. These observations became the foundation of a grander theory of self-control in social psychology that was over time expanded to accommodate a wide range of connected research questions (Baumeister et al., 2007, 2018). One obvious follow-up question was whether the decline in self-control can be prevented or compensated in some way. This was the question I tried to answer in my bachelor thesis (Frankenbach, 2015). I conducted a systematic, narrative review and collected more than a hundred studies that showed consistently how ego depletion was prevented or attenuated by diverse experimental treatments or personality traits. I spent most of early 2015 collecting and reading these studies, thinking hard about how to classify them theoretically under the umbrella of the resource theory of self-control, as well as one alternative motivational theory that was popular at the time (Inzlicht & Schmeichel, 2012). It was hard work, but also very rewarding and satisfying – a key experience that kindled my passion for doing science. Yet, as it turned out, the endeavor was entirely pointless. Around the same time, a meta-analysis showed that ego-depletion studies with less participants tended to have larger

effect sizes (Carter et al., 2015), presumably because they needed those larger effect sizes to pass the threshold of statistical significance, which is often an implicit requirement for studies to be published and thus enter the scientific record. This observation implied that the publication process filtered certain ego-depletion studies (those with larger effect sizes) and sliced off others (those with smaller effect sizes). Carter and colleagues argued that the patterns in the association between effect size magnitude and sample size indicated that there may in truth be no ego depletion effect. This and other developments regarding the validity of psychological science at the time led to systematic efforts to replicate the ego depletion effects without the biasing influence of the publication process (Dang et al., 2021; Hagger et al., 2016; Vohs et al., 2021). The results of these projects now suggest that the ego-depletion effect is miniscule or zero. Yet, somehow, hundreds of studies presumably showing this false effect entered the scientific record. Even more, as I found out when researching for my bachelor thesis, there were a great many studies that showed how this false effect can be attenuated, modulated, prevented, or counteracted. How could this be? How did it happen that thousands of words, thousands of hours spent thinking, and thousands of dollars were spilled on something that turned out to be only noise? Who were the people who wrote these pointless reports, and how do they view them now? What were the conditions that facilitated the emergence of a vast scientific literature, “rich” with theoretical ideas and a seemingly solid quantitative foundation, yet built solely on noise? And perhaps most importantly: Were these conditions exceptional, somehow unique to the ego depletion literature, or are they also present in other areas of psychology, perhaps even omnipresent? For me, these questions arose from the experience of doing passionate work that turned out to be nonsense, and they continued to accompany me during my subsequent scientific work.

This first experience of doing science had several tangible impacts on my future research. First, I became interested in meta-analyses (Borenstein et al., 2009). When I was first introduced to the idea of meta-analysis in an undergraduate lecture on industrial and organizational psychology, I considered them an intricate way of combining data from multiple studies that yielded definitive, final answers. Yet in the ego depletion literature, there was one meta-analysis that showed unanimous support (Hagger et al., 2010), while another questioned the existence of the effect (Carter et al., 2015). Surely, it was worth learning about the technique to understand how it can lead to vastly different conclusions. Also, when synthesizing more than one hundred studies on moderators of the ego depletion effect, I soon understood the limitations of narrative reviews and the value of a more objective approach, especially for understanding the impact of contextual factors. Second, I became highly aware of how easily quantitative research can be biased. I felt that it was important to understand these biasing processes in order to avoid them in my own work and when consuming the works of others. Third, I started wondering about the role of theory in meta-scientific work. One may argue that theories are irrevocably tied to primary studies (e.g., the sequential task paradigm was designed to test the ego depletion theory of self-control), and thus meta-analyses can only confirm (or reject) the theoretical claims of the primary studies. Then again, the role of the meta-analyst could also be perceived more broadly, such that meta-analysts can examine the evidence in relation to competing theories (as I did in my bachelor thesis when examining moderation of the ego-depletion effect from the lenses of a resource- versus motivation-oriented theory of self-control). One could go even further and grant the meta-analyst leeway to develop (and test) novel theoretical ideas. A counterargument to these “liberal” conceptions is that theoretical coherence is a prerequisite for synthesizing effect sizes that indicate relations of psychological variables. According to this view, psychological variables

are conceived within a fixed theoretical framework and lose their meaning if they are detached from their theory of origin. In any case, I realized that meta-analysts need to be clear about the role of theory in their work, or else theory, data, and methods will be scrambled into an indistinguishable mush.

### **The Present Research**

With these lessons in mind, I embarked on a journey that was to become my dissertation project. The project now consists of four papers, all published or accepted in peer-reviewed psychological journals. The papers are summarized in Table 1. It is worth noting that the dissertation project did not follow a predetermined project plan with topics and questions defined at the outset. Rather, I, together with my colleagues, followed my interests at the time and let ideas flow from one project to the next, going further down the rabbit hole. As a consequence, the three studies that examine substantive research questions are quite diverse, spanning three different subfields, but I will try my best to carve out the red line, which is meta-science and its validity. Although the papers were not initially conceived that way, I here present them as separated into two parts, which seems sensible from the retrospective. Each part consists of two papers, and the papers are ordered according to the timeline in which they were written and published. The work of part I raises more questions than it answers. It exemplifies challenges and limitations of doing meta-scientific work in social psychology and personality research and only begins to analyze them. Part II constitutes an attempt at developing some remedies. In the next section, I will briefly introduce the four papers and lay out how they tie in with the guiding question of this dissertation project: How can meta-scientific work advance social-psychological and personality theory despite an unknowable risk of bias in the literature?



Table 1

Overview of papers in the dissertation project

No.	Part	No. in part	Title	Role of the doctoral candidate	Published/accepted in
1	I	1	Does self-control training improve self-control? A meta-analysis	Shared first author	Perspectives on Psychological Science
2	I	2	<i>p</i> -hacking and publication bias interact to distort meta-analytic effect size estimates	Shared first author	Psychological Methods
3	II	1	Does neuroticism disrupt the psychological benefits of nostalgia? A meta-analytic test	First author	European Journal of Personality
4	II	2	Sex drive: Theoretical conceptualization and meta-analytic review of gender differences	First author	Psychological Bulletin

### Introduction to Part I: Challenges and Limitations of the Meta-Analysis Method

The first paper of part I, entitled “Does self-control training improve self-control? A meta-analysis”, reports a meta-analysis of experimental research on self-control training (Friese et al., 2017). The foundational paper of this line of research (Muraven et al., 1999) introduced the idea that people become better at exerting self-control, that is, overcoming dominant impulses, if they practice doing so. We conducted a systematic literature search that identified 33 studies in which participants trained self-control. Our meta-analysis suggested that self-control training increased self-controlled behavior by a small-to-medium effect size overall.

The paper exemplifies two lines of challenges that meta-analyses face. The first challenge is that meta-analysts must assert that there is sufficient conceptual coherence among the body of primary research to justify integration (the “apples and oranges” problem, AAO). My colleagues

and I adopted the common definition that self-control is “the ability to override or change one’s inner responses, as well as to interrupt undesired behavioral tendencies (such as impulses) and refrain from acting on them” (Tangney et al., 2004). As it turned out, the number of potential ways to train and measure self-control according to this definition seemed almost limitless. In other words, the study revealed considerable conceptual heterogeneity in the self-control training literature, with no two studies directly comparable in terms of outcome and treatment operationalization. On the treatment side, some studies asked participants to regulate their posture, do everyday task with their non-dominant hand, or train self-control in computerized inhibitory control tasks, among others. On the outcome side, measures spanned the domains of health behavior like smoking and alcohol consumption, aggression and emotion regulation, educational achievement and study behavior, computerized inhibitory control performance, and many more. Surely, they all reflected self-control to some extent according to the definition. However, the immense conceptual heterogeneity left my colleagues and I with the impression that self-control is a mere lowest common denominator of these studies, rather than a grand, unifying conceptual framework. Many outcome-treatment combinations allowed for alternative theoretical explanations. In the end, my colleagues and I concluded that there was insufficient evidence to definitively attribute the observed effects to the repeated overcoming of dominant responses (i.e., the training of self-control).

The second challenge is that meta-analysts must assess the risk of bias in the primary research, since biased primary research will lead to biased meta-analyses (the “garbage in, garbage out” problem, GIGO). The discussion in the paper focuses on publication bias as one specific form of bias where studies with certain characteristics have a higher probability of being selected for publication, while other studies remain hidden in the file-drawer. My colleagues and

I applied the same methods to detect publication bias that Carter and colleagues used to find bias in the original ego depletion literature (Carter et al., 2015; Stanley & Doucouliagos, 2014).

Unfortunately, there were unmistakable signs of publication bias in the self-control training literature. As in the study by Carter and colleagues, studies with smaller sample sizes had larger effect sizes. Additionally, we found that the unpublished studies which we included in the analysis had smaller effects than the published studies. In sum, we were unable to state with confidence that the observed effect was not pure bias.

Taken together, these limitations reduced the informative value of the meta-analysis considerably: We observed small effects, but they could be pure bias, and we were unable to attribute them to a theoretically grounded causal mechanism. This was quite a sobering conclusion, prompting me to take a step back and think some more about these biasing processes and the role of theory in meta-scientific work. The next paper of the dissertation project, paper II of part I, addresses the problem of bias in more detail.

The source of inspiration for this next paper, entitled “*p*-hacking and publication bias interact to distort meta-analytic effect size estimates” (Friese & Frankenbach, 2020), was our own puzzlement about bias in meta-analysis. At the time of this writing, scholars seem to have become more accustomed to the proposition that entire fields of research can be biased, but it is worth considering again how consequential this idea is. It means that something can arise from nothing, that researchers collectively “mine noise” and weave random patterns into coherent stories. This puzzled us immensely and sparked our interest to study this bias more systematically. Our method of choice was a computer simulation that explored how different forms of bias can add up and interact to distort the conclusions of meta-analyses. The simulation addresses the question whether, and under what conditions, different forms of bias can create

something from nothing. Put differently, we explore whether a meta-analysis showing significant effects can arise even if true effects are zero or very small. To this end, we employed a parameter-based simulation study that systematically varied two sources of bias, as well as various contextual factors, and explored how these factors work together to distort meta-analyses. The first source of bias we simulated was publication bias, a concept which has been introduced already. The second source of bias was “*p*-hacking”, a phenomenon where researchers tamper with their statistics to achieve significant results (Simmons et al., 2011). The study revealed interesting interaction patterns among the factors and demonstrated that something can indeed arise from nothing (or very little). Meta-analyses can be severely distorted by conditions that are likely present in many fields of research in social psychology, such as effect size heterogeneity, small or null effects, or the exploitation of researcher degrees-of-freedom (i.e., *p*-hacking).

Taken together, the two papers of part I exemplified and explored two of the key challenges of meta-analyses, namely conceptual heterogeneity and risk-of-bias, and they gave me a clearer idea of what it took to do valid meta-science. With this in mind, part II of the dissertation project aims to explore potential remedies, that is, answer substantial research questions in social psychology in a bias-free and theoretically coherent way. This second part also consists of two papers. Both papers report a separate meta-analysis each and aim to test innovations to the meta-analysis method that could potentially alleviate the problems of conceptual heterogeneity and risk-of-bias.

## **Introduction to Part II: Exploring Solutions to Conceptual Heterogeneity and Risk-of-Bias in Meta-Analyses**

The first paper of part II, entitled “Does neuroticism disrupt the psychological benefits of nostalgia? A meta-analytic test” (Frankenbach et al., 2020), is a meta-analysis of the effects of experimentally induced nostalgia (a “sentimental longing for the past”, Sedikides et al., 2015) on various psychological variables, specifically, the interaction of these inductions with trait neuroticism. The primary studies synthesized in this work are experiments in which participants enter a state of nostalgia, for example, by listening to nostalgic music or by writing about fond memories. Typically, being nostalgic has various positive effects on participants’ psychological state (e.g., enhanced self-esteem or more feelings of social connectedness). The main question of the meta-analysis was whether these positive effects are less pronounced for people high in trait neuroticism. This hypothesis was grounded in the observation that nostalgic memories also elicit some negative feelings (they are “bittersweet”), and that trait neuroticism tends to entail more sensitivity to negativity. In order to alleviate risk of bias, this study employed several strategies. First, we collaborated closely with primary authors in the field to identify unpublished data, resulting in 17 unpublished studies included in the analysis (out of 19 in total). Second, we obtained raw data for all studies, which allowed for more in-depth tests for bias, such as measurement unreliability or restricted variance. Third, the analysis focused on the nostalgia-neuroticism interaction, which was not focal in the original studies. This reduced the risk significantly that the discoverability of studies depended on the effect size. Results showed good psychometric properties of the included measurements. The main effects of nostalgia were significant (i.e., nostalgia had positive effects on self-oriented, existential, and social variables). However, the main hypothesis that trait neuroticism moderated these benefits was not supported.

Taken together, quality-assurance measures detailed above reduced the risk of bias dramatically, lending more confidence to the meta-analytic conclusions. This was perhaps of extra importance given that the main hypothesis was not supported, and researchers often must go to greater lengths when arguing for the informative value of null findings. The second key problem of the dissertation project, conceptual heterogeneity, was less pressing here, because the conceptual scope of the analysis was relatively limited (e.g., compared to the analysis of domain-general transfer effects of self-control training in paper I of part I).

Conceptual heterogeneity was, however, very much focal in paper II of part I, entitled “Sex drive: Theoretical conceptualization and meta-analytic review of gender differences” (Frankenbach et al., 2022). This paper reports a meta-analysis of gender differences in sex drive, specifically, average differences between men and women. In the literature, a plethora of definitions and conceptualizations of sex drive exist, which poses a considerable challenge for a meta-analysis. To address this, the study extends the analytic approach of the study reported in paper I of part II. The paper first develops a coherent, formalized conceptualization of sex drive. This framework was then employed to define a large set of questionnaire items that are valid indicators of sex drive according to the conceptualization. Data for this item set was then identified through a literature search and correspondence with primary authors. Wherever possible, we again obtained raw data to allow for more detailed tests for bias. For example, this approach enabled a meta-analytic investigation of convergent and discriminant validity. As with paper I of part II of the dissertation project, the analysis focused on associations that were not focal to the primary authors, significantly reducing the risk that there was publication bias with regard to the gender difference. Thus, the approach of selecting individual items from original studies based on a coherent theoretical rationale addressed both key problems of the dissertation

project. The analysis included more than 600,000 participants from 211 studies. Results showed that men have a consistently stronger self-reported sex drive. Detailed analyses of risk-of-bias from several sources (publication bias, response bias, lack of validity) lent confidence to this conclusion.

In summary, the two papers of part II of the dissertation project utilized methodological innovations that led to considerably more confidence in the results compared to paper I of part I. One key finding of the dissertation project is that the meta-analysis method can be utilized to address novel research questions using existing research data, while retaining theoretical coherence and reducing risk of bias. Obtaining new research data is expensive. Recent findings in methodology research on requirements for trustworthy psychological research have highlighted the need for replication, larger sample sizes, and effortful quality assurance procedures like pre-registration or registered reports (Nosek et al., 2018; Nosek & Lakens, 2014; Open Science Collaboration, 2015). These measures will further increase the costs and resource requirements for collecting new research data. In light of these developments, along with technological innovations in data sharing and management, secondary (re-)analyses of existing data are a promising avenue for efficient, impactful, and trustworthy psychological science. Naturally, this approach is not without limitations. For one, it is clear that not all research questions can be addressed using existing data. Innovative theoretical ideas often require innovative methods. Yet, whether a lack of new conceptual ideas, or a lack of theoretical coherence and an unbiased empirical basis is currently the most pressing concern in academic psychology is subject to debate. These questions will be examined in more detail in General Discussion section.

In the next section, the four papers of the dissertation project are reprinted as they were accepted for publication. The authoritative documents of record are the typeset versions as published in the respective journals. In paper I of part II, an error has been corrected in a table that was discovered after publication (and also corrected in a corrigendum). Note that papers I and II of part I and paper I of part II remain formatted in APA style version 6 (American Psychological Association, 2010) that was in effect at the time of publication. All manuscripts that are part of this dissertation have been prepared according to the principles of open science and reproducibility, including data sharing, open materials, and preregistration.



**Part I, Paper 1: “Does self-control training improve self-control? A meta-analysis.”**

# Does Self-Control Training Improve Self-Control? A Meta-Analysis

Malte Friese<sup>1</sup>, Julius Frankenbach<sup>1</sup>, Veronika Job<sup>2</sup>, and David D. Loschelder<sup>3</sup>

<sup>1</sup>Saarland University, <sup>2</sup>University of Zurich, and <sup>3</sup>Leuphana University of Lueneburg

Perspectives on Psychological Science  
 2017, Vol. 12(6) 1077–1099  
 © The Author(s) 2017  
 Reprints and permissions:  
[sagepub.com/journalsPermissions.nav](http://sagepub.com/journalsPermissions.nav)  
 DOI: 10.1177/1745691617697076  
[www.psychologicalscience.org/PPS](http://www.psychologicalscience.org/PPS)



## Abstract

Self-control is positively associated with a host of beneficial outcomes. Therefore, psychological interventions that reliably improve self-control are of great societal value. A prominent idea suggests that training self-control by repeatedly overriding dominant responses should lead to broad improvements in self-control over time. Here, we conducted a random-effects meta-analysis based on robust variance estimation of the published and unpublished literature on self-control training effects. Results based on 33 studies and 158 effect sizes revealed a small-to-medium effect of  $g = 0.30$ , confidence interval (CI<sub>95</sub>) [0.17, 0.42]. Moderator analyses found that training effects tended to be larger for (a) self-control stamina rather than strength, (b) studies with inactive compared to active control groups, (c) males than females, and (d) when proponents of the strength model of self-control were (co)authors of a study. Bias-correction techniques suggested the presence of small-study effects and/or publication bias and arrived at smaller effect size estimates (range:  $g_{\text{corrected}} = .13$  to  $.24$ ). The mechanisms underlying the effect are poorly understood. There is not enough evidence to conclude that the repeated control of dominant responses is the critical element driving training effects.

## Keywords

self-control training, intervention, meta-analysis, publication bias, robust variance estimation

Successful self-control is associated with a host of positive outcomes in life, including academic success, stable personal relationships, financial security, and good psychological and physical health. By contrast, poor self-control is associated with more aggression, substance use, and crime, among others (Duckworth & Seligman, 2005; Gottfredson & Hirschi, 1990; Tangney, Baumeister, & Boone, 2004). It is readily conceivable that how well people fare in these domains has not only important personal consequences but also consequences for society at large. Research shows that self-control assessed very early in life predicts a variety of important life outcomes (Daly, Delaney, Egan, & Baumeister, 2015; Moffitt et al., 2011). These findings seem to suggest that self-control is a stable trait being shaped early in life. However, other research perspectives highlight the possibility of self-control change by targeted interventions (e.g., Piquero, Jennings, Farrington, Diamond, & Gonzalez, 2016). Over the past 15 years, researchers have designed controlled psychological interventions that tested the effect of self-control training on self-control success

across diverse domains (Berkman, 2016). Given the importance of self-control in various life domains, there is a tremendous demand for such interventions that promise to reliably, appreciably, and enduringly improve self-control. The present article provides a meta-analysis of this self-control training literature.

## What Self-Control Is and Why It Should (Not) Be Possible to Improve It

One prominent conceptualization defines self-control as the “ability to override or change one’s inner responses, as well as to interrupt undesired behavioral tendencies (such as impulses) and refrain from acting on them” (Tangney et al., 2004, p. 274). In line with this definition,

### Corresponding Author:

Malte Friese, Department of Psychology, Saarland University, Campus A2 4, 66123 Saarbrücken, Germany  
 E-mail: [malte.friese@uni-saarland.de](mailto:malte.friese@uni-saarland.de)

the exertion of self-control is typically seen as deliberate, conscious, and effortful.

The main theoretical rationale for why training self-control should be beneficial comes from the strength model of self-control (Baumeister & Vohs, 2016b; Baumeister, Vohs, & Tice, 2007). This influential model proposes that all self-control efforts draw on a general capacity. This capacity is used and depleted regardless of in which domain a person exerts self-control (e.g., attention control, control of food intake, control of emotional expression). Because of its generality, improvements in the general self-control capacity should benefit all kinds of self-control behavior across various domains.

The strength model posits that the capacity to exert self-control works akin to a muscle. This assertion has two important implications: First, exerting self-control will lead to temporary exhaustion and make subsequent self-control failure more likely (ego depletion).<sup>1</sup> Second, repeated practice will strengthen the self-control muscle (training hypothesis). This will result in either a general increase in absolute muscle strength (i.e., improved self-control *strength*) and/or increased resistance to fatigue when confronted with demands (i.e., improved self-control *stamina*). Both increases in strength and stamina should benefit self-control in a broad range of domains in the laboratory and in everyday life.

From the perspective of the strength model, the crucial aspect of a training regimen lies in the repeated overriding of dominant responses. In typical self-control training studies that are examined in the present meta-analysis, participants are asked to complete everyday activities with the nondominant hand such as brushing teeth or using the computer mouse (Miles et al., 2016), to refrain from using highly prevalent slang words (Finkel, DeWall, Slotter, Oaten, & Foshee, 2009), or to work on computerized tasks requiring the control of dominant responses (Cranwell et al., 2014). After the training (typically 2 weeks long), laboratory or everyday-life indicators of self-control strength or stamina are compared to a control group. Training effects have been investigated on outcome variables such as success in quitting smoking (Muraven, 2010b), laboratory aggression (Denson, Capper, Oaten, Friese, & Schofield, 2011), or physical persistence (Cranwell et al., 2014).

The hypothesis that training self-control leads to broad improvements in self-control across domains is both intriguing and risky: It is *intriguing* because the trainability of self-control has implications for many subfields of psychology and is of high practical importance. Among other benefits, it would open the possibility of helping people deal with self-control problems in one domain by practicing self-control in a completely different domain. For instance, consider an obese person having gone through countless unsuccessful diets, still wishing to lose weight.

At this point, any new intervention directly concerned with restraining eating behavior may be difficult, because dieting is closely associated with frustration and feelings of personal failure. The self-control training hypothesis is intriguing in that it suggests this person could succeed at dieting by practicing self-control in unrelated and emotionally uncharged activities.

The self-control training hypothesis is a *risky* hypothesis because other literatures on training psychological capabilities are not very encouraging concerning appreciable and broad benefits in people's lives. Consider the literature on cognitive training of executive functions such as working memory capacity or task-shifting (Miyake & Friedman, 2012). This literature shows that the transfer of improvements in the specific training tasks to other tasks measuring the same construct (i.e., from one working memory task to the other) is sometimes found (near transfer). By contrast, transfer rarely emerges to related constructs (i.e., from working memory to task-shifting) or behaviors that should benefit from improving the focal construct (far transfer; Melby-Lervåg & Hulme, 2013; Melby-Lervåg, Redick, & Hulme, 2016; Owen et al., 2010; Shipstead, Redick, & Engle, 2012). The empirical studies that have been conducted to date to test the self-control training hypothesis have exclusively focused on far transfer—training took place in one domain (e.g., controlling speech and/or posture) and dependent variables were collected in different domains (e.g., persistence, aggression).

Within the self-control literature, related but distinct conceptualizations of self-control stress the importance of learning essential self-control skills early in life (Heckman, 2006; Mischel et al., 2011; Moffitt et al., 2011). For example, preschoolers can learn to conceive desired objects as less tempting by focusing on their nonconsummatory features (Mischel & Baker, 1975). Recent meta-analytic evidence suggests that teaching such self-control skills is effective in children and adolescents to improve self-control ( $g = 0.32$ ) and to reduce delinquency ( $g = 0.27$ ; Piquero et al., 2016). The self-control training interventions reviewed in the present meta-analysis focus on repeatedly overriding dominant responses without teaching strategies on how to do so. This approach might be less effective to appreciably and enduringly improve self-control.

## Previous Meta-Analyses

Two peer-reviewed meta-analyses have previously summarized evidence relating to the self-control training hypothesis. The first meta-analysis included a total of nine published studies and revealed a large average effect of  $d^+ = 1.07$  (Hagger, Wood, Stiff, & Chatzisarantis, 2010). Among these nine studies were three studies with exceptionally large effects sizes up to  $d^+ > 8$  (sic!) and

unclear methodology (Oaten & Cheng, 2006a, 2006b, 2007), leading to a very wide 95% CI for the estimated average effect size [0.10, 2.03]. A more recent meta-analysis excluded these 3 studies and included a total of 10 published studies (Inzlicht & Berkman, 2015). Inzlicht and Berkman used the recently introduced *p*-curve method (Simonsohn, Nelson, & Simmons, 2014) to compute two estimates of the meta-analytic self-control training effect size—one based on the first dependent variable reported for a given study, the other based on the last dependent variable reported. All other effects were discarded. The first estimate was  $d = 0.17$ ,  $CI_{95} [-0.07, 0.41]$ , a small effect not significantly different from zero. The second estimate was  $d = 0.62$ ,  $CI_{95} [0.13, 1.11]$ , a stronger but also more volatile effect size.<sup>2</sup>

## The Present Meta-Analysis

The present meta-analysis aims to deliver a comprehensive summary of the published and unpublished evidence and to considerably extend previous work. In particular, we pursued three goals: First, we aimed at estimating the average self-control training effect based on the most comprehensive data base possible. With 33 studies (23 published, 10 unpublished), we included more than three times as many studies than the Hagger et al. (2010) and the Inzlicht and Berkman (2015) meta-analyses. In addition, we based our estimates on *all* reported dependent variables, an issue of importance given that many of the original studies reported several dependent variables. In such cases, basing effect size estimates solely on the first and/or last reported effect (Inzlicht & Berkman, 2015) inevitably implies a loss of valuable information.

Second, we sought to conduct moderator analyses to elucidate boundary conditions of the self-control training effect. Moderator analyses can be crucially informative for both theory building and for applied purposes when designing self-control training procedures.

Finally, we sought to investigate the existence of small-study effects and publication bias. Publication bias accrues when studies with a statistically significant result are more likely to be published than studies with a null result. Because publishing almost exclusively significant results is how the field worked for many years (Bakker, van Dijk, & Wicherts, 2012; Fanelli, 2012), meta-analyses tend to overestimate population effect sizes (Ioannidis, 2008; Levine, Asada, & Carpenter, 2009).

## Methods

The present review followed reporting guidelines for meta-analyses outlined in the PRISMA statement (Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009).

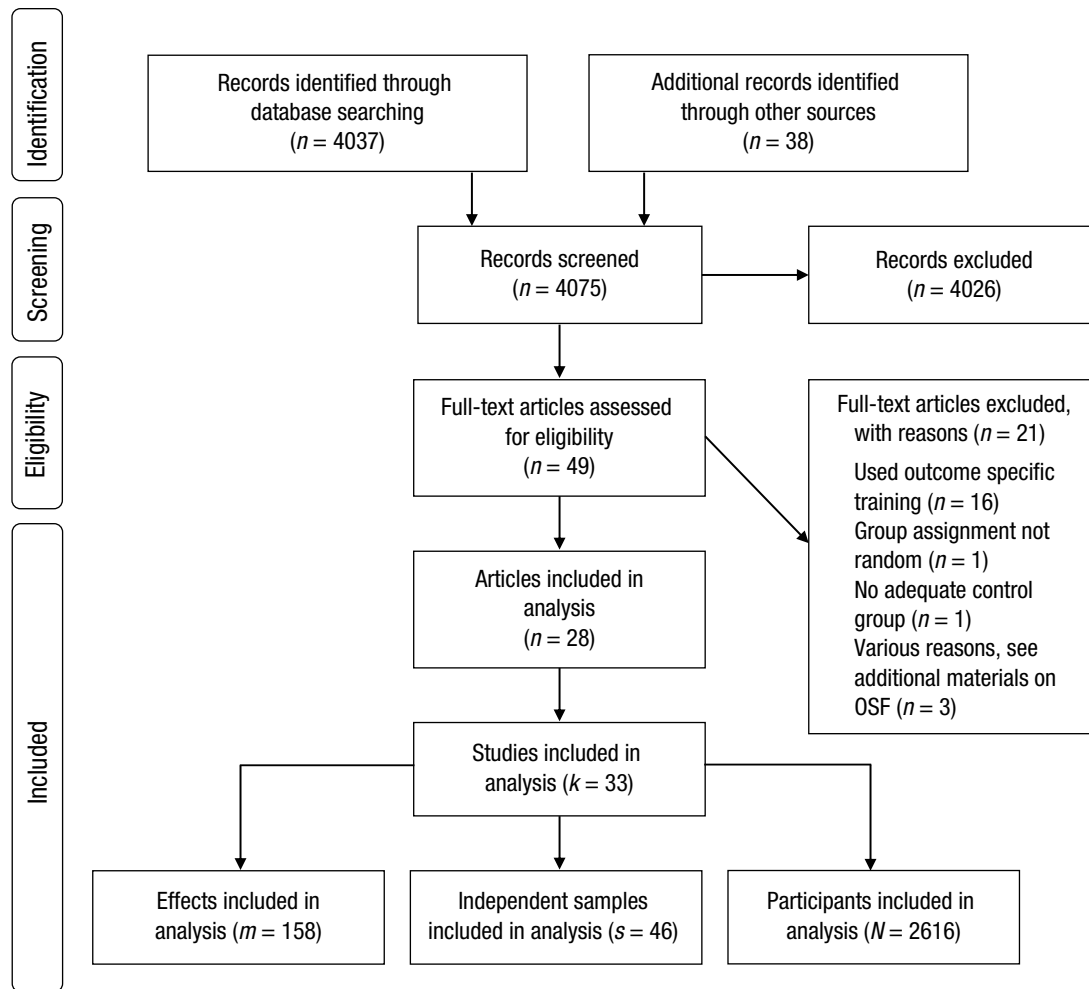
The study was preregistered under the international prospective register of systematic reviews (PROSPERO; registration number CRD42016033917, <http://www.crd.york.ac.uk/prospéro/>). Following recent recommendations for the reproducibility of meta-analyses (Lakens, Hilgard, & Staaks, 2016) and to facilitate future updates of this work, we made all data, code, full documentation of our procedures, and additional supplementary analyses available on the Open Science Framework (<https://osf.io/v7gxf/>).

## Inclusion criteria

Studies were eligible for inclusion if they (1) implemented at least one training procedure that contained the repeated control of dominant responses, (2) included at least one control group, (3) allocated participants randomly to conditions, (4) measured at least one self-control-related outcome variable in a different domain than the domain in which the training occurred, (5) assessed the outcome variable(s) at least 1 day after the last training session,<sup>3</sup> and (6) included samples of mentally healthy adults. We decided to only include studies with random allocation to conditions because only random allocation allows for a causal interpretation of training effects. For studies that contained conditions and/or outcomes irrelevant to our research question, we only included the conditions and/or outcome variables that matched all criteria. In case of ambiguity about the relevance of the chosen outcome variable(s), we generally followed the arguments of the original study authors. For a detailed documentation of all decisions that were made, see the documentation available on the Open Science Framework (<https://osf.io/v7gxf/>).

## Search strategy

We conducted a systematic literature search using three online citation-database providers—namely, EBSCO, ProQuest, and ISI Web of Science. In EBSCO, we searched the databases PsycINFO, ERIC, PsycARTICLES, and PSYINDEX, using the exact search term (TI self regulat\* OR TI self control OR TI inhibit\* OR TI willpower) AND (TI training OR TI practic\* OR TI exercis\* OR TI improv\*). For ISI Web of Science, the exact search term was TITLE: ([self regulat\* OR self control OR inhibition OR willpower] AND [training OR practic\* OR exercis\* OR improv\*]). This search was restricted to entries tagged as “psychology.” In ProQuest, we searched for ([“self regulat\*” OR “self control” OR “inhibition” OR “willpower”] AND [“training” OR “practice” OR “exercise”] AND “psychology”). All databases were searched from 1999 onward, the publication year of the first self-control training study (Muraven, Baumeister, & Tice, 1999). Additionally, we issued calls for unpublished data through the mailing lists of three scientific societies (SPSP, EASP, SASP) and personally corresponded with researchers that



**Fig. 1.** PRISMA flow chart of the literature search and study coding.

are active in the field. Finally, the literature search was complemented by unsystematic searches and reference harvesting from included studies and relevant overview articles.

### Screening

Titles and abstracts of 4,075 records were screened by the second author for relevance to the present work. Of these, 4,026 were excluded. Forty-nine full-text articles were assessed for eligibility according to the inclusion criteria. Twenty-eight were included in the final database. The PRISMA flow chart in Figure 1 provides details about these steps.

### Study coding

We coded several potential moderator variables of self-control training effects. One potential moderator pertains to the type of training that was implemented, some

pertain to the study level, and some pertain to level of the outcome used. For further potential moderators and respective analyses, please see the Supplemental Material available online. The second author and a research assistant coded all potential moderators explained in the remainder of this section (see documentation on the OSF for details). Interrater reliability was examined using intraclass correlation (ICC) for continuous moderators— $ICC(1,1)$  (Shrout & Fleiss, 1979) and Cohen's Kappa for categorical moderators (Cohen, 1968). Interrater reliability for the study coding was high by common standards (Cicchetti, 1994), mean  $\kappa = 0.83$ , mean  $ICC(1,1) = 0.92$ .

### Treatment-level moderator

*Type of training.* Some training procedures may be more effective than others. For example, training procedures that require more deliberate and effortful behavioral control (e.g., repeatedly squeezing a handgrip over several weeks) may differ in effectiveness from training procedures that require more frequent but less rigorous

behavioral control (e.g., using one's nondominant hand for everyday activities).

### **Study-level moderators**

*Length of training.* Longer training procedures may lead to stronger training effects. Length of treatment was coded in days. Length of training was coded as a study-level (instead of a treatment-level) moderator because in all studies with more than one treatment condition treatment length was equal across conditions.

*Publication status.* Studies with statistically significant results are more likely to be published, possibly leading to an overestimation of the average effect size. Published and in press studies were coded as published and all others as unpublished. (For a more comprehensive treatment of potential publication bias, see below.)

*Research group.* The self-control training hypothesis was derived from the strength model of self-control (Baumeister et al., 2007). Perhaps researchers from this group are more experienced and more skilled at operationalizing relevant variables than other researchers. Alternatively, they may also be more biased in favor of the self-control training hypothesis. Given the criticisms to the strength model, it is also possible that researchers from other research groups are biased against the hypothesis. Following Hagger et al. (2010), a study was coded "Strength model research group" if one of the authors or committee members of a dissertation or master's thesis was Roy Baumeister or one of his known collaborators (alphabetically: DeWall, Gailliot, Muraven, Schmeichel, Vohs). All other studies were coded "other."

*Control group quality.* Intervention effects that are based on comparisons of training conditions with inactive control groups can result from multiple different working mechanisms (e.g., demand effects, stronger engagement in the study in the intervention group, etc.). Active control groups narrow down the range of plausible working mechanisms and provide a more conservative test of the self-control training hypothesis. Control groups were coded as active when they worked on any task while the intervention group received treatment; all other control groups were coded as inactive.

*Gender ratio.* Meta-analytic evidence suggests that trait self-control is more strongly linked to the inhibition of undesired behaviors in males than in females (de Ridder, Lensvelt-Mulders, Finkenauer, Stok, & Baumeister, 2012). Thus, to the extent that self-control training improves trait self-control, training may show stronger effects in males than in females. We coded the gender ratio as the percentage of males in the sample.

### **Outcome-level moderators**

*Type of outcome.* Training effects on some outcome variables may be stronger than on others. We grouped outcome variables into clusters representing different content domains (e.g., physical persistence, health behaviors, academic behaviors).

*Lab versus real-world behavior.* For some outcomes, the relevant behavior is performed in the laboratory (e.g., computerized performance tasks). For others, the relevant behavior refers to real-world behavior performed outside the laboratory (e.g., "How often have you done X during the last week?") and may also be assessed outside the laboratory (e.g., daily diaries). Behavior assessed in the laboratory may provide more experimental control, and variables that reflect real-world behavior or experience may have higher external validity. Outcomes were coded as "lab behavior" or "real-world behavior."

*Stamina versus strength.* Some outcomes were assessed without a preceding effortful task, others after an effortful task. Outcomes were coded as "self-control stamina" (i.e., resistance to ego depletion) when they were preceded by an effortful task and as "self-control strength" when they were not preceded by an effortful task.

*Maximum versus realized potential.* Some dependent variables require the participant to perform as well as possible (i.e., realize their full self-control potential; e.g., Stroop task or keep hand in ice water for as long as possible). When not prompted, people may not always access their maximum potential but realize only a part of it in a given situation. Self-control training may differentially affect the maximum potential people *can* exert and the realized potential they *do* willingly exert.

*Follow-up.* Training effects may deteriorate with increasing time between the end of training and outcome measurement. Follow-up was coded as the number of days between the last day of training and the outcome measurement. If the outcome measurement spanned across a period of time, the middle of this time period was used to calculate follow-up.

### **Effect size coding**

We computed Hedges'  $g$  effect sizes and respective variances ( $\text{Var}_g$ ) for all effects (Hedges, 1981). Hedges'  $g$  is similar to Cohen's  $d$  but corrects for small sample bias. Two design types were prevalent: pretest-posttest-control designs (PPC) and posttest-only-control designs (POC). For continuous dependent variables, we first computed Cohen's  $d$  and its variance  $\text{Var}_d$  and then applied Hedges' correction factor for small sample bias to compute  $g$  and

Var<sub>g</sub>. For PPC designs, Cohen's  $d$  was defined as the difference of mean improvement between the training group and the control group, divided by the pooled pretest standard deviation ( $SD$ ):

$$d_{PPC} = \frac{(M_{Treat[POST]} - M_{Treat[PRE]}) - (M_{Ctrl[POST]} - M_{Ctrl[PRE]})}{\sqrt{\left( (n_{Treat} - 1) \times SD_{Treat[PRE]}^2 + (n_{Ctrl} - 1) \times SD_{Ctrl[PRE]}^2 \right)}} \times \frac{1}{n_{Treat} + n_{Ctrl} - 2} \quad (1)$$

Thus, the numerator in the Cohen's  $d$  fraction was a difference of differences—that is, the difference of the mean improvement ( $M_{post} - M_{pre}$ ) between the two conditions. Standardizing by pooled pretest  $SD$  rather than pooled posttest  $SD$  or pooled total  $SD$  has been shown to yield a more precise estimate of the true effect, as interventions typically cause greater variation at posttest (Morris, 2008).

For POC designs, Cohen's  $d$  was defined as the difference in means divided by the pooled posttest standard deviation.

$$d_{POC} = \frac{M_{Treat[POST]} - M_{Ctrl[POST]}}{\sqrt{\left( (n_{Treat} - 1) \times SD_{Treat[POST]}^2 + (n_{Ctrl} - 1) \times SD_{Ctrl[POST]}^2 \right)}} \times \frac{1}{n_{Treat} + n_{Ctrl} - 2} \quad (2)$$

For noncontinuous variables, appropriate effect sizes for the respective scale level were computed and then transformed to Hedges'  $g$  (Hedges, 1981). When possible, effect sizes were computed from descriptive statistics and sample sizes. We contacted the authors if required information was missing in the manuscript. Eighteen out of 23 responded to our inquiry. If authors did not respond or could not provide the required information, we approximated the effect size as closely as possible using the information provided in the original manuscript.

Some studies included more than one treatment group or control group (e.g., using self-control training tasks and/or control tasks from different domains). When multiple treatment and/or control groups were implemented, we compared each treatment group separately against each control group. For studies that included multiple outcomes, we computed one effect size per outcome for each comparison. For example, a study reporting two treatment groups, two control groups, and three outcomes would contribute a total of 12 effect sizes (2 treatments  $\times$  2 controls  $\times$  3 outcomes). Some studies reported multiple measurements of the same outcomes after training. In

these cases, we only included the measurement temporally most proximate to the training phase (exception: follow-up moderator analysis; see next paragraph).

For the moderator analysis "follow-up," we contrasted outcome variables measured directly after the training (posttraining, see above) with later measurement occasions (follow-up). If a study included both posttraining and follow-up measurements, we included effect sizes for both time points. When multiple training and/or control groups were implemented, we combined them, respectively, before computing the effect sizes, as type of training/control group was not of interest in this particular analysis.

### Meta-analytic procedure

We deviated from the path of data analysis outlined in the preregistration because we followed valuable reviewer suggestions made in the editorial process (i.e., reliance on the robust variance estimation, RVE, approach; see below). All analyses were conducted using random effects models because self-control training interventions, control groups, and outcome variables varied considerably between studies. Hence, it was unreasonable to expect one true, "fixed" population effect.

Conventional meta-analytical techniques assume that effect sizes are statistically independent. Including multiple effect sizes stemming from multiple outcomes or comparisons per study violates this assumption (Lipsey & Wilson, 2001). Several approaches have been proposed to address this issue and to arrive at a set of independent effect sizes (for an overview, see Marín-Martínez & Sánchez-Meca, 1999). One widely used approach averages and adjusts effect sizes based on the correlation of the combined effect sizes (Borenstein, Hedges, Higgins, & Rothstein, 2009). More specifically, the effect size variance estimate is more strongly reduced if the combined outcomes are weakly correlated compared to when they are highly correlated. This reflects the idea that uncorrelated outcomes contain broader informational value than highly correlated outcomes. One downside of this approach is that averaging effect sizes leads to a loss of information because analyses on the level of effect sizes are no longer possible. To illustrate, consider a study reporting treatment effects on reading and mathematics achievement. Averaging these effect sizes delivers one study summary effect. The single summary effect prohibits a moderator analysis investigating effects of the treatment on different outcomes such as reading versus mathematic achievement across several studies in the meta-analysis.

The recently developed RVE approach for meta-analysis (Hedges, Tipton, & Johnson, 2010) solves this issue. It permits conducting random effects meta-regression on dependent effect sizes, thus offering many advantages over the previously described averaging approach. Unfortunately,

there are some drawbacks to RVE as well. First, RVE estimates the correlation matrix of dependent effect sizes rather than accounting for it directly. It will therefore generally yield less precise results than approaches that incorporate the empirical correlation structure (e.g., the procedure proposed by Borenstein et al., 2009). Second, because the approach is relatively novel, the validity of some key meta-analytical techniques has not yet been validated in the RVE context, such as regression-based tests for small-study effects, Trim and Fill procedures, or power analyses. Third, although it is possible to calculate point estimates of true variance in the effect sizes in RVE (i.e.,  $I^2$ ), there are currently no significance tests of these estimates available. Hence, researchers must rely on conventions when interpreting the true variance of effect sizes (Higgins & Green, 2011).

Considering the respective (dis)advantages of the Borenstein approach and the RVE approach, we adopted the following threefold strategy for the present meta-analysis: First, we computed the global summary effect of self-control training based on RVE and provide the parallel estimate based on the Borenstein approach for converging evidence. Second, all moderator analyses were run based on RVE. Third, all tests to detect and correct for small study effects were run based on the Borenstein approach, as the validity of these procedures has not yet been investigated in the RVE context. We also ran these analyses within the RVE approach for converging evidence. These latter analyses should be interpreted with caution, however. Please refer to the Supplemental Material available online for details of the Borenstein approach. We relied on the *MAd* package to implement the approach (Del Re & Hoyt, 2014).

**RVE.** All RVE models were fitted using the *robumeta* package for *R* (Fisher, Tipton, & Hou, 2016). We ran the RVE analyses with the following specifications: First, standard RVE has been shown to perform satisfactorily with a minimum of 10 studies when estimating summary effects and with a minimum of 20–40 studies when estimating slopes in meta-regression (Hedges et al., 2010; Tipton, 2013). When the number of studies falls below these limits, significance tests tend to have inflated Type I error rates. We therefore implemented significance tests that incorporate small sample corrections for all RVE models (Tipton, 2015; Tipton & Pustejovsky, 2015). Specifically, we conducted Approximate Hotelling-Zhang tests for testing multiple parameters (Tipton & Pustejovsky, 2015, abbreviated HTZ in the *clubSandwich* package that we employed to run these analyses, Pustejovsky, 2016) and *t* tests for single parameters (Tipton, 2015). Both HTZ and *t* values had small-sample-corrected degrees of freedom and adjusted variance-covariance

matrices. It is important to note that the single-parameter *t* test (but not the multiple parameter HTZ test) may provide inaccurate results when degrees of freedom fall below  $df = 4$ . Consequently, we caution the reader to interpret the results when this was the case, and we refrained from reporting *p* values and confidence intervals in the figures depicting analyses with  $df < 4$ .

Second, meta-analysts using RVE need to decide how to weight the effect sizes. Following recent recommendations (Tanner-Smith & Tipton, 2014), we set the weights to account for the type of dependence that is likely to be most prevalent in the dataset (i.e., dependence due to correlated rather than hierarchical effects). Third, we estimated the average correlation of effect sizes by first averaging all Fisher *z*-transformed outcome correlations per study, averaging these means across all studies, and then transforming the value back to a Pearson correlation. This procedure returned a mean outcome correlation of  $r = .18$ . We additionally conducted sensitivity analyses for all models by varying the correlation estimate from  $r = 0$  to  $r = 1$  in steps of  $r = .2$ . This did not appreciably influence the conclusions drawn from the models. For example, the overall mean estimate of self-control training effectiveness only changed by  $\Delta g = 0.0002$  when going from  $r = 0$  to  $r = 1$ .

In order to compute the overall summary effect, we fitted an intercept-only random-effects RVE model to the set of dependent effect sizes. The regression coefficient of this model can be interpreted as the precision-weighted mean effect size of all studies, corrected for effect-size dependence. The corresponding significance test probes whether the estimate is significantly different from zero. To estimate the variance of true effects, we computed  $T^2$  (DerSimonian & Laird, 2015), which estimates the true heterogeneity of effects in the same metric as the original effect size. For a more interpretable measure of heterogeneity, we also computed  $I^2$  (Higgins, Thompson, Deeks, & Altman, 2003), which reflects the estimated proportion of true variance in the total observed variance of effect sizes.

To examine the convergence of RVE and the more conventional approach (Borenstein et al., 2009), we also computed an overall summary effect from the set of independent effect sizes by fitting a conventional random-effects model to the data using the *metafor* package (Viechtbauer, 2010, 2016). To test the dispersion of observed effect sizes for significance, we computed Cochran's *Q* (Cochran, 1954) that is defined as the ratio of observed variation to the within-study error. *Q* follows a  $\chi^2$  distribution. A significant *Q* value provides evidence that the true effects vary. We again computed  $T^2$  and  $I^2$  to estimate true heterogeneity. The summary effect was computed as the precision-weighted mean of all independent



effect sizes. Weights were set to the inverse of the sum of the respective effect size variance ( $\text{Var}_g$ ) and the estimated true heterogeneity ( $T^2$ ).

**Moderation analyses.** To test for moderation, we employed mixed-effects RVE models. RVE offers the advantage that several moderators can be analyzed simultaneously while taking dependence of predictors (moderators) and outcomes (effect sizes) into account. These models logically extend standard multiple regression to meta-analysis. Accordingly, methodological concerns relevant in multiple regression are also relevant to meta-regression, especially overfitting of models, confounding among predictor variables, and low power. The number of studies and effect sizes was not large enough to include all coded moderators in a single model. We therefore followed a stepwise procedure to analyze the effect of moderators on the summary self-control training effect. In a first step, we separately tested the bivariate relationship of each moderator with the effect sizes. Categorical predictors were dummy coded, and continuous predictors were entered without transformation. This step delivers evidence for moderators without accounting for the influence of other, potentially correlated moderators.

In a second step, we entered multiple predictors simultaneously into the model to control for possible confounds between moderators. To avoid overfitting of the model, it was necessary to preselect predictors. Because we had no a priori theoretical rationale for the relative importance of the various moderators, we examined the converging evidence of a twofold strategy to determine the most suitable set of predictors. The first strategy was to select all moderators with  $p$  values of .100 or smaller in the bivariate tests (see previous paragraph). The second strategy was to fit models for all possible combinations of predictor variables. From this set, we retrieved the 100 models that explained the largest amount of true heterogeneity in the effect sizes as indicated by  $I^2$ . Next, we scored the relative importance of each moderator according to the following rule: A moderator received a score of 100 if it was included in the best model (i.e., the model explaining the largest amount of true heterogeneity), a score of 99 if it was included in the second best model, and so forth. Scores per moderator were summed up to create indices of relative importance. Thus, the maximum importance score was 5,050 for a moderator that was included in all of the hundred most potent models. We then chose moderators to be included in the model based on their importance indices. This approach should be less susceptible to chance patterns in the data biasing the model than simply selecting the model with the single lowest  $I^2$  because relative importance across multiple models is taken into account. We developed this method of selecting predictors based

on the idea of all-subsets methods in multiple regression (Hocking, 1976), as there are currently no other methods for model building in meta-analysis available.

### ***Small-study effects and publication bias***

Publication bias results if studies with certain characteristics (e.g., significant effects, large effect sizes) are systematically more likely to be submitted for publication by authors and/or accepted for publication by journals than studies with nonsignificant or negligible effect sizes. If this happens, the published literature is not representative of the full body of research and overestimates the population effect size (Ioannidis, 2008). Publication bias is a pervasive problem in the social sciences including psychology (Bakker et al., 2012; Franco, Malhotra, & Simonovits, 2014, 2016).

When a given literature is affected by publication bias, there will likely be a negative relationship between studies' effect sizes and their precision (or sample size): More precise studies with larger samples yield smaller effect sizes. This relationship is found in many meta-analyses (Levine et al., 2009). Small studies are more likely than larger studies to be excluded from the published literature due to nonsignificance or to be influenced by questionable data analysis methods that lead to significant findings at the cost of a factually increased Type I error (e.g.,  $p$ -hacking; Simmons, Nelson, & Simonsohn, 2011). Therefore, several statistical methods to detect and correct for publication bias investigate the relation between effect size and precision. These assume that in an unbiased literature small studies (on average) should be no more likely to deliver strong effects than larger studies.

It is important to note that a negative relationship between effect size and precision may also result from unproblematic causes other than publication bias. For example, smaller studies may have used other populations that may be more strongly affected by the intervention. Further, it is possible that certain particularly effective interventions are more readily applied in small than in large studies. Also, experimental manipulations may be more rigorously (and therefore more effectively) applied in small than in large studies (Sterne et al., 2011). These kinds of small-study effects reflect true heterogeneity of effect sizes. This heterogeneity may be quantified and potentially explained by statistical analyses such as moderator analyses. Importantly, they are not a problematic sign of publication bias. In case of an empirically negative association of effect size and study precision, meta-analysts therefore need to reflect about possible reasons for this relationship with respect to the specific body of research that is being investigated.

We applied two methods to *detect* publication bias (Funnel plot, Egger's regression test) and two further

methods to *correct for* publication bias (Trim and Fill; Precision Effect Estimation With Standard Error, PEESE). In the way they have been developed and validated, these techniques require statistical independence, so we applied them to the set of independent effect sizes (Borenstein approach). However, the logic of Egger's regression test and PEESE can be readily extended to RVE. We report both approaches for these procedures, but caution is warranted in interpreting the RVE variants until the techniques have been thoroughly validated in RVE.

**Funnel plot.** A funnel plot provides a graphical depiction of the relation between effect size and study precision. Effect size is plotted on the  $x$  axis and precision (as indicated by the standard error of the study effect size) on the  $y$  axis with highest precision on top. Funnel plots feature a triangle that is centered on the empirical fixed effect estimate. The width of the triangle is 1.96 standard errors to either side such that 95% of studies would be expected to fall within the triangle in the absence of small study effects and heterogeneity. Studies are expected to spread symmetrically around the estimated effect and increasingly closer to the actual population effect as precision increases. Asymmetry of the funnel plot indicates small study effects that may be indicative of publication bias. Importantly, the funnel plot assumes homogeneous effect sizes—that is, all interventions share the same underlying population effect size. This is an assumption that is unlikely for research in the social sciences (Borenstein et al., 2009). Under the more realistic assumption of a random-effects model and true heterogeneity, funnel plots may overestimate small study effects and, ultimately, publication bias (Lau, Ioannidis, Terrin, Schmid, & Olkin, 2006).

**Egger's regression test.** Egger's regression test investigates whether there is a statistically significant relationship between effect sizes and study precision. The currently advocated variant is a random-effects meta-regression of study effect size on study standard error with an additive between-study error component (Sterne & Egger, 2005). A significant regression weight for the studies' standard error indicates the presence of small-study effects and potentially publication bias. Similar to other regression-based methods, Egger's regression test suffers from low statistical power when the number of studies is small (Kromrey & Rendina-Gobioff, 2006). The test also performs unsatisfactorily under conditions of heterogeneity. However, this downside is partly compensated for by the advantage that the approach can incorporate other study characteristics (that may account for heterogeneity). This allows investigating whether a possible relation between study precision (as indicated by

the study standard error) and effect size remains significant after controlling for other potential influences on effect sizes (Sterne & Egger, 2005). Extending the idea of the test, we additionally investigated the relationship of effect size and standard errors in a mixed-effects RVE meta-regression with dependent effect sizes.

**Trim and Fill.** The Trim and Fill method (Duval & Tweedie, 2000a, 2000b) investigates asymmetry in a funnel plot. The algorithm removes extreme studies until the funnel plot is symmetric, yielding (in theory) an unbiased overall effect size estimate. It then imputes mirror images of the trimmed studies to estimate the correct variance of the overall distribution of studies. The Trim and Fill method suffers from the funnel plot's problematic assumption of truly homogeneous studies and a fixed effect size. In fact, simulation studies showed that Trim and Fill may even adjust for publication bias when factually none exists; reversely, it may adjust insufficiently when in fact publication bias is strong—especially when a few precise studies diverge from the overall meta-analytic estimate (Inzlicht, Gervais, & Berkman, 2015; Moreno et al., 2009; Terrin, Schmid, Lau, & Olkin, 2003). Another problem is that the method assumes publication bias to be driven by weak effects, whereas indeed it is more likely that it is driven by statistical nonsignificance (Simonsohn et al., 2014). Large studies with significant results but weak effects are more likely to be published than smaller studies with big, but nonsignificant, effects.

**PEESE.** PEESE (Stanley & Doucouliagos, 2014) computes a meta-regression in which the squared standard errors of the effect sizes (an indicator of precision) predict the effect sizes. If there is a significant relationship, this may indicate small study effects and potentially publication bias. The intercept of this regression line is thought to indicate the effect size of a “perfect” study with a standard error of zero that is used as an indicator of the bias-corrected overall meta-analytic effect size. Because PEESE is based on linear regression, it works best in meta-analyses with large numbers of studies. We fitted an additive error random-effects model to derive the intercept for PEESE.<sup>4</sup> Additionally, we extended the logic of this test to RVE and investigated the intercept in a mixed-effects RVE model that regressed (dependent) effect sizes on squared standard errors.

## Results

### *Characteristics of included studies*

The search identified 4,075 articles, of which 28 were eligible for inclusion, contributing a total of 33 studies. See Figure 1 for a PRISMA flow chart and the documentation

on the Open Science Framework (<https://osf.io/v7gxf/>) for (a) a list of excluded studies with reasons for exclusion and (b) full references for all included studies. Out of these 33 studies, 10 were unpublished as of December, 14, 2016. Publication dates ranged from 1999 to 2016 ( $Mdn = 2014$ ). Self-control training was operationalized through a diverse set of training paradigms. For instance, participants were prompted to use their nondominant hand for everyday tasks, to complete multiple sessions of computerized inhibitory control tasks, or to control their diet. The majority of training procedures lasted 2 weeks ( $m = 19$  effect sizes). In total, the analysis included data from 2,616 participants who were on average 21.63 years old. The average total sample size per study was  $n = 79$ , comprising mostly student samples ( $k = 27$  studies) and females ( $M_{female} = 67\%$ ). A wide array of outcomes was used to measure self-control-related constructs after ( $k = 16$ ) or both before *and* after the training ( $k = 17$ ). Nine studies also included a follow-up measurement. Training effects were predominantly evaluated through inhibitory control tasks ( $m = 18$ ) or in the domains of physical persistence ( $m = 15$ ), health behavior ( $m = 16$ ), and affect and well-being ( $m = 29$ ).

## Main analyses

**Outlier treatment.** Initial examination of the data showed that no effect deviated markedly from the rest of the distribution ( $z_{min} = -2.80$ ,  $z_{max} = 2.78$ ). Leave-one-out analyses showed that sequential removal of each effect size, respectively, did not strongly influence the RVE point estimate or precision of the summary effect ( $\Delta g_{min} = -0.022$ ,  $\Delta g_{max} = 0.021$ ,  $\Delta I^2_{min} = -2.09\%$ ,  $\Delta I^2_{max} = 1.24\%$ ). We therefore did not replace any effect sizes for the RVE analyses.

Examination of the independent study-level effect sizes (Borenstein approach) showed that one study (Davisson, 2013;  $g = -0.67$ ) deviated markedly from the rest of the distribution as indicated by several influence statistics,  $z = -2.61$  (next closest:  $z = -1.27$ ),  $r_{student} = -3.53$  (next closest:  $r_{student} = -1.13$ ),  $DFFITs = -0.45$  (clearly detached from the distribution), Cook's  $D = 0.16$  (clearly detached from the distribution). The study also had a strong influence on the heterogeneity estimate ( $\Delta I^2 = -12.21\%$ ). We therefore replaced this outlier effect size with the next most extreme effect size ( $g = -0.16$ ) for all analyses based on independent study-level effect sizes.

**Summary effect.** The RVE random-effects mean effect size of self-control training was  $g = 0.30$ ,  $CI_{95} [0.17, 0.42]$ ,  $p < .001$ , a small to medium effect size according to the conventions by Cohen (1988). More than half of the variance in observed effect sizes was estimated to reflect true differences in effect sizes ( $I^2 = 59.13\%$ ,  $T^2 = 0.093$ ).

According to common conventions, this amount of heterogeneity can be classified as moderate-to-substantial (Higgins et al., 2003).

We also computed a summary effect from the set of independent effect sizes by fitting a conventional intercept-only random-effects model (Borenstein approach). This analysis largely replicated the results of the RVE model in terms of the point estimate ( $g = 0.28$ ,  $CI_{95} [0.19, 0.38]$ ,  $p < .001$ ) and heterogeneity ( $I^2 = 48.47\%$ ,  $Q[32] = 62.10$ ,  $p = .001$ ,  $T^2 = 0.032$ ). Study statistics and results of this analysis are depicted in Figure S1 in the Supplemental Material available online.

## Moderator analyses

Descriptive statistics, confidence intervals, and inferential statistics of all categorical moderator variables are provided in Table 1. Numbers of effect sizes per group ( $m$ ) are provided in parentheses. Results of the meta-regressions for continuous moderators are provided in Table 2.

### Treatment-level moderator

**Type of training.** Five types of training procedures were applied in at least five studies. The most effect sizes originated from studies that used repeated sessions of computerized inhibitory control training ( $g = 0.21$ ,  $m = 56$ ), followed by training procedures prompting participants to use their nondominant hand for everyday tasks ( $g = 0.42$ ,  $m = 49$ ). Other common procedures required participants to repeatedly press and squeeze a hand strength training device until failure ( $g = 0.37$ ,  $m = 21$ ), to continuously regulate their posture by sitting and walking upright ( $g = 0.23$ ,  $m = 11$ ), or to continuously regulate their diet ( $g = -0.01$ ,  $m = 8$ ). Despite substantial descriptive differences, the overall analysis between the subgroups was not significant,  $HTZ(7.37) = 1.11$ ,  $p = .421$  (Fig. 2).

### Study-level moderators

**Length of training.** The majority of studies used a training procedure with a duration of 2 weeks ( $m = 19$ ; 58%). Thus, there was little variability in training duration, precluding a meaningful test of this moderator. Consequently, there was no significant moderation effect of the length of the training duration,  $b_1 = 0.003$ ,  $t(4.01) = 0.44$ ,  $p = .682$  (Fig. S2 in the Supplemental Material available online).

**Publication status.** On average, effect sizes were almost three times larger for published ( $g = 0.37$ ,  $m = 131$ ) than for unpublished studies ( $g = 0.13$ ,  $m = 27$ ). This difference was close to conventional levels of statistical significance,  $t(16.47) = 1.76$ ,  $p = .098$  (Fig. S3 in the Supplemental Material available online).

**Table 1.** Results of Moderation Analyses for Categorical Moderators

Moderator	Summary effect and 95% CI							Test of moderation				
	<i>g</i>	LL	UL	<i>t</i>	<i>df</i>	<i>p</i>	<i>k<sub>study</sub></i>	<i>m<sub>effects</sub></i>	Statistic	<i>df</i>	<i>p</i>	<i>I</i> <sup>2</sup>
Treatment-level moderator												
Type of training								HTZ = 1.11 7.37 .421 54.85%				
Inhibitory control task	0.21	-0.02	0.44	2.04	9.41	.070	11	56				
Handgrip	0.37	—	—	5.21	3.66	—	5	21				
Nondominant hand	0.42	0.25	0.59	5.58	9.25	<.001	11	56				
Posture regulation	0.23	—	—	2.55	2.53	—	4	11				
Diet regulation	-0.01	—	—	-0.02	2.61	—	4	28				
Study-level moderators												
Publication status								<i>t</i> = 1.76 16.47 .098 56.48%				
Published	0.37	0.24	0.51	5.83	20.53	<.001	23	131				
Unpublished	0.13	-0.16	0.41	1.01	8.52	.338	10	27				
Research group								<i>t</i> = 2.49 12.53 .028 55.61%				
Strength model	0.51	0.29	0.74	5.42	7.20	<.001	9	22				
Other	0.22	0.08	0.36	3.19	21.81	.004	24	136				
Control group quality								<i>t</i> = 1.73 20.79 .099 57.43%				
Active control group	0.23	0.08	0.39	3.10	19.70	.006	22	79				
Inactive control group	0.43	0.23	0.64	4.68	11.02	<.001	13	79				
Outcome-level moderators												
Type of outcome								HTZ = 1.55 10.40 .259 62.76%				
Affect and well-being	0.30	-0.12	0.71	1.87	4.70	.124	6	29				
Health behavior	0.12	-0.21	0.45	1.01	4.01	.368	6	16				
Inhibition	0.17	-0.26	0.59	0.90	8.30	.395	11	18				
Inhibition after ego depletion	0.48	0.10	0.86	3.33	4.54	.024	6	9				
Physical persistence	-0.06	-0.42	0.29	-0.46	5.28	.665	8	15				
Subjectivity of outcome measurement								<i>t</i> = 0.30 26.07 .588 59.79%				
Other	0.32	0.13	0.51	3.50	21.90	.002	26	80				
Subjective	0.26	0.14	0.39	4.44	13.93	<.001	18	78				
Lab-based versus real-world behavior								<i>t</i> = -0.88 16.32 .392 59.35%				
Lab-based	0.32	0.16	0.48	4.18	24.35	<.001	29	79				
Real-world	0.23	0.05	0.40	2.93	10.00	.015	12	79				
Stamina versus strength								<i>t</i> = -2.84 17.52 .011 56.50%				
Stamina	0.60	0.33	0.87	4.83	11.79	<.001	16	29				
Strength	0.21	0.07	0.34	3.14	23.92	.004	28	129				
Maximum versus realized potential								<i>t</i> < 0.01 27.75 .997 59.36%				
Maximum	0.30	0.02	0.58	2.26	15.91	.038	21	54				
Realized	0.30	0.19	0.40	5.91	19.74	<.001	23	104				
Follow-up								<i>t</i> = 1.12 9.69 .291 61.22%				
Follow-up	0.18	-0.02	0.39	2.16	6.74	.069	28	9				
Posttraining	0.31	0.16	0.45	4.32	27.00	<.001	74	31				

Note: *df* = associated small sample corrected degrees of freedom; *g* = effect size; *k<sub>study</sub>* = number of studies that contributed to the respective moderator level; LL = lower limit of the 95% CI; *m<sub>effect</sub>* = number of effect sizes in the respective moderator category; *p* = *p* value associated with the *t* value and *df* in the same row; *t* = *t* value associated with the *g* value in the same row testing statistical significance in the respective moderator level; UL = upper limit of the 95% CI. *Statistic* (test of moderation): *t* value for single parameter tests or Hotelling-T-approximated (HTZ) test statistic for multiple parameter tests. Significant test statistics indicate significance of the overall model. *I*<sup>2</sup> reflects the proportion of true variance in the total observed variance of effect sizes after accounting for the respective moderator. For some moderator models the values for *I*<sup>2</sup> can become larger than for the global summary-effect model because of missing values or differences in effect size computation. Note that for three subgroups in the *type of training* analysis, degrees of freedom fell below 4. Significance tests for the summary effects should thus not be interpreted. Accordingly, we did not report CI<sub>95</sub> and *p* values for the respective subgroups.

**Table 2.** Results of Moderation Analyses for Continuous Moderators

Moderator	Meta-regression		Test of moderation			$I^2$
	Intercept	Slope	$t$	$df$	$p$	
<i>Study-level moderators</i>						
Length of training	0.25	0.003	0.44	4.01	.682	60.58%
Gender ratio	0.04	0.008	2.02	13.27	.064	55.83%

Note: Test of Moderation,  $t$  value and corresponding small-sample corrected degrees of freedom. Significant  $t$  values indicate significant moderation.  $I^2$  reflects the proportion of true variance in the total observed variance of effect sizes. For some moderator models the values for  $I^2$  can become larger than for the global summary-effect model because of missing values or differences in effect size computation.

**Research group.** Significantly larger effects were found by the “strength model research group” ( $g = 0.51$ ,  $m = 22$ ) compared to other research groups ( $g = 0.22$ ,  $m = 136$ ),  $t(12.53) = 2.49$ ,  $p = .028$  (Fig. 3).

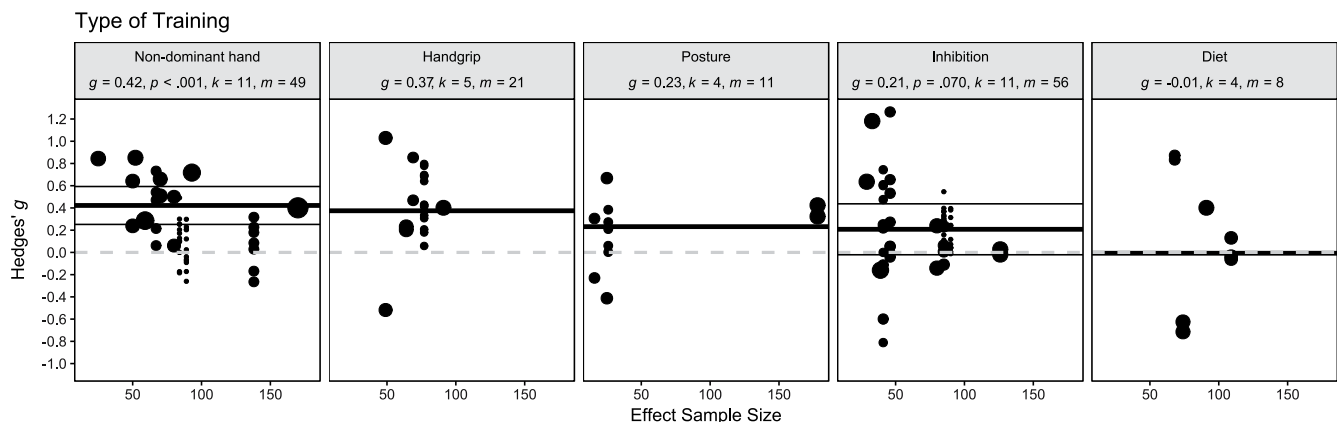
**Control group quality.** Descriptively smaller effects were evident in studies with active control groups ( $g = 0.23$ ,  $m = 79$ ) compared to studies with inactive control groups ( $g = 0.43$ ,  $m = 79$ ). The difference was close to statistical significance,  $t(20.79) = 1.73$ ,  $p = .099$  (Fig. S4 in the Supplemental Material available online).

**Gender ratio.** We imputed two missing values for this moderator by fitting the linear model based on all but the respective two effect sizes and then entering the two effect sizes in the regression equation, thus predicting the missing values from the effect sizes. The moderating

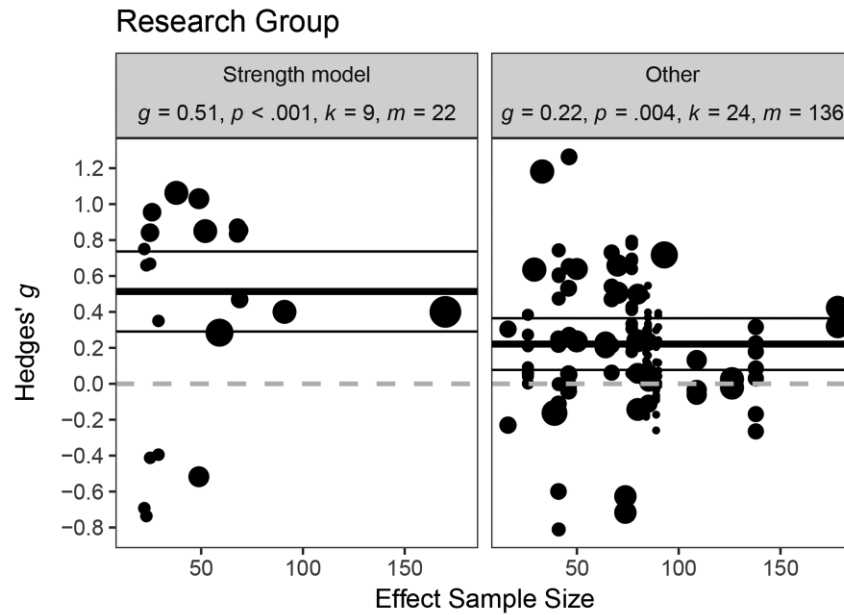
effect of the percentage of males in the study samples was close to statistical significance,  $b_1 = 0.008$ ,  $t(13.27) = 2.02$ ,  $p = .064$ , such that Hedges’  $g$  was predicted to increase by  $\Delta g = 0.08$  per 10% more males in the sample (Fig. 4). Percentages ranged from 0% to 64% across studies, so any interpretation of this slope should be limited to this range.

### Outcome-level moderators

**Type of outcome.** In total, the included studies featured 94 unique dependent variables. We grouped these variables into theoretically homogeneous clusters. Note that degrees of freedom for significance tests of subgroup summary effects are dependent on the number of studies and effect sizes within the respective cluster. Significance tests are only interpretable when  $df > 4$  (Tipton & Pustejovsky, 2015). Additionally, small clusters in subgroup analyses can bias tests of other clusters and the full model because



**Fig. 2.** Moderation by type of training,  $HTZ[7.37] = 1.11$ ,  $p = .421$ .  $g$  = Hedges’  $g$  summary effect within the respective subgroup;  $k$  = number of studies in a subgroup;  $m$  = number of effect sizes in a subgroup;  $p$  =  $p$  value testing Hedges’  $g$  against zero. Black dots represent individual effect sizes. The thick black horizontal lines represent the meta-analytic summary effects within the subgroups. The thin black horizontal lines represent the borders of the 95% CI around the subgroup summary effect. The dashed grey horizontal line represents the null effect at  $g = 0$ . For informational purposes, the sample size that was used to calculate the respective effect size is depicted on the  $x$  axis, but the moderating role of this attribute is not investigated in this analysis. Circle size represents the weight of the respective effect size in the meta-analytic RVE mixed-effects model depicted here. Diet: control one’s diet; handgrip: repeated use of a handgrip squeezer; inhibition: computerized inhibition control training procedures; non-dominant hand: use of non-dominant hand for everyday tasks; posture: keep an upright posture in everyday life. Note that for three subgroups in this analysis, degrees of freedom fell below 4. The corresponding significance tests for the summary effects should thus not be interpreted. Accordingly, we did not report  $CI_{95}$  and  $p$  values for the respective subgroups.



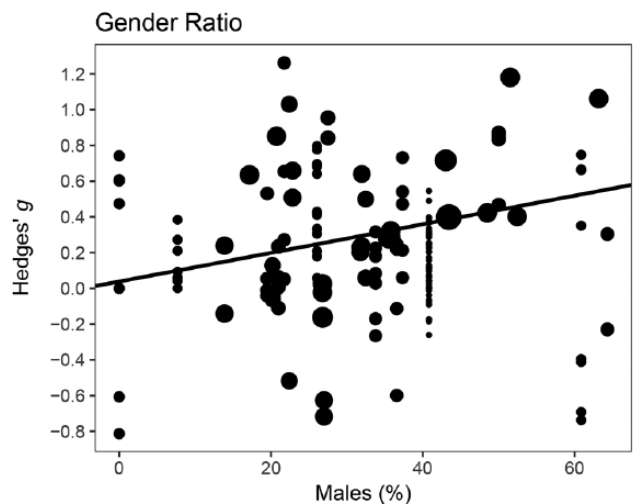
**Fig. 3.** Moderation by research group,  $t(12.53) = 2.49, p = .028$ .  $g$  = Hedges'  $g$  summary effect in subgroup;  $k$  = number of different studies within subgroup;  $m$  = number of effect sizes within subgroup;  $p$  =  $p$  value testing Hedges'  $g$  against zero. Black dots represent individual effect sizes. Thick black horizontal line, meta-analytic summary effect within subgroup; thin black lines, 95% CI; dashed grey line, null effect at  $g = 0$ . The associated sample size for each effect size is depicted on the  $x$  axis for informational purposes. Circle size represents effect size weight for the subgroup analysis.

they tend to increase imbalance in categorical predictors. Thus, it was necessary to exclude small clusters from the analysis to arrive at a model for which all parameters are interpretable. To do so, we sequentially removed clusters with the lowest degrees of freedom, until all degrees of freedom for the single parameter tests were four or larger. This procedure retained five outcome clusters in the final model. These were affect and wellbeing ( $g = 0.30, m = 29$ ), inhibitory control ( $g = 0.17, m = 18$ ), physical persistence ( $g = -0.06, m = 15$ ), health behavior ( $g = 0.12, m = 16$ ), and inhibitory control after depletion ( $g = 0.48, m = 9$ ). The difference between these outcome clusters was not significant,  $HTZ(10.40) = 1.55, p = .259$  (Fig. 5).

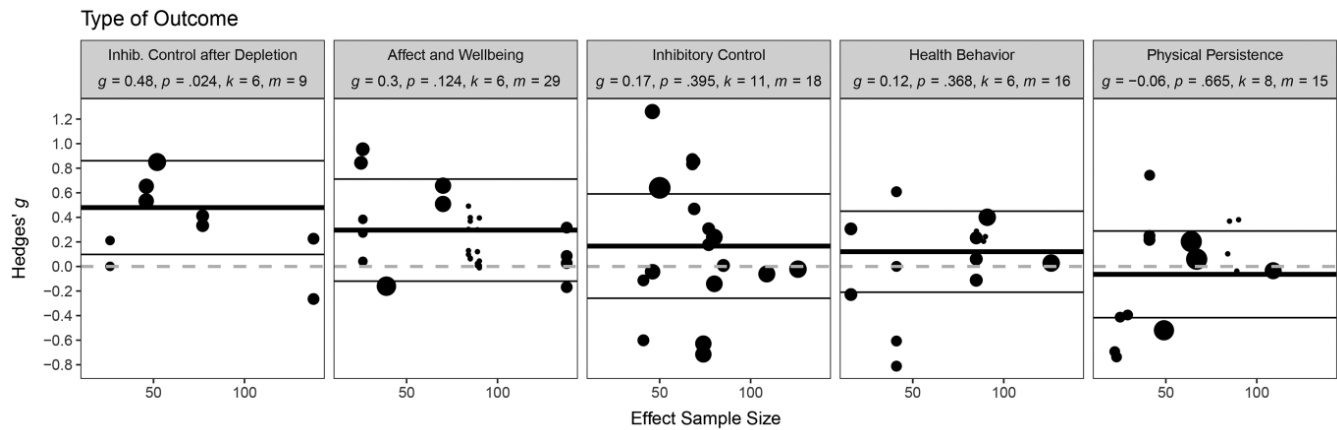
*Lab-based versus real-world behavior.* Effect sizes for outcomes that were measured in the lab ( $g = 0.32, m = 79$ ) were not significantly different from outcomes that reflect real-world behavior ( $g = 0.23, m = 79$ ),  $t(16.32) = -0.88, p = .392$  (Fig. S5 in the Supplemental Material available online).

*Stamina versus strength.* Effects for outcomes that were preceded by an effortful task (stamina;  $g = 0.60, m = 29$ ) were remarkably larger than for outcomes that were not preceded by an effortful task (strength;  $g = 0.21, m = 129$ ),  $t(17.52) = -2.84, p = .011$  (Fig. 6).

*Maximum versus realized potential.* Whether outcomes reflected maximum self-control potential ( $g = 0.30, m = 54$ ) or realized self-control potential ( $g = 0.30, m = 104$ ) had no effect on effect sizes,  $t(27.75) < 0.01, p = .997$  (Fig. S6 in the Supplemental Material available online).



**Fig. 4.** Moderation by gender ratio. The line represents the weighted RVE meta-regression of effect size on gender ratio,  $b_1 = 0.008, t(13.27) = 2.02, p = .064$ . Circle size represents effect size weight.

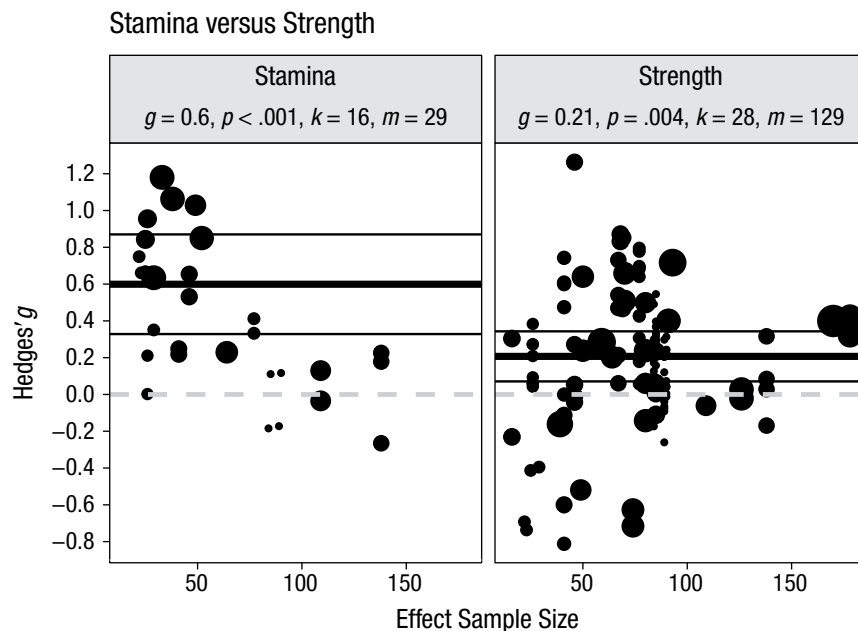


**Fig. 5.** Moderation by type of outcome,  $HTZ(10.40) = 1.55$ ,  $p = .259$ .  $g$  = Hedges'  $g$  summary effect in subgroup;  $k$  = number of different studies within subgroup;  $m$  = number of effect sizes within subgroup;  $p$  =  $p$  value testing Hedges'  $g$  against zero. Black dots represent individual effect sizes. Thick black horizontal line, meta-analytic summary effect within subgroup; thin black lines, 95% CI; dashed grey line, null effect at  $g = 0$ . The associated sample size for each effect size is depicted on the x axis for informational purposes. Circle size represents effect size weight for the subgroup analysis.

*Follow-up.* The distribution of the time-lags between the last day of the training and the time of outcome measurement was discontinuous with very large variance and therefore inept for a regression analysis. We therefore ran a categorical moderation analysis comparing post-test shortly after training with follow-up measurements (see *Effect Size Coding*). The follow-up measurements took place  $Mdn = 9.5$  days after the last day of training ( $M = 42$ ,  $SD = 65$ ,  $min. = 3.5$ ,  $max. = 184$ ). Outcome

measures that were assessed directly after the training yielded descriptively larger effect sizes ( $g = 0.31$ ,  $m = 74$ ) compared to outcomes measured at later time points ( $g = 0.18$ ,  $m = 28$ ). This difference was not significant,  $t(9.69) = 1.12$ ,  $p = .291$  (Fig. S7 in the Supplemental Material available online).

*Multiple moderators.* Testing multiple moderators simultaneously allows estimating the unique moderating



**Fig. 6.** Moderation by strength versus stamina,  $t(17.52) = -2.84$ ,  $p = .011$ .  $g$  = Hedges'  $g$  summary effect in subgroup;  $k$  = number of different studies within subgroup;  $m$  = number of effect sizes within subgroup;  $p$  =  $p$  value testing Hedges'  $g$  against zero. Black dots, individual effect sizes; thick black horizontal line, meta-analytic summary effect within subgroup; thin black lines, 95% CI; dashed grey line, null effect at  $g = 0$ . The associated sample size for each effect size is depicted on the x axis for informational purposes. Circle size represents effect size weight for the subgroup analysis.

**Table 3.** Summary of RVE Mixed-Effects Meta-Regression Model Predicting Effect Sizes From Multiple Moderators

Variable	<i>b</i>	<i>SE(b)</i>	<i>t</i>	<i>df</i>	<i>p</i>
Intercept	0.175	0.169	1.04	13.57	.317
Control group quality (inactive)	0.207	0.116	1.78	16.25	.094
Stamina versus strength (stamina)	-0.387	0.155	-2.50	13.20	.027
Research group (strength model)	0.205	0.114	1.80	12.17	.097
Self-control potential (realized)	0.174	0.146	1.20	16.84	.248
Gender ratio	0.006	0.004	1.45	13.71	.169

Note: Categorical predictors were dummy coded with 0 and 1. The moderator level coded as 1 is indicated in parentheses. *b* = regression coefficient; *df* = corresponding small-sample corrected degrees of freedom; *p* = *p* value associated with the *t* value and *df* in the same row; *SE(b)* = standard error of regression coefficient; *t* = *t* value testing whether the regression coefficient in the same row is significantly different from zero. The full model was significant,  $HTZ(13.46) = 3.32$ ,  $p = .036$ ,  $R^2 = 45.24\%$ .

role of each predictor while controlling for the overlap with other moderators. For this analysis, it was necessary to select a subset of moderators in order to avoid overfitting the model. Several moderators had to be excluded a priori from this process (e.g., due to missing values or restricted variance; please see the Supplemental Material available online for a full list of excluded moderators and reasons for exclusion).

As outlined in the *Methods* section, we employed two approaches to select the most appropriate moderators for this combined analysis: One approach relied on the findings from the bivariate moderator analyses; the second approach was a data-driven bottom-up approach seeking to explain a high degree of heterogeneity with a small number of predictors.

Results of the bivariate analyses suggested entering four moderators with *p* values of  $p = .100$  or smaller in the respective bivariate analysis into the combined model: control group quality, stamina versus strength, research group, and gender ratio. The data-driven bottom-up approach delivered converging evidence: We fitted multimoderator models for all possible combinations of predictor variables, resulting in  $2^9 = 512$  models, and retrieved the 100 models that explained the greatest amount of true heterogeneity (i.e., reduction in  $I^2$ ). Figure S8 in the Supplemental Material available online reports the relative importance of the nine examined moderators and can be interpreted akin to a Scree plot in factor analysis. There was a relatively large gap in importance between the fifth (*gender ratio*) and sixth (*subjectivity of outcome measurement*) most important moderators—suggesting entering the first five moderators in the combined analysis. Four of these five moderators match those identified in the bivariate analyses. *Maximum versus realized potential* emerged as an additional important

moderator despite being far from significance in the bivariate analysis ( $p = .996$ ). This suggests that this moderator binds residual variation in the other predictors and thereby contributes to explaining heterogeneity (suppression effect; Conger, 1974). In summary, the approach based on the bivariate analyses and the data-driven bottom-up approach provided converging evidence for the relevance of four moderators, and the latter approach unveiled the contribution of one additional moderator potentially acting as a suppressor variable.

The full model including all five predictors was significant,  $HTZ(13.46) = 3.32$ ,  $p = .036$  (Table 3). The model explained  $\Delta R^2 = 13.87\%$  more true effect size variance than the intercept-only model. The moderator *stamina versus strength* again emerged as significant ( $p = .027$ ). For *research group*, there still was a trend toward significance ( $p = .097$ ). The *p* value for *control group quality* was almost unchanged compared to the bivariate analysis ( $p = .097$ ). By contrast, *gender ratio* did not border on significance anymore ( $p = .169$ ). The alleged suppressor variable, *maximum versus realized potential*, was also not significant ( $p = .248$ ). These findings suggest that three of the four moderators that were at least marginally significant in the bivariate tests tended to explain unique portions of effect size heterogeneity, even when controlling for the influence of the other most potent moderators.

Note that in this regression, shared variance between predictors contributes to the overall model fit but is not assigned to any predictor specifically. Hence, to the extent that a predictor has a causal claim for parts of the nonassigned shared variance, even nonsignificant predictors may be important for the overall model. Nonsignificance of predictors should therefore not be overinterpreted as indicating that this predictor is unimportant in explaining heterogeneity.



## Small-study effects and publication bias

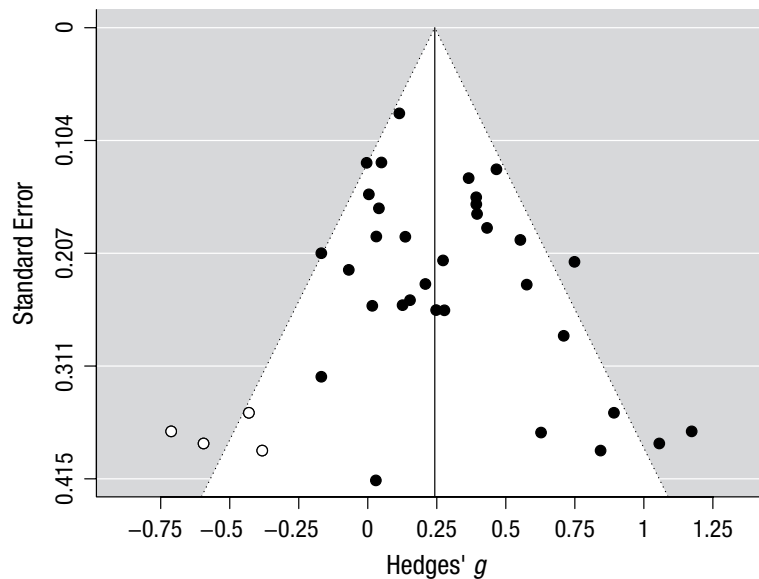
**Funnel plot.** Visual inspection of the funnel plot for the set of independent effect sizes (i.e., Borenstein approach, not RVE) revealed that the effect sizes were relatively symmetrically distributed around the summary effect (Fig. S9 in the Supplemental Material available online). For perfect symmetry, a set of studies with small-to-negative effect sizes and low precision was missing (see Trim and Fill below). Six studies fell out of the interval in that 95% of studies would be expected for any given level of precision. This analysis suggests a moderate degree of small-study effects and potentially publication bias.

**Egger's regression test.** The slope for the meta-regression of independent effect sizes on standard errors was significant,  $b_{se} = 1.51$ ,  $SE = 0.61$ ,  $z = 2.49$ ,  $p = .013$ , indicating a significant funnel plot asymmetry. We additionally entered covariates to examine whether standard errors had unique predictive value beyond other moderators (Sterne & Egger, 2005). We considered all moderators that were included in the multiple-predictor model reported above but could only enter *gender ratio* and *research group*. For the remaining moderators, several studies realized more than one moderator value, precluding this moderator from the analysis (e.g., featuring both an active and an inactive control condition). The effect of standard errors remained significant when controlling for *gender ratio* and *research group*,

$b_{SE} = 1.29$ ,  $SE = 0.62$ ,  $z = 2.08$ ,  $p = .038$ . Thus, Egger's regression test suggests a significant degree of small-study effects and potentially publication bias.

The RVE equivalent of Egger's regression test showed a similar yet nonsignificant relationship between standard errors and effect sizes,  $b_{SE} = 1.37$ ,  $SE = 0.80$ ,  $t(15.15) = 1.70$ ,  $p = .109$ . After reducing heterogeneity by controlling for all five moderators from the multiple moderator analysis reported above, the effect of standard errors was clearly not significant anymore,  $b_{SE} = 0.36$ ,  $SE = 0.70$ ,  $t(11.86) = 0.52$ ,  $p = .614$ . Follow-up analyses revealed that the notable change to the standard-error-only model in the  $p$  value was primarily due to the fact that effect sizes for self-control stamina (vs. strength) and effect sizes for inactive (vs. active) control groups tended to have greater standard errors. When these two moderators were not controlled for, the  $p$  value of the standard error predictor remained largely unchanged compared to the standard-error-only model ( $p = .136$ ).

**Trim and Fill.** After the previously reported *bias-detection* techniques, we turned to *bias-correction* techniques. The Trim and Fill method indicated that four studies were missing on the left of the mean meta-analytic effect size in order to obtain a fully symmetrical funnel plot (Fig. 7). Imputing these studies and adding them to the model delivered a bias-corrected random-effects summary estimate of  $g = 0.24$ ,  $SE_g = 0.051$ ,  $CI_{95} [0.14, 0.34]$ ,  $p < .001$ , that can be most adequately compared to the corresponding uncorrected summary effect size estimate



**Fig. 7.** Funnel plot after Trim and Fill bias correction. Note that this analysis is based on the study-level effect sizes (Borenstein approach). Compared to the original funnel plot (see the Supplemental Material available online), four studies were imputed to achieve symmetry (i.e., white circles). This resulted in a bias-corrected summary effect size of  $g = 0.24$ ,  $CI_{95} [0.14, 0.34]$  that is slightly smaller than the original (Borenstein approach) estimate of  $g = 0.28$ ,  $CI_{95} [0.19, 0.38]$ .

based on independent effect sizes ( $g = 0.28$ ). This analysis suggests a moderate degree of small-study effects and potentially publication bias.

**PEESE.** The meta-regression of independent effect sizes on squared standard errors was significant,  $b_1 = 3.41$ ,  $p = .008$ . The intercept that is thought to reflect the unbiased true meta-analytic summary effect was close to statistical significance,  $b_0 = 0.13$ ,  $SE_b = 0.07$ ,  $CI_{95} [-0.01, 0.27]$ ,  $z = 1.86$ ,  $p = .064$ . This corrected estimate is less than half of the size of the uncorrected summary effect ( $g = 0.30$  based on RVE,  $g = .28$  based on the Borenstein approach). The PEESE analysis suggests substantial small-study effects and potentially publication bias. Regressing dependent effect sizes on squared standard errors in an RVE mixed-effects model yielded a nonsignificant intercept,  $b_0 = 0.12$ ,  $SE_b = 0.11$ ,  $CI_{95} [-0.12, 0.36]$ ,  $t(16.31) = 1.08$ ,  $p = .295$ .

**Summary.** Both the funnel plot as well as Egger's regression test suggest that there are small-study effects in the dataset that may be indicative of publication bias. The Trim and Fill method delivered a moderately adjusted bias-corrected effect size estimate. By contrast, the bias-corrected PEESE estimate was less than half of the initial summary effect and only marginally significant. Extending the logic of Egger's regression test and PEESE to the RVE framework provided largely converging evidence, but the PEESE estimate for the summary effect was clearly nonsignificant. Taken together, all available evidence suggests that there are small-study effects that may at least partly reflect publication bias. Unfortunately, the severity of this bias is difficult to estimate based on currently available methods, especially because the available methods do not closely converge.

## Discussion

The present meta-analysis summarized studies testing the hypothesis that practicing self-control in one domain will lead to benefits in self-control performance in other domains. A random-effects meta-analysis based on 33 studies, 158 effect sizes, and more than 2,600 participants revealed an overall effect size of  $g = 0.30$ ,  $CI_{95} [0.17, 0.42]$ . Three comparisons help putting this effect size into perspective: First, it ranges between a small (0.2) and a medium (0.5) effect size according to the conventions by Cohen (1988), gravitating more toward a small than to a medium effect. Second, the effect size found here is a little larger than half of the average effect size found in a meta-analysis of 302 meta-analyses of a broad range of psychological, educational, and behavioral treatments ( $d = 0.50$ ,  $Mdn = 0.47$ ; Lipsey & Wilson, 1993). Third, the current effect size ranges between the fourth and fifth decile of effect sizes in social psychology according to a meta-analysis of

322 meta-analyses in social psychology that revealed a mean effect of  $d = .43$  ( $Mdn = .37$ ; Richard, Bond, & Stokes-Zoota, 2003). In sum, the present meta-analysis suggests that repeated practice improves self-control with an effect size that is somewhat smaller than common treatment effects in general and effects in social psychology in particular.

The analysis also revealed a moderate to high degree of heterogeneity, with about 60% of the variance estimated to be due to real differences in effect sizes. What are the underlying moderators that account for these differences? Training effects were stronger when they were assessed after performing an initial demanding self-control task, thus reflecting self-control stamina, as compared to assessments without such an initial task (reflecting self-control strength). This finding suggests that self-control training effects may be more pronounced when self-control demands accumulate (i.e., ego depletion).

Effects were also stronger when proponents of the strength model were involved compared to those conducted exclusively by other researchers. The origin of this effect is unclear. Possibly, proponents of the strength model operationalized treatments and instructed participants in particularly effective ways. Alternatively, strength model proponents may have been biased in favor of the hypothesis, or other researchers may have been biased against the hypothesis.

Effects also tended to be stronger in studies with inactive control conditions. This finding is plausible considering that inactive control conditions allow all kinds of mechanisms to drive training effects, while active control conditions narrow down the range of possible driving mechanisms. Finally, self-control training tended to be more effective in males than in females. One reason for this effect could be that men have stronger potentially problematic behavioral impulses, as has been suggested by previous research (Baumeister, Catanese, & Vohs, 2001; de Ridder et al., 2012). Men may therefore profit more from improved self-control through self-control training.

In an analysis that examined the most potent moderators simultaneously, *stamina versus strength*, *control group quality*, and *research group* remained at least marginally significant moderators. Gender ratio was no longer significant. Finally, it is noteworthy that even the comprehensive multimoderator model explained only a moderate amount of heterogeneity ( $\Delta I^2 = 13.89\%$ , remaining  $I^2 = 45.24\%$ ). This suggests that we either missed plausible moderating factors or that the bulk of variance in effect sizes is study-specific and not systematic.

In the course of working on this meta-analysis, we learned about another team of researchers working on a non-peer-reviewed analysis focusing on the effectiveness of self-control training to change health behavior (Beames, Schofield, & Denson, in press). Their work is related to

the present analysis, as the databases overlap. Yet there are notable differences between the two projects: They rely on different meta-analytic approaches (RVE vs. conventional random-effects meta-analysis), the calculation of effect sizes differs for some study designs, and they investigate different moderator variables. Despite these differences, it is noteworthy and reassuring that both analyses arrive at similar estimates for the uncorrected mean effectiveness of self-control training ( $g = .30$  in the present analysis vs.  $g = 0.36$  in the work by Beames et al., in press).

### ***Small-study effects and publication bias***

The Trim and Fill method indicated a moderate degree of bias and delivered a corrected effect size estimate of  $g = 0.24$ . By contrast, PEESE indicated a much greater degree of bias and delivered an estimate of  $g = 0.13$  that was not significantly different from zero. Note that an association between effect size and study precision (as detected by Trim and Fill and PEESE) can result from publication bias,  $p$ -hacking, and other biases, but it may also partly or completely be due to mundane reasons that cause small-study effects. For example, in the medical sciences, samples that are particularly receptive to an intervention due to a certain health condition may show particularly strong effect sizes. Such samples may also be difficult to recruit and therefore form smaller sample sizes than samples consisting of more readily available (and less susceptible) participants. Concerning the present database, we were unable to come up with analogous mundane reasons for small-study effects in the self-control training literature. Given how the field worked for many years (e.g., difficulty to publish nonsignificant findings), we deem it likely that there is publication bias in the investigated literature, but the severity of this bias is difficult to estimate. This is because none of the currently available techniques performs consistently well under conditions typical for (social) psychological literatures including heterogeneity and publication bias (Gervais, 2015; Inzlicht et al., 2015). Thus, the degree to which the bias-corrected estimates are biased themselves is unknown.

### ***Mechanisms underlying training effects***

The present meta-analysis suggests that self-control training may lead to slight improvements in self-control in other domains. The strength model postulates that the repeated control of dominant responses strengthens the “self-control muscle” (Baumeister & Vohs, 2016b). This metaphor is vivid and descriptive, but it is of limited explanatory value for the observed effects because it does

not specify the psychological mechanisms explaining training success. What do we know about mechanisms underlying training effects? One may approach this question from two perspectives. First, one may try to identify the crucial elements in a self-control training that make it effective. Second, one may think about the psychological processes that mediate self-control training effects.

The strength model claims that the repeated exertion of self-control by overcoming a dominant response is the driving “ingredient” of the self-control training. However, effect sizes stemming from studies with inactive control groups were almost 50% larger than those from studies with active control conditions. In studies with inactive control groups, various mechanisms besides the repeated control of dominant responses can cause an intervention effect (e.g., demand effects, greater engagement with the study by the active intervention group). What is more, even in the subset of studies with active control groups, few control groups closely matched the training condition, allowing for other than the focal mechanism to drive training effects. Thus, the net training effect due to the control of dominant responses may still be smaller than indicated by the training effect obtained for the studies employing active control groups ( $g_{\text{active}} = 0.23$ ,  $CI_{95} [0.08, 0.39]$ ).

With regard to the mediating psychological processes, surprisingly little is known. Some studies investigated changes in self-efficacy, awareness of the concept of self-control, and implicit theories about willpower as possible mechanisms but did not find evidence for mediation (Job, Friese, & Bernecker, 2015; Klinger, 2013; Muraven, 2010a, 2010b). In one study, self-control training reduced academic effort avoidance in university students, which partly mediated the effect of training on participants' grade point average (Job et al., 2015). This study suggests that motivational variables might play a mediating role. Future research has to test whether changes in effort avoidance may account for training effects in other domains than academic achievement.

One hitherto unexplored possibility is that training and control conditions differentially affect participants' expectations, thus allowing for placebo effects without actual changes in the trained constructs (Boot, Simons, Stothart, & Stutts, 2013; Foroughi, Monfort, Paczynski, McKnight, & Greenwood, 2016). Expectations regarding possible improvements on the dependent variables may differ between groups if they are not measured or, better, experimentally controlled—even in studies with active control groups. Hence, more knowledge is needed about how participants believe the (training or control) intervention is affecting them. What do participants believe their training regimen to be good for? What are their ideas about the researchers' goals for the study, and

which expectations about improvement on the measured constructs do participants hold?

In sum, little is known about the crucial elements of a training intervention. The literature to date does not deliver conclusive evidence that exerting self-control by repeatedly overriding dominant responses is the dominant *causal* mechanism that improves self-control over time and across domains. Even less is known about the psychological processes that are affected by a self-control training and lead to improved self-control performance.

### ***How to move forward?***

We will briefly discuss recommendations for future work concerning both methodological and theoretical developments. On the methodological level, future research should, first, conduct direct, high-powered, and preregistered replications. The set of the present 33 studies is very diverse, containing no close replications that would bolster confidence in obtained findings. Second, it will be important to more consistently use pre-post designs to increase statistical power. Based on the mean parameters evidenced by the current meta-analysis ( $g = .30$ ,  $N_{\text{average}} = 79$ ,  $\alpha = .05$ ,  $r_{\text{pre-post}} = .70$  within control groups), power for studies with pre-post designs is adequate ( $1-\beta = .92$ ). However, in post-only designs the same parameters result in a poor power of 37%, even with a one-tailed test. Note that it is possible that the true training effect is smaller than  $g = .30$ , which further increases demands on sample size. Third, future studies should employ (a) longer and (b) more varying training durations as well as (c) more consistently include follow-up measurements with (d) varying time lags. Only 9 of the analyzed 33 studies included a follow-up measurement (median time lag, 9.5 days). Effect sizes posttraining were considerably larger ( $g = 0.31$ ) than at follow-up ( $g = 0.18$ ). Although nonsignificant, this difference raises concerns about the practical utility of self-control training in the way it has been implemented to date. Researchers may want to consider ways to foster more sustainable self-control training, for example, by reminding participants of the training principles or implementing brief training refreshments after the main training period.

On the theoretical level, self-control training should only lead to performance improvements in activities that actually require self-control for a given person. This is not the case if a person has no goal to control a behavior. In this case, enacting such behaviors does not constitute a self-control failure. People who strive to achieve a certain goal or change a specific behavior—but are unsatisfied with their success in doing so (e.g., alcohol or nicotine consumption, eating behavior)—are the ones who are most likely to profit from a self-control training. For these people, a self-control training may constitute a

welcome means to work on the goal and provide a motivational boost by conveying the possibility that the training may help to achieve the respective goal (even if the person has no elaborate idea about how the training may do so). Ideally, a training sets in motion recursive motivational processes that help to build and keep up adaptive routines that may then contribute to lasting changes in behavior (Walton, 2014).

In addition, it will be important to control for differences in expectations about the consequences of a training regimen because different expectations may drive training effects (Boot et al., 2013). Such placebo effects are interesting in their own regard, but they limit researchers' ability to draw causal conclusions about a proposed working mechanism of *self-control* training. However, from the perspective of people who are interested in self-control improvements, making progress toward goal attainment is more pressing than identifying the underlying processes. If placebo effects do the trick and do so reliably, one may pragmatically advocate to let them do it. Researchers may interpret such, at first, poorly understood effects as an opportunity to investigate the underlying (motivational) processes in depth and apply this knowledge to new training interventions.

### ***Limitations***

The present work suffers from some limitations that future research may want to ameliorate. First, with 33 studies the available evidence on self-control training is still moderate. In light of the analyses presented here, it is premature to draw far-reaching conclusions. Several moderator analyses delivered substantial descriptive differences that did not reach significance, potentially due to low power.

A second limitation is that we could not calculate publication bias-corrected effect size estimates to the extent we had initially planned. Some techniques proved very unsatisfactory in simulation studies in that they severely underestimated true effect sizes under almost all realistic conditions (PET; Gervais, 2015; Inzlicht et al., 2015). Several other recently introduced techniques appear promising (Simonsohn et al., 2014; van Assen, van Aert, & Wicherts, 2015) but cannot be applied in a reasonable way to the current literature. These procedures rely exclusively on significant and published effect sizes with only one reported  $p$  value per study entering the computation. For the present meta-analysis, this would have led to an excessive loss of information (see Note 5). Also, they assume a homogeneous distribution of effect sizes, an assumption clearly not valid in the present literature. Future developments in meta-analytic techniques may be able to deliver valid publication bias-corrected effect size estimates for literatures with similar characteristics as the present one.

## Conclusion

Self-control is believed to be a domain-general capacity. The self-control training hypothesis suggests that practicing self-control in one domain improves self-control in other domains as well. The present random-effects meta-analysis found a small-to-medium self-control training effect. Bias-corrected estimates indicate a smaller effect. The working mechanisms underlying these far-transfer training effects are poorly understood and require further attention. We hope this meta-analysis will inspire researchers to further engage in this theoretically intriguing and practically relevant field of psychological research.

## Acknowledgments

The first and second author contributed equally to this work. We thank Alexander Hart for his assistance in coding the studies included in the meta-analysis; Joanne Beames, Tom Denson, and Martin Hagger for their valuable input and comments; and Zachary Fisher as well as Elizabeth Tipton for their advice with the implementation of the RVE approach to analyze the data. We are indebted to all primary authors who generously provided us with additional information about their studies.

## Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

## Funding

This work was supported by German Research Foundation (DFG) Grant FR-3605/3-1 (to M.F.).

## Supplemental Material

Additional supporting information may be found at <http://journals.sagepub.com/doi/suppl/10.1177/1745691617697076>. All data, code, full documentation of procedures, and additional analyses are available at <https://osf.io/v7gxf/>.

## Notes

1. Recently, there has been substantial debate about the magnitude of the ego depletion effect (Baumeister & Vohs, 2016a; Carter, Kofler, Forster, & McCullough, 2015; Hagger et al., 2016; Inzlicht et al., 2015). Details of this debate are beyond the scope of the present meta-analysis, which is primarily concerned with the second implication of the muscle analogy, the trainability of self-control.
2. Two further recent meta-analyses examined effects of computerized inhibitory control (a central component of self-control) training on health behavior (Allom, Mullan, & Hagger, 2016; Jones et al., 2016). However, studies included in these meta-analyses typically measured the outcome variable(s) directly after the training, leaving the possibility of short-term carryover and demand effects on the outcome measurement. In addition, many studies employed training-specific outcomes (e.g., effects of training the inhibition of food-related reactions

on subsequent eating behavior), whereas the current analysis focuses on far-transfer effects (i.e., practicing self-control in one domain and measuring effects in a different domain). In the studies included in the present analysis, these far-transfer effects were measured at least 1 day after the last training session. Thus, the overlap between these analyses and the present work is small due to the different aims and scopes.

3. This criterion was added to exclude studies that measured dependent variables only directly after the last training session, raising the possibility of short-term priming or demand effects. We made one exception from the rule for the following reasons: Lin, Miles, Inzlicht, and Francis (2016) measured various dependent variables *repeatedly* during a 30-day training period but not after the training period. We decided to include this study for two reasons: First, the study did not employ specific training sessions that would open the window for short-term priming and demand effects but employed a training procedure that instructed participants to use their nondominant hand for everyday life activities 5 days a week from 8 a.m. to 6 p.m. Second, the measurements (a) took place in a nonformalized context (online at home) and several dependent variables did not assess behavior or experience specific to the moment of assessment; instead, these outcome variables pertained to longer time spans (e.g., the previous week).

4. PEESE is often used together with a similar method called Precision Effect Test (PET; Stanley & Doucouliagos, 2014). Similar to Egger's regression test, PET uses the effect sizes' standard errors as predictors instead of the squared standard errors in case of PEESE. In Egger's regression test, the regression weight of the standard error predictor is interpreted. PET interprets the intercept as the bias-corrected true effect size. PET has been heavily criticized based on evidence that the algorithm performs particularly poorly and severely underestimates the true effect size under a range of conditions typical for social psychology (e.g., heterogeneity, small number of studies; Gervais, 2015, 2016; Inzlicht et al., 2015; Reed, 2015). We therefore refrained from using PET to correct for publication bias. Two other recently proposed methods to estimate true effect sizes in meta-analyses are *p-curve* and *p-uniform* (Simonsohn et al., 2014; van Assen et al., 2015). Both methods rely exclusively on significant and published effect sizes. Also, only one *p* value per study may enter the computation. For the present meta-analysis, these rules would have led to a substantial loss of information, because a considerable part of effect sizes were nonsignificant and/or unpublished. In addition, many studies included more than one dependent variable, of which we could have included only one. Of the total of 158 effect sizes, less than 20 would have been available for the computation of the effect size estimates based on *p-curve* and *p-uniform*. We therefore refrained from applying these methods.

## References

- Allom, V., Mullan, B., & Hagger, M. S. (2016). Does inhibitory control training improve health behaviour? A meta-analysis. *Health Psychology Review, 10*, 168–186.
- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543–554.

- Baumeister, R. F., Catanese, K. R., & Vohs, K. D. (2001). Is there a gender difference in strength of sex drive? Theoretical views, conceptual distinctions, and a review of relevant evidence. *Personality and Social Psychology Review*, *5*, 242–273.
- Baumeister, R. F., & Vohs, K. D. (2016a). Misguided effort with elusive implications. *Perspectives on Psychological Science*, *11*, 574–575.
- Baumeister, R. F., & Vohs, K. D. (2016b). Strength model of self-regulation as limited resource: Assessment, controversies, update. In M. O. James & P. Z. Mark (Eds.), *Advances in experimental social psychology* (Vol. 54, pp. 67–127). San Diego, CA: Academic Press.
- Baumeister, R. F., Vohs, K. D., & Tice, D. M. (2007). The strength model of self-control. *Current Directions in Psychological Science*, *16*, 351–355.
- Beames, J. R., Schofield, T. P., & Denson, T. F. (in press). A meta-analysis of improving self-control with practice. In D. T. D. de Ridder, M. A. Adriaanse, & K. Fujita (Eds.), *Handbook of self-control in health and well-being*. Abingdon, UK: Routledge.
- Berkman, E. T. (2016). Self-regulation training. In K. D. Vohs & R. F. Baumeister (Eds.), *Handbook of self-regulation: Research, theory, and applications* (3rd ed., pp. 440–457). New York, NY: Guilford.
- Boot, W. R., Simons, D. J., Stothart, C., & Stutts, C. (2013). The pervasive problem with placebos in psychology: Why active control groups are not sufficient to rule out placebo effects. *Perspectives on Psychological Science*, *8*, 445–454.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.
- Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, *144*, 796–815.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290.
- Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*, 101–129.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Conger, A. J. (1974). A revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement*, *34*, 35–46.
- Cranwell, J., Benford, S., Houghton, R. J., Golembewski, M., Fischer, J. E., & Hagger, M. S. (2014). Increasing self-regulatory energy using an internet-based training application delivered by smartphone technology. *Cyberpsychology, Behavior, and Social Networking*, *17*, 181–186.
- Daly, M., Delaney, L., Egan, M., & Baumeister, R. F. (2015). Childhood self-control and unemployment throughout the life span: Evidence from two British cohort studies. *Psychological Science*, *26*, 709–723.
- Davisson, E. K. (2013). *Strengthening self-control by practicing inhibition and initiation*. Unpublished dissertation thesis, Duke University, Durham, NC. Retrieved from <http://dukespace.lib.duke.edu/dspace/handle/10161/7258>
- Del Re, A. C., & Hoyt, W. T. (2014). *MAd: Meta-analysis with mean differences* (R package version 0.8-2) [computer software]. Retrieved from <http://cran.r-project.org/web/packages/MAd>
- Denson, T. F., Capper, M. M., Oaten, M., Friese, M., & Schofield, T. P. (2011). Self-control training decreases aggression in response to provocation in aggressive individuals. *Journal of Research in Personality*, *45*, 252–256.
- de Ridder, D. T. D., Lensvelt-Mulders, G., Finkenauer, C., Stok, F. M., & Baumeister, R. F. (2012). Taking stock of self-control: A meta-analysis of how trait self-control relates to a wide range of behaviors. *Personality and Social Psychology Review*, *16*, 76–99.
- DerSimonian, R., & Laird, N. (2015). Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials*, *45*, 139–145.
- Duckworth, A. L., & Seligman, M. E. P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, *16*, 939–944.
- Duval, S., & Tweedie, R. (2000a). A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*, 89–98.
- Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455–463.
- Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904.
- Finkel, E. J., DeWall, C. N., Slotter, E. B., Oaten, M., & Foshee, V. A. (2009). Self-regulatory failure and intimate partner violence perpetration. *Journal of Personality and Social Psychology*, *97*, 483–499.
- Fisher, Z., Tipton, E., & Hou, Z. (2016). *robumeta: Robust variance meta-regression* (R package version 1.8) [computer software]. Retrieved from <https://cran.r-project.org/package=robumeta>
- Foroughi, C. K., Monfort, S. S., Paczynski, M., McKnight, P. E., & Greenwood, P. M. (2016). Placebo effects in cognitive training. *Proceedings of the National Academy of Sciences, USA*, *113*, 7470–7474.
- Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*, 1502–1505.
- Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological & Personality Science*, *7*, 8–12.
- Gervais, W. M. (2015, June 16). *Putting PET-PEESE to the test*. Retrieved from <http://willgervais.com/blog/2015/6/25/putting-pet-peese-to-the-test-1>
- Gervais, W. M. (2016, March 3). *heavy PETting*. Retrieved from <http://willgervais.com/blog/2016/3/3/enough-heavy-petting>
- Gottfredson, M. R., & Hirschi, T. (1990). *A general theory of crime*. Stanford, CA: Stanford University Press.
- Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A., . . . Zwieneberg, M. (2016). A multi-lab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*, 546–573.
- Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin*, *136*, 495–525.
- Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*, 1900–1902.

- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128.
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*, 39–65.
- Higgins, J. P. T., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0) [updated March 2011]. Retrieved from <http://handbook.cochrane.org/>
- Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*, 557–560.
- Hocking, R. R. (1976). Analysis and selection of variables in linear regression. *Biometrics*, *32*, 1–49.
- Inzlicht, M., & Berkman, E. (2015). Six questions for the resource model of control (and some answers). *Social & Personality Psychology Compass*, *9*, 511–524.
- Inzlicht, M., Gervais, W. M., & Berkman, E. T. (2015). *Bias-correction techniques alone cannot determine whether ego depletion is different from zero: Commentary on Carter, Kofler, Forster, & McCullough, 2015*. Retrieved from <http://ssrn.com/abstract=2659409>
- Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–648.
- Job, V., Friese, M., & Bernecker, K. (2015). Effects of practicing self-control on academic performance. *Motivation Science*, *1*, 219–232.
- Jones, A., Di Lemma, L. C. G., Robinson, E., Christiansen, P., Nolan, S., Tudur-Smith, C., & Field, M. (2016). Inhibitory control training for appetitive behaviour change: A meta-analytic investigation of mechanisms of action and moderators of effectiveness. *Appetite*, *97*, 16–28.
- Klinger, J. (2013). *Examining mechanisms of self-control improvement*. Unpublished master's thesis, University of Waterloo, Ontario, Canada.
- Kromrey, J. D., & Rendina-Gobioff, G. (2006). On knowing what we do not know: An empirical comparison of methods to detect publication bias in meta-analysis. *Educational and Psychological Measurement*, *66*, 357–373.
- Lakens, D. D., Hilgard, J., & Staaks, J. J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, *4*, Article 24. doi:10.1186/s40359-016-0126-3.
- Lau, J., Ioannidis, J. P. A., Terrin, N., Schmid, C. H., & Olkin, I. (2006). Evidence based medicine: The case of the misleading funnel plot. *British Medical Journal*, *333*, 597–600.
- Levine, T., Asada, K. J., & Carpenter, C. (2009). Sample sizes and effect sizes are negatively correlated in meta-analyses: Evidence and implications of a publication bias against nonsignificant findings. *Communication Monographs*, *76*, 286–302.
- Lin, H., Miles, E., Inzlicht, M., & Francis, Z. (2016). *Mechanisms underlying self-control training*. Manuscript in preparation.
- Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*, 1181–1209.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- Marín-Martínez, F., & Sánchez-Meca, J. (1999). Averaging dependent effect-sizes in meta-analysis: A cautionary note about procedures. *The Spanish Journal of Psychology*, *2*, 32–38.
- Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology*, *49*, 270–291.
- Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of “far transfer”: Evidence from a meta-analytic review. *Perspectives on Psychological Science*, *11*, 512–534.
- Miles, E., Sheeran, P., Baird, H., Macdonald, I., Webb, T. L., & Harris, P. R. (2016). Does self-control improve with practice? Evidence from a six-week training program. *Journal of Experimental Psychology: General*, *145*, 1075–1091.
- Mischel, W., Ayduk, O., Berman, M. G., Casey, B. J., Gotlib, I. H., Jonides, J., . . . Shoda, Y. (2011). ‘Willpower’ over the life span: Decomposing self-regulation. *Social Cognitive and Affective Neuroscience*, *6*, 252–256.
- Mischel, W., & Baker, N. (1975). Cognitive appraisals and transformations in delay behavior. *Journal of Personality and Social Psychology*, *31*, 254–261.
- Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*, 8–14.
- Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., . . . Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences, USA*, *108*, 2693–2698.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*, e1000097. doi:10.1371/journal.pmed.1000097
- Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology*, *9*, Article 2. doi:10.1186/1471-2288-9-2
- Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*, 364–386.
- Muraven, M. (2010a). Building self-control strength: Practicing self-control leads to improved self-control performance. *Journal of Experimental Social Psychology*, *46*, 465–468.
- Muraven, M. (2010b). Practicing self-control lowers the risk of smoking lapse. *Psychology of Addictive Behaviors*, *24*, 446–452.
- Muraven, M., Baumeister, R. F., & Tice, D. M. (1999). Longitudinal improvement of self-regulation through practice: Building self-control strength through repeated exercise. *Journal of Social Psychology*, *139*, 446–457.
- Oaten, M., & Cheng, K. (2006a). Improved self-control: The benefits of a regular program of academic study. *Basic and Applied Social Psychology*, *28*, 1–16.
- Oaten, M., & Cheng, K. (2006b). Longitudinal gains in self-regulation from regular physical exercise. *British Journal of Health Psychology*, *11*, 717–733.
- Oaten, M., & Cheng, K. (2007). Improvements in self-control from financial monitoring. *Journal of Economic Psychology*, *28*, 487–501.

- Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., . . . Ballard, C. G. (2010). Putting brain training to the test. *Nature*, *465*, 775–778.
- Piquero, A. R., Jennings, W. G., Farrington, D. P., Diamond, B., & Gonzalez, J. M. R. (2016). A meta-analysis update on the effectiveness of early self-control improvement programs to improve self-control and reduce delinquency. *Journal of Experimental Criminology*, *12*, 249–264.
- Pustejovsky, J. (2016). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections* (R package version 0.2.1.9000) [computer software]. Retrieved from <https://github.com/jepusto/clubSandwich>
- Reed, W. R. (2015). A Monte Carlo analysis of alternative meta-analysis estimators in the presence of publication bias. *Economics*, *9*, 1–40.
- Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363.
- Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin*, *138*, 628–654.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*, 666–681.
- Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*, 60–78.
- Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99–110). New York, NY: Wiley.
- Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., . . . Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal*, *342*, d4002.
- Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, *72*, 271–324.
- Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods*, *5*, 13–30.
- Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, *22*, 2113–2126.
- Tipton, E. (2013). Robust variance estimation in meta-regression with binary dependent effects. *Research Synthesis Methods*, *4*, 169–187.
- Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*, 375–393.
- Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, *40*, 604–634.
- van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*, 293–309.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48.
- Viechtbauer, W. (2016). *metafor: Meta-analysis package for R* (R package version 1.9-9) [computer software]. Retrieved from <https://cran.r-project.org/package=metafor>
- Walton, G. M. (2014). The new science of wise psychological interventions. *Current Directions in Psychological Science*, *23*, 73–82.



**Part I, Paper 2: “P-Hacking and publication bias interact to distort meta-analytic effect size estimates.”**

*p*-Hacking and Publication Bias Interact to Distort Meta-Analytic Effect Size Estimates

Malte Friese &amp; Julius Frankenbach

Saarland University

Science depends on trustworthy evidence. Thus, a biased scientific record is of questionable value because it impedes scientific progress, and the public receives advice on the basis of unreliable evidence that has the potential to have far-reaching detrimental consequences. Meta-analysis is a valid and reliable technique that can be used to summarize research evidence. However, meta-analytic effect size estimates may themselves be biased, threatening the validity and usefulness of meta-analyses to promote scientific progress. Here, we offer a large-scale simulation study to elucidate how *p*-hacking and publication bias distort meta-analytic effect size estimates under a broad array of circumstances that reflect the reality that exists across a variety of research areas. The results revealed that, first, very high levels of publication bias can severely distort the cumulative evidence. Second, *p*-hacking and publication bias interact: At relatively high and low levels of publication bias, *p*-hacking does comparatively little harm, but at medium levels of publication bias, *p*-hacking can considerably contribute to bias, especially when the true effects are very small or are approaching zero. Third, *p*-hacking can severely increase the rate of false positives. A key implication is that, in addition to preventing *p*-hacking, policies in research institutions, funding agencies, and scientific journals need to make the prevention of publication bias a top priority to ensure a trustworthy base of evidence.

Word count: 220

*Keywords:* meta-analysis, *p*-hacking, publication bias, meta-science2020, *Psychological Methods*, 25, 456-471.

© 2020, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article is available via its DOI: 10.1037/met0000246

Science depends on trustworthy evidence. If the published scientific record is biased, its value is seriously compromised: Researchers are led to believe in phenomena that are frail or might not even exist at all. Theory development is led astray. The ability to explain the world to the public is undermined, and public trust in science is compromised. In short: If science fails to deliver trustworthy, reliable evidence, a society may wonder why it should invest in scientific endeavors at all.

In recent years, the trustworthiness of psychological science has been seriously questioned

(Lilienfeld & Waldman, 2017). One important reason for the doubt and criticism has been the observation that many published psychological studies cannot be replicated in a straightforward fashion (e.g., Nosek & Lakens, 2014; Open Science Collaboration, 2015). Several problems that may contribute to this lamentable status have been identified, including low statistical power (Bertamini & Munafò, 2012; Maxwell, 2004), the use of questionable research practices (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011), publication bias (Bakker, van Dijk, & Wicherts, 2012; Fanelli,

Malte Friese and Julius Frankenbach, Department of Psychology, Saarland University.

Both authors contributed equally to this work and share the first authorship. We thank Michael Inzlicht, David D. Loschelder, Dorota Reis, Simine Vazire, and an anonymous Reviewer for valuable comments on an earlier version of this article. All code is available at <https://osf.io/phwne/>.

Correspondence concerning this article should be addressed to Malte Friese or Julius Frankenbach, Department of Psychology, Saarland University, Campus A2 4, 66123 Saarbrücken, Germany. Email: malte.friese@uni-saarland.de or julius.frankenbach@gmail.com

2010), and hypothesizing after the results are known (HARKing; Kerr, 1998). Together, these problems may lead researchers to seriously overestimate the robustness of the cumulative evidence in a field of investigation. True effect sizes can be critically smaller and less stable than the available evidence suggests. As a consequence, Psychology has started to experience all of the detrimental consequences alluded to above.

The most important methodological tool that can be used to quantitatively summarize the available evidence in a given research literature is a meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2009; Gurevitch, Koricheva, Nakagawa, & Stewart, 2018; Johnson & Eagly, 2014). Meta-analyses summarize the results of multiple studies addressing the same research question to reach an overall understanding of the state of the evidence. Thus, the unit of analysis changes from the individual level to the aggregated level—ideally, the complete body of evidence that has been collected with respect to a particular research question (Murad & Montori, 2013).

Meta-analyses have several strengths. One salient strength is that due to the greater statistical power, meta-analyses can be conducted to reliably detect even small effects that are not as easy to detect with single primary studies. Meta-analyses can also be used to estimate (summary) effect sizes with greater precision (i.e., narrower confidence intervals) than single primary studies. Importantly, meta-analyses can also estimate variation in underlying true effects (e.g., when different populations are investigated across studies, different manipulations are employed, or different dependent variables are used) and shed light on moderating factors that may have been missed or were impossible to investigate in the primary studies. These and other properties make the meta-analysis a powerful tool that researchers can use to obtain a comprehensive overview of what is known and not yet known in a given field of research.

In times of doubt about the replicability and robustness of individual primary studies, researchers are even more likely to rely on meta-analyses to obtain a trustworthy picture of the state of the evidence. Importantly, the validity of meta-analyses may also be threatened by the problems that lead to a lack of replicability and robustness in primary studies. For example, the quality of a meta-analysis crucially depends on the quality of the primary studies it is composed of. In a field featuring many poorly conducted studies, a meta-analysis may be unable to level out the biases of primary studies if these biases are systematic rather than unsystematic (Borenstein et al., 2009). Thus, it is imperative to examine the impact that various sources of bias can have on meta-analytic effect size estimates.

In recent years, two problems in particular have received considerable attention as presumably the leading causes of deficient robustness in psychological science: Questionable research practices—often referred to as *p*-hacking—and publication bias (Bakker

et al., 2012; Munafò et al., 2017; Nelson, Simmons, & Simonsohn, 2018). It is widely assumed that both *p*-hacking and publication bias can seriously distort the cumulative evidence and consequently the meta-analyses that are conducted to summarize this evidence.

There has been an active meta-scientific debate about the *prevalence* of *p*-hacking and publication bias (e.g., Dubben & Beck-Bornholdt, 2005; Hartgerink, 2017; Head, Holman, Lanfear, Kahn, & Jennions, 2015; Kuhberger, Fritz, & Scherndl, 2014). What has been surprisingly neglected are the quantifiable *consequences* of *p*-hacking and publication bias with respect to cumulative knowledge formation. Some of the most important questions are: To what extent do different degrees of *p*-hacking and publication bias distort meta-analytic effect size estimates? What are the relative impacts of *p*-hacking and publication bias in bringing about these distortions? And how might the consequences of *p*-hacking and publication bias depend on the extent to which the other exists; that is, how might they interact to jointly distort cumulative scientific evidence? This knowledge is crucial: In order to implement the structural and procedural changes in research institutions, publishing, funding, and policy that promise the greatest progress for obtaining a realistic reflection of reality from the published literature, the field needs to know which problems cause the greatest harm under which circumstances.

Here, we addressed these important questions about the quantifiable consequences of *p*-hacking and publication bias for cumulative knowledge formation by conducting a large-scale simulation study. In this study, we made no assumptions about the prevalence rates of *p*-hacking and publication bias. Rather, we simulated their consequences using a broad range of potential severities, thus accounting for (a) potential realities across a diverse array of research and (b) diverging assumptions about these prevalence rates by different researchers.

## What are *p*-Hacking and Publication Bias?

### *Definition of p-hacking*

The concept of *p*-hacking refers to nonprincipled decisions during data analysis that are aimed at reducing the *p*-value of a significance test and thus make the data look more robust than they actually are. Examples are selectively excluding outliers, collecting additional data without controlling for inflated error rates, or selectively controlling for covariates (John et al., 2012; Simmons et al., 2011). Thus, although there are several different *p*-hacks, they all serve as functionally equivalent means to the same end: To reduce an originally nonsignificant *p*-value to significance. Such *p*-hacking can be caused by bad

intentions but may often be driven by good intentions to help the data reveal the insights that are presumably hidden in them and are otherwise not as clearly observable (Nelson et al., 2018). Also, it is likely that many researchers are not aware of the extent to which their data-analytic practices increase false-positive rates (Simmons et al., 2011).

The prevalence of *p*-hacking in (psychological) science is a subject of debate (Fiedler & Schwarz, 2016; John et al., 2012). Some researchers have argued that *p*-hacking is omnipresent and is so pervasive that it helps researchers get around a file drawer because they will *p*-hack (almost) any study into publishable significance (Nelson et al., 2018). Large-scale analyses have sought (and found) indirect evidence for *p*-hacking by examining empirical *p*-value distributions in the published literature that suggested a cluster of *p*-values just below .05 (e.g., Head et al., 2015; Masicampo & Lalande, 2012). These findings are consistent with the assumption that most *p*-hacking researchers stop once they reach an outcome that barely crosses the crucial .05 border. These analyses and their underlying assumptions have been criticized on methodological and logical grounds (e.g., Hartgerink, 2017; Lakens, 2015). They might also not be specific enough because they lump together all *p*-values reported across a large array of publications, including the many for which there was little publication pressure (e.g., manipulation checks, sanity checks, follow-up analyses, nonfocal hypothesis tests) with the few focal tests for which there was publication pressure and therefore the incentive to *p*-hack. In sum, the true prevalence of *p*-hacking is unknown (Bruns & Ioannidis, 2016) and most likely varies across the different literatures.

### ***Definition of publication bias***

Publication bias occurs when many studies that did not produce the desired outcomes are not published (Fanelli, 2012; Franco, Malhotra, & Simonovits, 2014). Authors are less likely to submit “failed” studies for publication, and if they do, reviewers and editors are less likely to support the publication of such studies compared with “successful” studies that produced statistically significant outcomes. As a result, most studies in Psychology that get published report hypotheses that “worked” (Fanelli, 2010; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995).

Publication bias is a major threat to the validity of meta-analytic results. To reflect the true state of the evidence, meta-analyses require access to the full evidence base, or at least a representative sample of this evidence. If studies with certain characteristics are more likely to be included in a meta-analysis than others, this introduces systematic bias that distorts the conclusions that will be drawn. Meta-analyses enjoy a good reputation and are particularly trusted by many researchers due to their often seemingly authoritative

data base. If they paint a misleading picture of the evidence because the evidence base is biased, scientific progress may be hampered because incorrect theories and beliefs will remain popular (Ferguson & Heene, 2012).

There is little disagreement in the literature that publication bias exists, but the actual prevalence of bias has been debated and tends to vary across subdisciplines and different areas of research (Fanelli, Costas, & Ioannidis, 2017). Some analysts have suggested that publication bias is pervasive and particularly so in the Social Sciences such as Psychology (Bakker et al., 2012; Fanelli, 2010, 2012; Ferguson & Brannick, 2012).

### **The Detrimental Impact of *p*-Hacking and Publication Bias**

There is a general consensus that both *p*-hacking and publication bias exist in the psychological literature. What is under debate and unknown is their factual prevalence rates in Psychology as a whole and its subdisciplines. For the present study, the actual prevalence rates of *p*-hacking and publication bias were not our primary interest (and we did not seek to determine the actual prevalence rates). Instead, we sought to model the *consequences* of *p*-hacking and publication bias in terms of meta-analytic effect size distortions *as a function of* wide ranges of potential severities of *p*-hacking and publication bias.

How do *p*-hacking, publication bias, and their interplay distort meta-analytic effect size estimates? This can be conveniently illustrated with a funnel plot. Consider Figure 1. Let us assume researchers suspect a difference between two conditions but do not know whether or not this difference actually exists. Panel A depicts 1,000 simulated studies with a true population effect of zero. (Hence, in this example, the suspected difference does not exist.) Larger studies are located toward the top of the funnel and are more closely distributed around the true effect size. By contrast, smaller studies are located toward the bottom of the funnel and are more widely distributed around the true effect size. By definition, only 2.5% of all studies produce significant effects in the expected direction (genuine false positives) when a two-tailed significance test with  $\alpha = .05$  is applied (i.e., studies that fall outside the funnel and to the right). These significant findings that fall in the expected direction have a high probability of getting published. All other studies have a lower probability of getting published (Bakker et al., 2012; Fanelli, 2012; Sterling, 1959; Sterling et al., 1995). This also includes 2.5% of all studies that produce significant effects in the unexpected direction (i.e., studies that fall outside the funnel and to the left).

The yellow-to-red colored dots on the right within the funnel represent studies that are “in danger of being *p*-hacked.” These studies revealed nonsignificant

effects in the expected direction that researchers might be able to push below the significance level through *p*-hacking. Panel B depicts the same funnel plot as Panel A with the exception that about 50% of the studies that were “in danger of being *p*-hacked” were hacked to significance with resulting *p*-values that fell between .05 and .001. For these studies, the effect sizes ended up being inflated. Such *p*-hacking may have occurred in a variety of ways (e.g., unplanned inclusion of covariates, flexible outlier treatment). For the present purposes, it is inconsequential which specific *p*-hacks were used. They all serve the same purpose: to reduce an originally nonsignificant *p*-value to significance. (Again, this does not necessarily imply intentionally inappropriate behavior but may occur when a researcher runs multiple analyses while searching for a coherent story that the data may tell and without a clear awareness of the extent to which this approach can increase the false positive rate.)

The two funnel plots in Figure 1 reveal three important insights: First, imagine that researchers only had access to studies that fell outside the funnel to the right. All other studies would be lost to the file drawer. Summarizing this subset of studies in a meta-analysis (i.e., the only evidence available: only significant studies in the expected direction) would lead to a vastly exaggerated meta-analytic effect size estimate. This is the consequence of publication bias. Second, in this case, when only significant studies in the expected direction are available to summarize, it seems that it would not matter much whether a large number of these significant studies were *p*-hacked to significance (Panel B, black dots plus colored dots) or not (Panel A, only black dots): Both subsets of studies that fell outside the funnel to the right would cluster relatively closely together and therefore yield similar summary effect sizes. In other words, *p*-hacking would seriously increase the rate of false positive studies in Panel B (i.e., all colored dots outside the funnel in Panel B are false positives). However, despite the larger number of false positives, the estimated summary effect would not increase to a notable extent. Third, imagine there was no publication bias, and researchers had access to *all* 1,000 studies in Figure 1. It would be obvious that this literature would reveal no effect. Again, it would not matter much whether *p*-hacking was absent (Panel A) or present (Panel B). In this case, there would simply be too many black-dot studies both inside and outside the funnel for the hacked colored-dot studies to make an appreciable impact on the meta-analytic effect size estimate.

In summary, the visual inspection of Figure 1 suggests that, perhaps surprisingly, *p*-hacking might not matter much for the estimation of meta-analytic effect sizes when the publication bias is close to 0% or close to 100%. Yet, what happens between these extremes is less clear. As the ratio of significant to nonsignificant studies changes, *p*-hacking may contribute additional bias.

## The Present Research

We set out to formally examine these provisional observations in a large-scale simulation study. In this study, we generated sets of simulated studies and systematically varied the degrees of both *p*-hacking and publication bias. More specifically, we varied the probability with which a study in danger of being *p*-hacked would actually be *p*-hacked to significance. This reflects the pervasiveness with which researchers in a field (intentionally or unintentionally) *p*-hack an originally nonsignificant *p*-value to significance if this is in principle possible. We also varied the degree of publication bias by moving different proportions of nonsignificant studies to the file drawer so that they would be unavailable for researchers interested in meta-analyzing the respective (simulated) literature. Finally, we conducted random-effects meta-analyses across the remaining (*p*-hacked and non-*p*-hacked) studies and calculated the meta-analytic effect sizes. In a real research literature, these meta-analytic effect sizes would be used to approximate the true sizes of the effects of interest in the population.

## Factors of influence

Of course, meta-analytic effect size estimates in the actual literature are influenced by many more factors than *p*-hacking and publication bias. To generalize the findings from varying levels of *p*-hacking and publication bias, we also systematically varied several such factors of influence that may be present in the actual literature:

- (1) Danger zone: How many studies are “in danger of being *p*-hacked”? Some researchers may believe that it is only possible to *p*-hack relatively small original *p*-values to significance (e.g.,  $p = .200$ ). Everything else may be unfeasible and reminiscent of the intentional fabrication of data. However, other researchers may believe that it is possible to *p*-hack even very large original *p*-values to significance (e.g.,  $p = .800$ ), for example, by employing complex combinations of various *p*-hacks (e.g., treatment of outliers, peeking at the data, inclusion of covariates).

The larger the danger zone for original nonhacked *p*-values is, the greater the influence of *p*-hacking on meta-analytic summaries. This is because a larger danger zone encompasses a larger number of studies that can be hacked and that make a more extensive horizontal movement toward significance in the funnel (i.e., they particularly distort the meta-analytic summary effect). The actual size of a danger zone in a given

literature is impossible to know. It is therefore important to examine the influence of the size of the danger zone across a broad range of possible values.

- (2) True effect size: When there is a true effect, nonhacked studies falling outside the funnel to the right will not represent false positives but will instead provide evidence of a real effect. The larger the true effect, the larger this proportion of studies (Simonsohn, Nelson, & Simmons, 2014a). Consequently, larger true effects should decrease the influence of *p*-hacking because the proportion of *p*-hacked studies in the set of all significant studies will be smaller than in a field with smaller true effects. Similarly, larger true effects will mean that publication bias will have less of an effect because, out of all the studies that were conducted, a larger proportion will be significant and will have a high probability of getting published.

We examined a broad range of true effect sizes to allow for a comprehensive understanding of how the true effect size impacts the biases that *p*-hacking and publication bias exert.

- (3) Heterogeneity: In the psychological research literature, there is not one true fixed effect size. Instead, true effects vary: One manipulation of a construct may be more effective than another manipulation; the same manipulation may be more effective in one population of participants than in another population; one dependent variable used to measure a construct of interest may be more sensitive to an experimental manipulation than another dependent variable, and so forth (Borenstein et al., 2009). The funnel plot (and a meta-analysis for that matter) specifies one mean effect across all studies. If heterogeneity is acknowledged (random-effects model), the effect size estimate reflects the mean of the underlying true effects. Thus, heterogeneity increases the variability of studies on the x-axis in the funnel depicted in Figure 1. This may lead to a larger number of genuinely significant studies. A recent analysis of between-study heterogeneity based on more than 700 meta-analyses provided evidence for substantial heterogeneity in Psychology and variability in the levels of heterogeneity across the various research literatures (van Erp, Verhagen, Grasman, & Wagenmakers, 2017). It is thus important to consider a broad range of

values of heterogeneity when examining the impact of *p*-hacking and publication bias.

- (4) Typical sample sizes: The more precise a study, the more accurately it can estimate the true underlying effect. Effect sizes based on smaller samples vary more strongly. Research literatures differ in how the sample sizes of individual studies are distributed: Some literatures typically feature larger, more precise studies than others (Marszalek, Barber, Kohlhart, & Cooper, 2011). This may influence the impact of *p*-hacking and publication bias, for example, because smaller studies require larger effect sizes to achieve statistical significance. Thus, it is important to consider the influence of various typical sample sizes when trying to understand the impact of *p*-hacking and publication bias on meta-analytic effect size estimates.

- (5) The probability that significant studies will be published: Studies with “positive” results (i.e., significant results in the expected direction) are more likely to be published than studies with “negative” (i.e., nonsignificant) results (Bakker et al., 2012; Fanelli, 2012; Sterling, 1959). However, not all studies that “worked” will be published. For example, authors may be reluctant to submit a study for publication if they feel the study did not provide strong enough evidence in support of the favored hypotheses (Giner-Sorolla, 2012). Also, reviewers and editors may be reluctant to advocate the publication of studies that might not extend previous knowledge far enough to warrant publication (Nosek, Spies, & Motyl, 2012). Thus, there may be variability in a significant study’s probability of getting published.

The lower the probability that significant studies will be published, the smaller the impact of *p*-hacking because, with a lower probability of publication, fewer of the *p*-hacked studies that could bias the meta-analytic estimate will be published. Also, the lower the probability of publication, the smaller the impact of publication bias because the distortion introduced by the nonpublication of nonsignificant results is offset to the extent that significant findings are also not published. (Everything else being equal, there would be no bias in the mean effect size estimate if the same proportions of significant and

nonsignificant findings were not published.)

The interplay of  $p$ -hacking, publication bias, and all factors of influence can conveniently be graphically examined by means of two freely accessible interactive online applications that will be discussed in the Method section. Although we focused on the meta-analysis of two-group comparison designs using Cohen's  $d$  as the measure of effect size (Cohen, 1988), we believe that the insights gained by our study can be readily applied to various kinds of meta-analyses using different effect sizes.

## Method

We simulated the effects of varying degrees of  $p$ -hacking and publication bias on the distortion of meta-analytic effect size estimates as a function of the five factors of influence identified in the Introduction: danger zone, true effect size, heterogeneity, typical sample sizes, and the probability that significant studies will be published. Essentially, this process involved simulating many different versions of the sets of studies depicted in the funnel plots in Figure 1, henceforth referred to as configurations. Figure 1 displays two of the many possible configurations. Simulating each configuration was a multistep process. In the first major step, studies were generated with varying levels of true effect sizes and heterogeneity and were based on different sample size distributions. In the second major step, varying levels of  $p$ -hacking and publication bias were introduced to the studies generated in the first step. In the third step, the meta-analytic summary effects of the configurations (a precision-weighted average of all studies in a configuration) were graphically depicted in outcome figures that illustrate how the five factors of influence changed the interplay between  $p$ -hacking and publication bias in distorting the cumulative evidence base in our simulations. All simulations were conducted using R (R Core Team, 2017). The meta-analytic models were fit using the *rmeta* package (Lumley, 2012).

In total, we simulated 282,240 different configurations. For reasons of clarity, we cannot report the results of all levels of the factors of influence (and their various combinations). However, we offer two interactive online applications that provide visual representations of the effects of the simulations: Interactive online application 1 (<https://bit.ly/2LIvRX7>) visually represents how the factors of influence impact the funnel plot depicted in Figure 1. Interactive online application 2 (<https://bit.ly/2Vno8gH>) visually represents effects of the factors of influence on the graphical displays of the results akin to Figure 2. Both applications offer the opportunity to examine the results as a function of additional values of the factors of influence not reported in the manuscript (e.g., additional values of

true effect sizes, danger zone, severity of  $p$ -hacking, heterogeneity).

## Study Generation

The first step was to simulate individual studies in which two independent groups were compared. The true between-group mean difference per study was set to the sum of a fixed effect  $\delta$  and a random effect  $\tau_i$ , where values for  $\tau_i$  were randomly drawn from a normal distribution with mean 0 and standard deviation  $\tau$ . Fixed-effects models are based on the assumption that there is one true effect size underlying all studies included in a meta-analysis. By contrast, random-effects models are based on the assumption that true effect sizes may differ across studies due to, for example, different effects in different populations or different experimental manipulations. In our simulation we simulated random-effects (by introducing heterogeneity) and accordingly used random-effects meta-analysis for modeling. We assume that for most of the psychological literature, a random-effects model is more plausible than a fixed-effects model (Borenstein et al., 2009).

The  $\tau$  and  $\delta$  values were varied across configurations. We entered 0, 0.2, or 0.5 for the  $\delta$  values (true effect size, Factor 2 listed above) and 0.10, 0.2, or 0.32 for the  $\tau$  values (heterogeneity, Factor 3). Selected values for  $\delta$  were based on Cohen's conventions for small and medium effects (Cohen, 1988). We assume that these cover the majority of effects in Psychology (e.g., Bosco, Aguinis, Singh, Field, & Pierce, 2015; Gignac & Szodorai, 2016; Richard, Bond, & Stokes-Zoota, 2003). According to a recent meta-analysis, our chosen values for  $\tau$  represent the 25%, 50%, and 75% quantiles in an empirical distribution of  $\tau$  estimates in Psychology (van Erp et al., 2017). For additional values, see interactive online application 2.

Samples sizes of individual studies were set to  $n_i = m_j + \chi_i^2 * k$ , where  $\chi_i^2$  was randomly drawn from a  $\chi^2$  distribution with three degrees of freedom,  $k$  was set to 8, and  $m_j$  was varied across configurations. Values more than 80 points above  $m_j$  were truncated (about 1.9% of the distribution). The default value for  $m_j$  was 20. The resulting distribution was right-skewed, with skewness = 1.03,  $Mdn = 39$ ,  $M = 42.61$ ,  $SD = 16.80$ ,  $Min = 20$ ,  $Max = 100$ . Hence, the distribution included both small and large sample sizes, but small sample sizes were more prevalent. In addition to the default of  $m_j = 20$ , we also realized configurations with  $m_j = 10$  and  $m_j = 50$  (typical sample sizes, factor of influence 4). Thus, we utilized sample size distributions with  $Mdn = 29, 39$ , and  $79$  per condition. This approach enabled us to shift the central tendency of the

distribution without changing its shape.<sup>1</sup> See Figure S1 in the supplemental materials for a graphical depiction of the three sample-size distributions. We preferred a synthetic sample-size distribution over an empirically derived distribution for two reasons. First, investigations of historical sample sizes in Psychology (e.g., Fraley & Vazire, 2014; Marszalek et al., 2011) typically do not report the study design, rendering it impossible to infer the typical sample size per condition. Second, sample sizes in Psychology are changing rapidly (Nelson et al., 2018; Sassenberg & Ditrich, 2019), and we aimed to make projections for the present (and future) rather than the past.

For each configuration, we started with a set of 1,000 simulated studies and computed the standardized mean difference ( $d$ , Cohen, 1988), the standard error of the standardized mean difference ( $SE_d$ ), and the  $p$ -value for each study.<sup>2</sup>

## Introduction of Biases

Next,  $p$ -hacking and publication bias were applied to the set of studies.

### *p*-hacking

Studies were defined as “in danger of being  $p$ -hacked” if  $d$  was positive and the  $p$ -value fell above .05 and below a predefined cut-off value. The upper border of this danger zone was linearly increased across the full range of standard errors so that the danger zone was smallest for studies with minimum  $SE_d$  (top of the funnel, Figure 1) and largest for studies with maximum  $SE_d$  (bottom of funnel, Figure 1). This approach resulted in the curved danger zone border visible in Figure 1. The danger zone is smaller for precise studies and larger for imprecise studies. For example, one danger zone we considered was .4/.6, such that studies with the lowest standard errors (i.e., the largest sample size in the set) were in danger if their  $p$ -values fell between .050 and .400, and studies with the highest standard errors (i.e., the smallest sample size in the set) were in danger if their  $p$ -values fell between .050 and .600. This reflects the fact that studies with small sample sizes are easier to  $p$ -hack compared with studies with larger sample sizes (Bakker et al., 2012). Note that the specific largest and smallest  $n$  per cell in a given

configuration depends on which 1,000 values (for 1,000 studies) are randomly selected from the sample size distribution and, of course, on the selected sample size distribution (small, standard, large). We report results for three levels of the danger zone factor: .2/.4, .4/.6, or .6/.8 (factor of influence 1). The effects of additional smaller and larger danger zones can be examined with interactive online application 2.

Any study identified as in danger of being  $p$ -hacked was then hacked with a certain probability. It did not matter which specific  $p$ -hacking technique was used to reduce the  $p$ -value of a study because different  $p$ -hacks serve as means to the same goal, that is, to lower the  $p$ -value of an originally nonsignificant study to significance. If a study was  $p$ -hacked, its  $p$ -value was replaced with a value randomly drawn from a triangular distribution with Max = 0.049, Mode = 0.049, and Min = 0.001, thus satisfying the assumptions that (a)  $p$ -hacking leads to a left-skewed distribution of significant  $p$ -values (more  $p$ -values close to .05 than expected given a true effect; Simonsohn et al., 2014a; Simonsohn, Nelson, & Simmons, 2014b), and (b) some studies nevertheless achieve quite low  $p$ -values after hacking (Simonsohn, Simmons, & Nelson, 2015). (For robustness checks of this procedure, see the Discussion section.) For  $p$ -hacked studies,  $d$  was then recomputed from the new  $p$ -value and the sample size. Average  $p$ -hacking probabilities were set to 0, 0.4 (only depicted in the illustrative example in Figure 2A) and 0.8 ( $p$ -hacking probability). Again, the choice of these values does not indicate that we deem these  $p$ -hacking probabilities particularly likely to reflect reality. Instead, we chose rather extreme values (no  $p$ -hacking at all, 80% of all studies in danger of being  $p$ -hacked were in fact hacked, respectively) and a moderate value to illustrate the potential range of influences that the probability of  $p$ -hacking can exert. (Although possible, we deemed  $p$ -hacking probabilities even higher than 0.8 unlikely to be representative of the psychological research literature.)

To account for the fact that  $p$ -values closer to  $p = .05$  are easier to push over the significance threshold than larger  $p$ -values, we introduced a linear gradient to the probability of  $p$ -hacking: Values that fell exactly on the (horizontal) middle of the danger zone received the nominal average  $p$ -hacking probability (i.e., 80% at the 0.8 level). Values that fell at either (horizontal) end of the danger zone, that is, almost exactly at the

<sup>1</sup> The specific shape of the sample size distribution was arbitrary and based on plausibility assumptions. In additional simulations, we systematically varied  $df$  and  $k$  to ensure that our results were not artifacts of the shape of the selected sample-size distribution. This was the case. Variations in the shape of the distribution affected the results to a negligible extent.

<sup>2</sup> Note that the results for summary effect sizes, the primary outcome of our simulation, are independent of the number of simulated studies. This is the case

because simulated studies in the meta-analysis are independent, such that studies already existing in the set have no impact on new studies being added. Every simulated configuration can be viewed as a data-generating mechanism with a specific underlying distribution of effect sizes. The presented results are estimates for the expected values (means) of the distributions. This expected value is the same, whether we draw  $k = 100$ ,  $k = 1,000$ , or  $k = 10,000$  from the distribution.



significant threshold of  $p = .05$  or almost exactly at the left-hand border of the zone, received *p*-hacking probabilities of 0.2 above and 0.2 below the nominal probability, respectively. The probability was linearly decreased from the left to the right border of the danger zone. In Figure 1, this is visually represented via the yellow-to-red color gradient. For simplicity, we refer to the *p*-hacking probabilities by their nominal (center) level. The gradient did not apply to *p*-hacking probabilities of 0 and 1 (only included in the online applications).

### **Publication bias**

We distinguished between two different kinds of publication bias: First, the nonpublication of studies that “did not work.” Second, the nonpublication of studies that “worked.” Following common conventions, we reserve the term publication bias for the former and call the latter “the probability of publishing significant studies.”

Publication bias was simulated by removing a random subset of studies that were either nonsignificant with positive *d*-values or had negative *d*-values. The severity of publication bias was operationalized by varying the percentage of studies removed in steps of 5% from 0% (no publication bias) to 100% (perfect publication bias).

The probability of the publication of significant studies was modeled analogously. Specifically, significant studies with positive *d*-values were included in the final set with probabilities of 100%, 90%, or 80% (factor of influence 4). For the effects of a smaller probability of the publication of significant studies see interactive online application 2.

### **Meta-analysis and graphical displays of the simulation results**

In total, we simulated 282,240 unique configurations. For each configuration, we fit random-effects meta-analysis models. Because the simulations are probabilistic, running the same configuration twice never yields perfectly identical results. To stabilize the results, we ran all configurations reported in this article 1,000 times and averaged the resulting estimates. All other configurations that are accessible in the interactive online application 2 were run 100 times. This procedure reduced the influence of chance to a negligible amount. Finally, the summary effects were retrieved, some of which are presented in graphical displays of the simulation results.

Figures 2B-2F depict summary effects of 126 configurations each (630 in total): Twenty-one levels of publication bias severity, two levels of *p*-hacking probability, and three levels of one selected factor of influence. All other factors of influence were held constant within each figure. In Figure 2A, all factors of influence were held constant at a set of default values.

## **Results**

In a first step, we will expand on one representative example that illustrates how varying severities of *p*-hacking and publication bias distort the meta-analytic effect size estimate of one hypothetical simulated literature (Figure 2A). In this example, the true effect was set to  $d = 0.2$ , heterogeneity was set to the median value in psychological science,  $\tau = 0.2$  (van Erp et al., 2017), the danger zone was set to .4/.6, and the probability that significant studies would be published was set to 90%. Although the true values from any given area of the literature are unknown, this configuration was intended to reflect one plausible approximation of reality for some research areas. Many additional configurations are reported in the following section, and other ones can conveniently be examined with interactive online application 2.

In a second step, we examined the unique effects of each of the five factors of influence—danger zone, true effect size, heterogeneity, typical sample sizes, and the probability of publishing significant studies—by separately varying the values of one factor while holding the other factors constant (Figures 2B-2F).

### **Step 1: Illustrative Example**

The three lines in Figure 2A reflect the (biased) estimated meta-analytic effect size when *p*-hacking was absent (*p*-hacking probability = 0, solid line), moderate (*p*-hacking probability = 0.4, dashed/dotted line), and when *p*-hacking was severe (*p*-hacking probability = 0.8, i.e., 80% of all studies in the danger zone were *p*-hacked to significance, dashed line).

The simulations revealed several interesting insights: First, in the absence of *p*-hacking (solid line), the effect size bias increased exponentially as the severity of publication bias increased. Even when 50% of all “failed” studies were lost to the file drawer, there was only a relatively small effect size bias (true effect  $d = 0.2$ , estimated effect  $d_{est} = 0.26$ , effect size bias  $d_{bias} = 0.06$ ). This pattern changed dramatically as publication bias became very severe (e.g., 95%). When this happened, the estimated effect was  $d_{est} = 0.50$  ( $d_{bias} = 0.30$ ), indicating that researchers would conclude that this literature represents a robust phenomenon with an average effect size that is more than twice the size of the true effect. An implication of this observation is that even modest reductions in publication bias can greatly reduce the effect size bias in the literature when researchers suspect very severe publication bias.

Second, the effect of moderate *p*-hacking (dashed/dotted line, *p*-hacking probability = 0.4) was only modest compared with no *p*-hacking (*p*-hacking probability = 0). Even assuming that 80% of all studies in the danger zone would be *p*-hacked to significance (dashed line, *p*-hacking probability = 0.8) did not dramatically increase the effect size bias. In the latter case of *p*-hacking probability of 0.8, the additional bias due to *p*-hacking (and its interaction with publication bias) remained below  $d_{hack} = 0.1$  across the various levels of publication bias ( $d_{hack}$  equals the difference between the solid and dashed lines; see Figure S2). At very high levels of publication bias (i.e., 90-100%), the additional bias due to *p*-hacking was negligible and even became negative (i.e., *p*-hacking led to a slight reduction in the overall degree of bias in these cases). Between 0% and approximately 80% of publication bias *p*-hacking had the greatest relative biasing effect (see Figure S2).

A different way to interpret Figure 2A is to examine what it takes to produce a certain effect. For example, let us assume that a meta-analysis revealed an average effect of  $d_{est} = 0.4$ . A researcher who is very experienced in the respective field may believe the true effect is  $d = 0.2$  at best. This researcher is interested in what degrees of *p*-hacking and publication bias may have produced the supposedly biased effect size estimate of  $d_{est} = 0.4$ . In the absence of publication bias (very left part of Figure 2A), *p*-hacking alone would be unable to even come close to a bias of  $d_{hack} = 0.2$ . The actual effect size bias in the absence of publication bias would be only  $d_{hack} = d_{bias} = 0.01$  (i.e.,  $d_{est} = 0.21$ ) if 40% of all studies in danger of being *p*-hacked were *p*-hacked, and only  $d_{hack} = d_{bias} = 0.04$  (i.e.,  $d_{est} = 0.24$ ) even when 80% of all studies in danger of being *p*-hacked were hacked. By contrast, in the absence of *p*-hacking, publication bias alone would be able to seriously bias the estimated effect: It would take a severity of 85% publication bias to double the true effect size to an estimated  $d_{est} = 0.40$ . Illustrating the interaction of *p*-hacking and publication bias, assuming a *p*-hacking probability of 80%, a publication bias of “only” 75% would also lead to a biased effect size of  $d_{est} = 0.40$ . In sum, to introduce serious meta-analytic effect size bias into the literature, publication bias is necessary, but *p*-hacking is not. However, in combination with publication bias *p*-hacking may exert considerable effect size bias.

## Step 2: Factors of Influence

Next, we examined the specific effects of each of the five factors of influence and how variations in these factors could affect the general conclusions drawn from the illustrative example. For ease of interpretation, we only considered the two most extreme values of *p*-hacking probability for the following analyses (0% and 80%).

### Danger zone

The illustrative example was based on a danger zone of  $p = .050$  to  $.400$  (larger, more precise studies) to  $p = .050$  to  $.600$  (smaller, less precise studies). It is possible that *p*-hacking is less or more feasible, and, consequently, fewer or more studies are in danger of being *p*-hacked. Figure 2B depicts the results when the danger zone was set to  $.2/.4$  (blue line),  $.4/.6$  (black line), and  $.6/.8$  (red line). Reducing the danger zone decreased the impact of *p*-hacking slightly; expanding the danger zone increased the impact of *p*-hacking slightly. However, even under the most severe circumstances (danger zone  $.6/.8$ , *p*-hacking probability = 0.8, red dashed line), the additional bias due to *p*-hacking (and its interaction with publication bias) remained smaller than  $\Delta d_{hack} = 0.10$  (difference between solid black and dashed red line, see also Figure S3). Overall, strongly varying the size of the danger zone from  $.2/.4$  to  $.6/.8$  had a modest impact on the effect size bias. Obviously, in the absence of *p*-hacking (solid line), the size of the danger zones did not impact the estimated effect size.

### True effect size

Figure 2C illustrates that in literatures with a true effect size of  $d = 0$ , the exponential relationship between publication bias and the estimated effect size was greatly amplified (Figure 2C, blue lines). At a publication bias of 90% and in the absence of *p*-hacking, the estimated biased effect size was  $d_{est} = d_{bias} = 0.21$ ; at a publication bias of 95%, it was already  $d_{est} = 0.31$ . In addition, publication bias and *p*-hacking interacted more strongly than when the true effect was  $d = 0.2$  as in the default example: The additional biasing effect of *p*-hacking was much stronger at high degrees of publication bias (e.g., 80% or 90%) than at low degrees of publication bias (e.g., 0% or 10%; difference between the solid and dashed blue lines; see also Figure S4).

By contrast, in literatures with a true effect size of  $d = 0.5$ , the exponential relationship between publication bias and the estimated effect size was greatly dampened (red lines). In the absence of *p*-hacking, extreme levels of publication bias (e.g., 90 or 95%) still biased the effect size by approximately  $d_{bias} \approx 0.15$ , but the distortion was much smaller than at lower true effect sizes. At high levels of publication bias (e.g., 90 or 95%), severe *p*-hacking even had a slightly dampening influence on the estimated effect size, effectively leading to *less* biased effect size estimates.

A true effect size of  $d = 0$  may be regarded as deserving special attention because in this case, publication bias and *p*-hacking may “produce” an effect out of thin air. Figure 2C illustrates that the overall level of bias may be particularly large when  $d = 0$ . Therefore, we additionally explored the effects of

all factors of influence specifically for the case of  $d = 0$  (Figure S5). Results revealed that the particularly pronounced interaction between  $p$ -hacking and publication bias at  $d = 0$  remained intact and the factors of influence exerted similar influences as in our illustrative example with  $d = 0.2$ . As before, the bias introduced by  $p$ -hacking alone was small, but the relative amount of bias attributable to  $p$ -hacking and its interaction with publication bias was noticeably increased for  $d = 0$  compared to  $d = 0.2$ .

### ***Heterogeneity***

Lower heterogeneity (i.e.,  $\tau = 0.10$ , 25% quantile) slightly reduced the effect size bias at high levels of publication bias (Figure 2D, blue lines). Higher heterogeneity (i.e.,  $\tau = 0.32$ , 75% quantile) slightly increased the effect size bias at high levels of publication bias (red lines). Both effects were rather modest.

### ***Typical sample size***

Entering smaller ( $Mdn = 29$  participants per condition, Figure 2E, blue lines) rather than standard sample sizes ( $Mdn = 39$  participants per condition, black lines) into the simulation led to slightly stronger biasing effects of  $p$ -hacking and publication bias, especially at high levels of publication bias ( $> 80\%$ ). This was plausible because smaller samples require larger effect sizes to achieve statistical significance. Conversely, larger sample sizes ( $Mdn = 69$  participants per condition, red lines) were associated with smaller effect size biases, especially at high levels of publication bias.

### ***Probability of publishing significant studies***

The assumption that 80% or 100% rather than 90% of all studies that “worked” would be published had only negligible effects on the estimated effect size (Figure 2F).

### ***Supplemental analyses***

We additionally explored the effects of publication bias and  $p$ -hacking on the number of studies in the meta-analysis ( $k$ , see Figure S6) and the precision of the summary effect (standard error, see Figure S7). Besides the switch of outcomes, Figures S6 and S7 are identical to Figure 2. The results for  $k$  are straightforward. The number of studies decreased linearly with increasing levels of publication bias, from 1,000 studies at 0% publication bias to about 200 studies at 100% publication bias in the default configuration. This effect was attenuated by  $p$ -hacking because studies were “saved” from ending up in the file

drawer. At 80%  $p$ -hacking and 100% publication bias, about 500 studies remained, depending on the configuration.

When standard errors were set as outcomes, an interesting pattern emerged. With increasing publication bias and no  $p$ -hacking, standard errors also increased. This is intuitive, because publication bias removes studies from the meta-analysis and smaller meta-analyses are less precise. However, when  $p$ -hacking was added, this effect was effectively canceled out. When  $p$ -hacking was at 80%, increasing publication bias did not decrease precision. Rather, precision remained approximately stable across the range of publication bias (0% - 100%). The cancellation effect occurred because  $p$ -hacking shifts effect sizes to a relatively narrow corridor outside the border of the funnel, thus creating a tightly-packed cluster of effects that results in a high-precision estimate.

When the true effect was zero (Figure S7C, solid, blue line) the impact of publication bias on precision was especially pronounced, because with no true effect, only 2.5 percent of studies reach significance by chance. Even here, adding  $p$ -hacking to the mix increased precision notably. By this way, publication bias and  $p$ -hacking were working in concert to create the illusion of a precise non-zero effect that was in truth zero.

## **Discussion**

Single studies are rarely conclusive. Therefore, scholars rely on meta-analyses that shift the focus from single studies to aggregated evidence: Different researchers contribute to the same research question through replication and extension. This collaborative approach to science holds great promise for the advancement of knowledge, but it is sensitive to distortions. If the cumulative evidence base is severely biased, researchers run the risk of drawing false conclusions. With the present simulations, we examined how a broad range of levels of severity with respect to  $p$ -hacking and publication bias could distort cumulative science as indicated by meta-analytic effect size estimates, considering broad variation in various factors of influence that are likely to exist in realistic research literatures. This study provides several key insights: First,  $p$ -hacking and publication bias interact: A high level of publication bias can greatly distort the available evidence base. At relatively low and high levels of publication bias, even severe  $p$ -hacking contributes little additional bias. At medium levels of publication bias, however,  $p$ -hacking can contribute considerable additional bias, especially when true effects are negligible or even approach zero.

The reason underlying this interaction is simple: When publication bias is low, there are so many studies that contribute to a valid estimation of effect sizes that  $p$ -hacking is not able to seriously distort such a robust

body of evidence. When publication bias is high, effect sizes are already considerably distorted. Hacking some of the few remaining nonsignificant studies to significance only adds to the body of significant studies that already dominate the meta-analytic effect size estimate, but such *p*-hacking does not necessarily change the already biased mean estimate. When publication bias is not particularly low or high, however, considerable *p*-hacking will inflate the effect sizes of the originally nonsignificant studies so much that the overall effect size estimate will be more affected than under other circumstances. This is particularly the case when there is only a negligible or no true effect.

Second, the factors of influence we considered impact the interplay of *p*-hacking and publication bias to varying degrees. No factor of influence fundamentally changes the general interaction pattern, but larger sample sizes, lower heterogeneity, and particularly larger true effects protect against the biasing influence of *p*-hacking and publication bias. We briefly elaborate on why this is the case.

Smaller sample sizes are associated with stronger effect size biases because, in small samples, large effect sizes are needed to achieve statistical significance, leading to greater deviations from the true effect. Everything else being equal, *p*-hacking therefore contributes particularly much effect size bias in studies with small samples. By contrast, in larger samples, smaller effect sizes can achieve statistical significance, which is usually the critical threshold for publication.

When heterogeneity is high, effect sizes are more widely dispersed around the mean effect. This implies that more results will achieve statistical significance (see interactive online application 1 for a graphical illustration). Of these, the ones that are significant in the expected directions will not be affected by publication bias. Instead, they will likely be published as evidence in support of the expected phenomenon. This leads to a greater effect size bias. Put differently, low heterogeneity offers partial protection against extreme effect size biases.

By far, the strongest protective role against effect size bias emerged for larger true effects. Both the maximum level of bias introduced by publication bias alone and the maximum additional bias from *p*-hacking and its interaction with publication bias were strongly reduced as the true effects increased. The reason for this is simple: The stronger the true effect, the greater the number of studies that will achieve (high levels of) statistical significance even without any *p*-hacking (Simonsohn et al., 2014b). In this case, *p*-hacked studies contribute to this pool of significant studies without seriously changing the effect size (see interactive online applications). The large numbers of significant studies with substantial true effects offer another beneficial consequence: Even when nonsignificant studies are sent to the file drawer, this will not have as severe an impact on the overall effect

size estimate because the large number of significant studies attests to the real effect size.

These observations about large true effects have a flip side: Maximum bias is greatest when there is no true effect at all ( $d = 0$ ). In this case, publication bias has the most leverage to distort meta-analytic effect size estimates, and *p*-hacking can add considerable additional bias when publication bias is high (e.g., higher than 50%). In other words, *p*-hacking and publication bias may produce seemingly robust meta-analytic effects out of thin air when in reality the effect is zero.

Taken together, the observations relating to variations in true effect sizes lead to a central and reassuring insight offered by the present simulations: Researchers working in areas involving phenomena with healthy, robust true effects need to worry much less about *p*-hacking and publication bias compared with those investigating small effects. In the case of strong effects, meta-analytic effect size bias will remain modest even under unfortunate conditions. Unfortunately, in many cases, researchers will not know with certainty whether or not an effect of interest has a robust or a negligible true effect before conducting a meta-analysis.

In supplemental analyses, we investigated how *p*-hacking and publication bias affect the precision of a summary effect. This analysis revealed that *p*-hacking can make meta-analytic estimates appear more precise than they actually are, because effect sizes are shifted into a narrow corridor just outside the significance border.

## The Role of *p*-Hacking

Some researchers have intensively argued that *p*-hacking—not publication bias—is *the* major threat to psychological science (e.g., Nelson et al., 2018; Simmons et al., 2011). Why do our simulations paint a more nuanced picture? The major difference between previous work on *p*-hacking and the present approach is that previous work focused on how dramatically *p*-hacking increased the rate of *false positives*, whereas, with the latter, we focused on the *distortion of meta-analytic effect sizes*. We argue that considering distortions of meta-analytic effect size estimates is important because not only false positives, but also substantial distortions in such estimates can impede scientific progress, lead researchers and practitioners astray, and result in a waste of resources if research is based on invalid inferences.

Is this meant to imply that researchers should not worry about *p*-hacking and false positives? Not at all. False positives are a major concern and can exert detrimental influences: For example, in small, emerging literatures, a few prominently published false positives may prematurely lead to the impression of a robust phenomenon. Likely, the more abundant a research literature, the less influence will a handful of

prominent studies exert on the researchers in a field (provided there is no extreme publication bias). However, if—for whatever reason—a literature never develops to a substantial size, the detrimental influence of a few prominent false positive findings may remain strong for a long time.

One reason why a literature might never develop to a substantial size is when the true effect is zero, and there is no *p*-hacking in that particular literature. In this case, researchers would not be likely to produce a large number of studies because it would quickly become clear that there is nothing to find. This would consequently affect the number of studies included in any meta-analysis of this literature or would even determine whether a meta-analysis gets conducted in the first place. If, however, there is *p*-hacking in a literature with a true null or a negligible effect then *p*-hacking may make this literature appear as if there is something there. This may encourage other researchers to conduct further studies on this effect – a lamentable waste of resources. Preventing *p*-hacking would be an important prophylactic against the risk of contributing further confidence in a potentially non-existing effect and ultimately against conducting meta-analyses of many false-positive studies, because the literature would never develop to a substantial size if researchers gave up on effects that would rather consistently spread around null. Precluding *p*-hacking is therefore also important to avoid a literature to accumulate when there is no interesting effect.

We emphasize that with the present work, we did not aim to question previous assertions about *p*-hacking and its deleterious effect on the rate of false positives. In fact, our work corroborates them: In our illustrative example (Figure 2A), *p*-hacking added a large number of false positives: About 55% (!) of all significant studies were false positives due to *p*-hacking. Nevertheless, in the absence of publication bias, *p*-hacking biased the effect size estimate only by  $d_{\text{bias}} = 0.04$ . This demonstrates that the “rate of false positives” and “effect size bias” are two different ways to examine the consequences of *p*-hacking and publication bias. Both are important, but a high rate of false positives does not necessarily imply a strong bias in cumulative effect sizes.

## Implications

The importance of preventing *p*-hacking has rightfully received considerable attention in recent years (Nelson et al., 2018), but the prevention of publication bias has not received as much attention. This needs to change. Both need to become top priorities in psychological research. Both an increase in false positives (the most tangible consequence of *p*-hacking) and distortions of meta-analytic effect sizes (the most tangible consequence of publication bias) have potentially deleterious consequences for psychological science.

Estimates of the prevalence of *p*-hacking vary widely (Fiedler & Schwarz, 2016; Hartgerink, 2017; Head et al., 2015; John et al., 2012; Nelson et al., 2018). The severity of publication bias likely varies considerably by research area, but evidence suggests it may be high in general in Psychology (and higher than in many other sciences; Bakker et al., 2012; Fanelli, 2010; Fanelli, 2012; Fanelli et al., 2017; Ferguson & Brannick, 2012). Therefore, the field may want to take a conservative position, assume drastic severities, and consider what should be done to curtail the consequences. The good news is: Even moderate reductions of a potentially strong publication bias will greatly reduce its biasing effects (see Figure 2). How can this be accomplished? We suggest four easy and cost-efficient solutions.

First, funding institutions have a strong interest in their money being used in the most efficient way to foster scientific progress. Therefore, they may enforce the transparent reporting of data and the results of all studies that have been paid for by a grant. In the case that not all studies end up being published, the remaining studies could mandatorily be placed in repositories such as the Open Science Framework and made public after a certain period of time. Funding institutions could even go so far as to make the allocation of future grants partly contingent on such transparent reporting for studies paid for by past grants. Second, more journals should require authors to clarify whether they have conducted any additional studies that addressed the same research question. These should be reported, and the results and data should be made available (e.g., in a supplement, via a link to a data repository). Journals could introduce standardized tables in which authors report all studies that have been conducted and identify them as pilot studies, actual tests of the hypotheses, and so forth. Even if not all authors responded truthfully to these requirements, this measure alone would unearth a considerable share of otherwise hidden studies and thereby considerably improve the robustness and validity of meta-analytic findings. Third, registered reports lessen both *p*-hacking and publication bias (Chambers, Dienes, McIntosh, Rotshtein, & Willmes, 2015; Jonas & Cesario, 2016). To the extent that more journals give this format ample journal space, publication bias will be reduced. Fourth, not only should preregistrations of studies (Nosek, Ebersole, DeHaven, & Mellor, 2018; van 't Veer & Giner-Sorolla, 2016) incentivize and regularly entail information about the study design, outcome variables, and analysis plans, but they should also encourage or even mandate to make the data public, even if the study remains unpublished. In this way, not only will preregistrations reduce *p*-hacking, but they will also make research discoverable for meta-analysts, even if the data are never published in a journal. More generally, preregistration and adherence to open science standards that emphasize the unbiased access to materials, data and code will improve the replicability not only of individual studies but also of

meta-analyses in the long run (Gurevitch et al., 2018; Nosek et al., 2018). In sum, the field needs a shift in culture. Researchers need to be aware not only of the detrimental consequences of *p*-hacking, but also publication bias, and they also need to be incentivized as well as required to minimize it. Minimizing publication bias comes with an extra premium: If all well-conducted studies are published independent of the outcome, incentives to *p*-hack will be drastically reduced. Preventing publication bias will indirectly reduce *p*-hacking as well.

The present simulations reveal how strikingly a strong severity of publication bias may distort meta-analytic effect size estimates even in the (albeit unrealistic) absence of *p*-hacking. One implication of these findings is that the field needs techniques that validly and reliably correct for the effects of publication bias under realistic circumstances (e.g., varying true effects, heterogeneity, scarcity of nonsignificant studies).

Various techniques have been proposed to correct for publication bias (e.g., trim and fill, Duval & Tweedie, 2000; *p*-curve, Simonsohn et al., 2014; *p*-uniform, van Assen, van Aert, & Wicherts, 2015; PET-PEESE, Stanley & Doucouliagos, 2014; selection models, Iyengar & Greenhouse, 1988). Two recent large-scale simulation studies compared the performance of several methods in correcting for publication bias under conditions that were realistic in various psychological research literatures (Carter, Schönbrodt, Gervais, & Hilgard, 2019; Renkewitz & Keiner, 2018). Using slightly different approaches, both studies concluded that no single method consistently performed well under diverse circumstances and no method consistently outperformed the others (but see van Aert & van Assen, 2018, for an advancement of the *p*-uniform method). Ideally, future versions of these techniques will also be suited for modern meta-analytic methods, such as robust variance estimation (Hedges, Tipton, & Johnson, 2010) or multilevel meta-analysis (Van den Noortgate & Onghena, 2003), which can account for effect size dependencies. These meta-analysis techniques are increasingly used with data sets that include more than one effect size from the same study, but as yet researchers using these methods have to resort to bias correction methods developed for traditional meta-analysis that only allows the inclusion of one effect size per study (e.g., Coles, Larsen, & Lench, 2019; Friese, Frankenbach, Job, & Loschelder, 2017).

For researchers who rely on meta-analyses, our findings provide a starting point from which to estimate the combinations of *p*-hacking, publication bias severity, and factors of influence that could produce a given meta-analytic effect size estimate. An expert in a given literature may be able to make informed guesses about realistic values of the factors of influence. Using interactive online application 2, a researcher may run sensitivity analyses of various values of the factors of

influence to check which combinations of *p*-hacking and publication bias may realistically produce a given meta-analytic estimate.

## Potential Objections to the Simulations

In this section, we address some arguments against some assumptions underlying the present study and the conclusions that can be drawn from it.

### *The severity of p-hacking was overestimated*

Critics may argue that it is unrealistic to *p*-hack initial *p*-values of .8 to significance without deliberately manipulating the data and that assuming a prevalence rate of .8 for *p*-hacking—80% of all studies in danger are in fact hacked—is far too pessimistic. In this case, the impact of *p*-hacking in our simulations would be overstated. In response, we refer to Figure 2, which illustrates that we simulated the impact of *p*-hacking across a broad range of severities in terms of both the probability of *p*-hacking and the size of the danger zone. Our simulations are thus informative with respect to a broad range of potential realities in different research areas.

### *The severity of p-hacking was underestimated*

One potential objection is that *p*-hacking was not modeled severely enough because researchers will basically do everything to avoid losing a study (and the resources they invested in it) to the file drawer (Nelson et al., 2018). Had *p*-hacking been modeled severely enough, the effects on meta-analytic effect size estimates would have been much stronger. The “severity of *p*-hacking” may refer to different aspects, either in isolation or together: (a) How many studies of those in danger of being *p*-hacked will in fact be *p*-hacked? (b) How large is the “danger zone” of studies that can be *p*-hacked? (c) If researchers *p*-hack, how far below the .05 threshold do they *p*-hack their studies? The smaller the final *p*-value, the more strongly the effect size of this particular study will be biased (given a constant sample size).

In response to the concern that *p*-hacking might not been modeled severely enough, we refer to Figure 2B, which (also) illustrates the effects of very severe *p*-hacking: 80% of studies with original *p*-values of up to .800 will be *p*-hacked to significance. The results (red line) reveal considerable distortions caused by this very severe *p*-hacking, but much less than that caused by severe degrees of publication bias. Interactive online application 2 allows interested researchers to simulate the effects of even more severe *p*-hacking.

A critic may go further and object that researchers will even *p*-hack studies to significance when they originally showed mean differences in the unexpected

direction. We argue that it is unlikely that this happens on a large scale. There will be both statistical and moral boundaries that prevent researchers from going this far. Even if it is technically possible in some cases, such behavior would more adequately be labeled fraud and data fabrication rather than *p*-hacking, and even pessimistic estimates have identified such misdeeds as very rare (John et al., 2012).

Alternatively, might researchers *p*-hack original findings that went in the “wrong” direction so that they become significant in this unexpected direction and later claim this result was expected (Nelson et al., 2018)? On the basis of individual studies, this may be possible. However, in taking the broader perspective of a cumulative literature, we deem this unlikely. Take studies on the contested stereotype threat effect as an example (Flore & Wicherts, 2015; Spencer, Logel, & Davies, 2016; Stoet & Geary, 2012). Any description of a study that was originally planned to examine stereotype threat will be identified as such with a decent probability. Any dependent variable included to indicate a stereotype threat effect will be difficult to defend as consistent with a priori assumptions if the result ends up going in the other direction. Dependent variables are selected for a purpose. We believe that in the long run, it is unlikely that *just any* effect on a dependent variable could be argued to be in line with a priori assumptions about a given phenomenon under investigation. After all, researchers do not only need to tell a coherent story across (potentially) several studies in a single manuscript, but their findings also have to be coherent over time with the authors’ other publications and the respective literature more generally.<sup>3</sup>

Perhaps researchers regularly *p*-hack original findings that went in the “wrong” direction to force them to achieve significance and then come up with a new post hoc explanation that is void of any reference to the literature that was the starting point for the study (e.g., stereotype threat). Again, on the level of individual studies, this could conceivably happen (Kerr, 1998). However, on the level of a cumulative literature, this is more difficult to envision. If this process were to happen on a large scale, this would lead to highly diverse publication lists of researchers in terms of topics. According to this argument, (almost) every study that “did not work” would be used to tell a completely different story to avoid making reference to the initial research question. In reality, most researchers conduct research on a few focused areas rather than a wide variety of different research areas. This is an indirect indication that studies that “did not work” often end up in the file drawer rather than being *p*-hacked to just *any* publishable finding for addressing a randomly and a posteriori fitted research question.

Finally, a critic may be concerned that if researchers *p*-hack, they will bring the *p*-value down to

even smaller levels than modeled here. Remember that in the present studies, *p*-hacked *p*-values lay between .049 and a minimum of .001 (a triangular distribution with the mode at .049). According to this distribution, the *p*-values are three times more likely to fall between .05 and .025 than between .025 and .001. The general notion that hacked *p*-values are more likely to fall just below  $p = .05$  is generally accepted and engrained in common bias detection methods such as a *p*-curve analysis (Simonsohn et al., 2014a, 2014b). To examine the consequences of this particular distribution of *p*-hacked *p*-values, we ran sensitivity analyses. First, we re-ran the simulations with the same triangular distribution, but we set the lower end of the *p*-values to .000001 instead of .001. The results were nearly identical. Second, we changed the distribution of *p*-hacked *p*-values from a triangular distribution to two uniform distributions ranging from .05 to .025 and from .025 to .000001, and we made the first uniform distribution twice as prevalent as the second. Even under these very conservative assumptions, the changes in the results were practically negligible and did not affect the conclusions in any way. Taken together, the presented results are remarkably robust even when assuming that *p*-hacked studies are very often *p*-hacked to excessively low *p*-values.

## Limitations

In our simulations, we considered the effects of several different factors of influence that plausibly could have seriously affected the results of the study. This approach resulted in a comprehensive set of findings. However, it is impossible to map every facet of reality. Modeling inevitably requires a reduction in complexity compared with reality. In the following, we discuss some plausible deviations from reality.

One limitation of the present work was that it was based on the assumption that all studies are valid observations of real (null) effects. In reality, low-quality study designs and executions may lead to systematically or unsystematically biased estimates of true effects. The possibility of unsystematic bias was partially captured by heterogeneity as one factor of influence that introduced the possibility of more than one true effect in a given literature, but we did not explicitly model systematic bias (e.g., high-quality studies are more likely to reveal true effects). In a similar vein, some studies that we modeled as independent might in fact be partly dependent in reality, for example, because they were conducted in the same lab or drew participants from the same participant pool. As we explained in the discussion section on *p*-hacking, social factors more generally may contribute to the emergence of literatures with partly dependent studies. For example, if *p*-hacked

<sup>3</sup> Note that we have no stakes in the debate about stereotype threat and do not take any position here.

This is only one of many examples for which this argument could be made.

studies appeared to elucidate an effect that in fact may not exist this may promote further studies on the effect that would likely not have been conducted had there been no  $p$ -hacking in the literature. These observations might not question the insights about the interplay of  $p$ -hacking and publication bias revealed by the present study, but they highlight that some literatures plagued by  $p$ -hacked false positives might never accumulate and be meta-analyzed if  $p$ -hacking could be effectively prevented.

A second limitation is that our results may be contingent on the assumption that the probability that “failed” studies will be published is independent of effect size. This assumption could be challenged. For example, it is reasonable to expect that studies that are significant in the opposite direction from what was predicted may be easier to publish than studies that found no difference whatsoever between conditions (Ioannidis & Trikalinos, 2005). The solutions to this problem mentioned earlier, such as registered reports or editors asking for additional studies, are focused on decoupling the probability of publication from the results, thereby ameliorating this problem.

Third, we assumed that the probability that nonsignificant studies will be published is independent of sample size. This assumption may be challenged on the basis of the argument that large nonsignificant studies may have a higher probability of getting published than small nonsignificant studies because researchers have greater incentives to publish studies in which they invested many (as compared with little) resources.

Finally, we operationalized  $p$ -hacking essentially as lowering an originally non-significant  $p$ -value to significance. However, how exactly researchers may end up with polished results that initially did not look as promising is not fully understood and could also entail processes not captured by our operationalization of  $p$ -hacking. For example, if multiple outcomes are assessed in a study, but only a subset of these is reported this does not fit our operationalization of  $p$ -hacking.

Mild forms of outcome selection, say, some researchers selecting from two or three outcomes, may be covered by the more extreme settings of the “danger zone” influence factor. With these settings,  $p$ -values shift from up to 0.8 to below 0.05. We may conceptualize the initial  $p$ -value as some average of multiple outcomes, while the final  $p$ -value stems only from a subset of outcomes. When outcome switching is more extreme, say, a substantial percentage of researchers consistently selected a small number of outcomes from an originally large number of outcomes, this go beyond the conceptual boundaries of our simulation. In this case, suppression of evidence in a literature would be so widespread that we would be hesitant to place the phenomenon under the conceptual umbrella of  $p$ -hacking. In fact, it would rather be a form of publication bias on the level of outcomes: Some outcomes are lost to the file drawer, others are reported.

In a possible next step, should the selected original  $p$ -values not be significant, they may undergo one or several treatments that let it drop under  $p = .05$  as is assumed in the current study (e.g, by transformation of variables, recoding or exclusion of outliers etc.). Future studies may want to add selective reporting of outcomes as an additional separate step in the simulations that determines which outcomes (and their  $p$ -values) then undergo  $p$ -hacking in the way operationalized here and examine the consequences on meta-analytic effect size biases as a function of the number of assessed and selected outcomes.

## Conclusion

Meta-analysis is an essential tool for scientific progress that is considered more trustworthy and robust against various biasing influences than individual studies (Gurevitch et al., 2018). However, the validity of meta-analyses is threatened by systematic sources of error. The present study highlighted how  $p$ -hacking and publication bias can interact to bias meta-analytic effect size estimates under a large number of circumstances. In recent years, the increase in false-positive findings due to  $p$ -hacking and, in turn, ways to prevent  $p$ -hacking have rightfully received considerable attention. The present results highlight that in order to increase the trustworthiness of psychological science, the reduction of publication bias also needs to become a primary objective. Otherwise researchers would run the risk of drawing vastly incorrect conclusions from bodies of evidence. Such a trend would have significant implications for theories, the robustness of practical interventions, the allocation of resources in both research and practice, and, last but not least, trust in our discipline.

## References

- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7, 543-554. <https://doi.org/10.1177/1745691612459060>
- Bertamini, M., & Munafò, M. R. (2012). Bite-size science and its undesired side effects. *Perspectives on Psychological Science*, 7, 67-71. <https://doi.org/10.1177/1745691611429353>
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: Wiley.
- Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, 100, 431-449. <https://doi.org/10.1037/a0038047>
- Bruns, S. B., & Ioannidis, J. P. A. (2016).  $p$ -curve and  $p$ -hacking in observational research. *Plos One*, 11, e0149144. <https://doi.org/10.1371/journal.pone.0149144>



- Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, 2, 115-144. <https://doi.org/10.1177/2515245919847196>
- Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered Reports: Realigning incentives in scientific publishing. *Cortex*, 66, A1-A2. <https://doi.org/10.1016/j.cortex.2015.03.022>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Vol. 2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Coles, N. A., Larsen, J. T., & Lench, H. C. (2019). A meta-analysis of the facial feedback literature: Effects of facial feedback on emotional experience are small and variable. *Psychological Bulletin*, 145, 610-651. <https://doi.org/10.1037/bul0000194>
- Dubben, H. H., & Beck-Bornholdt, H. P. (2005). Systematic review of publication bias in studies on publication bias. *British Medical Journal*, 331, 433-434. <https://doi.org/10.1136/bmj.38478.497164.F7>
- Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 455-463. <https://doi.org/DOI.10.1111/j.0006-341X.2000.00455.x>
- Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *Plos One*, 5, e10068. <https://doi.org/10.1371/journal.pone.0010068>
- Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90, 891-904. <https://doi.org/10.1007/s11192-011-0494-7>
- Fanelli, D., Costas, R., & Ioannidis, J. P. A. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences of the United States of America*, 114, 3714-3719. <https://doi.org/10.1073/pnas.1618569114>
- Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, 17, 120-128. <https://doi.org/10.1037/a0024445>
- Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, 7, 555-561. <https://doi.org/10.1177/1745691612459059>
- Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, 7, 45-52. <https://doi.org/10.1177/1948550615612150>
- Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology*, 53, 25-44. <https://doi.org/10.1016/j.jsp.2014.10.002>
- Fraley, R. C., & Vazire, S. (2014). The n-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *Plos One*, 9, e109019. <https://doi.org/10.1371/journal.pone.0109019>
- Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345, 1502-1505. <https://doi.org/10.1126/science.1255484>
- Friese, M., Frankenbach, J., Job, V., & Loschelder, D. D. (2017). Does self-control training improve self-control? A meta-analysis. *Perspectives on Psychological Science*, 12, 1077-1099. <https://doi.org/10.1177/1745691617697076>
- Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, 102, 74-78. <https://doi.org/10.1016/j.paid.2016.06.069>
- Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, 7, 562-571. <https://doi.org/10.1177/1745691612457576>
- Gurevitch, J., Koricheva, J., Nakagawa, S., & Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature*, 555, 175-182. <https://doi.org/10.1038/nature25753>
- Hartgerink, C. H. J. (2017). Reanalyzing Head et al. (2015): Investigating the robustness of widespread p-hacking. *PeerJ*, 5, e3068. <https://doi.org/10.7717/peerj.3068>
- Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, 13, e1002106. <https://doi.org/10.1371/journal.pbio.1002106>
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1, 39-65. <https://doi.org/10.1002/jrsm.5>
- Ioannidis, J. P. A., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, 58, 543-549. <https://doi.org/10.1016/j.jclinepi.2004.10.019>
- Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, 3, 109-117
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23, 524-532. <https://doi.org/10.1177/0956797611430953>
- Johnson, B. T., & Eagly, A. H. (2014). Meta-analysis of social-personality psychological research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of*

- research methods in social and personality psychology* (2nd ed., pp. 675-707). London: Cambridge
- Jonas, K. J., & Cesario, J. (2016). How can preregistration contribute to research in our field? *Comprehensive Results in Social Psychology, 1*, 1-7. <https://doi.org/10.1080/23743603.2015.1070611>
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196-217. [https://doi.org/10.1207/s15327957pspr0203\\_4](https://doi.org/10.1207/s15327957pspr0203_4)
- Kuhberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *Plos One, 9*. <https://doi.org/ARTN e105825> 10.1371/journal.pone.0105825
- Lakens, D. (2015). What p-hacking really looks like: A comment on Masicampo and LaLande (2012). *Quarterly Journal of Experimental Psychology, 68*, 829-832. <https://doi.org/10.1080/17470218.2014.982664>
- Lilienfeld, S. O., & Waldman, I. D. (2017). *Psychological science under scrutiny: Recent challenges and proposed solutions*. Chichester, UK: Wiley.
- Lumley, T. (2012). rmeta: Meta-analysis. R package version 2.16 [Computer Software]. Retrieved from <https://CRAN.R-project.org/package=rmeta>
- Marszalek, J. M., Barber, C., Kohlhart, J., & Cooper, B. H. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills, 112*, 331-348. <https://doi.org/10.2466/03.11.pms.112.2.331-348>
- Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *Quarterly Journal of Experimental Psychology, 65*, 2271-2279. <https://doi.org/10.1080/17470218.2012.711335>
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*, 147-163. <https://doi.org/10.1037/1082-989x.9.2.147>
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021. <https://doi.org/10.1038/s41562-016-0021>
- Murad, M. H., & Montori, V. M. (2013). Synthesizing evidence shifting the focus from individual studies to the body of evidence. *Jama-Journal of the American Medical Association, 309*, 2217-2218. <https://doi.org/10.1001/jama.2013.5616>
- Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511-534. <https://doi.org/10.1146/annurev-psych-122216-011836>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America, 115*, 2600-2606. <https://doi.org/10.1073/pnas.1708274114>
- Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*, 137-141. <https://doi.org/10.1027/1864-9335/a000192>
- Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615-631. <https://doi.org/10.1177/1745691612459058>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716. <https://doi.org/10.1126/science.aac4716>
- R Core Team. (2017). R: A language and environment for statistical computing [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Renkewitz, F., & Keiner, M. (2018, December 20). *How to detect publication bias in psychological research? A comparative evaluation of six statistical methods*. <https://doi.org/10.31234/osf.io/w94ep>
- Richard, F. D., Bond, C. F., & Stokes-Zoota. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*, 331-363. <https://doi.org/10.1037/1089-2680.7.4.331>
- Sassenberg, K., & Ditrich, L. (2019). Research in Social Psychology Changed Between 2011 and 2016: Larger Sample Sizes, More Self-Report Measures, and More Online Studies. *Advances in Methods and Practices in Psychological Science, 2*, 107-114. <https://doi.org/10.1177/2515245919838781>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359-1366. <https://doi.org/10.1177/0956797611417632>
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*, 666-681. <https://doi.org/10.1177/1745691614553988>
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve: A key to the file-drawer. *Journal of Experimental Psychology-General, 143*, 534-547. <https://doi.org/10.1037/a0033242>
- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (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, 1146-1152.  
<https://doi.org/10.1037/xge0000104>
- Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology*, 67, 415-437. <https://doi.org/10.1146/annurev-psych-073115-103235>
- Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5, 60-78. <https://doi.org/10.1002/jrsm.1095>
- Sterling, T. 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, 30-34
- Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited - the effect of the outcome of statistical tests on the decision to publish and vice-versa. *American Statistician*, 49, 108-112. <https://doi.org/10.2307/2684823>
- Stoet, G., & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology*, 16, 93-102.  
<https://doi.org/10.1037/a0026617>
- van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, 67, 2-12
- van Aert, R. C. M., & van Assen, M. A. L. M. (2018, October 2). *Correcting for publication bias in a meta-analysis with the p-uniform\* method*.  
<https://doi.org/10.31222/osf.io/zqjr9>
- van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, 20, 293-309.  
<https://doi.org/10.1037/met0000025>
- Van den Noortgate, W., & Onghena, P. (2003). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement*, 63, 765-790.  
<https://doi.org/10.1177/0013164402251027>
- van Erp, S., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E. J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in psychological bulletin from 1990–2013. *Journal of Open Psychology Data*, 5, 4

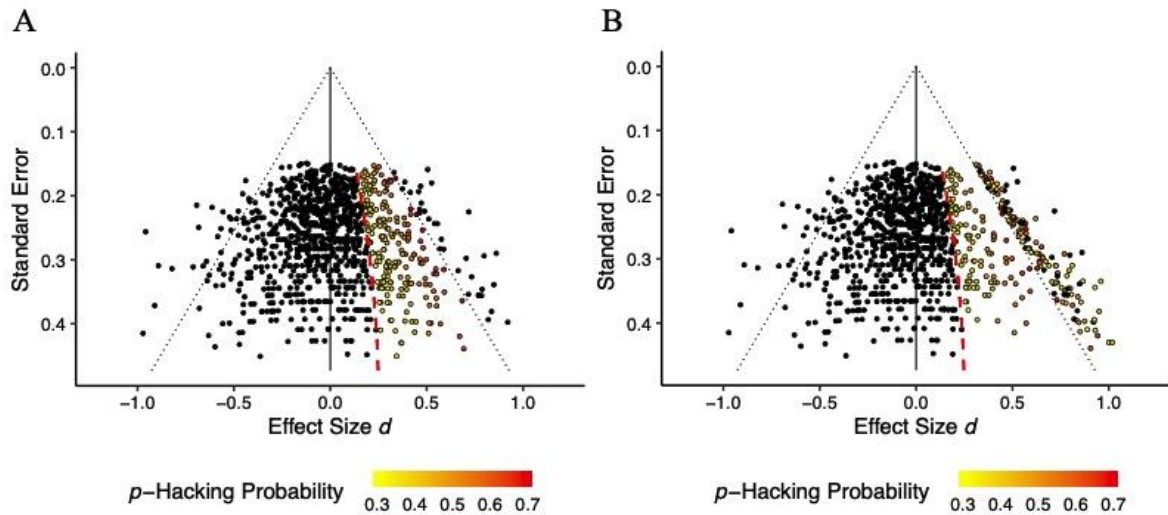
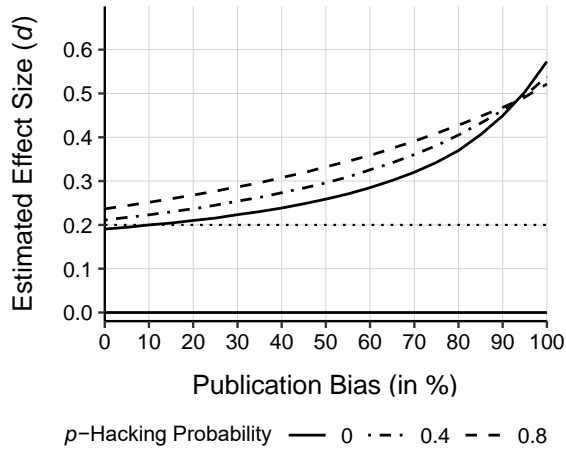
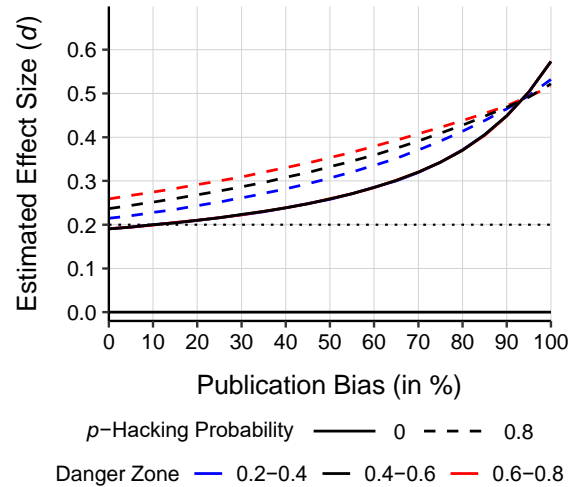


Figure 1. Funnel plots of 1,000 hypothetical studies with a true underlying effect size of zero. The outer dotted lines indicate the triangular region within which 95% of studies are expected to fall in the absence of *p*-hacking, publication bias, and heterogeneity. Effect sizes (Cohen’s *d*) are represented on the *x*-axis. Precision ( $SE_d$ ) is represented on the *y*-axis. The dashed red line indicates the left border of the *p*-hacking danger zone. Studies that fall between the dashed red line and the right border of the funnel are “in danger of being *p*-hacked.” The probability that a study that is in danger of being *p*-hacked is indeed *p*-hacked is indicated by color, such that the yellow colored studies are hacked with a probability of 0.3 and the red colored studies are hacked with a probability of 0.7. In Panel A, the probability of *p*-hacking is depicted for illustrative purposes only. In Panel B, the studies that are in danger have actually been hacked according to their assigned probability. Thus, most of the red studies but only a few of the yellow studies have been hacked.

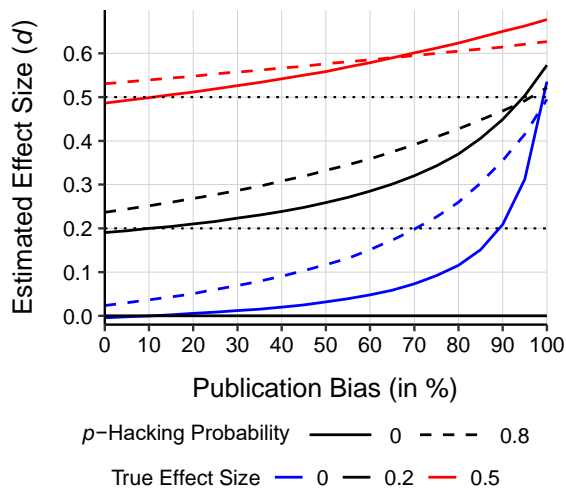
A – Default Settings



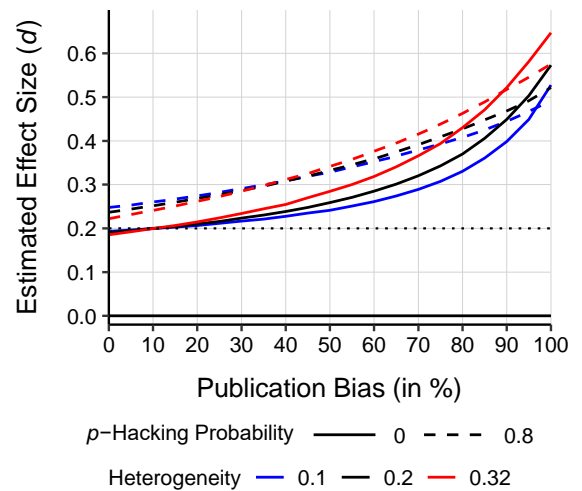
B - Danger Zone



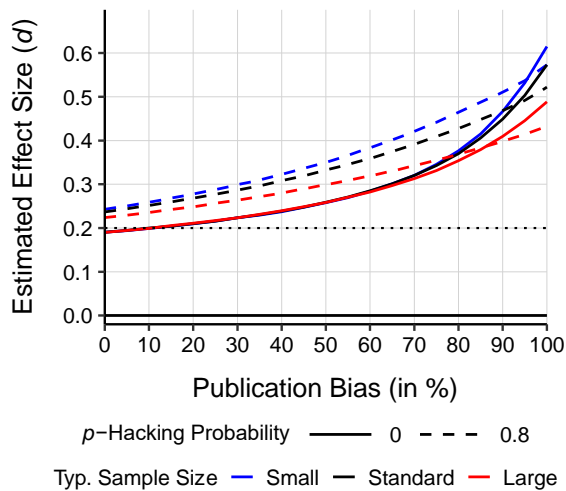
C - True Effect Size



D - Heterogeneity



E – Typical Sample Size



F – Publication Probability

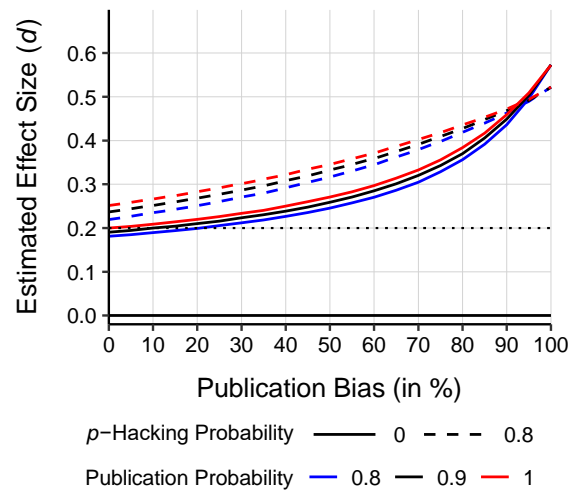


Figure 2. Results of the full simulation. Figure 2A displays the results of the default configuration. The settings for this configuration are  $d = 0.2$ ,  $\tau = 0.2$ , and danger zone =  $.4/.6$ . Typical sample sizes per condition were drawn from the standard distribution. The default probability of the publication of significant studies was  $.9$ . The thin

horizontal black line represents the null effect. The dotted black line represents the true effect. The curved solid line represents the estimated biased effect  $d_{est}$  at a *p*-hacking probability of 0. The curved dashed/dotted line represents the estimated biased effect  $d_{est}$  at a *p*-hacking probability of 0.4. The curved dashed line represents the estimated biased effect  $d_{est}$  at a *p*-hacking probability of 0.8. Figures 2B-2F display results when the five factors of influence were systematically varied. For these figures, all default settings were used and only the respective factor was varied. Different colors indicate different levels of the factors as described in the legend below each figure. Solid lines always indicate a *p*-hacking probability of zero, whereas dashed lines always indicate a *p*-hacking probability of 0.8. Hacking probabilities of 0.4 are not depicted.

Supplemental Materials

*p*-Hacking and Publication Bias Interact to Distort Meta-Analytic Effect Size Estimates

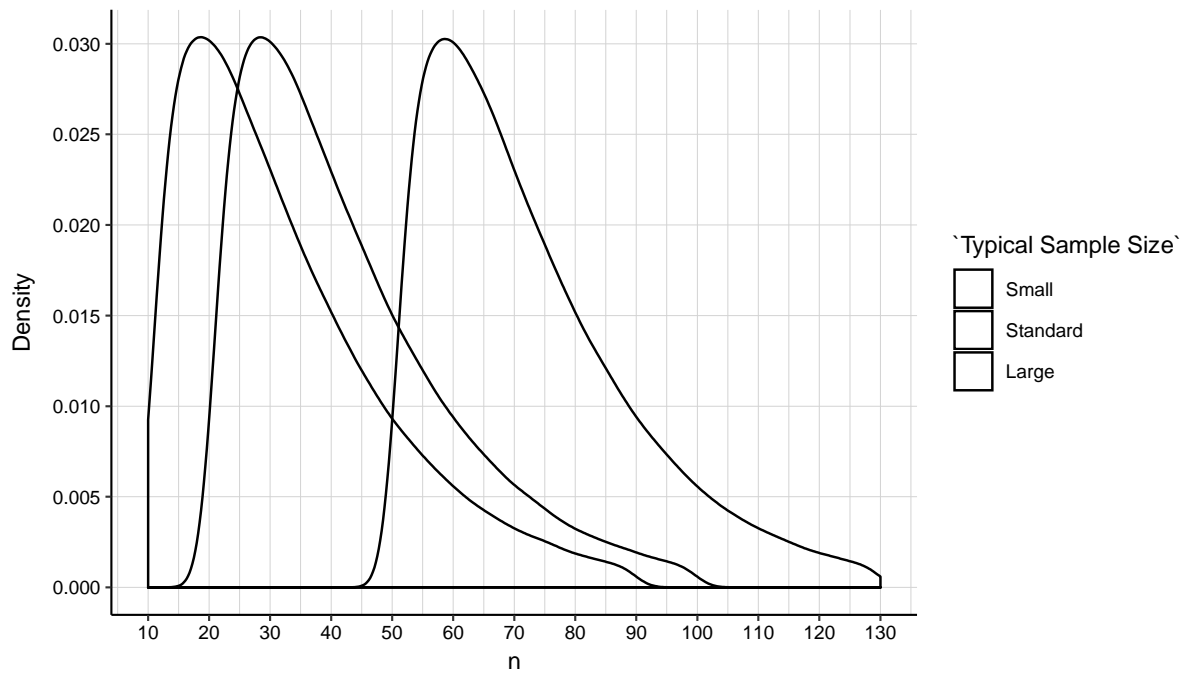


Figure S1. Density estimators of the sample size distributions (factor of influence 4, typical sample sizes). Per cell sample sizes for the simulation of each study were drawn from one of the depicted distributions. Measures of central tendency are  $M = 32.6$ ,  $Mdn = 29$  for Small,  $M = 42.6$ ,  $Mdn = 39$  for Standard, and  $M = 72.6$ ,  $Mdn = 69$  for Large.

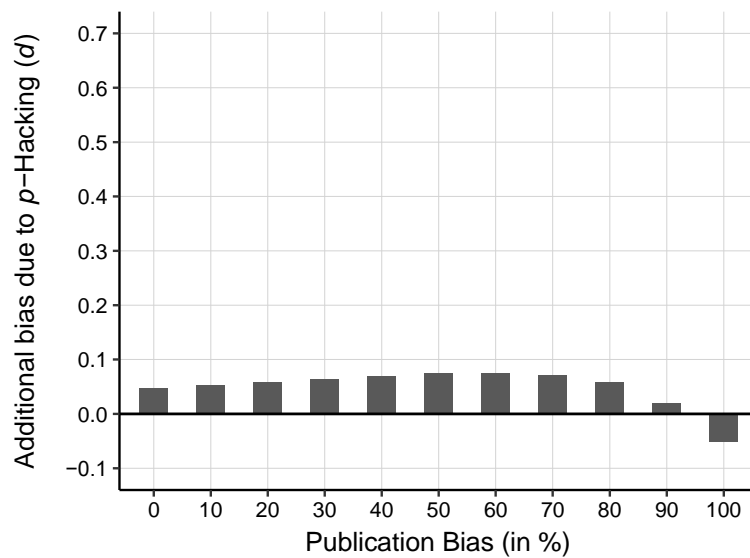


Figure S2. Additional bias due to *p*-hacking and its interaction with publication bias at different levels of publication bias. The *y*-axis depicts the change in the estimated effect size when the average nominal *p*-hacking probability is raised from 0 to 0.8. The *x*-axis depicts levels of publication bias, indicated by the percentage of non-significant or negative studies removed from the simulated set of studies. All influence factors are set to the default values.



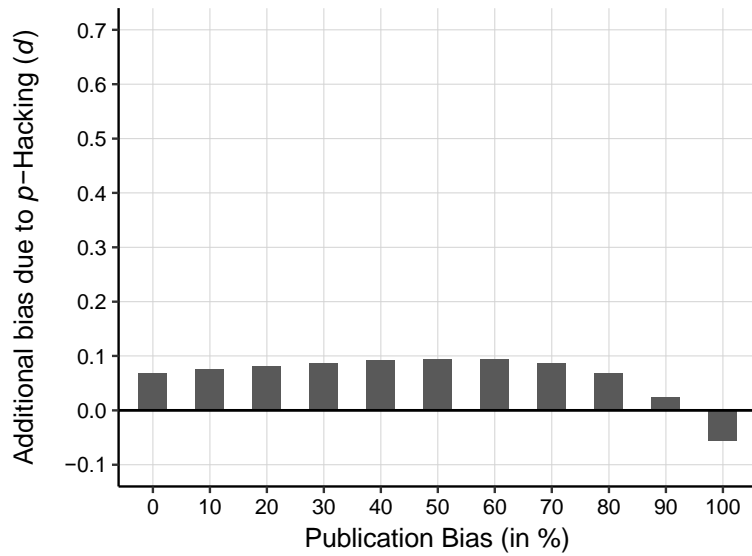


Figure S3. Additional bias due to *p*-hacking and its interaction with publication bias at different levels of publication bias. The *y*-axis depicts the change in the estimated effect size when the average nominal *p*-hacking probability is raised from 0 to 0.8. The *x*-axis depicts levels of publication bias, indicated by the percentage of non-significant or negative studies removed from the simulated set of studies. All influence factors are set to the default values, except that the danger zone is set to .6/.8 instead of .4/.6 (see Figure S1).

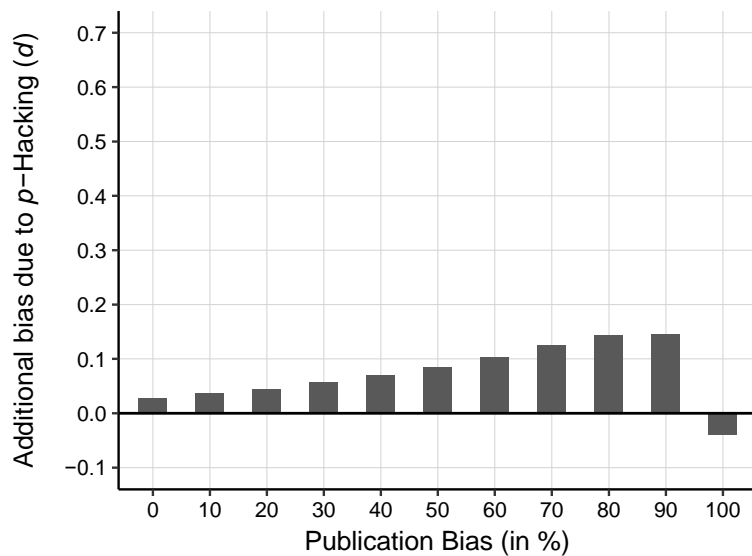
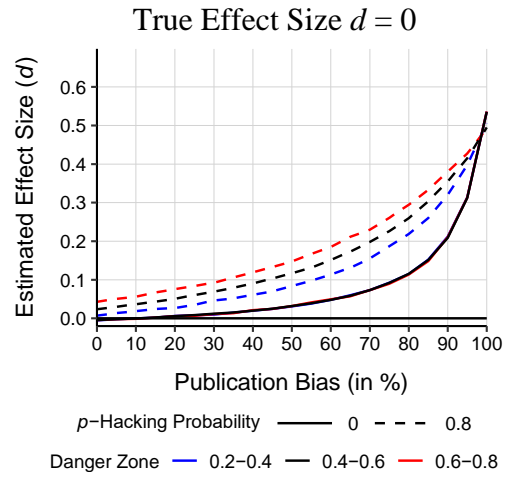
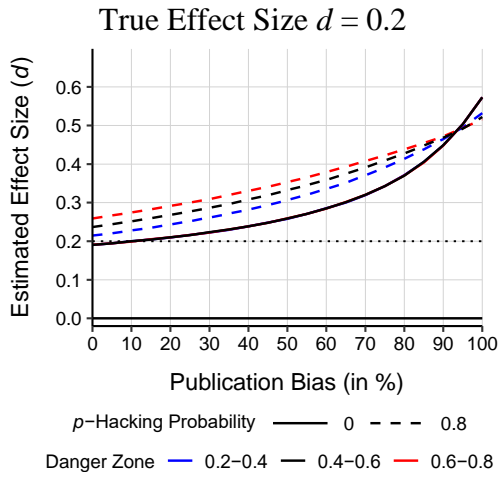
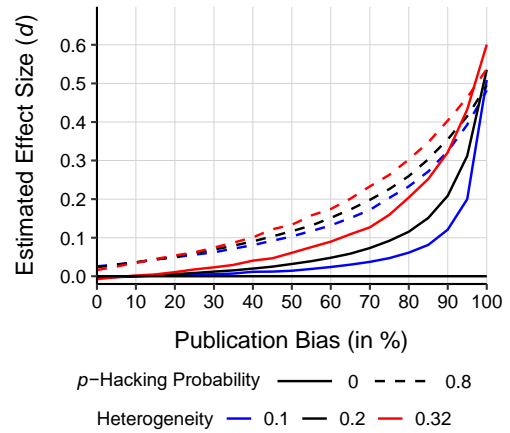
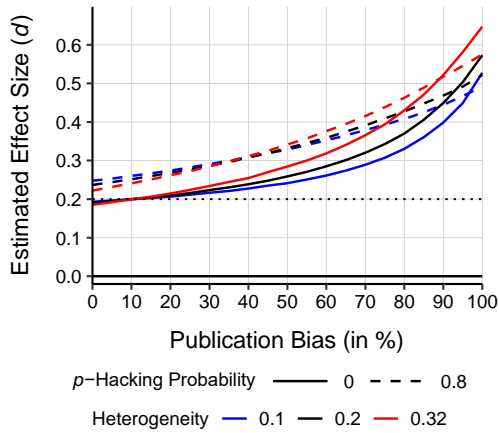


Figure S4. Additional bias due to *p*-hacking and its interaction with publication bias at different levels of publication bias. The *y*-axis depicts the change in the estimated effect size when the average nominal *p*-hacking probability is raised from 0 to 0.8. The *x*-axis depicts levels of publication bias, indicated by the percentage of non-significant or negative studies removed from the simulated set of studies. All influence factors are set to the default values, except that the true effect size is set to  $d = 0$  instead of  $d = 0.20$  (see Figure S1).

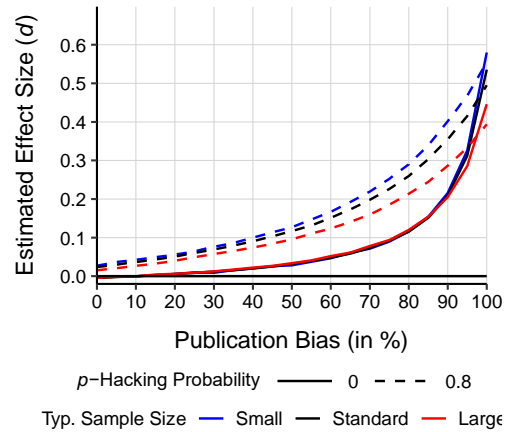
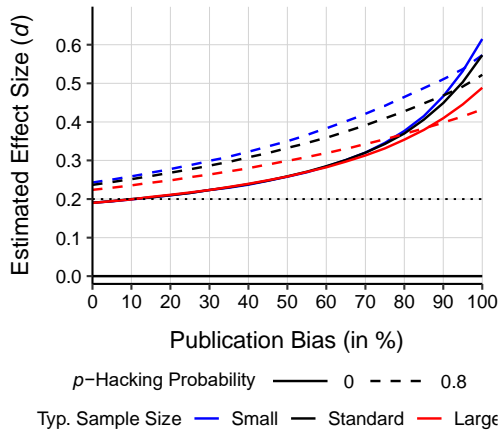
A – Danger Zone



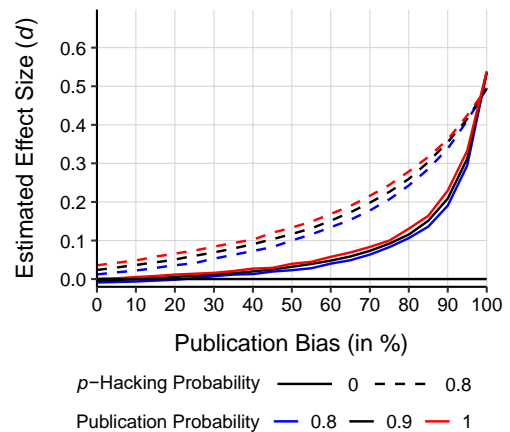
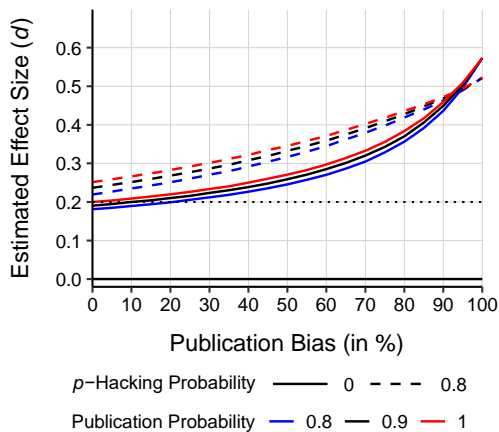
B – Heterogeneity



C – Typical Sample Size



D – Publication Probability



*Figure S5.* Results of the full simulation for default settings with true effect size  $d = 0.2$  (left column, taken from Figure 2) and  $d = 0$  (right column). The thin horizontal black line represents the null effect. The dotted black line represents the true effect (only left column). The curved solid lines represent the estimated biased effect  $d_{est}$  at a *p*-hacking probability of 0. The curved dashed lines represent the estimated biased effect  $d_{est}$  at a *p*-hacking probability of 0.8. Figures S5A-D display results when the remaining four influence factors are systematically varied. For these figures, all default settings were used, the respective factor was varied, and the true effect was either set to  $d = 0.2$  (left column) or  $d = 0$  (right column). Different colors indicate different levels of the factors, as described in the legend below each figure.

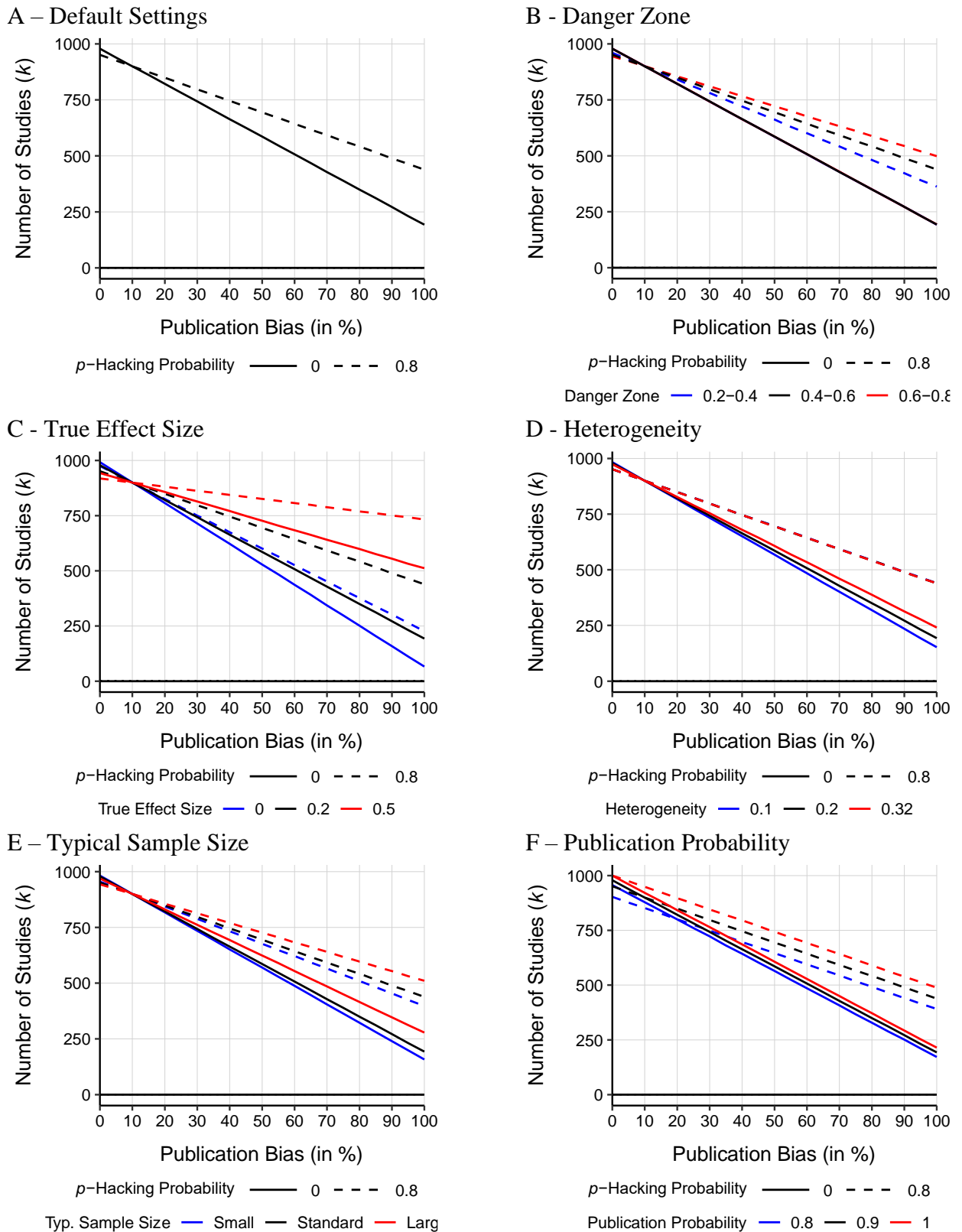
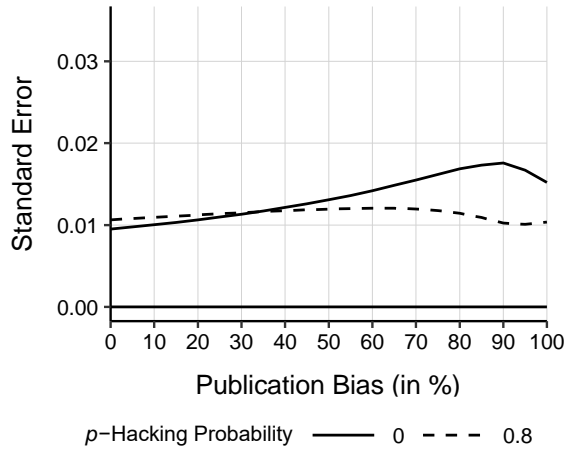


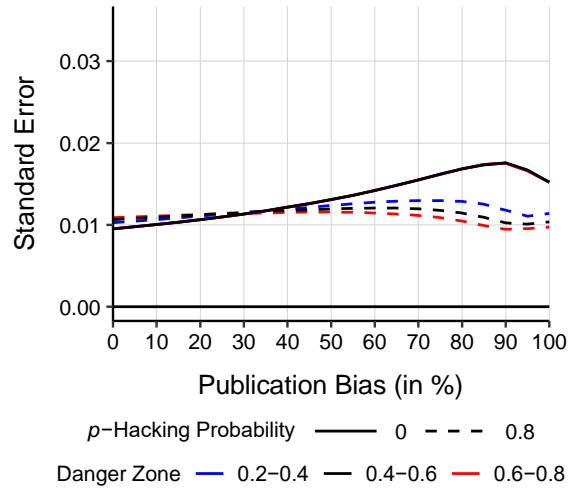
Figure S6. Results of the full simulation for the number of studies as main outcome variable. Figure S6A displays the results of the default configuration. The settings are identical to Figure 2. The thin horizontal black line represents the null effect. The dotted black line represents the true effect. The curved solid line represents the number of studies  $k$  at a  $p$ -hacking probability of 0. The curved dashed line represents the number of studies  $k$  at a  $p$ -hacking probability of 0.8. Figures S6B-S6F display results when the five factors of influence are

systematically varied. For these figures, all default settings were used and only the respective factor was varied. Different colors indicate different levels of the factors, as described in the legend below each figure. Solid lines always indicate a *p*-hacking probability of zero, dashed lines indicate a *p*-hacking probability of 0.8.

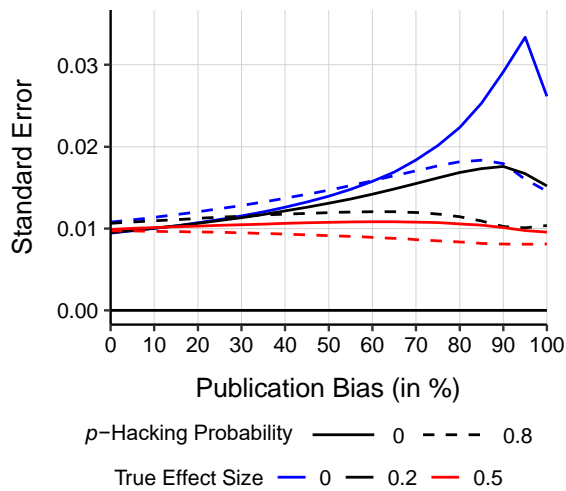
A – Default Settings



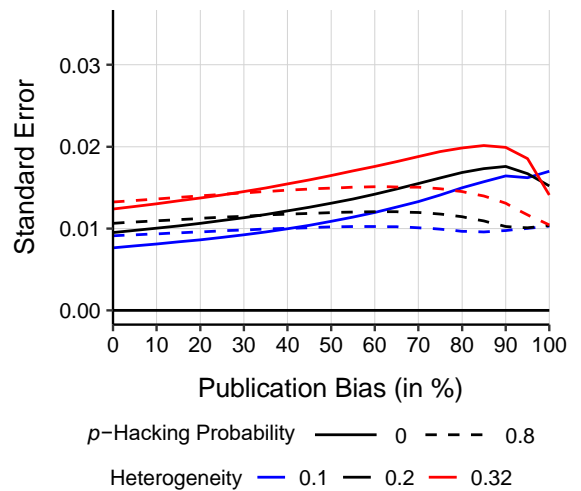
B - Danger Zone



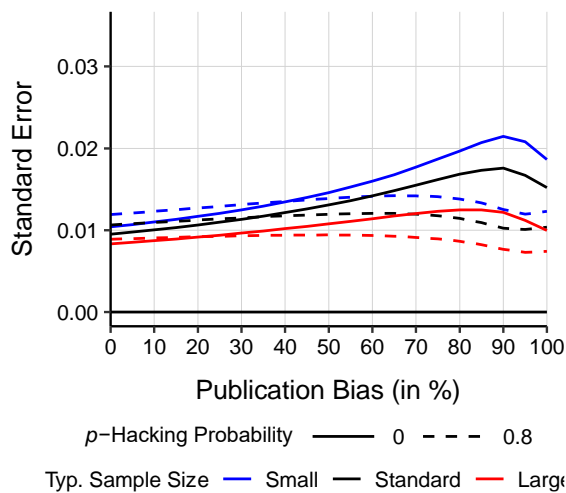
C - True Effect Size



D - Heterogeneity



E – Typical Sample Size



F – Publication Probability

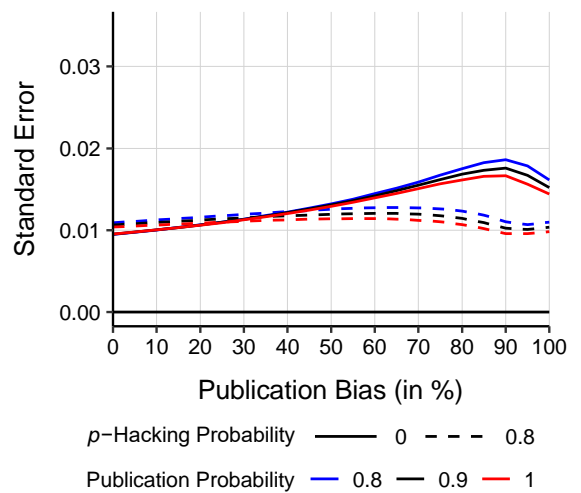


Figure S7. Results of the full simulation for standard errors as the main outcome variable. Figure S7A displays the results of the default configuration. The settings are identical to Figure 2. The thin horizontal black line represents the null effect. The dotted black line represents the true effect. The curved solid line represents the

number of studies  $k$  at a  $p$ -hacking probability of 0. The curved dashed line represents the number of studies  $k$  at a  $p$ -hacking probability of 0.8. Figures S7B-S7F display results when the five factors of influence are systematically varied. For these figures, all default settings were used and only the respective factor was varied. Different colors indicate different levels of the factors, as described in the legend below each figure. Solid lines always indicate a  $p$ -hacking probability of zero, dashed lines indicate a  $p$ -hacking probability of 0.8.

**Part II, Paper 1: “Does neuroticism disrupt the psychological benefits of nostalgia? A meta-analytic test.”**





# Does neuroticism disrupt the psychological benefits of nostalgia? a meta-analytic test

Julius Frankenbach<sup>1</sup>, Tim Wildschut<sup>2</sup>, Jacob Juhl<sup>2</sup> and Constantine Sedikides<sup>2</sup>

## Abstract

Nostalgia, a sentimental longing or wistful affection for the past, confers self-oriented, existential, and social benefits. We examined whether nostalgic engagement is less beneficial for individuals who are high in neuroticism (i.e. emotionally unstable and prone to negative affect). Specifically, we tested whether the benefits of experimentally induced nostalgia are moderated by trait-level neuroticism. To address this issue, we conducted a high-powered individual participant data meta-analysis ( $N = 3556$ ,  $k = 19$ ). We found that the benefits of nostalgia were not significantly moderated by neuroticism, as they emerged for both high and low neurotics. This finding upheld when the self-oriented, existential, and social benefits of nostalgia were analysed jointly and when they were analysed separately. Taken together, individuals high and low in neuroticism are equally likely to benefit psychologically from engagement in nostalgic reverie.

## Keywords

nostalgia, neuroticism, autobiographical memory, meta-analysis

Date received: 12 June 2019; revised: 17 May 2020; accepted: 20 May 2020

In the 17th century, Swiss physician Johannes Hofer coined the term nostalgia, a compound of the Greek words ‘nostos’ (meaning homecoming) and ‘álgos’ (meaning pain). He used this term to describe the adverse symptoms displayed by Swiss mercenaries serving abroad (e.g. fainting, high fever, indigestion, stomach pain, and insomnia; Sedikides, Wildschut, & Baden, 2004). Although the view that nostalgia is characterised by dysfunction and disorder prevailed for centuries, recent research has led to a reappraisal of the emotion as a useful resource that individuals recruit to counter adversity (Sedikides, Wildschut, Routledge, & Arndt, 2015). This research has shown that nostalgia is a ‘self-conscious, bittersweet but predominantly positive and fundamentally social emotion’ (Sedikides, Wildschut, Routledge, & Arndt, 2015, p. 190), which is prevalent in everyday life. Indeed, in a sample of British undergraduates, 79% reported experiencing nostalgia at least once a week (Wildschut, Sedikides, Arndt, & Routledge, 2006). Memories that evoke nostalgia are self-relevant, atypical, and positive and often include close others, important events, or time periods, and also locations, animals, or objects (Van Tilburg, Bruder, Wildschut, Sedikides, & Göritz, 2019; Wildschut et al., 2006). Nostalgia is often triggered by external stimuli, such

as music (Nash, 2012; Routledge et al., 2011), scents (Reid, Green, Wildschut, & Sedikides, 2015), or tastes (Supski, 2013), and by internal stimuli, such as negative affect (Wildschut et al., 2006), lack of meaning in life (Routledge et al., 2011), or loneliness (Zhou, Sedikides, Wildschut, & Gao, 2008). The emotion is observed cross-culturally (Hepper et al., 2014) and across ages (Hepper, Wildschut, Sedikides, Robertson, & Routledge, 2020; Madoglou, Gkinopoulos, Xanthopoulos, & Kalamaras, 2017).

Much of the literature reviewed above, and the rehabilitation of nostalgia, is due to an experimental approach. The emotion has been experimentally induced, for example, by instructing participants to recall and emotionally relive a nostalgic (vs. ordinary or positive) autobiographical episode (Stephan, Sedikides, & Wildschut, 2012; Wildschut et al., 2006), listen to nostalgic (vs. cheerful) music

<sup>1</sup>Department of Psychology, Saarland University, Saarbrücken, Germany

<sup>2</sup>Centre for Research on Self and Identity, Psychology Department, University of Southampton, Southampton, UK

### Corresponding author:

Julius Frankenbach, Department of Psychology, Saarland University, Campus A2 4, 66123 Saarbrücken, Germany.

Email: julius.frankenbach@gmail.com

(Routledge et al., 2011), or read nostalgic (vs. control) song lyrics (Cheung et al., 2013). The most frequently used induction method is the event reflection task (ERT; Sedikides, Wildschut, Routledge, & Arndt, 2015; Wildschut et al., 2006), which relies on targeted autobiographical recall. In the ERT, participants are randomly assigned to recall either a nostalgic or ordinary event from their past and to think about how it makes them feel. Then, they list keywords that capture the gist of the event and, more often than not, write a narrative account of their experience.

Experimental evidence reveals that nostalgia serves three vital intrapersonal and interpersonal functions (Sedikides, Wildschut, Routledge, & Arndt, 2015). First, nostalgia fulfils a self-oriented function by augmenting self-esteem (Hepper, Ritchie, Sedikides, & Wildschut, 2012; Wildschut et al., 2006), boosting optimism (Biskas et al., 2019; Cheung et al., 2013), and facilitating psychological growth and authenticity (Baldwin, Biernat, & Landau, 2015; Baldwin & Landau, 2014; Stephan et al., 2012). Second, nostalgia serves an existential function, as it sustains perceptions of meaning in life (Routledge et al., 2011; Routledge, Wildschut, Sedikides, Juhl, & Arndt, 2012; Sedikides & Wildschut, 2018; Van Tilburg, Igou, & Sedikides, 2013) and instils self-continuity (i.e. a sense of connection between one's past and present selves; Sedikides et al., 2015, 2016; Sedikides, Wildschut, & Stephan, 2018; Van Tilburg, Sedikides, Wildschut, & Vingerhoets, 2019). Third, nostalgia enhances sociality by promoting social connectedness (i.e. a sense of acceptance and belongingness; Sedikides & Wildschut, 2019; Turner, Wildschut, & Sedikides, 2012; Wildschut, Sedikides, Routledge, Arndt, & Cordaro, 2010; Zhou et al., 2008) and social action tendencies (i.e. a social approach orientation; Sedikides & Wildschut, 2020; Sedikides et al., 2018; Stephan et al., 2014; Zhou, Wildschut, Sedikides, Shi, & Feng, 2012). Indeed, a meta-analysis revealed robust main effects of nostalgia on self-oriented (self-esteem and optimism), existential (meaning in life and self-continuity), and sociality (social connectedness) functions, as well as on positive, but not negative, affect (Ismail, Cheston, Christopher, & Meyrick, 2020). Increasingly, however, researchers have been turning their attention to the role of personality: Are certain persons more capable of reaping the psychological benefits of nostalgia than others?

### Generality of nostalgia: the role of personality traits

The lion's share of attention has been enjoyed by a small number of traits. Two studies have examined the role of nostalgia proneness, the dispositional or trait-level tendency to experience nostalgia (Barrett et al., 2010; Wildschut & Sedikides, in press). In an ERT experiment, Cheung, Sedikides, and Wildschut

(2016) assessed nostalgia proneness prior to the nostalgia induction. Recalling a nostalgic (compared with ordinary) life event was more beneficial (i.e. increased self-esteem, social connectedness, and optimism) for participants who were high (compared with low) in nostalgia proneness. Layous, Kurtz, Wildschut, and Sedikides (2020) examined the role of nostalgia proneness (assessed at baseline) in a 6-week ERT-based intervention study. Well-being was assessed at the end of the 6 weeks and at a 1-month follow-up. The nostalgia intervention increased well-being at both time points for participants who were high in nostalgia proneness but decreased it for those who were low in nostalgia proneness. These findings are consistent with the person-activity fit principle in well-being interventions (Lyubomirsky & Layous, 2013). Individuals who experienced nostalgia regularly in their everyday lives (i.e. those who were relatively more nostalgia prone) benefitted the most from nostalgia inductions.

Two further studies examined individual differences that can be classified under the domain-level trait of neuroticism or emotional instability in the Big Five taxonomy of personality (John & Srivastava, 1999). Neuroticism is the enduring tendency to experience distress and negative emotions, such as fear, sadness, anxiety, loneliness, worry, self-consciousness, or dissatisfaction (John, Naumann, & Soto, 2008), and is considered a fundamental domain of human personality (McCrae & Costa, 2003). Verplanken (2012) assessed individual differences in habitual worrying (i.e. the tendency to engage repetitively and persistently in mental problem solving of uncertain or unresolved difficulties or challenges; Verplanken, Friborg, Wang, Trafimow, & Woolf, 2007) prior to an ERT-based nostalgia induction. Worry is a cognitive marker of neuroticism (Segerstrom, Tsao, Alden, & Craske, 2000) and is positively related to it (Muris, Roelofs, Rassin, Franken, & Mayer, 2005). Results revealed that nostalgia (compared with control) increased positive mood irrespective of habitual worrying. However, for participants scoring high (vs. low) on habitual worrying, nostalgia (compared with control) also increased feelings of anxiety and depression. In an ERT experiment among Syrian refugees residing in Saudi Arabia, Wildschut et al. (2019) assessed individual differences in resilience (i.e. the ability, when meeting adversity, to maintain psychological equanimity and cope adaptively with stress; Wagnild & Young, 1993) prior to the nostalgia induction. Vulnerability to stress is a core facet of neuroticism, and, accordingly, resilience is inversely related to neuroticism (Campbell-Sills, Cohan, & Stein, 2006). Compared with high-resilience refugees, those lacking resilience derived fewer psychological benefits, and suffered greater psychological costs, from the nostalgia induction.

There is a danger, when examining the moderating role of personality traits, of becoming mired in a

piecemeal and atheoretical exploration of ‘the thousands of particular attributes that make each human being individual and unique’ (John & Srivastava, 1999, pp. 102–103). To avoid this trap, we adopted a common, integrative framework that synthesises diverse systems of personality description—the Big Five taxonomy. To be precise, the specific findings for habitual worrying (Verplanken, 2012) and resilience (Wildschut et al., 2019) point to a particular role of neuroticism. Our key objective, then, was to examine if the psychological benefits of nostalgia inductions depend on trait-level neuroticism.

### The potential role of neuroticism

The postulated role of neuroticism as a suppressor of nostalgia’s benefits is based on two premises. *First*, despite being predominately positive, nostalgia commonly contains elements of negativity (Batcho, 2007; Hepper et al., 2012; Hertz, 1990) and has a unique bittersweet affective signature (Sedikides & Wildschut, 2016; Van Tilburg, Bruder, et al., 2019; Van Tilburg, Sedikides, & Wildschut, 2018). On the one hand, the content of nostalgic narratives is more positive than negative (Abeyta, Routledge, Roylance, Wildschut, & Sedikides, 2015; Madoglou et al., 2017; Wildschut et al., 2006), and nostalgia inductions typically (Hepper et al., 2012; Stephan et al., 2012; Wildschut et al., 2006, 2010; Zhou et al., 2008, 2012, Study 1), but not always (Turner, Wildschut, Sedikides, & Gheorghiu, 2013; Van Dijke, Wildschut, Leunissen, & Sedikides, 2015; Zhou et al., 2012, Studies 2–4), increase positive affect. On the other hand, nostalgia is not devoid of negative affect. Whereas nostalgia inductions tend to increase positive affect, they typically do not reduce negative affect (Cheung et al., 2013; Routledge et al., 2012; Routledge, Arndt, Sedikides, & Wildschut, 2008; Sedikides et al., 2016; Sedikides, Wildschut, Routledge, & Arndt, 2015; Wildschut et al., 2006, 2010; Zhou et al., 2012). Analyses of laypersons’ conceptualisation of nostalgia suggest that this negativity comes from missing or longing for aspects of the past. Specifically, laypersons regard ‘longing/yearning’, ‘missing’, and ‘wanting to return to the past’ as central features of the construct ‘nostalgia’ (Hepper et al., 2012, 2014). *Second*, neuroticism is linked to negative affectivity (Gray, 1981; Hamann & Canli, 2004; Rusting & Larsen, 1998) and so may undermine nostalgia’s benefits through several mechanisms. These are grounded in the availability, accessibility, and processing of emotional memories.

We start by considering the *availability* of nostalgic memories. Nostalgic memories can contain positive aspects (e.g. momentous life events or meaningful social interactions) and/or negative aspects (e.g. the loss of a loved one; Wildschut et al., 2006). The relative degree of positivity and negativity differs across memories and between individuals. Neuroticism is

associated with several negative life outcomes, such as lower subjective well-being (Steel, Schmidt, & Shultz, 2008), higher levels of psychopathology (Malouff, Thorsteinsson, & Schutte, 2005), and higher likelihood of criminal arrest (Huo-Liang, 2006). Thus, the pool of memories about which high neurotics (compared with low neurotics) could be nostalgic may be more negatively valenced on average. Indeed, high neurotics appear to report a larger proportion of negative autobiographical memories (Denkova, Dolcos, & Dolcos, 2012).

Additionally, irrespective of the availability of certain memories, neuroticism may entail a tendency to draw upon nostalgic memories that are more negatively valenced and thus have lower potential to convey psychological benefits. That is, there may be systematic differences between those high and low in neuroticism with respect to the *accessibility* of memories that they select for nostalgic reflection. Consistent with this, high neurotics (compared with low neurotics) are more likely to retrieve affectively negative content in cued or free recall tasks (Rusting & Larsen, 1998). Further, research on life stories (i.e. top-level narratives that people construct from personal experiences to derive and maintain a sense of self) indicates that high neurotics are more likely to include affectively negative content in their life stories (McAdams, Reynolds, Lewis, Patten, & Bowman, 2001; Raggatt, 2006; Thomsen, Olesen, Schnieber, & Tønnesvang, 2014) and to revive especially bitter memories (Cappeliez & O’Rourke, 2002).

Finally, in regard to the third and perhaps most important mechanism, individuals with elevated levels of neuroticism may *process* nostalgic memories differently. The same emotional memory may convey psychological benefits for someone low in neuroticism but may be costly for someone high in neuroticism. High neurotics (compared with low neurotics) may benefit less from nostalgic engagement, because their dispositional style of emotional processing could exacerbate the negatives inherent to the nostalgic experience that are otherwise reappraised or outweighed by the positives. That is, they may be particularly sensitive to the negative aspects of the nostalgia experience. Research on the functioning of neuroticism in the broader context of autobiographical memory indicates that high neurotics (compared with low neurotics) experience autobiographical memories as more emotionally and physiologically intense, rehearse them more, and see them as more central to their identity (Rubin, Boals, & Hoyle, 2014; Rubin, Dennis, & Beckham, 2011; Sutlin, 2008). Boelen (2009) found that high neurotics (compared with low neurotics) who lost a loved one are more likely to perceive the event as central to their identity and suffer more severe psychological harm. Similarly, Ogle, Siegler, Beckham, and Rubin (2017) reported that highly neurotic individuals suffer more serious consequences from traumatic events, because they



respond more emotionally to traumatic memories, rehearse them more, and perceive them as more central to their identity.

## Overview

We conducted a comprehensive meta-analysis to test whether neuroticism attenuates the psychological benefits of nostalgia. Although meta-analyses typically aim to summarise an existing literature, we relied on meta-analysis here to address a focused question. As such, we searched for studies (published or unpublished) that measured trait neuroticism and experimentally manipulated nostalgia. We derived the effect sizes of interest using raw data from these primary studies. This approach is sometimes referred to as two-step individual participant data meta-analysis (Riley, Lambert, & Abo-Zaid, 2010). We considered a wide range of dependent variables encompassed by the tripartite (self, existential, and social) taxonomy of nostalgia's psychological benefits. That is, we examined whether the effects of nostalgia on these three domains are smaller for high neurotics than low neurotics. Additionally, we explored whether the effects of nostalgia on positive affect and negative affect differ as a function of neuroticism.

## Method

We used the R environment for statistical computing (R Core Development Team, 2017) to process and analyse all data. We fit robust variance estimation (RVE) models using the *robumeta* package (Fisher, Tipton, & Zhipeng, 2017). Effect-size data and analysis scripts are publicly available at [osf.io/sfx6h](https://osf.io/sfx6h). The study was not preregistered.

### Inclusion criteria and data collection

Studies were eligible for inclusion in the meta-analysis if they (i) experimentally manipulated nostalgia, (ii) contained at least one control condition, (iii) randomly assigned participants to conditions, (iv) measured trait neuroticism, and (v) measured at least one outcome that could be classified as a self-oriented, existential, or social autobiographical-memory function. Some studies that met these criteria also contained positive and/or negative affect as outcomes. For these studies, we also analysed positive and negative affect. However, we excluded studies that assessed exclusively positive and negative affect, as this was not our focus. We only included studies for which we had access to the primary (or raw) data. To identify relevant studies, we contacted active researchers in the area of nostalgia. We further sent queries for data through mailing lists of the *Society of Experimental Social Psychology* and the *Society for Personality and Social Psychology*. Additionally, we conducted an electronic literature search of the Web

of Science Core Collection (in October, 2019), searching all fields for the terms 'nostalg\* AND (neurotic\* OR personality OR big five)'. For all relevant articles, we requested full data sets as well as any available materials and documentation. When information was missing or unclear, we consulted the primary authors to resolve ambiguities.

### Data preparation

We applied a standardised data-processing protocol to all studies to make effect sizes comparable. We coded the nostalgia manipulation as 0 for the control condition and 1 for the nostalgia condition. For studies that used multiple controls, we included the most neutral one. For example, if an experiment used both ordinary-memory and positive-memory control conditions, we calculated an effect size for the comparison between nostalgia and ordinary memory. We standardised neuroticism scores and all outcome variables by calculating  $z$  scores ( $M = 0$ ,  $SD = 1$ ). In supplementary analyses, we converted neuroticism scores to a 5-point scale to enable comparisons of the mean level and dispersion of neuroticism across studies. We reverse scored all dependent variables that reflected negative outcomes (except negative affect), so that higher scores indicated more beneficial outcomes. For example, we reverse scored the No Meaning in Life Scale (Kunzendorf, Moran, & Gray, 1995) for higher scores to reflect greater sense of meaning in life. Finally, we estimated scale reliability by computing Cronbach's alphas for neuroticism and all outcomes.

### Effect-size computation

We computed three effect sizes for each outcome per study: (i) nostalgia main effect, (ii) neuroticism main effect, and (iii) Nostalgia  $\times$  Neuroticism interaction. We used Cohen's  $d$  for all effect sizes. For the main effects of the nostalgia manipulation, we computed Cohen's  $d$  effect sizes as the mean difference between the nostalgia and control conditions divided by the pooled standard deviation. Higher values indicate higher means in the nostalgia condition. For neuroticism main effects, we calculated Pearson correlations ( $r$ ) between neuroticism and the respective outcome variable. We then transformed all correlations to Cohen's  $d$  (Borenstein, Hedges, Higgins, & Rothstein, 2009). For interactions, we fitted a multiple regression model for each outcome per study, predicting the respective outcome ( $z$ -standardised, Mean = 0,  $SD = 1$ ) from neuroticism ( $z$ -standardised), nostalgia (0 = control, 1 = nostalgia), and the Nostalgia  $\times$  Neuroticism interaction. We then retrieved the regression coefficients and standard errors of the interaction term from each analysis. The regression coefficient indicates the predicted change in the nostalgia main effect when levels of neuroticism in the sample increase by one standard

deviation. The metric of the nostalgia main effect is standard deviations, so the regression coefficient is also in the metric of Cohen's *d*.

We considered a range of outcomes. Analysing these diverse outcomes involves a trade-off between construct validity and statistical power. Power is maximised when all outcomes are synthesised into a single summary effect. However, this may entail combining psychologically distinct constructs. On the other extreme, construct validity is maximised when outcomes reflecting the exact same construct (e.g. self-esteem) are aggregated separately. This, though, may yield small subgroups of outcomes, and so statistical power to detect effects within these subgroups may be low. Taking this trade-off into account, we adopted a sequential procedure.

We started by synthesising all outcomes to arrive at a single summary effect (prioritising statistical power over construct validity). Next, we grouped outcomes in terms of the three previously established superordinate autobiographical-memory functions of nostalgia: self-oriented, existential, and social (Sedikides, Wildschut, Routledge, & Arndt, 2015). Subsequently, we calculated summary effects for these three superordinate categories (striking a balance between construct validity and power). Finally, we divided outcomes within the three superordinate categories into subcategories according to the psychological construct they reflected (yielding seven subcategories: self-esteem, optimism, inspiration, meaning in life, self-continuity, social connectedness, and social action tendencies—see below for details), and then we derived summary effects for these specific subcategories (prioritising construct validity over statistical power). We analysed positive and negative affect separately in subgroup analysis.

### Study coding

We coded for a range of study and outcome characteristics. We included some for descriptive purposes and others for examination as meta-moderators of the Nostalgia  $\times$  Neuroticism effect size in meta-regression analyses. We reasoned that these meta-moderators may account for variation in the magnitude of the Nostalgia  $\times$  Neuroticism effects across studies and outcomes.

**Type of nostalgia induction.** The magnitude of the Nostalgia  $\times$  Neuroticism interaction effect may depend on type of nostalgia induction. For instance, manipulations may differ in the degree of negativity they induce, and thus the degree to which their effects are moderated by neuroticism could differ. We coded whether nostalgia was induced by the ERT or music.

**Type of control condition.** Several control conditions have been used in the nostalgia literature. For the ERT, procedures that involve the recollection of

ordinary events are advantageous, because they provide a neutral reference point. Thus, the comparison of a neutral control condition and a nostalgia condition allows all psychologically active components of nostalgia to contribute to the effect. More stringent control conditions have also been implemented to isolate incremental effects of nostalgia manipulations. For example, in some studies participants in the control condition listened to happy music, which allowed researchers to examine the effects of nostalgia above and beyond positive mood. We coded whether the control condition was intended to be neutral or non-neutral.

**Type and reliability of neuroticism scale.** Neuroticism scales differ in several ways. First, measurement reliability may vary depending on number and type of items included in the scale. We expected for more reliable neuroticism scales to yield stronger interaction effects. Second, scales may assess distinct components of neuroticism, and some components may interact with nostalgia more strongly than others. We therefore coded for type of neuroticism scale and its reliability (indexed by Cronbach's alpha). We set the reliability of single-item scales to the minimum of all reliability estimates in the meta-analysis, as a conservative lower-bound estimate. Additionally, and in an effort to mark the relative length of neuroticism scales, we coded studies that used the Big Five Inventory (BFI—eight items; John et al., 2008) as 'long', and we coded studies that used either the Ten-Item Personality Inventory (TIPI—two items; Gosling, Rentfrow, & Swann, 2003) or the TIPI-Revised (TIPI-r—one item; Denissen, Geenen, Selfhout, & van Aken, 2008) as 'short'.

**Publication status.** We coded all studies that were published in peer-reviewed journals as 'published'. We coded the remaining studies as 'unpublished'. Two (out of 19) studies were published, and both were reported by Cheung et al. (2013).

**Mean sample age.** We calculated participants' average age, separately for each study. Doing so enabled us to examine whether focal effects varied as a function of the mean age within a sample.

**A autobiographical-memory functions and type of affect.** We only included studies reporting at least one outcome that was classifiable as self-oriented, existential, or social. Some of these studies also measured positive affect or negative affect as outcome variables. We coded all studies in terms of these five outcome categories. We tested whether the moderating role of neuroticism differed among the outcome categories.

**Outcome subcategory.** Within the three major outcome categories (self-oriented, existential, and social), effect sizes could be further classified into subcategories. For the self-oriented category, subcategories comprised

self-esteem (e.g. state version of the Rosenberg Self-Esteem Scale; Rosenberg, 1965; e.g. 'I feel that I'm a person of worth, at least on an equal basis with others'), optimism (e.g. Life Orientation Test-Revised; Scheier, Carver, & Bridges, 1994; e.g. 'In uncertain times, I usually expect the best'), and inspiration (e.g. Inspiration Scale; Thrash & Elliot, 2003; e.g. 'I feel inspired'). For the existential category, subcategories comprised meaning in life (e.g. Meaning in Life Questionnaire; Steger, Frazier, Oishi, & Kaler, 2006; e.g. 'I understand my life's meaning') and self-continuity (e.g. Self-Continuity Index; Sedikides, Wildschut, Routledge, Arndt, Hepper, & Zhou, 2015; e.g. 'There is continuity in my life'). For the social category, subcategories comprised social connectedness (e.g. 'Right now, I feel connected to loved ones'; Wildschut et al., 2006) and social action tendencies (e.g. 'Thinking about this nostalgic event makes me want to join a student group made up of a wide range of people I don't know'; Stephan et al., 2014). As we mentioned above, positive affect and negative affect were separate categories and were typically measured with the Positive and Negative Affect Schedule (Watson, Clark, & Tellegen, 1988).

**Outcome measurement reliability.** Interaction effects are dependent on the correlation of the predictors with the outcome, which in turn is dependent on the reliability of the outcome measurement. We computed Cronbach's alpha as estimates of reliability for all outcomes. For single-item measures, we entered the lowest reliability observed across all studies included in the meta-analysis. We expected that more reliable outcomes would register larger Nostalgia  $\times$  Neuroticism interaction effects.

### Meta-analytic procedure

**Meta-analytic modelling.** The analyses included various neuroticism scales, experimental procedures, and outcome variables. It is therefore unrealistic to treat the effect sizes as being drawn from the same population. Accordingly, we conducted all analyses using random-effects models. One central assumption of conventional random-effects meta-analytic models is statistical independence of effect sizes. This assumption is violated when multiple effect size from the same study are included. There are several approaches to addressing this issue. First, researchers often maintain independence by including only one effect size per study. However, this entails a considerable loss of information and comes with a risk of bias in the selection process. Second, researchers may aggregate all effect sizes stemming from the same study into a composite. One variant of this approach involves adjusting effect-size variances of the composite based on the correlation structure of the aggregated effect sizes (Borenstein et al., 2009). Specifically, variances are more strongly reduced if outcomes are

less correlated, reflecting the idea that less correlated outcomes provide more unique information, and consequently more precise estimates. Although this procedure reduces the risk of bias, it also entails a loss of information because different constructs are combined into a composite that may be difficult to interpret. Third, Hedges, Tipton, and Johnson (2010) recently proposed an RVE approach for meta-analysis. This approach permits fitting random-effects or mixed-effects meta-analytical models to sets of dependent effect sizes without a need for selection or aggregation. RVE estimates the covariance structure of effect sizes and adjusts standard errors accordingly. This approach, however, has two drawbacks. To begin, although it is possible to derive point estimates for true effect-size heterogeneity in RVE ( $I^2$ ), significance tests for this estimate are currently unavailable. Moreover, procedures for power analysis in RVE have not yet been developed. Considering the (dis)advantages of the three approaches outlined above, we implemented RVE for all analyses. To evaluate the magnitude of true effect-size heterogeneity, we resorted to rules of thumb (Higgins & Green, 2011). We estimated statistical power by applying power analysis for conventional meta-analysis as an upper bound estimate.

**Robust variance estimation.** Before conducting RVE, we considered three issues (TannerSmith & Tipton, 2014). First, we needed to determine if the number of studies sufficed to obtain accurate model estimates. Standard RVE performs satisfactorily with a minimum of 10 studies when estimating summary effects and with a minimum of 40 studies when estimating slopes in meta-regression (Hedges et al., 2010; Tipton, 2013). However, when the number of studies falls below these limits, significance tests are plagued by inflated Type I error rates. Recently, small sample corrections have been developed for single and multiple parameter tests in RVE that account for inflated error rates (Tipton, 2015; Tipton & Pustejovsky, 2015). We implemented these corrections for all RVE models. Specifically, we computed regression coefficients using adjusted covariance matrices. We tested single regression coefficients using *t*-tests with Satterthwaite-adjusted degrees of freedom (Tipton, 2015) and multiple regression coefficients with the approximate Hotelling-Zhang test (AHZ; Tipton & Pustejovsky, 2015). Second, we needed to decide how to weigh the effect sizes in the summary effect. Following relevant recommendations (Tanner-Smith & Tipton, 2014), we set the weights to account for dependence due to correlated, rather than hierarchical, effects, because this type of dependence was likely to be more prevalent in the data set. Third, we needed to estimate the average correlation between effects sizes. We estimated this value by averaging all outcome correlations per study and then averaging these means across studies. This procedure returned a mean



outcome correlation of  $r = 0.45$ . We conducted a sensitivity analysis for all models by varying this estimate from .10 to .90. In no case did  $r$  considerably influence any conclusions drawn from the models.

**Meta-moderation analyses.** To examine whether the magnitude of the Nostalgia  $\times$  Neuroticism interaction is moderated by study characteristics (e.g. type of nostalgia induction, mean sample age), we entered these characteristics as predictors in meta-regression. Meta-regression is analogous to linear regression in primary studies, with the exception that effect sizes (rather than participant-level outcomes) are regressed on predictors. The meta-moderation analyses focused on accounting for variation in the Nostalgia  $\times$  Neuroticism interaction effect—the main focus of this meta-analysis. We report meta-moderation analyses for the nostalgia and neuroticism main effects in the Supporting Information. Given that all music-induction studies used a non-neutral control condition, and all but one ERT-induction studies used a neutral control condition, the type of nostalgia induction and type of control condition are confounded. Therefore, results for type of nostalgia induction and type of control condition are similar or, in most cases, identical.

**Results**

We identified  $k = 19$  eligible studies and obtained raw data for all of them, totalling  $m = 155$  effect sizes and

$N = 3556$  participants (Figure S1). One hundred sixteen effect sizes related to the three autobiographical-memory functions and 39 related to positive and negative affect. Sample sizes ranged from 48 to 647 ( $Md = 121$ ), and studies contributed between three and 17 outcomes ( $Md = 9$ ). Seventeen studies were unpublished as of June 2019 (89%). The most prevalent nostalgia induction was the ERT ( $k_{ERT} = 16$ ,  $k_{music} = 3$ ). Control conditions were mostly neutral ( $k_{neutral} = 15$ ,  $k_{non-neutral} = 4$ ). Neuroticism was typically measured by the BFI ( $k = 12$ ), followed by the TIPI ( $k = 4$ ) and the TIPI-r ( $k = 3$ ). Among the three superordinate autobiographical-memory functions, outcomes measuring the self-oriented function were overrepresented (self-oriented, 43%; existential, 30%; social, 27%). All but two studies measured positive affect and negative affect. The total sample comprised 62% women, and the median age was 22 years ( $M = 29.94$ ,  $SD = 15.45$ ,  $min = 14$ ,  $max = 85$ ). Figure S2 displays a histogram of the age distribution. We summarise key information about the included studies in Table 1.

**Nostalgia functions**

**Nostalgia main effect.** The overall nostalgia effect across self-oriented, existential, and social functions was significant,  $d = 0.284$ ,  $SE = 0.044$ ,  $p < .001$ ,  $CI_{95}[0.190, 0.377]$ . Nostalgia manipulations induced an average increase of 0.284 standard deviations across the three superordinate autobiographical-memory functions.

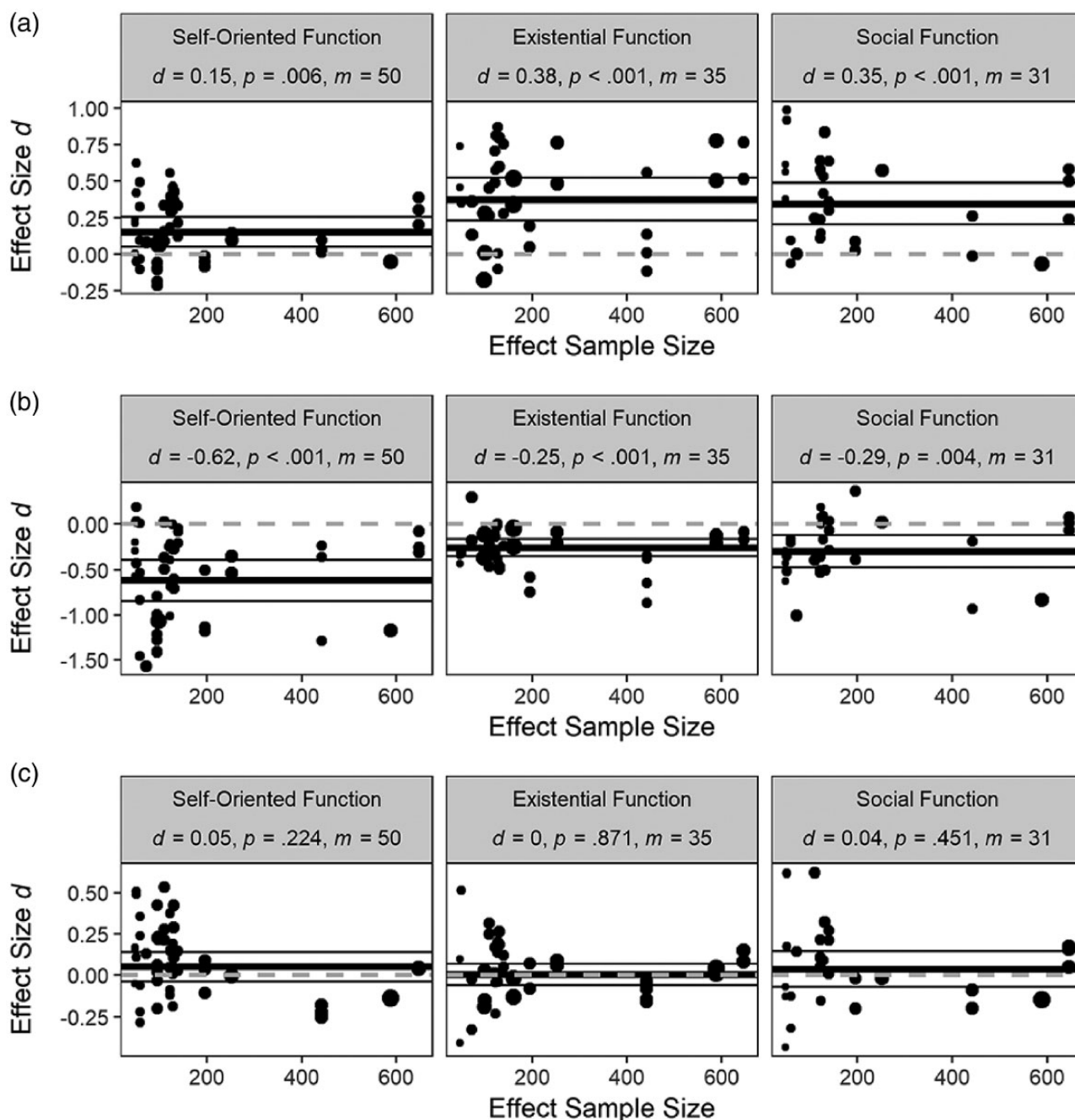
**Table 1.** Study overview

ID	N	m	Scale	Published	Induction	Control condition	Outcomes	Mean age	Date	Country	Corresponding author
1	59	7	BFI	No	ERT	Neutral	1, 2, 3, 4	23.83	2005	UK	T. Wildschut
2	122	12	BFI	No	ERT	Neutral	1, 2, 3, 4, 5, 6, 7, 8, 9	19.73	2007	UK/USA	T. Wildschut
3	442	12	BFI	No	ERT	Neutral	1, 2, 5, 6, 7, 8, 9	50.13	2008	UK	E. G. Hepper
4	127	12	BFI	Yes	ERT	Neutral	1, 2, 4, 5, 6, 7, 8, 9	18.95	2012	USA	W.-Y. Cheung
5	95	8	BFI	No	ERT	Neutral	1, 2, 8, 9	17.02	2004	UK	T. Wildschut
6	98	4	BFI	No	ERT	Neutral	2, 5, 8, 9	24.35	2005	UK	T. Wildschut
7	50	9	BFI	No	ERT	Neutral	1, 2, 4, 5, 6, 8, 9	19.94	2006	UK	C. Routledge
8	195	9	TIPI	No	ERT	Neutral	2, 3, 5, 6, 7, 8, 9	20.14	2014	USA	J. D. Green
9	100	2	BFI	No	ERT	Non-neutral	7	19.28	2012	UK	J. Juhl
10	121	10	TIPI	No	ERT	Neutral	2, 5, 6, 7, 8, 9	20.21	2014	UK	J. Juhl
11	252	7	BFI	No	ERT	Neutral	1, 2, 5, 6, 7, 8, 9	36.82	2014	USA	J. Juhl
12	161	4	TIPI	No	ERT	Neutral	5, 7, 8, 9	19.31	2016	UK	J. Juhl
13	130	9	TIPI	No	ERT	Neutral	1, 2, 3, 5, 6, 7, 8, 9	19.78	2014	UK	J. Juhl
14	647	10	TIPI-r	Yes	Music	Non-neutral	1, 2, 3, 5, 6, 7, 8, 9	36.68	2012	Netherlands	W.-Y. Cheung
15	139	10	TIPI-r	No	Music	Non-neutral	1, 2, 3, 5, 6, 7, 8, 9	37.58	2013	Netherlands	T. Wildschut
16	48	10	TIPI-r	No	Music	Non-neutral	1, 2, 3, 5, 6, 7, 8, 9	41.02	2015	Netherlands	T. Wildschut
17	72	6	BFI	No	ERT	Neutral	2, 5, 6, 7, 8, 9	18.72	2012	Ireland	W. A. P. Van Tilburg
18	589	6	BFI	No	ERT	Neutral	2, 5, 6, 7, 8, 9	26.36	2012	Ireland	W. A. P. Van Tilburg
19	109	8	BFI	No	ERT	Neutral	1, 2, 3, 5, 6, 7, 8, 9	19.91	2014	UK	E. G. Hepper

Note: Outcomes: 1, optimism; 2, self-esteem; 3, inspiration; 4, social action tendencies; 5, self-continuity; 6, social connectedness; 7, meaning; 8, positive affect; 9, negative affect. Date, date of data collection. Both published studies were reported by Cheung et al. (2013), as Study 2 (ID 4) and Study 3 (ID 14), respectively. N for ID 14 (647) is lower than reported by Cheung et al. (664), because some participants did not complete the TIPI-r. ID, study identification number; m, number of effects per study; Scale, neuroticism scale; BFI, Big Five Inventory; ERT, event reflection task; TIPI, Ten-Item Personality Inventory; TIPI-r, TIPI-Revised.

There was a substantial amount of effect-size heterogeneity,  $I^2 = 72.85$ ,  $\tau^2 = 0.063$ .  $I^2$  is interpreted as the percentage of true effect-size variance in the total variance.  $\tau^2$  reflects the true variance of effect sizes in the metric of the effect size (i.e. one standard deviation). To test if the magnitude of the nostalgia effect differed between the three superordinate functions, we dummy coded the functions (self-oriented,

existential, and social) and entered them as predictors in the model. Results of this analysis revealed that the nostalgia effect differed significantly among the three domain-level functions,  $AHZ(14.57) = 9.02$ ,  $p = .003$ , remaining  $I^2 = 67.23$  (Figure 1a). Although the nostalgia effect was larger for the existential and social functions than the self-oriented function, it was statistically significant for all three of them (self-



**Figure 1.** Nostalgia main effects (a), neuroticism main effects (b), and Nostalgia  $\times$  Neuroticism interaction effects (c) for autobiographical-memory functions.  $d$ , summary effect size;  $m$ , number of effect sizes per autobiographical-memory function. Effect-size magnitude is depicted on the y-axis, and the associated sample size for each effect size is depicted on the x-axis. Larger points indicate more weight. The thick black horizontal line represents the summary effect for the given autobiographical-memory function. Thin black horizontal lines represent the boundaries of the 95% confidence interval of the summary effect. The dashed grey line represents the null effect.



**Table 2.** Summary effects by nostalgia functions.

Type of effect	Function	<i>d</i>	SE	<i>P</i>	CI95 [LL, UL]	<i>k</i>	<i>m</i>
Nostalgia main effect	Self-oriented	0.15	0.05	.006	[0.05, 0.26]	17	50
	Existential	0.38	0.07	<.001	[0.23, 0.52]	17	35
	Social	0.35	0.06	<.001	[0.20, 0.49]	15	31
Neuroticism main effect	Self-oriented	-0.62	0.11	<.001	[-0.85, -0.39]	17	50
	Existential	-0.25	0.04	<.001	[-0.35, -0.16]	17	35
	Social	-0.29	0.08	.004	[-0.47, -0.12]	15	31
Nostalgia × Neuroticism	Self-oriented	0.05	0.04	.224	[-0.04, 0.14]	17	50
	Existential	0.00	0.03	.871	[-0.06, 0.07]	17	35
	Social	0.04	0.05	.451	[-0.07, 0.15]	15	31

Note: *d*, summary effect size; *p*, *p*-value testing the respective summary effect against zero; SE, standard error; CI95, limits of the 95% confidence interval of the summary effect; *k*, number of studies in the subgroup; *m*, number of effect sizes in the subgroup.

**Table 3.** Summary effects for the main effects of the nostalgia manipulation by outcome subcategory.

Subcategory	<i>d</i>	SE	LL	UL	<i>t</i>	<i>df</i>	<i>P</i>	<i>k</i>	<i>m</i>
Self-esteem	0.08	0.04	-0.00	0.17	2.16	12.11	.052	17	25
Optimism	0.22	0.07	0.07	0.36	3.34	9.58	.008	12	17
Inspiration	0.29	0.07	0.13	0.46	4.24	6.39	.005	8	8
Meaning in life	0.32	0.06	0.18	0.46	5.21	8.19	.001	15	17
Self-continuity	0.44	0.10	0.21	0.67	4.23	10.94	.001	16	18
Social action tendencies	0.37	0.18	-0.23	0.97	2.09	2.66	.139	4	6
Social connectedness	0.34	0.07	0.19	0.49	5.03	9.74	.001	14	25
Negative affect	0.22	0.06	0.09	0.35	3.62	14.02	.003	17	17
Positive affect	0.22	0.06	0.09	0.35	3.73	14.37	.002	17	22

Note: *d*, summary effect size; LL, lower limit of the 95% confidence interval (CI); UL, upper limit of the 95% CI; *t*, *t*-value associated with the *d*-value in the same row testing statistical significance in the respective subcategory; *p*, *p*-value associated with the *t*-value in the same row; *df*, degrees of freedom associated with the *t*-value in the same row; *k*, number of studies in the respective subcategory; *m*, number of effect sizes available for the respective subcategory.

oriented, *d* = 0.15, *p* = .006; existential, *d* = 0.38, *p* < .001; social, *d* = 0.35, *p* < .001). We summarise results of the subgroup analysis for autobiographical-memory functions in Table 2.

Next, we partitioned the superordinate functions into subcategories (e.g. self-oriented partitioned into self-esteem, inspiration, and optimism) and again applied subgroup analysis. The nostalgia main effect differed significantly across the subcategories, *AHZ*(7.89) = 10.03, *p* = .002, remaining *I*<sup>2</sup> = 65.15. We present the nostalgia main effects within subcategories in Table 3. The nostalgia effect was significant for each outcome subcategory, except self-esteem (marginal) and social action tendencies. The latter subcategory was very small (*m* = 6).

**Neuroticism main effect.** The overall neuroticism effect across self-oriented, existential, and social functions was significant, *d* = -0.405, *SE* = 0.060, *p* < .001, *CI*<sub>95</sub>[-0.530, -0.279]. High (vs. low) neuroticism decreased scores across the three superordinate autobiographical-memory functions. Results revealed considerable effect-size heterogeneity, *I*<sup>2</sup> = 84.30, *τ*<sup>2</sup> = 0.135. To examine if the magnitude of the neuroticism effect varied among the self-oriented, existential, and social domains, we entered these superordinate functions in the model as dummy-coded predictor

variables. The neuroticism effect differed significantly among the domain-level functions, *AHZ*(15.15) = 5.43, *p* = .017, remaining *I*<sup>2</sup> = 81.72 (Figure 1b). Neuroticism was most negatively related to the self-oriented function, yet all neuroticism effects were significant (self-oriented: *d* = -0.62, *p* < .001; existential: *d* = -0.25, *p* < .001; social: *d* = -0.29, *p* = .004; Table 2).<sup>1</sup>

Partitioning the functions further into subcategories again revealed significant differences among the subcategories, *AHZ*(8.12) = 5.96, *p* = .012, remaining *I*<sup>2</sup> = 78.59. We present the neuroticism main effects within subcategories in Table 4. The neuroticism effect was significant (and negative) for each outcome subcategory, except inspiration (marginal) and social action tendencies. The null effects of nostalgia and neuroticism on social action tendencies stand in contrast to the robust and consistent effects on other outcomes, pointing to idiosyncrasies in this particular outcome subcategory.

**Nostalgia × Neuroticism interaction.** We now turn to our primary objective: the meta-analysis of Nostalgia × Neuroticism interaction coefficients. We found no evidence for a Nostalgia × Neuroticism interaction effect across the self-oriented, existential, and social autobiographical-memory functions, *d* = 0.030, *SE* = 0.033, *p* = .382, *CI*<sub>95</sub>[0.101, -0.042]. Hence,

**Table 4.** Summary effects for the main effects of neuroticism by outcome subcategory

Subcategory	<i>d</i>	SE	LL	UL	<i>t</i>	<i>df</i>	<i>P</i>	<i>k</i>	<i>m</i>
Self-esteem	-0.80	0.11	-1.04	-0.56	-7.19	12.52	<.001	17	25
Optimism	-0.50	0.11	-0.76	-0.25	-4.39	9.76	.001	12	17
Inspiration	-0.16	0.07	-0.33	0.01	-2.32	6.62	.055	8	8
Meaning in life	-0.27	0.05	-0.39	-0.16	-5.48	7.63	.001	15	17
Self-continuity	-0.23	0.05	-0.34	-0.12	-4.45	11.24	.001	16	18
Social action tendencies	-0.10	0.13	-0.54	0.34	-0.74	2.73	.517	4	6
Social connectedness	-0.33	0.09	-0.53	-0.12	-3.57	10.04	.005	14	25
Negative affect	0.67	0.13	0.39	0.96	5.02	15.35	<.001	17	17
Positive affect	-0.38	0.07	-0.52	-0.24	-5.78	15.53	<.001	17	22

Note: *d*, summary effect size; LL, lower limit of the 95% confidence interval (CI); UL, upper limit of the 95% CI; *t*, *t*-value associated with the *d*-value in the same row testing statistical significance in the respective subcategory; *p*, *p*-value associated with the *t*-value in the same row; *df*, degrees of freedom associated with the *t*-value in the same row; *m*, number of effect sizes available for the respective subcategory.

**Table 5.** Summary effects for the Nostalgia × Neuroticism interaction effects by outcome subcategory

Subcategory	<i>d</i>	SE	LL	UL	<i>t</i>	<i>df</i>	<i>P</i>	<i>k</i>	<i>m</i>
Self-esteem	0.06	0.05	-0.06	0.18	1.10	10.70	.294	17	25
Optimism	0.06	0.04	-0.04	0.15	1.40	8.93	.195	12	17
Inspiration	0.02	0.05	-0.11	0.14	0.32	5.18	.759	8	8
Meaning in life	0.01	0.03	-0.07	0.09	0.33	8.50	.746	15	17
Self-continuity	-0.00	0.04	-0.08	0.08	-0.06	9.71	.953	16	18
Social action tendencies	-0.04	0.07	-0.29	0.20	-0.63	2.65	.580	4	6
Social connectedness	0.05	0.05	-0.07	0.17	0.91	7.83	.389	14	25
Negative affect	0.03	0.04	-0.06	0.11	0.69	12.19	.502	17	17
Positive affect	0.05	0.05	-0.07	0.16	0.84	12.53	.414	17	22

Note: *d*, summary effect size; LL, lower limit of the 95% confidence interval (CI); UL, upper limit of the 95% CI; *t*, *t*-value associated with the *d*-value in the same row testing statistical significance in the respective subcategory; *p*, *p*-value associated with the *t*-value in the same row; *df*, degrees of freedom associated with the *t*-value in the same row; *m*, number of effect sizes available for the respective subcategory.

there is no general support for the idea that individuals who are high (vs. low) in neuroticism derive less psychological benefit from nostalgia inductions.

Effect-size heterogeneity was small to moderate,  $I^2 = 25.69$ ,  $\tau^2 = 0.008$ . To test if the Nostalgia × Neuroticism effect size differed among the self-oriented, existential, and social functions, we again entered these superordinate functions as dummy-coded predictor variables. The size of the Nostalgia × Neuroticism interaction did not differ significantly among functions,  $AHZ(11.06) = 0.69$ ,  $p = .522$ , remaining  $I^2 = 26.80$ . Furthermore, the Nostalgia × Neuroticism interaction was not significant within any of the three superordinate functions ( $ps > .224$ ; Table 2, Figure 1c).

Partitioning the superordinate functions into subcategories revealed no significant differences among the subcategories,  $AHZ(6.53) = 0.34$ ,  $p = .895$ , remaining  $I^2 = 28.60$ . We present the Nostalgia × Neuroticism interaction effects within subcategories in Table 5. The interaction effect was not significant for any of the subcategories ( $ps > .195$ ). In light of the strong and consistent main effects of nostalgia and neuroticism, these unequivocal null results for the Nostalgia × Neuroticism interaction cannot be attributed simply to methodological issues (e.g. failed experimental manipulations, and unreliable or invalid measurement).

### Positive and negative affect

The nostalgia manipulations significantly increased both positive affect ( $d = 0.220$ ,  $p = .002$ ) and negative affect ( $d = 0.220$ ,  $p = .003$ ). Neuroticism was negatively associated with positive affect ( $d = -0.380$ ,  $p < .001$ ) and positively associated with negative affect ( $d = 0.670$ ,  $p < .001$ ). Finally, the Nostalgia × Neuroticism interaction effect was not significant for either positive affect ( $d = 0.050$ ,  $p = .414$ ) or negative affect ( $d = 0.030$ ,  $p = .502$ ). In summary, nostalgia manipulations increased both positive affect and negative affect, whereas high (vs. low) neuroticism predicted decreased positive affect and increased negative affect. We again obtained null results for the Nostalgia × Neuroticism interaction.

### Meta-moderation by study characteristics

Next, we conducted meta-moderation analyses to examine if the Nostalgia × Neuroticism interaction varied as a function of study characteristics. (We report meta-moderation analyses for the nostalgia and neuroticism main effects in Tables S1 and S2.) We tested the association between the Nostalgia × Neuroticism effect size and the following study characteristics: (i) type of nostalgia induction, (ii) type of control condition, (iii) type of neuroticism

**Table 6.** Summary of significance tests for meta-moderation of the interaction effects

Outcome category	Type of nostalgia manipulation			Type of neuroticism scale			Type of control condition			Mean sample age		
	<i>t</i>	<i>df</i>	<i>P</i>	<i>t</i>	<i>df</i>	<i>P</i>	<i>t</i>	<i>df</i>	<i>P</i>	<i>t</i>	<i>df</i>	<i>P</i>
Self-related	0.58	2.19	.617	1.02	8.04	.337	0.58	2.19	.617	-2.48	3.75	.072
Existential	1.68	2.10	.229	1.34	6.91	.223	1.03	2.81	.382	-0.70	3.76	.527
Social	0.33	2.55	.764	0.56	9.56	.587	0.33	2.55	.764	-2.00	4.21	.113
Positive affect	-0.56	2.37	.626	0.71	11.34	.493	-0.56	2.37	.626	-2.58	4.01	.061
Negative affect	1.88	2.00	.201	0.07	8.94	.942	1.88	2.00	.201	0.99	3.50	.387

Note: The values *t*, *df*, and *p* denote statistical significance tests testing whether interaction effect sizes of the respective outcome category vary as a function of a study-level meta-moderator. Positive *t*-values indicate smaller effects in the reference category (coded as 0). Reference categories were 'ERT' for type of nostalgia manipulation (versus 'music'), 'long' (versus 'short') for type of neuroticism scale, and 'neutral' (versus 'non-neutral') for type of control condition.

scale (BFI, TIPI, and TIPI-r), and (iv) mean sample age. There were too few published studies ( $k = 2$ ) to examine publication status as a meta-moderator. We found no evidence that the magnitude of the Nostalgia  $\times$  Neuroticism interaction depended on type of nostalgia induction, type of control condition, type of neuroticism scale, or mean sample age for any of the outcome subcategories (Table 6).

### Sensitivity analyses

A common concern in meta-analysis is the presence of publication bias. Meta-analyses may overestimate effects, because studies reporting small, non-significant effect sizes are less likely to be submitted to, and published by, scientific journals (Ioannidis, 2008). We think it is unlikely that publication bias affected our findings, because only two of the included studies (out of 19) were published as of June 2019. For completeness, we applied a test for detecting small-study effects in the dataset (Sterne & Egger, 2005). For this test, effect sizes are regressed on standard errors of effect sizes in metaregression. A significant, positive slope indicates that effects are larger for smaller studies, which is often, but not always, due to publication bias. The test was non-significant for nostalgia main effects ( $b = -0.05$ ,  $p = .945$ ), neuroticism main effects ( $b = -1.44$ ,  $p = .173$ ), and Nostalgia  $\times$  Neuroticism interaction effects ( $b = 0.60$ ,  $p = .337$ ). These results should, however, be treated with caution. Although the underlying logic is applicable, tests for small-study effects have not yet been validated within the RVE framework. Finally, we concluded the analysis with a visual inspection of the scatter plots for the autobiographical-memory functions (Figure 1). There were no signs of anomalies in the data. As would be expected, effects were more variable, but not consistently larger, for smaller studies.

Another potential source of bias is low quality in the primary studies. Not all studies included in our analysis have undergone peer review, so potential errors in experimental design and psychometric measurement may have gone unnoticed. We address four

potential quality issues in the primary studies. (i) It is possible that the experiments were inadequately designed and conducted. However, we observed reliable main effects of the nostalgia inductions, which corresponded to those reported in the peer-reviewed literature (Sedikides, Wildschut, Routledge, & Arndt, 2015). (ii) It is possible that psychometric measurement of the outcomes was inadequate. Yet, across all studies, outcome measurements were highly reliable ( $M_{\alpha} = 0.87$ ,  $Md_{\alpha} = 0.89$ ,  $SD_{\alpha} = 0.14$ ) and sensitive to nostalgia inductions. (iii) Neuroticism measurements may have been inadequate. Still, neuroticism measures had adequate reliability (BFI:  $M_{\alpha} = 0.80$ ,  $Md_{\alpha} = 0.79$ ,  $SD_{\alpha} = 0.05$ ; TIPI:  $M_{\alpha} = 0.68$ ,  $Md_{\alpha} = 0.66$ ,  $SD_{\alpha} = 0.04$ ) and were robustly associated with the outcome variables. (iv) Primary studies could have inadvertently recruited samples that were uncommonly high or low in neuroticism (i.e. producing ceiling or floor effects, respectively). Overall, however, neuroticism scores (on a scale from 1 to 5) fell close to the scale midpoint ( $M = 2.81$ ,  $Md = 2.81$ ,  $SD = 0.14$ ), and there were no signs of range restriction. The overall standard deviation within studies ( $M_{SD} = 0.79$ ,  $Md_{SD} = 0.77$ ) was comparable with standard deviations reported in the literature (e.g.  $SD = 0.82$  in a large study by Srivastava, John, Gosling, & Potter, 2003). It is thus unlikely that neuroticism levels in the included samples were too extreme to detect moderation effects. In summary, we found no reason to suspect that Nostalgia  $\times$  Neuroticism interaction effects were systematically masked or attenuated owing to poor data or study quality.

Finally, the analysis may have insufficient statistical power. Accepting the null hypothesis is only warranted when the power to detect theoretically or practically relevant effect sizes is sufficient. Meta-analyses typically have higher power than primary studies (Borenstein et al., 2009) and should have a high probability of detecting even small effects. Methods to estimate power for RVE meta-analysis are currently unavailable, but we can make an approximation under certain assumptions. Power in

**Table 7.** Statistical power for small, medium, and large effects at the level of nostalgia functions

Type of effect	Function	Power ( $d = 0.2$ )	Power ( $d = 0.5$ )	Power ( $d = 0.8$ )
Nostalgia main effect	Self-oriented	0.99	>0.99	>0.99
	Existential	0.85	>0.99	>0.99
	Social	0.87	>0.99	>0.99
Neuroticism main effect	Self-oriented	0.47	>0.99	>0.99
	Existential	>0.99	>0.99	>0.99
	Social	0.69	>0.99	>0.99
Nostalgia $\times$ Neuroticism	Self-oriented	>0.99	>0.99	>0.99
	Existential	>0.99	>0.99	>0.99
	Social	0.99	>0.99	>0.99

Note: Power is calculated based on two-sided tests using standard errors from RVE models.

conventional meta-analysis model is based on a test statistic  $Z$  for the summary effect, computed as the summary effect divided by the standard error of the summary effect (Borenstein et al., 2009, p. 268). If we assume that  $Z$  follows a standard normal distribution when standard errors from RVE models are entered, we can compute a priori power for small ( $d = 0.2$ ), medium ( $d = 0.5$ ), and large ( $d = 0.8$ ) effects.<sup>2</sup> For example, the standard error for the interaction summary effect for the existential function is 0.03 (Table 2). For a small effect ( $d = 0.2$ ), the corresponding  $Z$  value is  $Z = 6.83$ , and power is  $1 - \beta > .99$ . We summarise results for power analyses at the level of autobiographical-memory functions in Table 7. Power was consistently high. Crucially, power was very high even for small interaction effects.

In addition to power analysis, we conducted an equivalence test for meta-analysis to probe whether the interaction effect is practically equivalent to zero (Rogers, Howard, & Vessey, 1993), where ‘practically equivalent with zero’ was defined as effects that fall in the range between  $d = -0.2$  and  $d = 0.2$  (small effects). The hypothesis of non-equivalence is rejected if the 90% confidence interval around the summary effect includes either the lower ( $d = -0.2$ ) or upper ( $d = 0.2$ ) boundary of this range. For the summary effect of the interaction across all functions ( $d = 0.030$ ), the confidence interval  $CI_{90}[0.084, -0.024]$  does not include either boundary. We therefore conclude that the effect is practically equivalent to zero. These results and the findings from the power analysis are consistent with the conclusion that neuroticism does not moderate the beneficial effects of nostalgia inductions.

## Discussion

Scrutinising the interplay between traits and experimentally induced states is promising in advancing

theory and understanding of person-situation interactions. Yet comprehensive meta-analyses of study-level interactions are rare owing to inherent difficulties in comparing interactive patterns across different studies. We aimed to test the generalisability of nostalgia’s psychological benefits by examining whether they are qualified by trait-level neuroticism. More precisely, we examined whether individuals high (vs. low) on neuroticism derived fewer psychological benefit from nostalgia. In a high-powered meta-analytic test ( $N = 3556$ ,  $m_{\text{functions}} = 116$ ,  $m_{\text{affect}} = 39$ ), we found that neuroticism did not moderate the experimental influence of nostalgia on autobiographical memory functions (i.e. self-oriented, existential, and social) or on positive and negative affect. High statistical power, careful examination of potential bias, and high data quality lend confidence to this conclusion.

Beyond turning to the possibility that the psychological benefits of nostalgia are contingent upon neuroticism, we provided a synthesis of nostalgia’s main effects on said benefits. Although the synthesis was incomplete, as it was limited to studies that included a measure of neuroticism, it was nevertheless consistent with the literature (Ismail et al., 2020; Sedikides, Wildschut, Routledge, & Arndt, 2015). As per our findings, nostalgia’s self-oriented (inspiration and optimism), existential (meaning in life and self-continuity), and social (social connectedness) benefits were small to medium in magnitude and statistically significant. The influence of nostalgia on social action tendencies was small to medium, but not significant. However, this estimate was imprecise owing to the small number of pertinent effect sizes.

We note two other findings. First, the effect of nostalgia on self-esteem was small and marginal ( $d = 0.08$ ,  $p = .052$ ). This is surprising in light of prior evidence for nostalgia’s positive impact on self-esteem (Cheung et al., 2013; Hepper et al., 2012; Stephan et al., 2014; Wildschut et al., 2006, 2010) but suggests that this effect is less robust than previously thought or is highly qualified. Consistent with the latter possibility, an ERT experiment by Cheung et al. (2016) showed that nostalgia increased self-esteem only among individuals who were high in dispositional nostalgia proneness, but not among those low in nostalgia proneness. Second, the meta-analytic effect of nostalgia on negative affect was significant ( $d = 0.220$ ,  $p = .002$ ). This was partly due to three large studies in which participants listened either to a nostalgic or happy song (Table S1); the nostalgic song gave rise to more negative affect than the happy song ( $d_{\text{music}} = 0.51$ ,  $p_{\text{music}} < .001$ ). In ERT studies with a neutral control condition, the nostalgia-induced rise in negative affect was smaller but also significant ( $d_{\text{ERT}} = 0.11$ ,  $p_{\text{ERT}} = .005$ ). To achieve 80% power for detecting an effect of this magnitude (two-tailed,  $\alpha = .05$ ), 2597 participants are required. It is therefore unsurprising that such a small effect would remain undetected in primary studies.



Finally, we conducted a meta-analysis of the neuroticism main effects (i.e. the bivariate correlations of trait-level neuroticism scores with the state-level outcomes that were assessed following the nostalgia manipulation). High neurotics (compared with low neurotics) reported significantly lower self-esteem, inspiration, optimism, self-continuity, meaning in life, and social connectedness. Correlations were the strongest with constructs pertaining to the self-oriented function and weaker for the existential and social functions. Further, neuroticism was associated with less positive affect and more negative affect. These findings should be interpreted with caution, however, because they are based exclusively on studies that experimentally manipulated nostalgia and pertain exclusively to state-level (i.e. transient or momentary) outcomes.

### *Limitations and future directions*

In recent years, the question of generality in the nostalgia literature has attracted increasing attention: Are nostalgia inductions more beneficial to some individuals than to others due to systematic variation in personality traits? Our decision to focus on the Big Five trait of neuroticism was, in part, predicated on prior evidence for the moderating roles of habitual worrying (Verplanken, 2012) and resilience (Wildschut et al., 2019). Worry is a cognitive marker of neuroticism and is positively associated with it (Muris et al., 2005; Segerstrom et al., 2000). Resilience entails reduced vulnerability to stress and, given that such vulnerability is a core facet of neuroticism, resilience is inversely related to neuroticism (Campbell-Sills et al., 2006). Yet whereas previous research directly implicated neuroticism, past findings involving worry and resilience seemingly misalign with ours. This apparent discrepancy has several implications for future research.

First, the large bandwidth of the Big Five traits comes at the cost of fidelity; information is lost as one moves up to hierarchy from specific traits (e.g. habitual worrying and resilience) to domain-level traits (John & Srivastava, 1999). Perhaps, then, the generality of nostalgia's benefits should be explored at lower, more specific levels in the hierarchy of personality descriptors. For example, resilience, rather than merely reflecting the absence of neuroticism, captures flexible and successful adaptation to stress and trauma (Bonanno, 2004; Rutter, 1987). Stressful and traumatic events thus represent trait-expressive situations (Fleeson, 2007) that catalyse the manifestation of trait-level resilience in an individual's thoughts, feelings, and actions. Highlighting the differences between neuroticism and resilience in this regard, Campbell-Sills et al. (2006) demonstrated that high (compared with low) resilience attenuated the link between childhood emotional neglect and current psychiatric symptoms, whereas low (compared with

high) neuroticism did not. Resilient individuals' ability to withstand adversity may derive in part from their capacity to harness positive autobiographical memories so as to self-generate positive emotions in the context of experiences that induce sadness and anxiety (Philippe, Lecours, & Beaulieu-Pelletier, 2009). The capacity, under challenging circumstances, to draw strength from one's memories may explain why a nostalgia induction was more beneficial (and less costly) to forcibly displaced Syrian refugees who were high (compared with low) in resilience (Wildschut et al., 2019). The implication is that, to achieve maximum precision, future research should be concerned not only on specific (rather than domain-level) traits but, simultaneously, with the specific trait-expressive situations in which they are manifested most clearly.

Alternatively, rather than being too general, perhaps our focus was not general enough. Research on the interrelations among the Big Five traits indicates that they are subordinate to two higher-order meta traits: the Big Two (DeYoung, 2006; Digman, 1997). The first, labelled *stability*, captures the Big Five traits of neuroticism (reversed), agreeableness, and conscientiousness. The second, labelled *plasticity*, includes extraversion and openness. They refer, respectively, to the ability 'to maintain stability and avoid disruption in emotional, social, and motivational domains', and 'to explore and engage flexibly with novelty, in both behavior and cognition' (DeYoung, 2006, p. 1138). Although our unequivocal finding that neuroticism did not moderate the benefits of nostalgia inductions casts doubt on a potential role for the higher-order stability factor, it does not rule out this possibility. Still, the plasticity factor may offer a more promising target for future research, for two reasons. First, habitual worrying is indicative of a repetitive and automatic cognitive process (Verplanken et al., 2007), pointing to an inverse relation with plasticity. Resilience, in contrast, reflects flexibility in enhancing and suppressing emotional expression (Bonanno, Papa, Lalande, Westphal, & Coifman, 2004) and is positively associated with extraversion and openness—the constituent domain-level traits of plasticity (Campbell-Sills et al., 2006). Thus, prior evidence pertaining to the dependence of nostalgia effects by habitual worrying (Verplanken, 2012) and resilience (Wildschut et al., 2019) implicates plasticity. Second, examining plasticity may shed light on the finding that nostalgia inductions are more beneficial (and less costly) for individuals who are high (compared with low) in nostalgia proneness. Nostalgia proneness has also been linked with higher levels of both plasticity components: extraversion (Stephan et al., 2014) and openness (Newman, Sachs, Stone, & Schwarz, 2020). The plasticity meta trait, then, offers a tantalising prospect of broad theoretical and empirical integration.

An unanswered question relates to the availability, accessibility, and processing mechanisms that provided the theoretical foundation for the postulated

Nostalgia  $\times$  Neuroticism interaction effect. On the one hand, our failure to detect evidence for this interaction effect casts doubt on the proposed mechanisms. On the other hand, the highly robust neuroticism main effects lend them support, if one assumes (as the data indicate) that high neurotics (compared with low neurotics) were equally impaired when recalling nostalgic and ordinary autobiographical events. Future research could offer a more definitive answer by assessing the three mechanisms—for example, by coding the content and/or emotional tone of retrieved memories.

Our work is not without limitations. To begin, all participants were members of Western cultures. Despite the panculturality of nostalgia *per se* (Hepper et al., 2014), future research will need to test the generalisability of our findings in non-Western cultures. Also, our meta-analysis included mostly younger participants (in total: 30% over 33 years old, 20% over 44, 10% over 53, and 5% over 62; Figure S2). Our findings revealed that age did not moderate the effects of nostalgia or nostalgia's interactive effect with neuroticism, and prior research has suggested that psychological benefits of nostalgia (e.g. well-being) generalise across age (Hepper et al., 2020). Still, follow-up work will need to provide a more fine-grained analysis as to whether our findings are equally applicable to older and younger persons.

Our meta-analysis focused exclusively on studies that implemented experimental inductions of nostalgia. Irrespective of neuroticism, these brief nostalgia inductions had positive immediate effects, but a question arises about the duration of such effects. Recently, researchers have begun to address this question by focusing on implications of nostalgia in naturalistic settings (Kersten, Cox, & Van Enkevort, 2016; Iyer & Jetten, 2011; Newman et al., 2020; Van Dijke, Leunissen, Wildschut, & Sedikides, 2019; Wohl et al., 2018). For example, in a longitudinal study of students entering university, Iyer and Jetten (2011) showed that perceived identity continuity moderated the effects of nostalgia. Students who experienced high identity continuity ('I have maintained strong ties with the same groups I belonged to before coming to university') perceived fewer academic obstacles when nostalgia for their community was high (compared with low). However, when students experienced low identity continuity, they perceived more academic obstacles when nostalgia was high (compared with low). Future work would do well to test systematically moderation hypotheses in experimental and naturalistic contexts for safeguarding both internal and external validity.

### Coda

Nostalgia comprises negative components, such as longing, loss, and wanting to return to the past. Neuroticism entails sensitivity to negativity and is

strongly linked with psychopathology. Nonetheless, nostalgia yields key psychological benefits even for individuals high in neuroticism.


### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed no financial support for the research, authorship, and/or publication of this article.

### Data accessibility statement

 This article earned Open Data and Open materials badges through Open Practices Disclosure from the Center for Open Science: <https://osf.io/tvyxz/wiki>. The data are permanently and openly accessible at <https://osf.io/sfx6h/>. Author's disclosure form may also be found at the Supporting Information in the online version.

### Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Table S1.** Summary of Meta-Moderation Analyses for the Nostalgia Main Effects

**Table S2.** Summary of Meta-Moderation Analyses for the Neuroticism Main Effects

**Table S3.** Meta-Analytic Summary Effects of Partial  $\beta$  at the Level of Nostalgia Functions

**Figure S1.** Flow chart of the study selection process. Additional studies from other sources were identified through personal correspondence, personal archives, and calls through mailing lists.

**Figure S2.** Distribution of participant age. Bin width was set to 5. The two dashed lines represent the 33% (20) and 67% (30) quantiles.

### Notes

1. We repeated the meta-analyses for main effects of neuroticism and nostalgia using partial beta coefficients from the interaction models. We specified the models as described in the Method section, except that we coded the nostalgia manipulation as  $-1$  (*control condition*) and  $1$  (*nostalgia condition*), rather than 0 and 1. We again *z*-standardised neuroticism and outcome variables. We summarise the results in Table S3.
2. See Tipton (2015) for a discussion why this assumption may sometimes be violated.

### References

- Abeyta, A. A., Routledge, C., Roylance, C., Wildschut, T., & Sedikides, C. (2015). Attachment-related avoidance and the social and agentic content of nostalgic memories. *Journal of Social and Personal Relationships*, 32, 406–413. <https://doi.org/10.1002/0265407514533770>.
- Baldwin, M., Biernat, M., & Landau, M. J. (2015). Remembering the real me: Nostalgia offers a window

- to the intrinsic self. *Journal of Personality and Social Psychology*, *108*, 128–147. <https://doi.org/10.1037/a0038033>.
- Baldwin, M., & Landau, M. J. (2014). Exploring nostalgia's influence on psychological growth. *Self and Identity*, *13*, 162–177. <https://doi.org/10.1080/15298868.2013.772320>.
- Barrett, F. S., Grimm, K. J., Robins, R. W., Wildschut, T., Sedikides, C., & Janata, P. (2010). Music-evoked nostalgia: Affect, memory, and personality. *Emotion*, *10*, 390–403. <https://doi.org/10.1037/a0019006>.
- Batcho, K. I. (2007). Nostalgia and the emotional tone and content of song lyrics. *The American Journal of Psychology*, *120*, 361–381.
- Biskas, M., Cheung, W.-Y., Juhl, J., Sedikides, C., Wildschut, T., & Hepper, E. G. (2019). A prologue to nostalgia: Savoring creates nostalgic memories that foster optimism. *Cognition and Emotion*, *33*, 417–427. <https://doi.org/10.1080/02699931.2018.1458705>.
- Boelen, P. A. (2009). The centrality of a loss and its role in emotional problems among bereaved people. *Behaviour Research and Therapy*, *47*, 616–622. <https://doi.org/10.1016/j.brat.2009.03.009>.
- Bonanno, G. A. (2004). Loss, trauma, and human resilience. *American Psychologist*, *59*, 20–28. <https://doi.org/10.1037/0003-066X.59.1.20>.
- Bonanno, G. A., Papa, A., Lalande, K., Westphal, M., & Coifman, K. (2004). The importance of being flexible: The ability to both enhance and suppress emotional expression predicts long-term adjustment. *Psychological Science*, *15*, 482–487. <https://doi.org/10.1111/j.0956-7976.2004.00705.x>.
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009) *Introduction to meta-analysis*. John Wiley & Sons, Ltd: Chichester, DOI: <https://doi.org/10.1002/9780470743386>.
- Campbell-Sills, L., Cohan, S. L., & Stein, M. B. (2006). Relationship of resilience to personality, coping, and psychiatric symptoms in young adults. *Behaviour Research and Therapy*, *44*, 585–599. <https://doi.org/10.1016/j.brat.2005.05.001>.
- Cappeliez, P., & O'Rourke, N. (2002). Personality traits and existential concerns as predictors of the functions of reminiscence in older adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *57*, P116–P123. <https://doi.org/10.1002/geronb/57.2.P116>.
- Cheung, W.-Y., Sedikides, C., & Wildschut, T. (2016). Induced nostalgia increases optimism (via social-connectedness and self-esteem) among individuals high, but not low, in trait nostalgia. *Personality and Individual Differences*, *90*, 283–288. <https://doi.org/10.1016/j.paid.2021.11.028>.
- Cheung, W.-Y., Wildschut, T., Sedikides, C., Hepper, E. G., Arndt, J., & Vingerhoets, A. J. J. M. (2013). Back to the future: Nostalgia increases optimism. *Personality and Social Psychology Bulletin*, *39*, 1484–1496. <https://doi.org/10.1002/0146167213499187>.
- Denissen, J. J. A., Geenen, R., Selfhout, M., & van Aken, M. A. G. (2008). Single-item big five ratings in a social network design. *European Journal of Personality*, *22*, 37–54. <https://doi.org/10.1002/per.662>.
- Denkova, E., Dolcos, S., & Dolcos, F. (2012). Reliving emotional personal memories: Affective biases linked to personality and sex-related differences. *Emotion*, *12*, 515–528. <https://doi.org/10.1037/a0026809>.
- DeYoung, C. G. (2006). Higher-order factors of the big five in a multi-informant sample. *Journal of Personality and Social Psychology*, *91*, 1138–1151. <https://doi.org/10.1037/0022-3514.91.6.1138>.
- Digman, J. M. (1997). Higher-order factors of the big five. *Journal of Personality and Social Psychology*, *73*, 1246–1256. <https://doi.org/10.1037/0022-3514.73.6.1246>.
- Fisher, Z., Tipton, E., & Zhipeng, H. (2017). *robumeta: Robust Variance Meta-Regression*. R package version 2.0. <https://CRAN.R-project.org/package=robumeta>.
- Fleeson, W. (2007). Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of Personality*, *75*, 825–861. <https://doi.org/10.1111/j.1467-6494.2007.00458.x>.
- Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality*, *37*, 504–528. [https://doi.org/10.1016/S0092-6566\(03\)00046-1](https://doi.org/10.1016/S0092-6566(03)00046-1).
- Gray, J. A. (1981A) A critique of Eysenck's theory of personality. In H. J. Eysenck (Ed.), *A model for personality*. (pp. 246–276). Springer: Berlin, DOI: [https://doi.org/10.1007/978-3-642-67783-0\\_8](https://doi.org/10.1007/978-3-642-67783-0_8).
- Hamann, S., & Canli, T. (2004). Individual differences in emotion processing. *Current Opinion in Neurobiology*, *14*, 233–238. <https://doi.org/10.1016/j.conb.2004.03.010>.
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*, 39–65. <https://doi.org/10.1002/jrsm>.
- Hepper, E. G., Ritchie, T. D., Sedikides, C., & Wildschut, T. (2012). Odyssey's end: Lay conceptions of nostalgia reflect its original Homeric meaning. *Emotion*, *12*, 102–119. <https://doi.org/10.1037/a0025167>.
- Hepper, E. G., Wildschut, T., Sedikides, C., Ritchie, T. D., Yung, Y.-F., Hansen, N., ..., Zhou, X. (2014). Pancultural nostalgia: Prototypical conceptions across cultures. *Emotion*, *14*, 733–747. <https://doi.org/10.1037/a0036790>.
- Hepper, E. G., Wildschut, T., Sedikides, C., Robertson, S., & Routledge, C. (2020). The time capsule: Nostalgia shields wellbeing from limited time horizons. *Emotion*. Advance online publication. <https://doi.org/10.1037/emo0000728>.
- Hertz, D. G. (1990). Trauma and nostalgia: New aspects on the coping of aging holocaust survivors. *Israel Journal of Psychiatry and Related Sciences*, *27*, 189–198.
- Higgins, J. P., & Green, S. (2011) *Cochrane handbook for systematic reviews of interventions 4*. Hoboken: John Wiley & Sons.
- Huo-Liang, G. (2006). Personality and crime: A meta-analysis of studies on criminals' personality. *Chinese Mental Health Journal*, *20*, 465–468.
- Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–648. <https://doi.org/10.1097/EDE.0b013e31818131e7>.
- Ismail, S. U., Cheston, R., Christopher, G., & Meyrick, J. (2020). Nostalgia as a psychological resource for people with dementia: A systematic review and meta-analysis of evidence of effectiveness from experimental studies. *Dementia*, *19*, 330–351. <https://doi.org/10.1002/1471301218774909>.
- Iyer, A., Jetten J. (2011). What's left behind: Identity continuity moderates the effect of nostalgia on well-being and life choices.. *Journal of Personality and Social Psychology*, *101*, 1037–1051. <https://doi.org/10.1037/a0026809>.



- Psychology*, 101, (1), 94–108. <https://doi.org/10.1037/a0022496>.
- John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research*. (3rd ed., pp. 114–158). New York: Guilford.
- John, O. P., & Srivastava, S. (1999). The big-five trait taxonomy: History, measurement, and theoretical perspectives. In L. A., Pervin, & O. P., John (Eds.), *Handbook of personality: Theory and research*. (pp. 102–138), 2. New York: Guilford Press.
- Kersten, M., Cox, C. R., & Van Enkevort, E. A. (2016). An exercise in nostalgia: Nostalgia promotes health optimism and physical activity. *Psychology & Health*, 31, 1166–1181. <https://doi.org/10.1080/08870446.2016.1185524>.
- Kunzendorf, R. G., Moran, C., & Gray, R. (1995). Personality traits and reality-testing abilities, controlling for vividness of imagery. *Imagination, Cognition and Personality*, 15, 113–131. <https://doi.org/10.2190/B76E-MJ9E-07AV-KAKK>.
- Layous, K., Kurtz, J. L., Wildschut, T., & Sedikides, C. (2020). The effect of a multi-week nostalgia intervention on well-being: Mechanisms and moderation. *Manuscript under review*. , California State University, East Bay.
- Lyubomirsky, S., & Layous, K. (2013). How do simple positive activities increase well-being? *Current Directions in Psychological Science*, 22, 57–62. <https://doi.org/10.1002/0963721412469809>.
- Madoglou, A., Gkinopoulos, T., Xanthopoulos, P., & Kalamaras, D. (2017). Representations of autobiographical nostalgic memories: Generational effect, gender, nostalgia proneness and communication of nostalgic experiences. *Journal of Integrated Social Sciences*, 7, 60–88
- Malouff, J. M., Thorsteinsson, E. B., & Schutte, N. S. (2005). The relationship between the five-factor model of personality and symptoms of clinical disorders: A meta-analysis. *Journal of Psychopathology and Behavioral Assessment*, 27, 101–114. <https://doi.org/10.1007/s10862-005-5384-y>.
- McAdams, D. P., Reynolds, J., Lewis, M., Patten, A. H., & Bowman, P. J. (2001). When bad things turn good and good things turn bad: Sequences of redemption and contamination in life narrative and their relation to psychosocial adaptation in midlife adults and in students. *Personality and Social Psychology Bulletin*, 27, 474–485. <https://doi.org/10.1002/0146167201274008>.
- McCrae, R. R., & Costa, P. T. (2003) *Personality in adulthood: A five-factor theory perspective*. Guilford: New York, DOI: <https://doi.org/10.4324/9780203428412>.
- Muris, P., Roelofs, J., Rassin, E., Franken, I., & Mayer, B. (2005). Mediating effects of rumination and worry on the links between neuroticism, anxiety and depression. *Personality and Individual Differences*, 39, 1105–1111. <https://doi.org/10.1016/j.paid.2005.04.005>.
- Nash, J. E. (2012). Ringing the chord: Sentimentality and nostalgia among male singers. *Journal of Contemporary Ethnography*, 41, 581–606. <https://doi.org/10.1002/0891241611429943>.
- Newman, D. B., Sachs, M. E., Stone, A. A., & Schwarz, N. (2020). Nostalgia and well-being in daily life: An ecological validity perspective. *Journal of Personality and Social Psychology*, 118, 325–347. <https://doi.org/10.1037/pspp0000236>.
- Ogle, C. M., Siegler, I. C., Beckham, J. C., & Rubin, D. C. (2017). Neuroticism increases PTSD symptom severity by amplifying the emotionality, rehearsal, and centrality of trauma memories. *Journal of Personality*, 85, 702–715. <https://doi.org/10.1111/jopy.12278>.
- Philippe, F. L., Lecours, S., & Beaulieu-Pelletier, G. (2009). Resilience and positive emotions: Examining the role of emotional memories. *Journal of Personality*, 77, 139–175. <https://doi.org/10.1111/j.1467-6494.2008.00541.x>.
- R Core Development Team. (2017). R: A language and environment for statistical computing(), Computer software. Vienna, Austria R Foundation for Statistical Computing.
- Raggatt, P. (2006). Putting the five-factor model into context: Evidence linking big five traits to narrative identity. *Journal of Personality*, 74, 1321–1348. <https://doi.org/10.1111/j.1467-6494.2006.00411.x>.
- Reid, C. A., Green, J. D., Wildschut, T., & Sedikides, C. (2015). Scent-evoked nostalgia. *Memory*, 23, 157–166. <https://doi.org/10.1080/09658211.2013.876048>.
- Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *BMJ*, 340, c221. <https://doi.org/10.1136/bmj.c221>.
- Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, 113, 553–565. <https://doi.org/10.1037/0033-2909.113.3.553>.
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press, <https://doi.org/10.1515/9781400876136>.
- Routledge, C., Arndt, J., Sedikides, C., & Wildschut, T. (2008). A blast from the past: The terror management function of nostalgia. *Journal of Experimental Social Psychology*, 44, 132–140. <https://doi.org/10.1016/j.jesp.2006.11.001>.
- Routledge, C., Arndt, J., Wildschut, T., Sedikides, C., Hart, C. M., Juhl, J., Vingerhoets, A. J. M. (2011). The past makes the present meaningful: Nostalgia as an existential resource. *Journal of Personality and Social Psychology*, 101, 638–652. <https://doi.org/10.1037/a0024292>.
- Routledge, C., Wildschut, T., Sedikides, C., Juhl, J., & Arndt, J. (2012). The power of the past: Nostalgia as a meaning-making resource. *Memory*, 20, 452–460. <https://doi.org/10.1080/09658211.2012.677452>.
- Rubin, D. C., Boals, A., & Hoyle, R. H. (2014). Narrative centrality and negative affectivity: Independent and interactive contributors to stress reactions. *Journal of Experimental Psychology: General*, 143, 1159–1170. <https://doi.org/10.1037/a0035140>.
- Rubin, D. C., Dennis, M. F., & Beckham, J. C. (2011). Autobiographical memory for stressful events: The role of autobiographical memory in posttraumatic stress disorder. *Consciousness and Cognition*, 20, 840–856. <https://doi.org/10.1016/j.concog.2011.03.015>.
- Rusting, C. L., & Larsen, R. J. (1998). Personality and cognitive processing of affective information. *Personality and Social Psychology Bulletin*, 24, 200–213. <https://doi.org/10.1002/0146167298242008>.
- Rutter, M. (1987). Psychosocial resilience and protective mechanisms. *American Journal of Orthopsychiatry*, 57, 316–331. <https://doi.org/10.1111/j.1939-0025.1987.tb03541.x>.



- Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the life orientation test. *Journal of Personality and Social Psychology*, *67*, 1063–1078. <https://doi.org/10.1037/t01038-000>.
- Sedikides, C., & Wildschut, T. (2016). Nostalgia: A bitter-sweet emotion that confers psychological health benefits. In A. M., Wood & J., Johnson (Eds.), *Wiley handbook of positive clinical psychology*. (pp. 25–36). Wiley: Hoboken, NJ, DOI: <https://doi.org/10.1002/9781118468197.ch9>.
- Sedikides, C., & Wildschut, T. (2018). Finding meaning in nostalgia. *Review of General Psychology*, *22*, 48–61. <https://doi.org/10.1037/gpr0000109>.
- Sedikides, C., & Wildschut, T. (2019). The sociality of personal and collective nostalgia. *European Review of Social Psychology*, *30*, 123–173. <https://doi.org/10.1080/10463283.2019.1630098>.
- Sedikides, C., & Wildschut, T. (2020). The motivational potency of nostalgia: The future is called yesterday. *Advances in Motivation Science*, *7*, 75–111. <https://doi.org/10.1016/bs.adms.2019.05.001>.
- Sedikides, C., Wildschut, T., & Baden, D. (2004). Nostalgia: Conceptual issues and existential functions. In J., Greenberg, S., Koole, & T., Pyszczynski (Eds.), *Handbook of experimental existential psychology*. (pp. 200–214). New York, NY: Guilford Press.
- Sedikides, C., Wildschut, T., Cheung, W.-Y., Routledge, C., Hepper, E. G., Arndt, J., . . . , Vingerhoets, A. J. (2016). Nostalgia fosters self-continuity: Uncovering the mechanism (social connectedness) and consequence (eudaimonic well-being). *Emotion*, *16*, 524–539. <https://doi.org/10.1037/emo0000136>.
- Sedikides, C., Wildschut, T., Routledge, C., & Arndt, J. (2015). Nostalgia counteracts self-discontinuity and restores selfcontinuity. *European Journal of Social Psychology*, *45*, 52–61. <https://doi.org/10.1002/ejsp.2073>.
- Sedikides, C., Wildschut, T., Routledge, C., Arndt, J., Hepper, E. G., & Zhou, X. (2015). To nostalgize: Mixing memory with affect and desire. *Advances in Experimental Social Psychology*, *51*, 189–273. <https://doi.org/10.1016/bs.aesp.2014.10.001>.
- Sedikides, C., Wildschut, T., & Stephan, E. (2018). Nostalgia shapes and potentiates the future. In J. P., Forgas & R. F., Baumeister (Eds.), *The social psychology of living well*. (pp. 181–199). Routledge: New York, NY, DOI: <https://doi.org/10.4324/9781351189712-11>.
- Segerstrom, S. C., Tsao, J. C. I., Alden, L. E., & Craske, M. G. (2000). Worry and rumination: Repetitive thought as a concomitant and predictor of negative mood. *Cognitive Therapy and Research*, *24*, 671–688. <https://doi.org/10.1023/A:1005587311498>.
- Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change?. *Journal of Personality and Social Psychology*, *84*, 1041–1053. <https://doi.org/10.1037/0022-3514.84.5.1041>.
- Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin*, *134*, 138–161. <https://doi.org/10.1037/0033-2909.134.1.138>.
- Steger, M. F., Frazier, P., Oishi, S., & Kaler, M. (2006). The meaning in life questionnaire: Assessing the presence of and search for meaning in life. *Journal of Counseling Psychology*, *53*, 80–93. <https://doi.org/10.1037/0022-0167.53.1.80>.
- Stephan, E., Sedikides, C., & Wildschut, T. (2012). Mental travel into the past: Differentiating recollections of nostalgic, ordinary, and positive events. *European Journal of Social Psychology*, *42*, 290–298. <https://doi.org/10.1002/ejsp.1865>.
- Stephan, E., Wildschut, T., Sedikides, C., Zhou, X., He, W., Routledge, C., . . . , Vingerhoets, A. J. J. M. (2014). The mnemonic mover: Nostalgia regulates avoidance and approach motivation. *Emotion*, *14*, 545–561. <https://doi.org/10.1037/a0035673>.
- Sterne, J. A. C., & Egger, M. (2005) Regression methods to detect publication and other bias in meta-analysis. In H. R., Rothstein, A. J., Sutton, & M., Borenstein (Eds.), *Publication bias in metaanalysis: Prevention, assessment and adjustments*. (pp. 99–110). Wiley: New York, DOI: <https://doi.org/10.1002/0470870168.ch6>.
- Supski, S. (2013). Aunty Sylvie's sponge: Foodmaking, cookbooks and nostalgia. *Cultural Studies Review*, *19*, 28–49. <https://doi.org/10.5130/csr.v19i1.3074>.
- Sutin, A. R. (2008). Autobiographical memory as a dynamic process: Autobiographical memory mediates basic tendencies and characteristic adaptations. *Journal of Research in Personality*, *42*, 1060–1066. <https://doi.org/10.1016/j.jrp.2007.10.002>.
- Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods*, *5*, 13–30. <https://doi.org/10.1002/jrsm.1091>.
- Thomsen, D. K., Olesen, M. H., Schnieber, A., & Tonnesvang, J. (2014). The emotional content of life stories: Positivity bias and relation to personality. *Cognition & Emotion*, *28*, 260–277. <https://doi.org/10.1080/02699931.2013.815155>.
- Thrash, T. M., & Elliot, A. J. (2003). Inspiration as a psychological construct. *Journal of Personality and Social Psychology*, *84*, 871–889. <https://doi.org/10.1037/0022-3514.84.4.871>.
- Tipton, E. (2013). Robust variance estimation in meta-regression with binary dependent effects. *Research Synthesis Methods*, *4*, 169–187. <https://doi.org/10.1002/jrsm.1070>.
- Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*, 375–393. <https://doi.org/10.1037/met0000011>.
- Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, *40*, 604–634. <https://doi.org/10.3102/1076998615606099>.
- Turner, R. N., Wildschut, T., & Sedikides, C. (2012). Dropping the weight stigma: Nostalgia improves attitudes toward persons who are overweight. *Journal of Experimental Social Psychology*, *48*, 130–137. <https://doi.org/10.1016/j.jesp.2011.09.007>.
- Turner, R. N., Wildschut, T., Sedikides, C., & Gheorghiu, M. (2013). Combating the mental health stigma with nostalgia. *European Journal of Social Psychology*, *43*, 413–422. <https://doi.org/10.1002/ejsp.1952>.
- Van Dijke, M., Leunissen, J. M., Wildschut, T., & Sedikides, C. (2019). Nostalgia promotes intrinsic

- motivation and effort in the presence of low interaction justice. *Organizational Behavior and Human Decision Processes*, 150, 46–61. <https://doi.org/10.1016/j.obhdp.2018.12.003>.
- Van Dijke, M., Wildschut, T., Leunissen, J. M., & Sedikides, C. (2015). Nostalgia buffers the negative impact of low procedural justice on cooperation. *Organizational Behavior and Human Decision Processes*, 127, 15–29. <https://doi.org/10.1016/j.obhdp.2014.11.005>.
- Van Tilburg, W. A. P., Bruder, M., Wildschut, T., Sedikides, C., & Göritz, A. S. (2019). An appraisal profile of nostalgia. *Emotion*, 19, 21–36. <https://doi.org/10.1037/emo0000417>.
- Van Tilburg, W. A. P., Igou, E. R., & Sedikides, C. (2013). In search of meaningfulness: Nostalgia as an antidote to boredom. *Emotion*, 13, 450–461. <https://doi.org/10.1037/a0030442>.
- Van Tilburg, W. A. P., Sedikides, C., & Wildschut, T. (2018). Adverse weather evokes nostalgia. *Personality and Social Psychology Bulletin*, 44, 984–995. <https://doi.org/10.1002/0146167218756030>.
- Van Tilburg, W. A. P., Sedikides, C., Wildschut, T., & Vingerhoets, A. J. J. M. (2019). How nostalgia infuses life with meaning: From social connectedness to self-continuity. *European Journal of Social Psychology*, 49, 521–532. <https://doi.org/10.1002/ejsp.2519>.
- Verplanken, B. (2012). When bittersweet turns sour: Adverse effects of nostalgia on habitual worriers: Habitual worrying and nostalgia. *European Journal of Social Psychology*, 42, 285–289. <https://doi.org/10.1002/ejsp.1852>.
- Verplanken, B., Friborg, O., Wang, C. E., Trafimow, D., & Woolf, K. (2007). Mental habits: Metacognitive reflection on negative self-thinking. *Journal of Personality and Social Psychology*, 92, 526–541. <https://doi.org/10.1037/0022-3514.92.3.526>.
- Wagnild, G. M., & Young, H. M. (1993). Development and psychometric evaluation of the resilience scale. *Journal of Nursing Measurement*, 1, 165–178. <https://doi.org/10.1186/1756-0500-4-509>.
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54, 1063–1070. <https://doi.org/10.1037/0022-3514.54.6.1063>.
- Wildschut, R. T., & Sedikides, C. (in press). The measurement of nostalgia. In W., Ruch, A. B., Bakker, L., Tay, & F., Gander (Eds.), *Handbook of positive psychology assessment*. Hogrefe: Göttingen.
- Wildschut, T., Sedikides, C., Arndt, J., & Routledge, C. (2006). Nostalgia: Content, triggers, functions. *Journal of Personality and Social Psychology*, 91, 975–993. <https://doi.org/10.1037/0022-3514.91.5.975>.
- Wildschut, T., Sedikides, C., Routledge, C., Arndt, J., & Cordaro, F. (2010). Nostalgia as a repository of social connectedness: The role of attachment-related avoidance. *Journal of Personality and Social Psychology*, 98, 573–586. <https://doi.org/10.1037/a0017597>.
- Wildschut, T., Sedikides, C., & Alowidy, D. (2019). Hanin: Nostalgia among Syrian refugees. *European Journal of Social Psychology*, 49, (7), 1368–1384. <https://doi.org/10.1002/ejsp.2590>.
- Wohl, M. J., Kim, H. S., Salmon, M., Santesso, D., Wildschut, T., & Sedikides, C. (2018). Discontinuity-induced nostalgia improves the odds of a self-reported quit attempt among people living with addiction. *Journal of Experimental Social Psychology*, 75, 83–94. <https://doi.org/10.1016/j.jesp.2017.11.011>.
- Zhou, X., Sedikides, C., Wildschut, T., & Gao, D. G. (2008). Counteracting loneliness: On the restorative function of nostalgia. *Psychological Science*, 19, 1023–1029. <https://doi.org/10.1111/j.1467-9280.2008.02194.x>.
- Zhou, X., Wildschut, T., Sedikides, C., Shi, K., & Feng, C. (2012). Nostalgia: The gift that keeps on giving. *Journal of Consumer Research*, 39, 39–50. <https://doi.org/10.1086/662199>.

# Corrigendum to: Does neuroticism disrupt the psychological benefits of nostalgia? a meta-analytic test

European Journal of Personality  
2021, Vol. 35(5) 772  
© The Author(s) 2021  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/08902070211026136  
journals.sagepub.com/home/ejop  


Frankenbach J, Wildschut T, Juhl J and Sedikides C (2021) Does neuroticism disrupt the psychological benefits of nostalgia? a meta-analytic test. *European Journal of Personality*, 35: 249–266. DOI: 10.1002/per.2276

Tables 3, 4, and 5 in the above article contained errors in the printed and original online issue version.

The row labels for ‘Positive Affect’ and ‘Negative Affect’ were reversed owing to an error in the code that generated the tables. The error only affects the presentation in the tables, the authors report and interpret the correct values in the text of the article. The corrected tables appear below.

**Table 3.** Summary effects for the main effects of the nostalgia manipulation by outcome subcategory.

Subcategory	<i>d</i>	SE	LL	UL	<i>t</i>	<i>df</i>	<i>P</i>	<i>k</i>	<i>m</i>
Self-esteem	0.08	0.04	−0.00	0.17	2.16	12.11	.052	17	25
Optimism	0.22	0.07	0.07	0.36	3.34	9.58	.008	12	17
Inspiration	0.29	0.07	0.13	0.46	4.24	6.39	.005	8	8
Meaning in life	0.32	0.06	0.18	0.46	5.21	8.19	.001	15	17
Self-continuity	0.44	0.10	0.21	0.67	4.23	10.94	.001	16	18
Social action tendencies	0.37	0.18	−0.23	0.97	2.09	2.66	.139	4	6
Social connectedness	0.34	0.07	0.19	0.49	5.03	9.74	.001	14	25
Negative affect	0.22	0.06	0.09	0.35	3.62	14.02	.003	17	17
Positive affect	0.22	0.06	0.09	0.35	3.73	14.37	.002	17	22

Note: *d*, summary effect size; LL, lower limit of the 95% confidence interval (CI); UL, upper limit of the 95% CI; *t*, *t*-value associated with the *d*-value in the same row testing statistical significance in the respective subcategory; *p*, *p*-value associated with the *t*-value in the same row; *df*, degrees of freedom associated with the *t*-value in the same row; *k*, number of studies in the respective subcategory; *m*, number of effect sizes available for the respective subcategory.

**Table 4.** Summary effects for the main effects of neuroticism by outcome subcategory

Subcategory	<i>d</i>	SE	LL	UL	<i>t</i>	<i>df</i>	<i>P</i>	<i>k</i>	<i>m</i>
Self-esteem	−0.80	0.11	−1.04	−0.56	−7.19	12.52	<.001	17	25
Optimism	−0.50	0.11	−0.76	−0.25	−4.39	9.76	.001	12	17
Inspiration	−0.16	0.07	−0.33	0.01	−2.32	6.62	.055	8	8
Meaning in life	−0.27	0.05	−0.39	−0.16	−5.48	7.63	.001	15	17
Self-continuity	−0.23	0.05	−0.34	−0.12	−4.45	11.24	.001	16	18
Social action tendencies	−0.10	0.13	−0.54	0.34	−0.74	2.73	.517	4	6
Social connectedness	−0.33	0.09	−0.53	−0.12	−3.57	10.04	.005	14	25
Negative affect	0.67	0.13	0.39	0.96	5.02	15.35	<.001	17	17
Positive affect	−0.38	0.07	−0.52	−0.24	−5.78	15.53	<.001	17	22

Note: *d*, summary effect size; LL, lower limit of the 95% confidence interval (CI); UL, upper limit of the 95% CI; *t*, *t*-value associated with the *d*-value in the same row testing statistical significance in the respective subcategory; *p*, *p*-value associated with the *t*-value in the same row; *df*, degrees of freedom associated with the *t*-value in the same row; *m*, number of effect sizes available for the respective subcategory.

**Table 5.** Summary effects for the Nostalgia × Neuroticism interaction effects by outcome subcategory

Subcategory	<i>d</i>	SE	LL	UL	<i>t</i>	<i>df</i>	<i>P</i>	<i>k</i>	<i>m</i>
Self-esteem	0.06	0.05	−0.06	0.18	1.10	10.70	.294	17	25
Optimism	0.06	0.04	−0.04	0.15	1.40	8.93	.195	12	17
Inspiration	0.02	0.05	−0.11	0.14	0.32	5.18	.759	8	8
Meaning in life	0.01	0.03	−0.07	0.09	0.33	8.50	.746	15	17
Self-continuity	−0.00	0.04	−0.08	0.08	−0.06	9.71	.953	16	18
Social action tendencies	−0.04	0.07	−0.29	0.20	−0.63	2.65	.580	4	6
Social connectedness	0.05	0.05	−0.07	0.17	0.91	7.83	.389	14	25
Negative affect	0.03	0.04	−0.06	0.11	0.69	12.19	.502	17	17
Positive affect	0.05	0.05	−0.07	0.16	0.84	12.53	.414	17	22

Note: *d*, summary effect size; LL, lower limit of the 95% confidence interval (CI); UL, upper limit of the 95% CI; *t*, *t*-value associated with the *d*-value in the same row testing statistical significance in the respective subcategory; *p*, *p*-value associated with the *t*-value in the same row; *df*, degrees of freedom associated with the *t*-value in the same row; *m*, number of effect sizes available for the respective subcategory.

The authors apologise for this error.

**Part II, Paper 2: “Sex drive: Theoretical conceptualization and meta-analytic review of gender differences.”**

## Sex Drive: Theoretical Conceptualization and Meta-Analytic Review of Gender Differences

Julius Frankenbach<sup>1</sup>, Marcel Weber<sup>1</sup>, David D. Loschelder<sup>2</sup>, Helena Kilger<sup>1</sup>, and Malte Friese<sup>1</sup>

<sup>1</sup> Department of Psychology, Saarland University

<sup>2</sup> Institute of Management & Organizations, Leuphana University of Lüneburg

Few spheres in life are as universally relevant for (almost) all individuals past puberty as sexuality. One important aspect of sexuality concerns individuals' sex drive—their dispositional sexual motivation. A vigorous scientific (and popular) debate revolves around the question of whether or not there is a gender difference in sex drive. Several theories predict a higher sex drive in men compared to women, with some theories attributing this difference to biased responding rather than true differences. Currently, there is little consensus on how to conceptualize sex drive, nor does a quantitative summary of the literature exist. In this paper, we present a theory-driven conceptualization of sex drive as the density distribution of state sex drive, where state sex drive is defined as momentary sexual motivation that manifests in sexual cognition, affect, and behavior. We conduct a comprehensive meta-analysis of gender differences in sex drive based on 211 studies, 856 effect sizes, and 621,463 persons. The meta-analysis revealed a stronger sex drive in men compared to women, with a medium-to-large effect size ( $g = 0.69$ ,  $CI_{95} [0.58, 0.81]$ ). Men more often think and fantasize about sex, more often experience sexual affect like desire, and more often engage in masturbation than women. Adjustment for biased responding reduced the gender difference ( $g = 0.54$ ). Moderation analyses suggest that the effect is robust and largely invariant to contextual factors. There was no evidence of publication bias. The discussion focuses on validity considerations, limitations, and implications for psychological theory and people's everyday lives.

Word count: 249

Keywords: sexual motivation, individual differences, sexual thoughts, sexual desire, masturbation

This paper has been accepted for publication in *Psychological Bulletin*.

© 2022, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1037/bul0000366

Julius Frankenbach <https://orcid.org/0000-0002-1627-8803>

Marcel Weber <https://orcid.org/0000-0002-0507-0070>

David D. Loschelder <https://orcid.org/0000-0001-5818-0559>

Helena Kilger <https://orcid.org/0000-0003-4165-3748>

Malte Friese <https://orcid.org/0000-0003-0055-513X>

We have no conflicts of interest to disclose. The study was registered with PROSPERO ([https://www.crd.york.ac.uk/prospéro/display\\_record.php?RecordID=72894](https://www.crd.york.ac.uk/prospéro/display_record.php?RecordID=72894)). Data, code, and materials are openly available at <https://osf.io/h4jbx/>. We would like to thank our editor Daphna Oyserman as well as Roy Baumeister, Paul Eastwick, and an anonymous reviewer for their stimulating comments that helped to improve this article.

Correspondence concerning this article should be addressed to Julius Frankenbach or Malte Friese, Department of Psychology, Saarland University, Campus A2 4, 66123 Saarbrücken, Germany. E-mail: [julius.frankenbach@gmail.com](mailto:julius.frankenbach@gmail.com) or [malte.friese@uni-saarland.de](mailto:malte.friese@uni-saarland.de)

Past puberty, few spheres in human life are as universally relevant as sexuality. Sexual experiences can bring about intense emotions, positive and negative alike. They can create intense intimacy, feelings of lust and love, but also sadness and anger. Sex can deepen or destroy romantic relationships. In short, sexuality impacts people's everyday lives in myriad ways by influencing their thinking, feeling, and behavior.

A crucial aspect of human sexuality concerns individuals' sex drive. People arguably differ in their dispositional sexual motivation. As a result, scientific research and party conversations alike have long been drawn to the question of whether there is a gender difference in human sex drive.<sup>1</sup> In fact, this question has spurred a vigorous debate, with some authors claiming that there is compelling evidence that men have a stronger sex drive than women (for a review, see Baumeister et al., 2001). Others doubt the validity of this evidence and assume that empirical differences are due to various biasing factors. For example, some have argued that empirically observed gender differences in sex drive may be attributable to factors that are not specific to sex drive, such as socially desirable self-presentation tendencies or a problematic choice of sex drive indicators (Conley et al., 2011; Dawson & Chivers, 2014).

The current research seeks to make two contributions. First, we provide a conceptualization of sex drive grounded in psychological theorizing that has clear implications for what constitutes an adequate indicator of sex drive and what does not. Second, based on this theoretical conceptualization, we conduct a comprehensive meta-analytic review of gender differences in sex drive. Our aims with this meta-analysis are threefold: First, we seek to quantify the overall effect: Do men and women differ in sex drive, and if so, in which direction, and how strongly? Second, we address critical validity concerns. Could potential gender differences in sex drive be the result of biased responding or methodological artifacts (e.g., how participants were compensated or whether the study was advertised as a study on sexuality)? If bias is present, to what degree can it account for potential gender differences? Finally, we investigate the issue of generality. If gender differences exist, do they vary depending on other individual characteristics, such as age, sexual orientation, or relationship status?

The question of whether there is a gender difference in sex drive has substantial practical and theoretical implications. In monogamous relationships, differences in sex drive within a

couple may manifest in sexual desire discrepancy, leading to an interdependence dilemma in which both partners may feel that they need to leave their comfort zone. Partners with a lower sex drive may engage in sex more often than they would like; partners with a stronger sex drive may end up having sex less often than they would like. As a result, both partners may question their compatibility with their partner in a potentially key aspect of their relationship. Not all couples manage to successfully resolve this interdependence dilemma and avoid its negative consequences (Day et al., 2015). Although research on differences in sex drive and resultant desire discrepancies in intimate relationships is still scant, preliminary findings suggest that it may lead to negative consequences such as increased conflict, reduced relationship satisfaction, and lower relationship stability (Mark, 2015). Given that many relationships are heterosexual, knowledge about a potential gender difference in sex drive is of great concern.

If an average gender difference in sex drive large enough for people to discern in their everyday lives were to exist, societies would be likely to pick up this difference and incorporate it into their typical gender roles, which are socially shared (Eagly & Wood, 1999). Also, people may form beliefs about characteristics that they perceive to typically go along with a stronger versus weaker sex drive and make corresponding inferences about members of each gender. In this way, gender roles may shape expectations about gender-typical communication, interaction patterns, and behavior in interpersonal relationships, thereby potentially reinforcing and bolstering the gender difference beyond factually existing differences.

Aside from these and other practical implications, on which we elaborate in the discussion, the question of whether there is a true gender difference in sex drive also has pronounced theoretical implications: Whole theories are built on the assumption that such a gender difference exists. For example, a fundamental premise of sexual economics theory is that men have a stronger sex drive compared to women (Baumeister & Vohs, 2004). In an analogy to an economic market, men's stronger sex drive places them in the societal role of 'buyers', who invest resources to acquire sex from women, who take on the role of 'sellers' in this theory. If the premise that men have a stronger sex drive than women have does not hold, the theory no longer has ground to stand on and would need to be abandoned or considerably revised.

<sup>1</sup> We note that the term "gender" typically refers to whether people self-identify as men, women, or other, whereas "sex" typically refers to biological sex assigned at birth. Using these terms with precision is especially important when making a distinction between acculturative versus biological factors. For sex drive, we presume that both biological and cultural influences

may be at play. Thus, both the term "sex differences" and the term "gender differences" may apply. For simplicity, we refer to differences between men and women as "gender differences", although this is not meant to imply that any alleged differences are solely caused by cultural influences.

## Previous Empirical Evidence for Gender Differences in Sex Drive

Twenty years ago, a widely received narrative review addressed the question whether men and women differ in their sex drive (Baumeister et al., 2001). The authors narratively reviewed a large number of outcomes indicative of sex drive, including thoughts and fantasies, spontaneous arousal, desired frequency of sex, the desired number of sex partners, masturbation, willingness to forego sex, initiating versus refusing sex, enjoyment of various sexual practices, sacrificing resources to get sex, favorable attitudes toward sex, the prevalence of low sexual desire, and self-rated sex drive. Men showed evidence of stronger sexual motivation on each of these indicators. The authors concluded that men have a stronger sex drive than women, a pivotal finding that has since been incorporated into theorizing in related fields (e.g., de Ridder et al., 2012; Schmitt, 2005).

The review by Baumeister and colleagues (2001) strongly focuses on directional evidence to shed light on the question of whether one gender has a stronger sex drive than the other. With the present work, we aim to reconsider these findings quantitatively and extend them by addressing some questions left open by the review's narrative nature, especially regarding the extent of gender differences as well as moderating factors that may influence when the gender difference is more versus less pronounced. In addition, 20 years have passed since the publication of this review. Cultural changes in some societies may have altered average levels of sex drive, how women and men typically respond to questions indicative of sex drive, or both.

Despite a host of research providing directional evidence on possible gender differences in sex drive, studies that explicitly seek to quantify this difference are relatively rare. For instance, Ostovich and Sabini (2004) proposed a four-item scale to assess sex drive, asking participants how often they experienced sexual desire, orgasmed, masturbated, and how they would compare their sex drive to the average person of the same age and gender. They examined gender differences and found a large effect, according to common conventions (Cohen, 1988), indicating a stronger sex drive in men than in women ( $d = 1.20$ ). In three studies (total  $N > 3,600$ ), Lippa (2006) found gender differences in the same direction ranging from  $d = 0.58$  to  $d = 0.84$ . In two of these studies, participants responded to a single item ("I have a strong sex drive"), while a third study included four additional items ("I frequently think about sex"; "It doesn't take much to get me sexually

excited"; "I think about sex almost every day"; "Sexual pleasure is the most intense pleasure a person can have"). The largest single published study ( $N > 200,000$  across 53 nations; Lippa, 2009) asked two questions to assess sex drive ("I have a strong sex drive; It doesn't take much to get me sexually excited") and found an average gender difference of  $d = 0.62$ .

Taken together, then, evidence from these and other studies seeking to quantify a potential gender difference in sex drive suggests a moderate-to-large gender difference, with men having a stronger sex drive than women.

An important observation is that the previous review of the literature (Baumeister et al., 2001) included a diverse array of outcomes as indicators of sex drive. These outcomes cover a considerably broader theoretical scope than the rather focused attempts to directly assess sex drive in the just reviewed studies. At the same time, not all of these outcomes may be equally valid. Even the comparatively focused studies that sought to measure sex drive directly did so in quite different ways. Sometimes a scale comprising several items was used, sometimes only two items or even a single item. In some studies, cognitive aspects such as thoughts and fantasies about sex were focal, in others affective aspects such as sexual desire. These domain-specific estimates were occasionally accompanied by global self-ratings of sex drive strength or self-rated comparisons of one's own sex drive with that of other people. Both domain-specific estimates and global self-ratings left open what precisely was meant by the term 'sex drive'.

One reason for the large variability in employed indicators for sex drive may be that they were not derived from a coherent theoretical conceptualization of the construct. Instead, they appear to have been created based on face validity considerations. In each case, it is difficult to know why a particular indicator was (not) chosen. What seems missing from the literature is a coherent theoretical conceptualization of sex drive that also has clear implications for the question of which outcomes are best suited to indicate the strength of an individual's sex drive.

## Theoretical Conceptualization: What is Sex Drive?

Sex drive is an individual's intrinsic sexual motivation—the driving force to obtain sexual experiences and pleasure (Baumeister et al., 2001).<sup>2</sup> Although momentary sexual motivation, or state sex

<sup>2</sup> When we talk about sex drive, we are referring to motivation for sexual experiences and sexual pleasure as an end in itself. The goal is sexual experience and pleasure, not other, potentially (un)related goals. There are many reasons why people may seek out sexual experiences besides the sexual experience

itself (e.g., stress relief, procreation, emotional closeness to another person, see Meston & Buss, 2007). We are not referring to these instances, in which sex is a means to achieve other ends. Instead, our definition is confined to intrinsic sexual motivation where sexual experiences are an end in themselves.



drive, clearly varies within persons over time, the present research concerns stable individual differences between persons. Some people are consistently more eager for sexual experiences than others. We are thus interested in sex drive as a trait, where traits are understood as (inter-)individual differences in tendencies to show relatively consistent patterns of thoughts, feelings, and behaviors (Johnson, 1997; McCrae & Costa, 2003; Roberts, 2009).<sup>3</sup> People high in trait sex drive think about sex more often (thoughts/cognition), desire sex more often (feelings/affect), and are sexually active more often (behavior) compared to people lower in trait sex drive.

The seeming conundrum between intraindividual state variability on the one hand and temporal stability of interindividual trait differences on the other can be elegantly solved by understanding traits as ‘density distributions of states’ (Fleeson, 2001, 2004). This perspective assumes that for any personality trait, a corresponding personality state exists with the same cognitive, affective, and behavioral content as the corresponding trait, thereby constituting the personality an individual manifests from moment to moment (Fleeson & Jayawickreme, 2015). States vary over time within an individual as situational cues interact with disposition, but they vary in consistent, predictable patterns. Specifically, the distribution they form over time is stable with regard to its central tendency and dispersion (Fleeson, 2001). Traits are thus dispositions, not absolute determinants. Someone high in trait extraversion, for instance, does not act in an extraverted way all the time, highlighting the influence of situational circumstances. However, over a longer period, an extraverted person will more often act in an extraverted way compared to someone less extraverted and this interindividual difference will reliably emerge across several such extended periods. This illustrates the influence of the trait. In sum, this understanding of traits explains the temporal consistency of psychological patterns that vary between individuals, but also explicitly incorporates the notion of cross-situational variability within individuals (Fleeson, 2001; Fleeson & Jayawickreme, 2015; Johnson, 1997; Roberts, 2009).

<sup>3</sup> We use the term sex drive and sexual motivation interchangeably. The term ‘sex drive’ has been criticized as problematic by some theorists (Beach, 1956; Singer & Toates, 1987), especially the notion that an innate need for sex arises independently of external stimuli and builds up over time. In contrast to food, water, or sleep, deprivation of sex is not fatal or directly harmful. We acknowledge these shortcomings of the term. Our use of it is not intended to reflect biological drives akin to those for food, water, or sleep. The reason we retain the term sex drive is that it is widely used and understood in the literature and the general population. Although failure to satisfy one’s sex drive is not harmful to an individual, it appears appropriate to say that people may have a drive to pursue certain (sexual) goals or activities.

Based on this understanding, we can define (trait) sex drive more explicitly as the central tendency of the distribution of state sex drive, or momentary sexual motivation, across time and situations. Put more simply, a person’s sex drive is their average sexual motivation over time. We conceptualize state sex drive as a latent concept that is manifested or reflected in how often people experience three kinds of events: sexual cognitions (e.g., thoughts, fantasies), sexual affect (e.g., desire to have sex), and sexual behavior (e.g., masturbation). The probability that a manifest sexual event occurs at a given point in time will depend on the level of state sex drive at that time (which in turn depends on the interplay between a person’s trait sex drive and situational cues).

We illustrate this conceptualization of sex drive in Figure 1. Let us assume that state sex drive varies along an arbitrary scale, roughly between -4 and +4, where 0 represents the average population level. The data presented in Figure 1 have been randomly generated under our assumptions for a hypothetical person whose (trait) sex drive is 0, that is, exactly average. This is illustrated in Panel A1. The density distribution (dashed line) for our hypothetical individual centers directly on zero. Panel A2 illustrates how this person’s state sex drive fluctuates over time, while still centering around zero. The “observed” distribution (grey histogram in Panel A1) will never perfectly match the expected distribution for a finite sample of observations, but the correspondence is evident. Panel C illustrates the occurrence of sexual events. Yellow rectangles illustrate that a sexual cognition occurred at a certain point in time, blue rectangles illustrate sexual affect, and red rectangles illustrate sexual behavior. The association of sexual events and state sex drive is visible: After about half the time (x-axis), there is a noticeable dip in state sex drive and correspondingly, fewer sexual events occur. The relationship between state sex drive and the frequency of sexual events, or Panels A1 and C, is depicted in Panel B. The probabilities for the occurrence of sexual events displayed in Panel B are a direct function of the corresponding level of state sex drive at that time.<sup>4</sup> When state sex drive is relatively low, the probability of sexual events is also low, and fewer events occur. The reverse is true

<sup>4</sup> For illustration purposes, we assumed that probabilities for sexual behavior are generally lower than for cognition and affect, and that probabilities for sexual affect are generally lower than for sexual cognition. This, as well as the exact nature of the relationship between state sex drive and the probability of sexual events, currently remains subject of speculation, albeit, in our view, useful speculation. While this is beyond the scope of the present work, future research could seek to devise ways to test and parametrize this model. For example, a recent longitudinal study found that sexual cognition occurs more frequently than sexual affect, and that sexual affect occurs more frequently than sexual behavior, providing tentative evidence for the differential average probabilities we assumed (Weber et al., 2022a).



when state sex drive is high. In line with recent calls to incorporate more formal modelling into psychology (Guest & Martin, 2020), we enclosed a preliminary mathematical definition of our conceptualization of sex drive in the supplemental materials.

To summarize what we have described in the previous sections and illustrated in Figure 1, we conceptualize sex drive as average sexual motivation over time, formally described by density distributions of state sex drive. State sex drive, or momentary sexual motivation, manifests in sexual events, specifically sexual cognition, affect, and behavior, that occur more or less frequently, depending on the level of state sex drive. According to this definition, persons high in trait sex drive experience events characterized by sexual thoughts and fantasies, feelings such as desire, and behaviors (or any combination thereof) more often in their daily lives compared to people low in trait sex drive. In other words, people high in trait sex drive think about sex more often (cognition), feel a desire for sexual pleasure more often (affect), and are more often sexually active (behavior) compared to people lower in trait sex drive. This notwithstanding, even someone with a strong sex drive will not constantly think about sex, desire to have sex, or engage in sexual activity. Sex drive strongly varies within persons over time due to stress, time of day, availability of a partner, presence of other persons, conversations, media depictions, sexual satisfaction, and many other factors. Thus, while often influenced by sex drive, sexual experience and behavior can also be driven by other factors. For instance, the relative role of sex drive compared to other contextual variables may be attenuated in romantic relationships, where sexual manifestations such as sexual desire may become a function of each partner's characteristics and characteristics of the relationship itself (e.g., Impett et al., 2008; Regan, 2000). Someone high in sex drive will not always have consistently strong desire for their partner(s), and conversely, a person with a below-average sex drive is not generally incapable of developing sexual desire in a relationship.

Note that our definition of trait sex drive says nothing about the origin of these individual differences and potential gender differences in sex drive. Sex drive (and gender differences in sex drive) may be the result of a complex interplay of various cultural and biological influences. Social roles (Eagly & Wood, 1999) and social learning experiences (Bussey & Bandura, 1999) may contribute to these differences alongside genetic influences. The presence of genetic influences would in turn suggest that sex drive may also be heritable to some extent, similar to other traits (Polderman et al., 2015). We are agnostic towards the respective contributions of these and other possible origins of (gender differences in) trait sex

drive. While discussions of the etiology of gender differences in sex drive, that is, how they are shaped by biological and/or social factors, can be found elsewhere (Lippa, 2009), the present work is concerned with the phenomenology of the trait. We neither intend to nor are we able to elucidate the underlying causes of sex drive variability across persons in the present analysis.

## Sex Drive versus Sexual Desire

We understand sexual desire as an emotion, a feeling of wanting sex or sexual pleasure. Sexual desire takes a primary role among the affective manifestations of sex drive. Some scholars have adopted broader definitions that emphasize motivational aspects. Levine (2003) defines sexual desire as “the sum of the forces that lean us toward and away from sexual behavior” (p. 280). Spector and colleagues (1996) proposed “interest in sexual activity” (p. 178) as a working definition of sexual desire. Breznsnyak and Wishman (2004) define sexual desire as “a motivation to seek out, initiate, or respond to sexual stimulation or the pleasurable anticipation of such activities in the future” (p. 199). Sexual desire according to these views seems closely related to our understanding of sex drive as intrinsic motivation for sexual experiences and pleasure. The definition of Diamond (2004), “a need or drive to seek out sexual objects or to engage in sexual activities” (p. 116), includes an explicit reference to “drive”. Notably, these definitions do not define sexual desire as either a trait or a state, yet previous research has suggested that this differentiation is important in regard to gender differences (Dawson & Chivers, 2014). However, our conceptualization of sex drive (Figure 1) which rests on established frameworks specifying the relation between states and traits (e.g., Fleeson, 2001; Roberts, 2009) proposes that this is not an either-or-question, but that trait sex drive manifests in patterns of states of sex drive that are variable across time, but consistent when considering longer periods of time. Our view classifies sexual desire into the affective facet of the triad of sexual cognition, affect, and behavior, and reserves the terms sex drive/sexual motivation for the superordinate construct.

## Indicators of Sex Drive

Previous work on sex drive has used a large number of indicators. Without a coherent and theoretically grounded conceptualization of the construct, it is difficult to decide what may or may not qualify as a suitable indicator. The psychological conceptualization of trait sex drive put forward here has clear implications for what does and does not constitute a suitable indicator of sex drive.

According to the present conceptualization, the frequency of sexual cognitions (e.g., thoughts, fantasies, daydreams), sexual feelings (e.g., desire, craving, lust), and sexual behavior (e.g., masturbation, self-stimulation) constitute valid indicators of sex drive. These indicators can thus be directly derived from our sex drive conceptualization. All of them have been used as indicators of sex drive before (Baumeister et al., 2001). Throughout the manuscript, we collectively refer to these three sex drive indicators, that is, the triad of sexual cognition, affect, and behavior, as “facets” of sex drive.

For several reasons, we confine the behavioral facet to solitary sexual activities (i.e., masturbation, self-stimulation) and do not include sex with a partner. First, sex with a partner depends not only on a person’s intrinsic sexual motivation, but also strongly depends on other influences such as the availability of a partner, their sexual motivation, or interpersonal dynamics between the partners (e.g., desire to feel close and connected to the partner, desire to please the partner). Second, by simple arithmetic, there cannot be a true, objective gender difference in this variable for heterosexual persons on the population level. Every time a woman has sex, a man also has sex, and vice versa. Any appreciable reported difference is likely due to (motivated or otherwise) biased responding. Thus, in line with the present conceptualization, only solitary events can be a meaningful behavioral indicator of gender differences in intrinsic sexual motivation.

The conceptualization of sex drive directly suggests the frequency of sexual cognition, affect, and behavior as suitable indicators of sex drive, because they are measurable manifestations of latent, momentary sexual motivation. We refer to these three indicators as “manifestations of sex drive” throughout the manuscript. Our conceptualization also suggests another group of indicators, namely, measures that may directly reflect the latent level. In our survey of the literature, we identified two sets of questions that could serve this function: self-rated sex drive (e.g., “I have a strong sex drive”) and intensity of sexual affect (e.g., “My desire for sex is strong”). Self-rated sex drive may indicate latent sex drive at the highest level of abstraction (Panels A1 and A2 in Figure 1). Intensity of sexual affect may indicate latent sex drive at a more intermediate level as a latent state that is already somewhat differentiated towards sexual affect. We refer to these two indicators collectively as “indicators of latent sex drive” in this study. In our analyses, we will prioritize the sex drive manifestations over the indicators of latent sex drive, because the former are directly suggested by our conceptualization (Figure 1). For the indicators of latent sex drive, more detailed psychometric analyses may be needed before concluding that these

measures do indeed reflect latent variables according to our conceptualization.

The present conceptualization of sex drive also identifies concepts that have been used as indicators of sex drive in the past, but do not align with our view. For example, favorable attitudes toward sex may or may not be influenced by sex drive, but they are evaluations of sex that are likely influenced by all sorts of cultural and social influences. The frequency of sexual cognitions, affect, and behavior constitute manifestations of sex drive, not individuals’ subjective evaluation of sexuality. Second, previous research has utilized the desired number of sex partners as an indicator of sex drive, with higher numbers of desired partners indicating a stronger sex drive. However, in addition to a true impact of latent sex drive, the (reported) number of desired partners may also be influenced by self-verification motives, a desire for social status, or again cultural and social influences. What is more, a person who frequently thinks about and desires sex with only one partner would unequivocally be considered someone with a high sex drive according to the present conceptualization. Desiring sex with many different partners is not a defining aspect of a high sex drive according to this understanding (although empirically the two may be correlated). Third, and in a similar vein, enjoying a large variety of sexual practices may or may not be influenced by a strong sex drive. Someone who frequently fantasizes about and desires sex always in the same way would clearly have a strong sex drive according to the present conceptualization. A large variety of sexual practices is not a defining element of the current conceptualization and from this perspective unsuited to serve as an indicator of sex drive.

There are also other extant indicators that seem reasonable downstream consequences of sex drive, but are not directly derivable from the current conceptualization and therefore not considered valid indicators (e.g., unwillingness to forego sex, sacrificing resources to get sex, subjective importance of sex). Finally, because the present meta-analysis is interested in sex drive as a psychological force to obtain sexual experiences and pleasure, we do not regard capacity for physiological reactions as indicative of sex drive (e.g., capacity for sexual arousal or orgasm). Note that we do not exclude the possibility that the constructs discussed in this section may in some way be related to or influenced by sex drive, but rather maintain that they are of subordinate importance as indicators of the construct compared to the frequency of sexual cognition, affect, and behavior.

## Theoretical Approaches

This section reviews theoretical approaches relevant to (gender differences in) sex drive, namely

sexual strategies theory (Buss & Schmitt, 1993), the sexual double standards hypothesis (Crawford & Popp, 2003), social role theory (Eagly & Wood, 1999), social learning theory (Bussey & Bandura, 1999), the gender similarity hypothesis (Hyde, 2005), and sexual economics theory (Baumeister & Vohs, 2004).

### ***Sexual Strategies Theory***

Rooted in the larger evolutionary psychology framework, sexual strategies theory proposes that humans have evolved a variety of short-term and long-term strategies for passing their genes on to the next generation (Buss, 1998; Buss & Schmitt, 1993, 2019). According to the theory, these strategies differ between men and women, for example, due to the minimum parental investment both sexes have to make to produce a child. The number of offspring women can have is more limited compared to men, and women incur higher biological costs in terms of energy needs, risks during pregnancy, and effort in infant care. In short-term mating contexts, women should therefore be more selective while men should, on average, seek to engage in more casual sexual activities. In long-term mating contexts, men's and women's preferences will be largely similar, and both will be selective. However, according to the theory, women will prefer men who possess resources and/or have qualities that make the future acquisition of resources more likely. Men, by contrast, will be particularly attracted to cues of youth and health in women, both of which are linked to fertility (Buss, 2012).

Sexual strategies theory does not speak directly about gender differences in sex drive. However, it makes some predictions which indicate that the theory assumes a stronger sex drive in men compared to women. For example, the theory predicts that men will be particularly upset when their female partners decline or delay opportunities to have sex, or desire sex less frequently than themselves (Buss, 1998). This implies that, on average, men want sex more often than women do. In addition, some authors have argued that evolution may have favored a weaker sex drive in women compared to men. The higher the sex drive, the more likely a woman will become pregnant, which is associated with higher parental investment costs compared to men (Baumeister et al., 2001). A key tenet of sexual strategies theory is that in evolutionary history, it was likely adaptive for women to withhold sex under certain circumstances. A high sex drive would interfere with this tendency. This tentatively suggests that a higher sex drive in men is more plausible according to sexual strategies theory, and this gender difference would reflect genuine differences on the construct level rather than merely differences on the measurement level.

### ***Sexual Double Standards Hypothesis***

The sexual double standards hypothesis (Crawford & Popp, 2003) suggests that men are viewed positively and socially rewarded for sexually permissive behaviors, whereas women are viewed negatively and socially punished for the same behaviors (for a meta-analysis, see Endendijk et al., 2020). Awareness of sexual double standards may lead men to exaggerate their reports of sexual permissiveness and women to underreport their sexual permissiveness. This suggests that some gender differences in reported sex drive may emerge on the measurement level due to biased responding in line with gender roles that in fact may be smaller or even nonexistent on the construct level.

In line with this idea, some studies have found reduced or erased gender differences in reported sexual experiences when participants were connected to a fake lie detector (encouraging truthful responding) compared to conditions in which participants were led to believe that their responses might be seen by a peer or in which participants were assured anonymity (Alexander & Fisher, 2003). Further indirect evidence comes from findings that men report having had more opposite-sex sexual partners than women (e.g., Mitchell et al., 2019). In heterosexual populations, substantive differences are impossible because every time a woman has sex, a man also has sex, and vice versa.

### ***Social Learning Theory***

Social learning theory (Bandura, 1986; Bussey & Bandura, 1999, 2004) suggests that behaviors that are rewarded are more likely to be repeated and behaviors that are punished are less likely to be repeated. This is true both for one's own behaviors as well as behaviors an individual observes other people perform. Learning is particularly likely if the individual perceives similarity to or identifies with the acting person (for example, because the person is powerful, successful, or admirable). According to this theory, gender differences emerge because boys and girls (a) observe different behaviors in men and women, and (b) observe that men and women are rewarded and punished for different behaviors. Boys and girls pick up these different standards for gender-appropriate behavior and learn to behave in accordance with these gender norms.

To the extent that boys and girls learn (in real life or through media) that men are rewarded and/or women are punished for behaviors indicative of a strong sex drive, they may learn to behave accordingly and adopt corresponding attitudes (see the sexual double standard in the previous section). Thus, social learning theory makes clear predictions about openly expressed sexual attitudes and behaviors, which reflect genuine differences in this

sex drive facet.<sup>5</sup> In other words, boys and girls may learn to have a “gender-appropriate” sexuality (indicating true differences on the construct level). At the same time, they may also learn to express their sexuality in a norm-conforming way (indicating differences on the measurement level, i.e., biased responding). Whether the theory predicts gender differences in sex drive in the sense of sexual cognition and affect, which are non-observable for anyone other than the person themselves and arguably more difficult to control than overt behavior, is less clear.

### *Social Role Theory*

Social role theory (also referred to as the biosocial model or sociocultural theory) focuses on social processes instead of evolutionary selection processes to explain gender differences in behavior (Eagly & Wood, 1999; Wood & Eagly, 2012). Specifically, social role theory acknowledges evolved physical differences between the genders such as size, strength, and the capacity to bear and nurse children. In many societies, these differences led to a division of occupational and family labor. Both men and women tended to take on those tasks that aligned with their unique physical properties (i.e., men more often engaged in physically demanding tasks such as hunting and warfare, women more often engaged in less physically demanding tasks requiring care for others). In any given society, people observe the activities typically carried out by each gender and infer that these genders possess not only the physical requirements, but also the corresponding psychological characteristics that allow them to excel in gender-typical tasks. This is how gender stereotypes develop and in turn reinforce and perpetuate a gender-stereotypical division of labor, according to social role theory. Gender stereotypes cause gender-typical behaviors because (a) individuals tend to conform to their gender identities, and (b) other people encourage gender-typical behavior. Role-incongruent behavior is more likely to be societally sanctioned. Therefore, role-congruent behavior is perpetuated unless the anticipated benefits of gender-incongruent behavior outweigh the anticipated costs.

Evidence for social role theory comes from observations that typical gender differences in interests, preferences, and even personality characteristics such as agency have decreased over time as the division of labor has become increasingly less polarized and women’s social role has shifted (Wood & Eagly, 2012). In a similar vein, a meta-analysis found that typical gender differences are

smaller in countries in which gender equality is greater, including differences in the domain of sexuality such as masturbation (Petersen & Hyde, 2010), although recent studies found no support for a moderation of typical gender differences in mate preferences by gender equality (Walter et al., 2020; Zhang et al., 2019).

Social role theory assumes that socialization processes shape individuals’ behaviors, beliefs, typical emotional responses, competencies, and personality traits to conform with societal stereotypes about how men and women are. Thus, to the extent that a society views a high sex drive as more typical and normative for men than women, social role theory predicts a higher sex drive in men than in women reflecting genuine differences on the construct level. These differences should be smaller in societies with greater gender equality. Also, gender differences in sex drive should have become smaller over time as gender equality has increased in many societies in recent decades. Differences on the measurement-level due to self-presentation tendencies (i.e., biased responding) are also plausible under the assumptions of social role theory: People may adopt self-presentational tendencies on their own accord in order to conform with societal stereotypes, or they may learn them directly during the gendered socialization processes.

### *Gender Similarities Hypothesis*

The gender similarities hypothesis (Hyde, 2005, 2014) is not exactly a theory, but rather a set of observations based on several meta-analyses of gender differences in psychological variables. The hypothesis states that gender differences in most, but not all psychological variables are small or negligible (with ‘small’ being defined as everything up to a meta-analytic effect size of Cohen’s  $d \leq 0.35$ , considered a small-to-moderate effect size according to common conventions). Hence, contrary to many stereotypes and public portrayals, women and men may not be vastly different in many spheres (Hyde, 2014). However, there are exceptions to this general rule. Hyde (2005) reports non-trivial gender differences in physical aggression, cognitive variables such as mental rotation and spatial perception (men score higher), and indirect aggression and some language or verbal skills (women score higher). Relevant to the present purposes, men report masturbating and watching pornography more often than women ( $d_s > 0.5$ , Petersen & Hyde, 2010). Thus, the gender similarities hypothesis, based on previous meta-analytical observations, suggests a higher sex drive in men compared to women with respect to the

<sup>5</sup> In line with social role theory (Wood & Eagly, 2012), smaller gender differences in sexual attitudes and behavior should occur in societies where such gender-specific rewards and

punishments occur less (e.g., in countries with greater gender equality).

behavioral facet of the construct. This difference is assumed on the construct rather than purely on the measurement level, thus reflecting genuine gender differences. The hypothesis makes no direct predictions with respect to the cognitive and affective facets.

### ***Sexual Economics Theory***

Sexual economics theory (SET, Baumeister et al., 2017; Baumeister & Vohs, 2004) posits that sex in heterosexual couples is negotiated in an economic marketplace. In this market exchange, women (the ‘sellers’) give sex and, in return, receive sex by men (the ‘buyers’) plus a negotiable amount of nonsexual resources, because female sex is – according to the theory – inherently more valuable than male sex. The theory assumes a real gender difference in sex drive on the construct level, not merely on the measurement level. Gender differences in sex drive are therefore a fundamental premise rather than a prediction of SET, and considerable theoretical revision would be needed should it turn out that there is no such gender difference. SET does, however, provide prediction and explanation for biased responding. According to the principles of economic exchange in sexuality posited by the theory, female sex is at risk of diminishing in value when distributed freely (or appearing so); hence, women should be motivated to underreport sexual interest and activity. Men, by contrast, should be motivated to exaggerate reports of past sexual activity, since these reflect that they can exchange ample resources to obtain sex. For men’s sexual interest, predictions are somewhat less clear. To some degree, interest in a resource may also signal an ability to obtain it, yet this would run counter to the age-old principle of hiding one’s true interest in a negotiation. In sum, SET predicts some degree of response bias, since women should understate and men may (or may not) exaggerate.

### ***Summary***

Although some of the theoretical approaches reviewed here differ greatly in their core assumptions, they largely converge in the prediction that men have a stronger sex drive compared to women, at least on the measurement level. The distinction between the measurement level and the construct level underlines the importance of considering the possibility of bias as a result of systematically distorted responding. We note that the psychological mechanisms by which the social environment influences sex drive on the measurement or construct level are largely left open

by the theories. This influence may manifest in conscious self-presentation tendencies—men overreport to gain social status and women underreport to avoid loss of social status (Jonason, 2008; Mitchell et al., 2019). However, the effect of social influence may also manifest in more subtle ways, for example memory biases in the form of gender differences in estimating versus actually counting sexual events (Brown & Sinclair, 1999; Mitchell et al., 2019; Wiederman, 1997).

### **The Present Meta-Analysis**

We conducted a preregistered, comprehensive meta-analysis of gender differences in sex drive. Based on the theoretical conceptualization of sex drive presented above, we primarily investigated gender differences in the frequency of sexual cognitions, affect, and behavior. We additionally included analyses on two potential indicators of latent sex drive: self-rated sex drive and intensity of sexual affect.

We put particular emphasis on the possibility that gender differences in sex drive may be (partly) due to biased responding. To this end, a separate meta-analysis examined gender differences in responses to ‘bias indicators’—that is, questions that logically cannot exhibit a substantive gender difference in heterosexual populations (e.g., total number of sex partners). The meta-analytic gender difference in these indicators may be interpreted as an indicator of the extent to which gender differences in sex drive in the main analyses may have been driven by biased responding.

Finally, a series of moderation analyses examined the potential impact of a number of either theoretically derived or methodological factors on the magnitude of gender differences in sex drive. These analyses are suited to both test theoretical predictions and examine the stability of potential gender differences.

### **Method**

Effect sizes were drawn from manuscripts, provided by authors, or computed from raw primary study data (two-step individual participant data meta-analysis, see Riley et al., 2010). The unit of analysis is individual questionnaire items, with each effect size indicating the mean gender difference on one particular item.

### **Inclusion Criteria**

Studies were eligible for inclusion in the meta-analysis if: (a) they measured frequency of sexual cognitions, sexual affect, or sexual behavior (sex drive manifestations), or sex drive self-ratings or

intensity of sexual affect (indicators of latent sex drive); (b) the sample included male and female participants; (c) participants were at least 14 years of age; and (d) the sample included at least 20 male and female participants each. Studies were excluded if: (a) the sample was drawn from a clinical population, an asexual population, or residents of long-term care facilities; (b) the study included an experimental manipulation or other intervention procedure; (c) it took place in the context of pregnancy or abortion; or (d) was published before 1997. The last criterion was imposed post hoc when it proved unrealistic to attain raw data that was necessary to compute item-level effect sizes for older research (see sub-section “Meta-Analysis of Item-Level Effect Sizes” for details). Note, however, that data collection may have taken place before 1997.

Questionnaire items were eligible for inclusion if they reflected a sex drive manifestation or indicator of latent sex drive. In addition, we also included items that reflected a bias indicator (see Table 1). Note that we did not search for and did not include studies reporting only bias indicator items. Items were excluded if they framed sexuality in a negative or clinical way, or invoked perceived social norms (e.g., “I suffer from a lack of desire”, “I think about sex more often than I should”, “Masturbation sometimes gets in the way of my daily activities”). Another exclusion criterion on the item level was imposed post hoc: items were not eligible for inclusion if participants were asked to report their sex drive compared to other individuals of their own sex. Such items are designed to eliminate gender differences between men and women—the very purpose of our meta-analysis—and hence are not suitable to address the present research question.

## Literature Search

We used different strategies to identify relevant studies. First, we conducted an electronic literature search using Web of Science (Indices: Science Citation Index Expanded, Social Sciences Citation Index, and Emerging Sources Citation Index), EBSCO (Indices: PsycARTICLES, PsycINFO, and PSYINDEX), and PubMed (Indices: primarily MEDLINE). All databases were searched on the abstract/title level. Searches were done separately for sexual affect, cognition, and behavior. Search terms were constructed with the following pattern: one of ‘term 1’ AND one of ‘term 2’ NOT one of ‘term 3’. Term 1 established the link to sexuality (e.g., “erotic”, “sexual”), Term 2 evoked the construct (e.g., “thought”, “fantasy”, “desire”), and Term 3 excluded clinical studies (“disorder”, “dysfunction”). See Table S1 in the supplemental materials for the complete set of search terms. Second, we used Google Scholar to screen all publications that cited relevant psychometric

inventories. Relevant inventories were identified by searching the Handbook of Sexuality Related Measures (T. D. Fisher, 2011) and through unstructured electronic searches (see Tables S2 and S3 for a list of the inventories). Third, we submitted calls for data through the mailing lists of the International Academy of Sex Research, the European Association of Social Psychology, and the Society for Personality and Social Psychology. Fourth, we asked all authors with whom we corresponded for unpublished data. The literature search was restricted to articles in English or German and was completed in 2018. Data collection was completed in 2019.

## Screening and Requests for Data

All identified records were screened by one of the authors or a research assistant. First, studies were screened on the title level for eligibility. If studies were deemed potentially eligible, their abstracts were retrieved and screened again. During this screening on the abstract and title level, we adopted a maximally inclusive stance, such that records were only discarded when there was a clear indication that an inclusion criterion was violated or an exclusion criterion fulfilled. Next, full texts were obtained for all studies that passed the initial screening phase and screened again for suitability. For technical reasons, duplicates between different searches were not removed prior to screening. The full results of the search procedure are summarized in the flow chart depicted in Figure 2.

During the full text screening, we checked if studies reported sufficient item-level statistics to compute effect sizes. If this was not the case, the study’s corresponding author was contacted per email with a request for data. Each email also included a request for unpublished data on the subject. A reminder was sent after two weeks.

## Effect Size Computation

We computed Hedges’  $g$  for all outcomes (Hedges, 1981). This effect size indicates the average gender differences in the sample in the metric of the pooled standard deviation. Positive values for  $g$  indicate higher values in men. For each relevant outcome, means, standard deviations, and sample sizes for men and women were retrieved from the paper or computed from the raw data. Hedges’  $g$  was then computed from these summary statistics. If means or standard deviations were missing, Hedges’  $g$  was computed from  $p$ -values,  $t$ -values, and degrees of freedom for  $t$ -tests comparing the groups. When measurements were taken multiple times, all time points were averaged prior to effect size computation. For raw data, all values

that deviated more than 3.5 standard deviations from the mean were classified as outliers and removed.<sup>6</sup> To put findings into perspective, we additionally report a range of natural language interpretations that are more easily interpreted than standardized mean effect sizes (Mastrich & Hernandez, 2021). First, Cohen's  $U_3$ , a measure of non-overlap, indicates which percentage of population A is surpassed by the upper half of population B (Cohen, 1988). Second, the overlapping coefficient  $OVL$  indicates the overlap between two distributions (Reiser & Faraggi, 1999). Third, the common language effect size  $CL$  indicates the probability that an observation drawn at random from population B surpasses an observation drawn at random from population A (McGraw & Wong, 1992; Ruscio, 2008).

These effect size statistics are computed as follows:

$$U_3 = \Phi(g) \quad (1)$$

$$OVL = 2\Phi\left(\frac{-|g|}{2}\right) \quad (2)$$

$$CL = \Phi\left(\frac{g}{\sqrt{2}}\right) \quad (3)$$

where  $\Phi$  is the cumulative distribution function of the standard normal distribution.

## Meta-Analysis of Item-Level Effect Sizes

A body of research can only be subjected to a meta-analysis when there is a sufficient level of coherence in theorizing and research methodology. The key challenge for any meta-analysis is to determine if integration is warranted and how it can be achieved. When first reviewing the literature on sex drive, we encountered considerable conceptual heterogeneity. This heterogeneity was reflected in the wide variety of psychometric inventories used to gauge the construct, which rendered conventional meta-analysis of complete inventories unfeasible. To solve this, we developed a new approach that draws on (and separates) the individual inventory items in line with the underlying psychological theorizing. To that end, we first developed the new conceptualization of sex drive outlined in the introduction. This conceptualization served as the theoretical foundation for the integration of previous research. We then selected individual items from existing inventories or ad-hoc measurements that reflected the theoretically derived indicators of sex drive. For example, the item "During the last month, how often have you had sexual thoughts involving a

partner?" was selected (among others) from the Sexual Desire Inventory (Spector et al., 1996) and classified as frequency of sexual cognition. All inventories from which individual items were retrieved are listed in Tables S2 and S3. We then collected and synthesized data for gender differences with respect to these items. This novel approach thus combines the advantages of meta-analysis (high generalizability, high statistical power) with high conceptual coherence by exclusively including a set of selected items that adequately fit the theoretical conceptualization of the construct of interest.

## Coding

We coded several characteristics of (a) the publication, (b) the study design, (c) the sample, and (d) the outcomes. Some of these were used in statistical tests for moderation of gender differences in sex drive, others serve descriptive purposes. For some moderators, outliers were removed according to cut-off values prior to analyses. Outliers were determined by visual inspection of the data. We list all such cut-off values when discussing the codings in the following section.

For some characteristics, we derived tentative hypotheses regarding moderation effects based on previous research, but we generally consider these analyses exploratory in nature. Unless noted otherwise, characteristics were coded as "yes" or "no", or "NA" if the information was not available. For categorical characteristics with more than two possible codings, we list all possibilities, except for country of data collection and the outcome-level codings. For the outcome-level codings, possible categories were generated inductively during the coding process and consolidated after coding.

The coding work was shared among two of the authors. A sample of twenty-one studies was coded by both coders. Results for interrater reliability are summarized in Table S7. We computed Cohen's  $\kappa$  (Cohen, 1960) for categorical codings with a known number of categories, Pearson's correlation for numerical codings, and percent agreement for categorical codings with an unknown number of categories. Overall, coder agreement was good (mean of all  $\kappa$  values: 0.87) and is classified as an "almost perfect" strength of agreement according to conventions (see Landis & Koch, 1977).

## Publication Characteristics

The first set of coded characteristics on the publication level related to the intent and topic of the

<sup>6</sup> Robustness analyses showed that results did not vary when a smaller (2.5 standard deviations) or larger threshold (4.5 standard deviations) was chosen.

manuscript, authors' gender (distribution), and the focus of the journal the manuscript was published in (Codings 1 to 6). The journal focus was inferred by considering the abstract and journal title only. The authors' gender was inferred from their names and institutional web pages. These codings allow for sensitivity analyses for potential biasing effects on the part of researchers. With Coding 7 (focus on anonymity), we aimed to capture whether the authors expressed awareness of the need to create a private and secure environment for participants to respond truthfully to questions about sexuality in order to maximize chances of truthful responses. In Coding 8, we coded the manuscript's publication status to test for potential publication bias, that is, smaller or larger effect sizes for unpublished studies. For unpublished data sets with no manuscript, publication characteristics were coded as missing except for author gender. Thus, the codings for publication characteristics were as follows:

1. Focus on gender differences: Did the authors focus on gender differences in the study?
2. Focus on gender differences in sex drive: Did the authors focus on gender differences in sex drive?
3. Aim to find gender differences in sex drive: Did the authors state that they were aiming to find gender differences in sex drive?
4. First author gender: What was the gender of the first author (male/female/non-binary)?
5. Mean author gender: Female and male authors were coded as 0 and 1, respectively, and non-binary excluded.
6. Sexuality journal: Does the journal publish research specifically on sexuality?
7. Focus on anonymity: Was there any general or specific statement about participants' anonymity, confidentiality, or privacy anywhere in the paper?
8. Publication status: Had the manuscript been published in a peer-reviewed journal as of October 2020?

### *Study Characteristics*

Codings on the study level were mostly intended to gauge how privacy-preserving the study situation and experience was for participants (Codings 1 to 5). From a theoretical perspective, this seems promising, as empirical sex drive differences may be more pronounced under study situations with less privacy or less subjectively perceived security and

anonymity. This is because a lack of perceived privacy, security, and anonymity may promote biased responding, which may manifest as more restrictive responding in women and more liberal responding in men.

We also coded how the study was advertised and how participants were compensated to probe potential selection bias effects (Codings 6 and 7). Studies on sexuality may suffer from volunteer bias, with people willing to participate differing systematically from people not willing to participate. Some evidence suggests that volunteers tend to be more sexually experienced and hold more positive attitudes toward sexuality (Strassberg & Lowe, 1995; Wiederman, 1997), have higher levels of education, are less conservative, and more novelty-seeking (Dunne et al., 1997). The year of the study was coded to probe for potential changes in gender differences over time (Coding 8). Social norms change over time, and attitudes toward sexuality are becoming less restrictive (Mercer et al., 2013). This would suggest that, if the results are affected by biased responding, observed gender differences may have decreased over time. Thus, the codings for study characteristics were as follows:

1. Face-to-face interview: Were the questions asked in person by an interviewer?
2. Personal contact: Did participants have personal contact with anyone affiliated with the research team?
3. Group assessment: Were participants tested in groups or not (or a combination thereof)?
4. Electronic data collection: Was the data collection electronic versus not electronic (or a combination thereof)?
5. Participant anonymity: Were participants reassured about anonymity, privacy, or confidentiality? Example for positive coding: "Participants were assured that no IP addresses would be saved to ensure anonymity."
6. Sexuality study: Was the study advertised as a study on sexuality?
7. Compensation: Did participants receive material compensation (money, coupons, etc.), course credit, a combination of both, or nothing in return for participation?
8. Year of study: If the information about year of data collection was missing for published studies, we entered the year the study was published minus two.



### *Sample Characteristics*

We collected several data points characterizing the sample. For some of these, associations with sexuality have been previously established in the literature. Others were included solely for descriptive purposes or sensitivity analyses. Most codings related to common demographic characteristics (Codings 1 to 9). For descriptive purposes, codings for mean age, standard deviation of age, percent heterosexual, and percent single were taken separately for men and women (whenever possible) to collect additional information on potential within-sample differences between men and women on these characteristics. Country-level sex ratio (Coding 12), country-level gender inequality (Coding 13), and country-level gender development (Coding 14) were coded based on the country of data collection (Coding 11). These sets of codings were included to gauge how the social and cultural context may shape gender differences in sex drive. Thus, the codings for sample characteristics were as follows:

1. Mean sample age (in years): Participants' average age was coded. Samples with an average age above 70 were classified as outliers based on visual inspection of the data and removed from the respective moderation analysis (3 effect sizes from 3 studies removed, next closest average age = 51.28).
2. Sexually active: Some studies restricted sampling to sexually active participants, others did not. This was usually defined as having had sex with a partner recently. The definition of "recently" varied across studies.
3. Percent religious: What percentage of participants are religious? Participants were counted as non-religious if they responded with "none" or equivalent to questions assessing religiosity, faith, etc.
4. Percent single: What percentage of participants are single? All participants indicating any sort of romantic affiliation were counted as not single.
5. Average partnership duration (in weeks): Relationship length was coded for the subset of participants in relationships. Manifestations of sex drive, including sexual desire, are well documented to fluctuate and in some cases decrease over the course of a long-term relationship (Klusmann, 2002). Recent work has shown sexual desire to decline particularly in wives but less so in husbands during the first couple of years of marriage (McNulty et al., 2019).
6. Percent White: What percentage of participants are of White/European/American ethnicity? This served as a proxy for the percentage of respondents with minority status in the sample for most studies in the database. More fine-grained coding of ethnicity was complicated by varying definitions across studies. We had no prior hypotheses regarding the association between ethnicity and gender differences in sex drive.
7. Percent heterosexual: What percentage of participants are heterosexual?
8. Percent university students: What percentage of participants are university students?
9. Percent parents: What percentage of participants are parents? Parenthood, especially early parenthood, can impact sexual desire in couples, and may affect new fathers and mothers differently (Ahlborg et al., 2005).
10. Contraceptive use: What percentage of the female sample used hormonal contraceptives?
11. Country of data collection: For studies that took place in multiple countries, we retrieved percentages per country.
12. Country-level sex ratio: Previous research suggests that sexual desire is influenced by the number of potential partners available, and that this influence unfolds differently for men and for women (Gebauer et al., 2014b). For each country, we coded the number of males per 100 women in the 25-49 age bracket (Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat, 2019), since this bracket was most representative for our data. Values were retrieved for the year closest to the year of the study (see previous section on study characteristics). For studies that spanned multiple countries, we entered a weighted score. This coding was not preregistered.
13. and 14. Country-level gender inequality and country-level gender development: Social norms regarding the expression of sexuality may differ for men and women (see the previous sections on social role theory and the sexual double standard hypothesis). In order to capture how participants may

be exposed to differing social norms, we retrieved data for gender inequality (Gender Inequality Index, GII) and gender development (Gender Development Index, GDI). Both indices are produced by the United Nations (United Nations Development Programme, 2019). The GII captures discrimination against girls and women in the areas of health, education, political participation, and labor market opportunities. Higher GII values indicate greater gender inequality, with values ranging from 0 to 1. The GDI measures gender differences in human development achievement in health, knowledge, and living standards. GDI values of 1 indicate gender parity, values below 1 indicate discrimination against females, and values above 1 indicate discrimination against males. We again entered the value from the year closest to the year of the study and computed weighted scores for studies spanning multiple countries. For the GDI, values below 0.90 were classified as outliers and removed (6 effect sizes from 3 studies removed, next closest GDI value = 0.94). For context, a value of 1.00 implies equality. In 2018, the country closest to gender parity (Norway) was rated at 0.99, while the world's GDI was estimated at 0.94.

### *Outcome Characteristics*

For outcomes, we coded several characteristics of the item and the response scale. When items were taken from a common psychometric inventory, but no further information was given, we assumed that the item wording and response scale corresponded to the original publication of the inventory. The codings for outcome characteristics were as follows:

1. Item content: What was the content or target of a sexual thought or affect, that is, who or what does one think about or feel desire for? Possible codings were: 'no target', typically just containing general references to sex (e.g., "How often do you think about sex?"); 'unspecified partner', when a partner is mentioned but not further specified (e.g., "How often do you think about sex with a partner?"); 'own partner' for references to one's own partner specifically; 'extra-pair partner', when asking about a partner outside of the current relationship (e.g., "How often do you fantasize about having sex with

someone other than your current dating partner?"), and 'masturbation' (e.g., "How often do you feel desire to masturbate?").

2. Item context: What was the context in which the cognition, affect, or behavior occurred (e.g., sexual thoughts while bored at work, sexual desires in romantic situations)?
3. Item wording: How was the construct labelled (e.g., 'self-stimulation' vs. 'masturbation', 'sexual daydream' vs. 'fantasy', 'sexual need' vs. 'sexual desire')?
4. Aggregation span: For outcomes indicating frequency of a sexual event: What was the period (in weeks) across which frequency was aggregated? For example, the item "How often did you think about sex in the past four weeks?" would be coded as 4. Values above 60 weeks were removed (10 effect sizes from 9 studies removed, next closest value = 30).
5. Type of response scale: Was the response scale open or closed (e.g., Likert-type scale)?
6. Scale range: For closed response scales, what was the scale range (scale maximum minus scale minimum)?

### **Statistical Analyses**

We aggregated effect sizes using meta-analytic models. In the primary analysis, effect sizes for the sex drive manifestations were modelled as a function of sex drive facet (frequency of sexual cognition, sexual affect, or sexual behavior), akin to a one-way analysis of variance in primary studies. This model estimates the summary effects within subgroups (i.e., per facet) and enables testing for between-group differences. Dependency due to the inclusion of multiple effects per study was handled using robust variance estimation (RVE) meta-analysis (Hedges et al., 2010; Tipton, 2015). Indicators of latent sex drive and bias indicators were analyzed separately using the same model, that is, in a subgroup analysis by type of indicator. To derive a global, cross-indicator estimate for gender differences in sex drive, we fitted a random-effects meta-analysis model with equal weights assigned to the group-wise summary effects for the sex drive manifestations. The same procedure was applied to the bias indicators to obtain a global estimate of biased responding.

We also applied univariate moderation analyses to probe how effect sizes for the sex drive manifestations (i.e., sex drive facets) varied as a function of publication, study, sample, or item characteristics. Models were fitted separately for each type of sex drive manifestation.

## Robust Variance Estimation

Due to our approach of meta-analyzing item-level effect sizes, many studies contributed multiple effect sizes. This creates dependence among effect sizes, which constitutes a violation of the assumptions of standard meta-analysis models (Lipsey & Wilson, 2001). Conventional approaches for solving this problem involve either selecting one effect size per study or manually aggregating multiple effect sizes prior to modelling (Borenstein et al., 2009). Both approaches, however, entail a loss of information and severely complicate meta-moderation analyses. To illustrate the latter point, consider a study reporting participants' frequency of sexual thoughts on an open scale and their frequency of sexual fantasies on a closed scale. These two effects would need to be averaged manually to satisfy the independence assumption. However, this aggregation would preclude the meta-moderation analysis of the effect of closed versus open scales, so effect sizes need to be left unaggregated or be dropped from the analysis. In the former case, a new data set needs to be created for every single moderation analysis, while in the latter case valuable information is lost. Both options are unsatisfactory. RVE meta-analysis elegantly solves the problem of effect size dependency by estimating a "working" model for the variance-covariance matrix of effect sizes (Hedges et al., 2010), thereby allowing dependent effects to be modelled as a function of one or more predictor variables while minimizing loss of information.

Modelers have a choice between a "hierarchical" and a "correlated" effects model for the dependency structure—the "hierarchical" model is more appropriate when dependence arises predominantly from identifiable clusters of estimates (e.g., multiple studies by the same laboratories or authors), while the "correlated" model is more appropriate when dependence arises predominantly from multiple outcomes per study. Additionally, modelers have to decide on a default value for the correlations between effect sizes, although this usually has no discernable impact on the model estimates. For all models, we selected the "correlated" effects model and a default correlation of 0.8. To test the sensitivity and robustness of the results, we varied the latter correlation value between 0 and 1 in steps of 0.1 for the primary analyses. In no case did this correlation assumption considerably influence the results. We employed small sample corrections when testing for meta-regression by adjusting the degrees of freedom (Tipton, 2015) and using an approximate Hotelling test (AHZ, see Tipton & Pustejovsky, 2015).

## Heterogeneity

We report two measures of effect size heterogeneity,  $\tau$  and  $I^2$  (Borenstein et al., 2009). Although both  $\tau$  and  $I^2$  are indicators of heterogeneity, they serve different purposes. First,  $\tau$  is the standard deviation of the true effects. It answers the question of how much the true effects vary independent of variation due to sampling error. Because  $\tau$  reflects the absolute amount of true variation, it says nothing about the proportion of the observed variation that is due to true variation of effects and not mere sampling error. To facilitate the interpretation of  $\tau$  estimates, they can be examined relative to  $\tau$  estimates in other meta-analyses. A recent study examined between-study heterogeneity estimates published in *Psychological Bulletin* between 1990 and 2013 (Van Erp et al., 2017). For studies reporting  $d$  or  $g$  effect sizes, the 25%, 50%, and 75% quantiles of  $\tau$  were 0.12, 0.20, and 0.32. These may serve as reference points for small, medium, and large heterogeneity.

Second,  $I^2$  indicates the proportion of the variation in observed effects that is due to variation in true effects rather than sampling error (Borenstein et al., 2017). Because  $I^2$  is the *proportion* of variance that is true, it says nothing about the *absolute* amount of variation, as  $\tau$  does. If all variation in observed effect sizes were only due to sampling error, both  $\tau$  and  $I^2$  would approach zero.

## Meta-Analytic Correlation Analyses

In addition to the summary effects for gender differences, we also investigated the correlations between outcomes. If studies reported multiple outcomes, the Pearson correlations between outcomes (their respective variances) were retrieved from the manuscript or computed from the raw data and labelled according to the indicators they represented (e.g., CF-AF for a correlation between one cognition frequency item and one affect frequency item). We then aggregated correlations for all available outcome pairs using the meta-analytic models described previously to create a meta-analytic correlation table. We expected notable correlations among the sex drive manifestations and indicators of latent sex drive (convergent validity). Correlations between sex drive indicators and bias indicators were expected to be lower, but still positive. (On the individual level, a stronger sex drive may well be associated with higher responses to questions that indicate bias on the level of gender differences.) All Pearson correlations were transformed to Fisher's  $Z$  prior to analysis and then back to Pearson correlations for interpretation. Variances for the Fisher's  $Z$  values were computed from the sample size (Borenstein et al., 2009).

## Publication Bias

Publication bias occurs when studies that did not produce the desired outcomes are less likely to be published (Fanelli, 2012; Franco et al., 2014). Authors are less likely to submit “failed” studies for publication, and if they do, reviewers and editors are less likely to favor publication compared to “successful” studies that produced significant outcomes. As a result, most published studies in psychology report hypotheses that “worked” (Fanelli, 2010; Sterling, 1959; Sterling et al., 1995). There is widespread agreement that publication bias exists, that it may bias meta-analytic effect size estimates, and that its prevalence varies across different bodies of research literature (Bakker et al., 2012; Fanelli, 2010; Ferguson & Brannick, 2012; Friese & Frankenbach, 2020).

For the present meta-analysis, publication bias is unlikely to play a role for several reasons. First, the majority of studies providing relevant data primarily investigated a research question unrelated to the one examined in the present meta-analysis. Therefore, the decision to submit and publish the respective studies did not depend on outcomes regarding gender differences in indicators of sex drive. Second, to adequately test our research question, we resorted to comparing responses to individual items instead of complete inventories. These fine-grained data are rarely reported in any manuscript and are therefore unlikely to influence the decision to publish a study.

Despite these reasons to believe a priori that publication bias is rather unlikely to affect the present meta-analysis, we applied two statistical approaches to detect publication bias (Iyengar & Greenhouse, 1988; McShane et al., 2016; Sterne & Egger, 2005). Unfortunately, the toolset for detecting publication bias for meta-analyses with multiple, dependent effect sizes is still limited (Friese et al., 2017). A recent simulation study suggested two approaches—Egger’s test and a three-parameter selection model; neither, however, is without drawbacks (Rodgers & Pustejovsky, 2020).

The first approach is a variant of Egger’s test for funnel-plot asymmetry (Egger et al., 1997). In this test, effect sizes are regressed on their standard errors in an RVE meta-regression. The underlying logic is that studies with smaller sample sizes, and thus larger standard errors, require larger effect sizes to achieve statistical significance than studies with larger samples sizes and smaller standard errors. Consequently, studies with larger standard errors are expected to have larger effect sizes on average, if effect sizes are selected based on statistical significance. Note that these so-called small-study effects can arise from publication bias, but also from legitimate sources (e.g., systematically different populations in smaller compared to larger studies).

This approach exhibits nominal Type I error rates, but can have little statistical power when the number of included effect sizes is small (Rodgers & Pustejovsky, 2020).

The second approach is the so-called three-parameter selection model (Iyengar & Greenhouse, 1988; McShane et al., 2016; Vevea & Hedges, 1995). This model does not handle dependency due to multiple outcomes natively, so this needs to be addressed beforehand by aggregating effect sizes per study or randomly selecting one effect size per study. The 3PSM approach compares an unadjusted meta-analysis baseline model to an adjusted model in a likelihood-ratio test. The unadjusted model is the standard meta-analysis model and can be any fixed-, random-, or mixed-effects model. We fitted an intercept-only random-effects model. In the adjusted model, the selection process, that is, the process of selecting studies for inclusion based on statistical significance, is explicitly modelled by estimating weights for pre-specified  $p$ -value intervals of interest. We set two intervals:  $0 < p < 0.025$  for significant studies (i.e., a two-tailed  $p$ -value of 0.05) and  $0.025 < p < 1$  for non-significant studies. Publication bias is assumed to be present if the inclusion of the selection process significantly improves the baseline model, as indicated by the likelihood-ratio test.

We tested for publication bias using both approaches separately for each sex drive manifestation. For the 3PSM, we created sets of independent effect sizes by randomly drawing one effect size per study. We then fitted the adjusted and unadjusted models, performed the likelihood-ratio test, and retrieved the results. The process was repeated 100 times to reduce the impact of chance during sampling. We report the average  $p$ -values across repetitions.

## Statistical Software

Data handling and analysis were done with the *R* language for statistical computing (R Core Team, 2020). We relied on the *robumeta* package for robust variance estimation (Z. Fisher & Tipton, 2015), the *metafor* package for random-effects meta-analysis (Viechtbauer, 2010), and the *tidyverse* package for data preparation (Wickham et al., 2019). Figures, tables, and the results text were produced programmatically for increased reproducibility.

## Transparency and Openness

We adhered to the PRISMA reporting guidelines for systematic reviews (Moher et al., 2009). We made the PRISMA checklist, effect size data, and computer code available in an open online repository at <https://osf.io/h4jbx/>, following recent

recommendations to increase the reproducibility of meta-analyses (Lakens et al., 2016). Note that we are not permitted to share the raw data we collected from the primary authors but do share all aggregate statistics computed from these raw data. Approval by a research ethics committee was not required for this review. The meta-analysis was preregistered on PROSPERO. We included an annotated copy of the protocol in the open repository to denote all deviations. The original protocol can be accessed at [https://www.crd.york.ac.uk/prospero/display\\_record.php?RecordID=72894](https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=72894). The supplemental materials to this article contain additional figures and tables, as well as complimentary information on our conceptualization of sex drive.

## Results

### Search Results

In total, the search team screened 20,397 titles, 3,784 abstracts, and 1,715 full texts (all including duplicates, see flow chart in Figure 2). 483 publications containing eligible studies were identified. Out of these, 460 did not report all necessary information. We contacted all 314 corresponding authors with requests for data (some contributed more than one publication), out of whom 144 responded and 118 provided data. Those who did not provide data cited lack of time, no access to data, data loss due to hardware failure, not keeping the data, or inability to find the data as reasons. Data could not be obtained for 297 eligible publications. Thus, overall, data could be obtained for 39% of eligible publications. In total,  $n = 621,463$  participants were included in the analysis. We retrieved  $m = 856$  effect sizes from  $k = 211$  studies. About half of the effect sizes were sex drive manifestations ( $m = 439$ ,  $k = 195$ ,  $n = 225,102$ ), one quarter were indicators of latent sex drive ( $m = 173$ ,  $k = 54$ ,  $n = 444,530$ ), and one quarter were bias indicators ( $m = 244$ ,  $k = 123$ ,  $n = 102,634$ ).

### Study and Sample Characteristics

We next present available demographic information on the subset of participants who reported on sex drive manifestations (i.e., frequency of sexual cognition, affect, and behavior).<sup>7</sup> Their average age was 30 years old (information available for 84% of the samples), 89% were heterosexual (50% of samples with information), 76% were White (29% with information), 53% were religious

(11% with information), 38% were single (56% with information), and 59% were university students (45% with information). Data for sex drive manifestations was collected between 1992 and 2019 (*mean* = 2011, *median* = 2012) and 90% of effect sizes were computed from raw data.

For some samples, codings for sexual orientation, age, and relationship status were available separately for men and women, allowing us to calculate weighted differences scores (i.e.,  $\Delta$ s): Male participants were more likely to be heterosexual ( $\Delta = 1.74\%$ , 36% with information), older ( $\Delta = 1.81$  years, 36% with information), and more likely to be single ( $\Delta = 1.84\%$ , 36% with information). Across all studies that included a sex drive manifestation, 50% were published in sexuality journals (100% with information), 54% had female first authors (98% with information), 88% were published (100% with information), 60% used electronic data collection (85% with information), 34% documented reassuring participants about privacy (88% with information), 49% focused on gender differences in sex drive (87% with information), 35% rewarded participants materially (25% course credit, 9% mixed, 31% no reward; 57% with information), and 80% were advertised as studies on sexuality (47% with information). The studies were mostly conducted in North America (52% total; United States: 79%, Canada: 13%, Mixed: 5%, Mexico: 2%, Costa Rica: less than 1%) and Europe (43% total; Germany: 54%, Portugal: 11%, United Kingdom: 9%, Spain: 9%, Norway: 4%, Croatia: 2%, Estonia: 2%, Italy: 2%, Others: 10%), with some from Asia (4% total; Japan: 38%, China: 31%, Israel: 13%, Turkey: 13%, India: 6%), Oceania (1% total; Australia: 100%), and Africa (less than 1% total; Cameroon: 100%). Further information on the codings is summarized in Table 5 for the sex drive manifestations and Table S6 for the indicators of latent sex drive.

### Correlation Structure of Outcomes

Figure 3 depicts the meta-analytic correlation table. As expected, the sex drive manifestations and indicators of latent sex drive formed a coherent cluster (box with solid line in Figure 3), providing evidence for convergent validity. Summary effects for Pearson correlations ranged from  $r = 0.28$  to  $r = 0.65$ . Correlations between sex drive indicators and bias indicators were lower and less consistent, but still positive.

<sup>7</sup> We note that in some cases, whether demographic information was reported was likely correlated with the demographic information itself. For example, studies that reported the percentage of college students likely included a

larger percentage of students than studies that did not report this information. Consequently, we should be cautious in generalizing these statistics to the full sample.

## Outlier Analysis and Treatment

We conducted leave-one-out analyses for the sex drive manifestations to detect outliers with a notable influence on the summary effects. We repeatedly fitted the model for predicting effect sizes from sex drive facet (frequency of sexual cognition, affect, and behavior) in an RVE meta-regression while dropping each effect size once. We then examined the change in estimated summary effects and standard errors resulting from dropping the effect size. Results for the leave-one-out analyses are depicted in Figure S2. There were no notable outliers for cognition frequency (CF),  $(\Delta g)_{\min} = -0.0043$ ,  $(\Delta g)_{\max} = 0.0038$ , nor for behavior frequency (BF),  $(\Delta g)_{\min} = -0.0194$ ,  $(\Delta g)_{\max} = 0.0206$ . For affect frequency (AF), however, one study had an outsized influence,  $\Delta g = -0.0436$ ,  $\Delta SE = 0.0169$ . This study examined older couples (average age = 74.60 years) and found a medium-to-large effect size indicating higher frequency of sexual affect in women,  $g = -0.64$ . This outlier is also clearly visible in the corresponding funnel plot (see Figure 5, middle panel, effect farthest left). We removed this outlier from all further analyses. We applied the same procedure to the indicators of latent sex drive and bias indicators, respectively. We removed one effect size for affect intensity and one for sexual intercourse frequency (see Figure S2). Some effect sizes were additionally removed for the moderation analyses based on visual inspection of the scatter plots. These are reported in the Method section.

## Gender Differences in Sex Drive

Full results for sex drive manifestations, indicators of latent sex drive, and bias indicators are displayed in Table 2. We first analyzed the sex drive manifestations. We found significant, medium-to-large gender differences in sexual cognition frequency,  $g = 0.76$ ,  $CI_{95} [0.71, 0.80]$ , sexual affect frequency,  $g = 0.58$ ,  $CI_{95} [0.49, 0.66]$ , and sexual behavior frequency,  $g = 0.75$ ,  $CI_{95} [0.66, 0.84]$ . The difference between the three facets was significant,  $AHZ(68.53) = 7.26$ ,  $p = .001$ . The absolute amount of heterogeneity was medium in magnitude,  $\tau = 0.21$  (Van Erp et al., 2017). Most of the variation in observed effects was estimated to be due to variation in true effects rather than sampling error,  $I^2 = 91.03$ .

For the indicators of latent sex drive, there were significant small-to-medium and medium-sized gender differences for affect intensity,  $g = 0.40$ ,  $CI_{95} [0.35, 0.45]$ , and for self-rated sex drive,  $g = 0.63$ ,  $CI_{95} [0.35, 0.92]$ . The difference between these two indicators was not significant,  $AHZ(4.00) = 5.75$ ,  $p = .074$ . Again, the absolute amount of heterogeneity was medium-sized,  $\tau = 0.15$ , and overall variation was estimated to be due to variation in true effects rather than sampling error,  $I^2 = 90.45$ .

Out of the 612 effect sizes relating to sex drive manifestations or indicators of latent sex drive, only 17 (2.8%) showed a descriptively larger sex drive in women (indicated by an effect size of  $g < 0$ ).

## Gender Differences in Potentially Biased Responding

Next, we analyzed the bias indicators (see Table 2). There was no significant gender difference for sexual intercourse frequency,  $g = 0.04$ ,  $CI_{95} [-0.09, 0.17]$ . In contrast, gender differences were significant for total one-night stands,  $g = 0.21$ ,  $CI_{95} [0.18, 0.25]$ , total sexual partners in the last year,  $g = 0.15$ ,  $CI_{95} [0.11, 0.19]$ , and total sex partners,  $g = 0.19$ ,  $CI_{95} [0.02, 0.36]$ . The difference between these indicators was significant,  $AHZ(27.75) = 7.49$ ,  $p < .001$ . The heterogeneity was comparable to sex drive manifestations and indicators of latent sex drive,  $\tau = 0.16$ ,  $I^2 = 80.41$ .

Logic implies that there should be practically no gender differences on any of these indicators, given the premises that (a) the participants included in the primary studies constitute a representative sample of the heterosexual population and (b) all participants responded truthfully. If these premises hold, an empirical gender difference could emerge only if on average men overreported and/or women underreported (or vice versa) due to any motivational and/or cognitive biases that may have influenced responses in an invalid way. However, approximately 11% of the participants in our sample were homosexual. For this subsample (and consequently to a lesser extent for the overall average estimate), valid positive gender differences on the bias indicators in favor of men (e.g., suggesting more one-night stands by men than women) are plausible if homosexual women were less promiscuous and had sex less often than homosexual men (and vice versa for valid negative gender differences).

As a preliminary, posthoc test of this possibility, we conducted meta-regression analyses for the bias indicators, regressing each indicator on the percentage of heterosexual participants in the sample. If the gender differences on the bias indicators are driven by differences in sexual behavior between homosexual men and homosexual women, the effect sizes should become larger if there are less heterosexual (and hence more homosexual) participants in the sample. Contrary to the expectation, all slopes were descriptively positive, indicating larger gender differences if the sample included more heterosexual participants. There was insufficient data for the bias indicators total partners and sexual frequency to conduct significance tests for the slopes ( $df < 4$ ). For number of one-night-stands and number of partners during the previous year, both tests were not significant ( $p$ 's

> 0.183). These tests were not pre-registered. We cautiously interpret them as evidence against the possibility that the gender differences we obtained for the bias indicators are driven by gendered same-sex sexuality.

### Global Summary Effect, Adjustment for Response Bias, and Natural Language Interpretation

In the previous sections, we reported summary effects separately for each sex drive indicator and bias indicator, respectively. To estimate a global summary effect of the gender difference in sex drive, we fitted a random-effects meta-analysis model with equal weights to the summary effects of the sex drive manifestations (frequency of sexual cognition, affect, and behavior). The results are displayed in Table 2. The global summary effect was  $g = 0.69$ ,  $CI_{95} [0.58, 0.81]$ .

We also computed a summary effect for all four bias indicators that may indicate a gender difference in (potentially) biased responding. This summary effect was  $g = 0.15$ ,  $CI_{95} [0.08, 0.22]$ . Assuming that the size of this summary effect is completely driven by men's and/or women's biased responding, then subtracting this bias effect estimate from the summary effect of sex drive differences should establish a global summary effect adjusted for response bias. This bias-adjusted global summary effect was of moderate size:  $g = 0.54$ . We note, however, that in fact the bias indicators may be more strongly affected by response bias than the sex drive indicators. For example, reporting the number of sexual partners (a bias indicator) may be more prone to self-presentation tendencies than reporting the number of sexual thoughts (a sex drive indicator), and responses to the former may be afflicted with stronger forms of normative pressures. If this reasoning is correct, subtracting the complete bias indicator summary effect constitutes an over-correction. We therefore view the summary effect of  $g = 0.54$  as a lower bound for the true bias-adjusted gender difference in sex drive.

Thus, based on the available evidence, we estimate male sex drive to be 0.69 standard deviations stronger than female sex drive on average. Out of this difference, up to 0.15 standard deviations may be attributable to biased responding, such that the true difference may lie between 0.54 and 0.69 standard deviations (not considering the respective confidence intervals around these point estimates).

Standardized effect sizes are well-suited to compare effects across different studies, but it can be difficult to comprehend what a standardized effect size actually means in more intuitive terms. To make this summary effect more easily

interpretable, we now report natural language interpretations. Corresponding values for the fully adjusted summary effect are presented in parentheses. An effect of  $g = 0.69$  (adjusted:  $g = 0.54$ ) means that 76% (adjusted: 71%) of all men will have a stronger sex drive than the average sex drive among women (Cohen's  $U_3$ ). It also indicates that 73% (adjusted: 78%) of men's and women's sex drive distributions overlap (overlapping coefficient  $OVL$ , also see Figure S5). Finally, a  $g = 0.69$  (adjusted:  $g = 0.54$ ) indicates that the probability of a randomly picked man having a higher sex drive than a randomly picked woman picked is 69% (adjusted: 65%, common language effect size  $CL$ ). For  $U_3$  and  $CL$ , switching from the men's to the women's perspective provides a different, yet also worthwhile angle on the statistics: 24% (adjusted: 29%) of women have a larger sex drive than the average man, and the probability of a random woman having a higher sex drive than a random man is 31% (adjusted: 35%). We note that these interpretations remain relative in nature and do not speak to whether the difference is practically relevant in absolute terms.

### Publication Bias

Funnel plots for cognition, affect, and behavior frequency appeared highly symmetric in visual inspection, revealing no indication of bias (Figure 5, see also Figures S3 and S5 for funnel plots of the indicators of latent sex drive and bias indicators, respectively). This impression was confirmed by both Egger's regression tests and the bootstrapped 3PSM tests, which found no indication of publication bias or small-study effects for cognition (Egger:  $p = .149$ ; 3PSM:  $p = .812$ ), affect (Egger:  $p = .940$ ; 3PSM:  $p = .706$ ), or behavior frequency (Egger:  $p = .629$ ; 3PSM:  $p = .271$ ).

### Moderation Analyses

#### *Sex Drive Manifestations*

For the sex drive manifestations, tests for moderation by various characteristics of the publication, study, sample, and outcome are summarized in Table 3. Corresponding regression tables are summarized in Table 4. Selected analyses are graphically displayed in Figure 4. Small-sample corrections were employed for all statistical tests (Tipton, 2015; Tipton & Pustejovsky, 2015). These provide reliable results when degrees of freedom are larger than 4. We refrain from reporting  $p$ -values when this threshold of  $df > 4$  is not reached. We did not conduct moderation analyses for contraceptive use and religiosity, as insufficient information was available for these codings.



**Cognition frequency.** There was one very strong moderation pattern for frequency of sexual cognition. Specifically, gender differences were notably larger when the item captured sexual cognitions about extra-pair partners (i.e., others outside of one's current relationship),  $g = 0.82$  (item content, e.g., "How often do you have fantasies about having sex with someone you are not in a committed romantic relationship with?"), as opposed to smaller effects for sexual cognitions about a non-specific partner (e.g., "How often do you think about sex with a partner?"),  $g = 0.58$ , or non-specific sexual cognitions without mentioning any partner (e.g., "How often do you think about sex?"),  $g = 0.57$ , test for difference:  $AHZ(49.54) = 21.00$ ,  $p < .001$ . Closer examination of the data revealed that this item content coding was correlated with other codings. For example, studies using items about extra-pair partners were more often conducted by male first authors and more often focused on gender differences in sex drive specifically. We consequently repeated all moderation analyses while statistically controlling for this characteristic, collapsing cognitions about a non-specific partner and non-specific sexual cognitions into one category to achieve a binary control variable. We report the controlled tests in Tables 4 and 5. The uncontrolled tests are reported in the supplemental materials, Tables S4 and S5. This was not anticipated and therefore not preregistered.

After controlling for item content (extra-pair vs. other), there were five significant moderation tests. Gender differences were larger when participants were asked to aggregate frequency of sexual cognitions across a larger period (e.g., "Over the past month, how often have you fantasized about sex?") compared to smaller periods (e.g., "How often do you think about sex on a typical day?"),  $AHZ(5.57) = 8.46$ ,  $p = .029$ . Two analyses suggest that not having access to a sexual partner may lead to increases in sexual cognitions for men, decreases for women, or both—in any case, gender differences in sex drive were more pronounced: First, studies that did not restrict sampling to sexually active participants reported larger differences,  $AHZ(20.04) = 4.99$ ,  $p = .037$ . Second, gender differences were more pronounced when the sample contained a larger percentage of single participants,  $AHZ(26.18) = 7.21$ ,  $p = .012$ .

Further, gender differences were larger when studies used either group assessment,  $g = 0.63$ , or individual assessment,  $g = 0.58$ , compared to studies that used both types of assessment,  $g = 0.38$ , test for difference:  $AHZ(20.13) = 3.86$ ,  $p = .038$ . However, this moderation finding is not straightforward to interpret, as one would expect the results for the "both" coding to fall between the other two if the pattern were meaningful. Finally, gender differences were slightly larger in studies that were not advertised as studies on sexuality,  $g = 0.66$ ,

compared to studies that were,  $g = 0.55$ , test for difference:  $AHZ(37.96) = 4.37$ ,  $p = .043$ .

**Affect frequency.** There were four significant moderation tests. The gender difference was larger when there was no 'content' or target of sexual desire specified (e.g., "How often do you feel sexual desire?") compared to items that mentioned an unspecified partner (e.g., "How often do you feel desire for sex with a partner?"),  $AHZ(39.15) = 4.24$ ,  $p = .046$ . Further, studies by female first authors revealed larger gender differences,  $AHZ(45.05) = 4.36$ ,  $p = .043$ . In the same vein, research teams with a larger percentage of female authors found larger gender differences in affect frequency,  $AHZ(23.18) = 9.22$ ,  $p = .006$ . Further, the gender difference decreased when a larger percentage of participants were single,  $AHZ(12.72) = 5.75$ ,  $p = .033$ . Four tests did not reach the threshold of  $df > 4$  due to low number of studies and effect sizes.

**Behavior frequency.** For behavioral frequency, only the percentage of university students in the sample moderated gender differences significantly,  $AHZ(7.80) = 9.54$ ,  $p = .015$ , such that the gender difference was more pronounced when the sample included more university students. Six tests did not reach the threshold of  $df > 4$ .

### *Indicators of Latent Sex Drive*

The results are summarized in Tables S4 and S5. For self-reported sex drive, there were too few studies and effect sizes to conduct meaningful moderation analyses. For sexual affect intensity, three moderation patterns emerged. Gender differences were larger when the aggregation span for the response scale was larger (e.g., two weeks versus two days),  $AHZ(11.37) = 7.07$ ,  $p = .022$ . Item content also had a significant influence,  $AHZ(23.05) = 15.76$ ,  $p < .001$ , such that gender differences were larger for desire for sex when no target was mentioned, (content = 'no target':  $g = 0.45$ ), and desire for masturbation (content = 'masturbation':  $g = 0.49$ ), and smaller for desire for sex with an unspecified partner (content = 'unspecified partner':  $g = 0.27$ ), or specifically one's own partner (content = 'own partner':  $g = 0.27$ ).

The context in which desire occurred was also relevant,  $AHZ(14.16) = 21.41$ ,  $p < .001$ : Gender differences were very small for sexual desire in romantic situations,  $g = 0.09$ , small for desire while having sexual thoughts,  $g = 0.23$ , small-to-medium for non-specified contexts,  $g = 0.43$ , medium-sized for while spending time with an attractive person,  $g = 0.50$ , and medium-to-large for when first seeing an attractive person,  $g = 0.67$ .



### *Interim summary*

The comparably small number of significant moderation analyses despite the multitude of theory-driven and methodological moderator candidates coded (see Table 3) suggests that the gender differences in sex drive facets are remarkably robust. This view is further corroborated by a different perspective on the moderator analyses. Up to this point, we have discussed the moderator analyses as a function of sex drive facet (cognition, affect, behavior). To examine the robustness of a moderator, it is also informative to inspect whether a significant moderator in one facet also moderates gender differences in one of the other facets. The only moderator for which this was the case was the percentage of participants who were single. As this percentage increased, gender differences increased for sexual cognition frequency and decreased for sexual affect frequency. All other moderators were significant for only one facet despite the facets being substantially positively correlated (Figure 3). No moderator was significant for all three sex drive facets. This further suggests that there are few substantial moderating factors of gender differences in sex drive.

### **Discussion**

Sex drive and particularly the notion of gender differences in sex drive have sparked considerable debate. This debate has been afflicted by underdeveloped conceptualizations and heterogeneous measurements of sex drive, making it difficult to structure and compare the diverse findings. The present article seeks to make two substantial contributions—first, a theory-driven coherent conceptualization of sex drive, and second, a comprehensive meta-analysis of gender differences in sex drive that adheres to current best-practice standards for quality, reproducibility, and transparency (Lakens et al., 2016; Moher et al., 2009).

We understand sex drive as an individual's intrinsic motivation to obtain sexual experiences and pleasure. This latent motivation is expected to manifest in the psychological triad of sexual cognition, affect, and behavior, and to vary both within and between individuals. Building upon modern and integrative concepts of personality (Fleeson, 2001; Fleeson & Jayawickreme, 2015; Johnson, 1997; Roberts, 2009), we propose that individuals differ in their typical (trait) level of sex drive, without questioning intraindividual (state) variability. This conceptualization is not only rooted in seminal understandings of the nature of personality traits (McCrae & Costa, 2003; Roberts, 2009), it also provides a clear rationale for deriving

suitable indicators of sex drive: the frequency of sexual cognitions, affect, and behaviors.

The meta-analysis includes a total of 621,463 persons from 211 studies and 856 effect sizes. Overall, we found a stronger sex drive in men compared to women with a moderate-to-large effect size ( $g = 0.69$ ,  $CI_{95} [0.58, 0.81]$ ), confirming previous findings (Baumeister et al., 2001). Summary effects varied across sex drive facets – that is, the three sex drive manifestations – from moderate for affect ( $g = 0.58$ ) to moderate-to-large for cognition ( $g = 0.76$ ) and behavior frequency ( $g = 0.75$ ). A meta-analysis of within-study correlations between sex drive manifestations and indicators of latent sex drive provided evidence for our conceptualization's convergent validity. We also examined variables that should logically not reveal any substantive gender differences (e.g., total sex partners or one-night stands), and thus may be indicative of biased responding. Across multiple of these response bias indicators, we found small gender differences on average ( $g = 0.15$ ). We then subtracted the effect size for potential bias ( $g = 0.15$ ) from the meta-analytic gender difference in sex drive ( $g = 0.69$ ) to arrive at an estimate of the lower-bound of bias-adjusted gender differences in sex drive:  $g = 0.54$ , a medium-sized effect. Since this may or may not constitute an over-correction, we argue that a range of point estimates of  $g = 0.54$  to  $g = 0.69$  best represents our main finding (see the next section for a discussion of possible biased responding).

To put this finding into perspective, we relied on natural language interpretations for this effect size range: overlap, non-overlap, and probability of superiority. These interpretations indicated that, assuming normality, the distributions of male and female sex drive greatly overlapped (73-78%), that the average man has a lower sex drive than 24-29% of women, and that the probability of a random woman having a higher sex drive than a random man is 31-35%. Particularly the latter interpretation is quite intuitive: When a woman with an unknown sexual motivation walks down the street, she will on average exceed every third man she encounters in her drive to pursue sexual gratification.

We also applied the bias correction procedure to the summary effects within the subcategories to attain lower-bound estimates for each indicator. After correction, gender differences were medium-to-large for cognition frequency ( $g = 0.61$ ) and behavior frequency ( $g = 0.60$ ), medium-sized for affect frequency ( $g = 0.43$ ) and self-rated sex drive ( $g = 0.49$ ), and small for affect intensity ( $g = 0.25$ ).

Analyses of effect size heterogeneity ( $I^2$ ) showed consistently that 80% or more of the observed variation in effect sizes was not due to sampling error, but rather variation in the true effects. This is not surprising given that our analyses included very large studies, some with thousands of participants. There should be little sampling error in such large

studies, so any excess variability will be attributed to true effects. When considering the absolute variation in true effects ( $\tau$ ) rather than the proportion of variation due to true effects ( $I^2$ ), heterogeneity was average compared to other meta-analyses in psychology (Van Erp et al., 2017).

Apart from natural language interpretations of the summary effect, it can also be informative to compare empirical effects with benchmarks to put them in perspective (Funder & Ozer, 2019). In terms of common statistical effect sizes ( $g$ , Hedges, 1981), the obtained gender differences are considerably larger than many other gender differences in the domain of sexuality (Petersen & Hyde, 2010) and gender differences from a broad variety of other domains (Hyde, 2014), but of similar magnitude as some domains known to exhibit reliable gender differences, such as spatial cognition and physical aggression (Hyde, 2014). Even after conservatively correcting for potential gender-specific response bias, the effect sizes are also larger than most effect sizes in social psychology and research on individual differences (Gignac & Szodorai, 2016; Richard et al., 2003). Broadening the perspective to domains other than psychology, the effects are in a similar range as the gender difference in weight for U.S. adults ( $d = 0.54$ ), but less than half the size of the gender difference in height for U.S. adults ( $d = 1.81$ ; Meyer et al., 2001).

Although these comparisons of statistical effect sizes help to situate the present effects in the context of other bodies of literature, they leave the substantial question unanswered what effect sizes of this magnitude really mean in everyday life. For example, it is unclear how these observed gender differences influence heterosexual dating behavior or the dynamics of heterosexual long-term relationships in the context of various other influences—such as socially-learned behavioral patterns and expectations, the partners' impression management considerations, or the distribution of gender differences in sex drive across heterosexual couples. After all, we analyzed facets of sex drive that are usually not readily observable to others (cognitions, affect, masturbation behavior). Does sex drive manifest in observable behaviors in everyday life? And if so, how? How accurate are women's and men's perceptions of other's sex drive? These questions are pivotal, but they cannot be answered based on the current data. It is up to future research to answer these questions and to disentangle the actual effects of gender differences in sex drive from perceived gender differences in order to reveal the real-world implications of the present findings.

One key feature of the present meta-analysis is that it revealed gender differences in relative rather than absolute terms. On any absolute scale, it may be that both men and women have a high sex drive, and that men's is merely a little higher. Similarly,

both men and women could be regarded as relatively low in sex drive on an absolute scale, women just somewhat lower than men. The key insight behind this observation is that the present findings by no means imply that women generally have a low sex drive or that men generally have a high sex drive. It is impossible to come to an absolute conclusion based on the present analysis (e.g., that men's sex drive is  $X$  times higher than women's sex drive).

## Biased Responding

Sexuality is a sensitive topic, which begs the question as to what extent reporting biases may have influenced our results. Some evidence suggests that women tend to underreport and men tend to overreport permissive sexual attitudes and behaviors (e.g., Alexander & Fisher, 2003; Jonason, 2008; Mitchell et al., 2019), possibly due to different social punishments and rewards for these behaviors (Endendijk et al., 2020). In the case of the behavioral facet of our sex drive conceptualization, some evidence suggests that masturbation can be associated with shame and guilt for women (Kılıç Onar et al., 2020). To the extent that this is the case, this may bias reports about gender differences in masturbation. In light of these considerations, one may wonder: How likely are biased response tendencies to drive the gender differences found in the present analysis?

First, we argue that some of the sex drive facets derived from our conceptualization are less prone to biased reporting than other constructs for which bias has been previously documented (e.g., sexual attitudes or number of sex partners). For instance, for a woman who is concerned with not appearing too sexually permissive, it may be easier to report frequent sexual thoughts than to report having had many different sex partners. In addition, men may stand to gain little social status by reporting that they think about sex frequently and masturbate a lot. With respect to masturbation specifically, a meta-analysis revealed no significant gender differences in attitudes toward masturbation ( $d = 0.02$ , Petersen & Hyde, 2010), suggesting that attitudes toward masturbation will not affect the genders differently. For men, masturbating a lot may be seen as nothing to brag about, because it may indicate that a man cannot fulfill his sexual needs with actual sexual intercourse, but has to resort to masturbation. This would argue against a strong bias (or any for that matter) toward larger gender differences in masturbation that originated from biased responding. Consistent with the notion that masturbating a lot is not necessarily a desirable characteristic for men, a recent experimental study demonstrated a reversed sexual double standard for masturbation, such that men received social punishment for masturbating; they were seen as

lower quality partners than women who masturbated (Haus & Thompson, 2020). In a similar vein, one reviewer suggested that in recent decades the public discussions have tended to encourage female sexuality, while (strong) male sex drive has frequently been viewed more critically, pointing to the possibility that sexual double standards may be shifting, at least in western societies. This could even lead females to overreport and males to underreport sexual thoughts, desires, and behaviors. Despite these preliminary findings, the question if and to what extent the sex drive indicators used in the context of the present conceptualization are prone to bias measured gender differences towards larger or smaller values than warranted on the construct level is an important question for further research.

Second, our moderator analyses found no evidence for moderation by characteristics of the primary studies that should affect perceived privacy, such as group assessment or personal contact with the research team. Similarly, there was no evidence that gender differences have decreased over time. Had these differences been driven by biased responding, a decrease would have been plausible considering societal changes toward less restrictive social norms and attitudes toward sexuality.

Finally, we examined gender differences in several bias indicators that should theoretically exhibit little to no substantive gender differences in heterosexual populations (e.g., number of total sex partners or one-night stands). These analyses suggest that biased responding may have indeed played a role, but that this effect was small ( $g = 0.15$  at most). The effect may be driven by social norms through unconscious or subconscious influences, such as memory errors, different estimation strategies, or differential accounting for “edge cases” of having had sex, but they may also at least partly be driven by self-presentation tendencies for men to overreport and/or women to underreport their sexual experiences. There are arguments to be made that subtracting this estimate of response bias from the gender difference in sex drive could be an overcorrection. Due to their characteristics (i.e., all behavioral; all but intercourse frequency typically found in the literature on Sexual Double Standards, Endendijk et al., 2020), these bias indicators may be even more prone to (gender-specific) biased responding than the sex drive manifestation items. Also, it could be that these measures do not indicate pure bias, but that they partly reflect true differences due to undersampled sub-populations such as sex workers (versus consumers of sex work) or gender-specific responses among homosexual persons (i.e., homosexual men may have sex more frequently and may have more sexual partners compared to homosexual women). We found no association between the percentage of homosexual participants in the sample and gender differences on the bias

indicators. This speaks against the possibility that some of the gender difference we obtained for the bias indicators is valid, rather than pure bias, but does not rule it out. Yet, even when taking the full mean gender difference of these bias indicators as a proxy for the extent of motivated response bias and correcting for bias in the main analyses—a quite conservative approach—a substantial gender difference of approximately medium size (Cohen, 1988) remains. This indicates that the identified gender difference in sex drive is unlikely to solely be the result of biased responding.

## Publication Bias

The academic incentive structure of recent decades has strongly favored the file-drawering of findings that did not reveal the hoped-for outcome (Nosek et al., 2012). As a result, publication bias is widespread in the social sciences (Fanelli, 2010, 2012). This is concerning, given that severe publication bias can strongly distort meta-analytic effect size estimates (Friese & Frankenbach, 2020). For several reasons, publication bias was unlikely to affect present meta-analysis. First, for maximum fit with our theoretical conceptualization, we extracted individual items from a diverse array of larger inventories. Thus, we analyzed a different subset of data than the primary researchers. Second, gender differences in sex drive were not focal to many of the original studies. This means that whether (and to what extent) gender differences in sex drive emerged was likely not relevant for many authors when deciding how to proceed with their projects once the data were analyzed. Third, we included unpublished data, which counteracts publication bias.

## Implications for Theory

In the introduction, we reviewed a set of psychological theories that either make predictions about a gender difference in sex drive or rely on its existence as a theoretical pre-requisite. In this section, we discuss the implications of our findings for these theories.

Sexual strategies theory (Buss & Schmitt, 1993) posits that women have evolved to show more sexual restraint and selectivity than men, because for them, the evolutionary stakes are much higher in sexual encounters (i.e., women bear the biological risks and opportunity costs of pregnancy, childbirth, and infant care). The theory does not speak directly to gender differences in sex drive, but a stronger generalized motivation to pursue sex among men seems more plausible under the assumptions of sexual strategies theory than vice versa. Our results are consistent with this perspective.

Social role theory (Eagly & Wood, 1999) and social learning theory (Bussey & Bandura, 1999) state that men and women experience different social role expectations and social reward patterns, respectively. Empirical observation suggests that such differences in social context do indeed exist, such that the expression of sexuality tends to be encouraged for men but sanctioned for women (e.g., sexual double standard hypothesis, Crawford & Popp, 2003). Notably, both theories predict gender differences both on the construct level and on the measurement level. In other words, men and women may actually think, feel, and act in ways consistent with gender specific roles and reward patterns (i.e., they may truly have different sex drives), but they may also “just” self-present in different ways in order to conform with their social context (i.e., they exhibit response bias). Our results provide support for both of these possibilities: Men and/or women may not answer fully truthfully to questions regarding their sex drive, but we also found a substantial true gender difference in sex drive above and beyond biased responding.

Sexual economics theory (Baumeister & Vohs, 2004) is rooted in the assumption that men are more interested in sex than women and posits that, as a result, the negotiation and exchange of heterosexual sexuality follows the pattern of an economic marketplace, in which men offer resources to obtain sex from women. An empirical refutation of this assumption would have rendered the theory void of its first and most central tenet. Despite providing support for this particular tenet, we note that our findings on sex drive neither prove nor disprove sexual economics theory itself. Our results confirmed a prediction that can be derived from the theory, namely that there should be gender differences in biased responding regarding the tallies of past sexual partners and sexual engagements due to the differential signaling implications for men and women (low tallies for women signal higher value of sex deserving greater male investment of resources; high tallies for men signal the ability to obtain sex through other resources).

Finally, the gender similarity hypothesis (Hyde, 2005) states that similarity between men and women is the norm and dissimilarity the exception. Our results suggest that, in addition to previously documented exceptions like physical aggression, mental rotation, or spatial perception, sex drive is another notable exception where robust gender differences exist.

### Some Evidence for Moderation

Uncorrected gender differences were large for frequency of sexual cognition ( $g = 0.76$ ) and behavior ( $g = 0.75$ ), moderate for frequency of

sexual affect ( $g = 0.58$ ) and self-rated sex drive ( $g = 0.63$ ), and yet somewhat smaller for intensity of sexual affect ( $g = 0.40$ ). These differences could be rooted in the underlying temporal sequence of psychological processes that might mediate the emergence of sexual events: A sexual episode may start with some fleeting sexual affect or impulse, triggered by internal or external stimuli. This impulse may lead to more developed cognitions about sex, a sexual fantasy perhaps, which is then later enacted in solitary or partnered sexual behavior. Along this process, men and women may differ in their ability and/or motivation to inhibit sexual experience and behavior. It could be that processes further downstream are easier to regulate, that is, sexual cognitions and behavior are easier to regulate than affect. Accordingly, gender differences may be exacerbated for cognition and behavior compared to affect if, on average, women are more motivated or men less able (and motivated) to inhibit sexuality. Note, however, that this is speculative at this point. The temporal sequence of events could also typically start with a fleeting thought or fantasy (cognition) that sometimes develops into a sexual desire (affect). This would be in line with a recent experience-sampling study that found more frequent sexual cognitions than sexual affect, and more frequent sexual affect than sexual behavior (Weber et al., 2022a). Addressing such process-related questions would require more fine-grained data that allows examining the temporal sequence of the occurrence of events in everyday life.

Apart from the differences between the sex drive indicators, there was relatively little reliable evidence for any of the many theoretically derived and methodological moderator variables. The overall gender differences were remarkably stable. Nevertheless, despite this general impression of remarkable effect size consistency, a noteworthy pattern emerged for sexual cognitions: Gender differences were considerably more pronounced when sexual cognitions pertained to an extra-pair partner (i.e., a person the respondent is not in a relationship with; large gender difference) compared to an unspecified partner (medium-sized gender difference). This result is in line with previous findings on gender differences in sociosexuality (Lippa, 2009; Simpson & Gangestad, 1991). Other moderation findings were smaller in magnitude, and in several cases, tests barely crossed the significance threshold of  $p = .05$ . Type 1 error rates may be inflated due to multiple testing of moderators, so caution should be exercised in interpreting these findings.

Moderation patterns arose relating to the phrasing of questions. Items with a larger aggregation span (e.g., daily frequency of sexual fantasies over 30 days versus 3 days) yielded larger gender differences for cognition frequency and

affect intensity. A natural explanation for this effect is that more aggregation leads to more precise estimates and hence larger effect sizes. Alternatively, this pattern may also point to a previously undiscussed source of response bias. When participants retrospectively report how often or how intensively psychological states occurred over a period of time, longer time periods may involve more uncertainty, guesswork, and ultimately more response bias due to reliance on stereotypes: In the face of uncertainty, people may draw more heavily on perceived societal norms, which may reward disclosure of sexuality for men and punish it for women (Crawford & Popp, 2003). Larger observed gender differences for longer aggregation spans may thus either be closer to the true difference due to more accurate measurement or, instead, farther away from the true difference due to more response bias in line with societal norms. This is left for future primary research to find out.

There was an inconsistent association between gender differences in sex drive and the percentage of singles in the sample. A larger percentage of singles in the sample was associated with larger gender differences in sexual cognitions. For sexual affect, though, the pattern was reversed: A larger share of singles in the sample was associated with a smaller gender difference. This inconsistent pattern may result from some or all the following processes: Being single may (a) increase sexual cognitions in men, (b) decrease sexual affect in men, (c) decrease sexual cognitions in women, or (d) increase sexual affect in women. However, while theoretically interesting, the underlying processes remain speculative, and this potential three-way interaction pattern should be replicated and further illuminated in dedicated primary research.

Gender differences in sexual affect intensity were strongly dependent on the content and context of desire evoked by the questionnaire item. The difference was comparatively smaller in romantic situations ( $g = 0.09$ ), and when a partner ( $g = 0.27$ ) or the participant's long-term partner ( $g = 0.27$ ) were referenced. The largest difference was obtained for items that gauged sexual desire "when first seeing an attractive person" ( $g = 0.67$ ). Taken together, these observations point to the possibility that gender differences in sexual motivation may be larger when intimate relationships are not yet established and may decrease after the relationship has been initiated. However, future research needs to consolidate these possibilities with recent longitudinal evidence showing that gender differences in sex drive increase over the course of a long-term relationship (McNulty et al., 2019, see the discussion of partnership duration as a moderator below).

Apart from these isolated findings, no consistent moderation patterns were found. Some of the non-emergent moderator effects can be cautiously

interpreted as evidence for the robustness of the results. For example, similar gender differences in sex drive emerged whether or not the study focused on gender issues and whether or not it focused on gender differences in sex drive (with the exception of a slightly less pronounced gender difference for cognition frequency when studies focused on gender differences in sex drive or aimed to find them). Likewise, similar effects emerged whether a study was advertised as a sexuality study to participants or as a study primarily concerned with a different domain (again apart from a slightly smaller gender difference for cognition frequency found in 'sexuality' studies). As mentioned in the previous section, a range of other characteristics relating to perceived privacy in the study context (e.g., group assessment, personal contact with the research team) did not emerge as moderators.

For other factors, it was more surprising that moderation effects did not emerge. Previous longitudinal evidence suggested that gender difference in sex drive increase over the course of a relationship (McNulty et al., 2019), yet partnership duration did not emerge as a moderator in the present meta-analysis. We are hesitant to over-interpret this null-finding as meta-analytic analyses on the sample level have much lower resolution than dedicated participant-level work. We note, however, that the finding is consistent with a recent machine learning study showing that it is difficult to predict differential changes over the course of a relationship from baseline variables like participant sex (Joel et al., 2020). Neither age, year of study, or gender inequality exerted a significant moderating effect: Gender differences in sex drive remain relatively stable across the life span, across countries with different gender inequalities, as well as across time, indicating that previous findings on gender differences in sex drive continue to hold true today (Baumeister et al., 2001). It is worth noting that there seem to be no changes in effect size over time during the covered period. This could be tentatively interpreted as supporting an evolutionary perspective on gender differences. If, in contrast, the gender difference was a primarily cultural product, the effect size should have changed (and become smaller) with a changing culture. Then again, progress towards gender equality (in terms of educational and economic attainment) has slowed down since the turn of the millennium (England et al., 2020). The lack of change in gender differences over time may reflect this development.

It would also have been plausible to expect that gender differences in sex drive vary across the life span. Moreover, both social learning theory (Bussey & Bandura, 1999) and social role theory (Wood & Eagly, 2012) predict that gender differences have decreased over time as gender stereotypes and gender inequality decreased. These theories also predict that gender differences are less pronounced

in countries with less gender inequality. Our results are consistent with a meta-analysis of gender differences in sexual behaviors and attitudes, which found no moderation by age and year of study for masturbation (Petersen & Hyde, 2010). The findings are inconsistent with evidence from the same meta-analysis indicating a larger gender difference in masturbation in countries with greater gender inequality. Corroborating the present findings, a large-scale study across 53 nations also found no moderation by gender inequality for self-rated sex drive (Lippa, 2009).

Although the non-emergence of these moderating effects is theoretically surprising, they may have a mundane methodological explanation: range restriction. Year of the study only ranged from 1992 to 2019, which may not have been long enough to capture long-term cultural changes. Similarly, most data stemmed from countries with relatively high levels of gender equality, rendering tests for moderation by country-level GDI and GII less informative than desired.

One variable may potentially impact gender differences in sex drive but can unfortunately not be tested in our data: women's menstrual cycle phase. Women may be less likely to experience sexual cognitions, affect, and desire for masturbation during certain phases of their cycle. If this were to be the case, frequency-based indicators of sex drive may be biased towards lower estimates for women compared to men. By contrast, men's sexuality does not fluctuate along a stable monthly cycle. Relevant to this concern, a recent large-scale diary study based on more than 26,000 self-reports by more than 1,000 women found hardly any changes in both in-pair and extra-pair desire across 40 days for women using hormonal contraceptives. The study also found generally stronger in-pair desire among hormonal contraception users, but more pronounced in-pair and extra-pair desire around ovulation in naturally cycling women (Arslan et al., 2018; cf. Huang et al., 2020). These results thus suggest an increase in desire around ovulation rather than a decrease around menstruation. In the present meta-analysis, we coded the percentage of women using hormonal contraceptives, but this information was unfortunately not available for most studies, which precluded a formal moderation analysis. In any case, the fluctuations documented by Arslan and colleagues (2018) were rather small (around 0.2 on a 6-point scale for extra-pair desire and less than 0.2 on a 5-point scale for in-pair desire). Thus, at this point, the available evidence suggests that the gender differences found in the present analysis are unlikely to result from changes in the menstrual cycle. However, we do deem it important to examine the role of the menstrual cycle further for all three sex drive facets.

## Future Directions for Conceptualizing Sex Drive

Our theoretical rationale for defining sex drive had two central pillars. The first pillar was that traits are relatively enduring patterns of thoughts, feelings, and behavior (McCrae & Costa, 2003; Roberts, 2009). The second pillar was the understanding of traits as intraindividual density distributions of psychological states (Fleeson, 2001, 2004; see Figure 1). One pathway to further develop this conceptualization would be to broaden the perspective beyond frequencies and consider further dimensions such as the intensity or even duration of sexual events. Such a perspective would call for a more fine-grained theoretical position on what characterizes a person with a stronger versus weaker sex drive. For example, some researchers may argue that a person with a stronger sex drive should not only experience sexual events more often, but also more intensely and more enduringly. Other researchers may argue that a stronger sex drive will manifest itself in the more frequent experience of sexual events, but when a sexual event is experienced, there is no reason to believe that this should be more intense compared to a person with a lower sex drive. We leave it to future theoretical work to develop coherent and specific positions on these and similar questions regarding intensity and duration.

Such advancements in theory can also improve psychometric practices in sexuality research when measurement is guided by theoretical work. In the present meta-analysis, almost no study provided definitions of sex (or sex drive), so there was little control over what participants had in mind when responding. This can be a validity concern. For example, it could be that most people think of sex primarily in terms of penile-vaginal intercourse. One could argue that due to physical differences this experience is not equivalent for women and men, implying that the genders may have partly different experiences in mind when responding to questions indicative of sex drive.

Another intriguing issue pertains to the relative weight of each sex drive facet for the overall construct. The relative importance of cognition, affect, and behavior may differ between traits (Pytlík Zillig et al., 2002). Historically, traits have been defined primarily in terms of overt behavior (Pervin, 1994), suggesting a particularly strong weight for behavior. More recent definitions of traits have emphasized cognition and affect as additional central facets (Johnson, 1997; McCrae & Costa, 2003; Roberts, 2009).

In the current meta-analysis, we weighed each facet equally to compute the overall gender difference in sex drive. Arguments for other weights could be made: Sexual cognition (e.g., thoughts) is the most frequently used indicator of sex drive

(Conley et al., 2011), suggesting a stronger weight for this facet. Conversely, some people may see an inherent association between sexual affect (e.g., desire) and sex drive, and indeed, sexual desire is often in the center of academic discussions around the concept (Dawson & Chivers, 2014). Nevertheless, behavior is also pivotal, since a person who thinks about or desires to feel sexual pleasure but never engages in solitary or partnered sexual behavior would hardly qualify as someone high in sex drive (although it should be noted that some people freely abstain from sexual behavior, for example for religious reasons). Finally, one could argue that there is a funnel-shaped hierarchy inherent to the conceptualization of sex drive proposed here: A sexual episode may often start with a cognition, perhaps only a fleeting thought. If time and circumstances allow, this thought may develop into a sexual desire. Again, only a subset of desires will eventually lead to actual behavior, because a variety of reasons preclude individuals from putting every sexual desire into practice. One implication of this view could be to regard sexual cognition as the purest indicator of sex drive and consequently assign the largest weight to this facet, followed by affect and behavior.

As a final note, we did not connect the present conceptualization to extant research on sex drive in clinical contexts, and clinical studies were explicitly excluded from the analysis. It stands to reason that clinical phenomena such as hypoactive or hyperactive sexuality (Kafka, 2010; Kaplan, 1977) can be placed at the extremes of the sex drive continuum suggested by our conceptualization, but explicating this link is left to future research.

## Limitations

In this section, we discuss questions that could be raised about the conclusiveness and implications of the finding that men's sex drive is on average stronger than women's. Some common methodological concerns have been already discussed in previous sections (for publication bias, see Discussion section; for effect size dependency, see Methods section).

## Response Bias

The present analysis employed various means to address the possibility of response bias due to gendered stigma regarding sexuality, including a correction procedure based on additional meta-analytic estimates. These considerations suggest that it is unlikely that the documented gender difference in sex drive is solely due to response bias, yet some uncertainty regarding the presence of biased responding remains and should be addressed in future dedicated primary research.

## Limitations of moderator analyses

Despite the large number of participants and studies included in the review, some moderation analyses suffered from methodological limitations. For some analyses, unavailable codings reduced statistical power. Some moderators were subject to range restriction (e.g., most studies stemmed from countries with relatively low gender inequality), which can compromise regression analyses. Finally, the sample-level analyses we employed for some hypotheses (e.g., the association of sex drive and mean sample age) can have lower resolution than participant-level analyses.

## Rate of Responses to Data Requests

Our method relied on data for individual questionnaire items, which was not directly available for most publications eligible for inclusion in the analyses. Missing data was requested from the original authors but could only be obtained for 39% of eligible publications. It is possible that mean gender differences in sex drive in the unavailable data are systematically different from the differences we observed in the available data.

## Specificity of Sexual Cognitions

We found more frequent sexual cognitions in men compared to women. But how specific are these gender differences? One previous study found that men reported not only more sexual cognitions than women, but also more other need-based cognitions referring to sleep and food (T. D. Fisher et al., 2012). Is it possible that the observed gender difference in cognition frequency is general and *not* specific to men's and women's sex drive? We deem this possibility unlikely. First, gender differences in cognition frequency were particularly pronounced for extradyadic cognitions, consistent with the ample evidence for gender differences in sociosexuality (Lippa, 2009). Second, a recent experience sampling study in more than 200 young adults in committed relationships also found gender differences in sexual cognition, but not for other needs, including sleep and food (Weber et al., 2022a).

## Generalizability

Psychological studies often examine people who are not representative of the world population, such as undergraduate students in Western countries (Henrich et al., 2010). This was also reflected in the present work. Most (but not all) studies were conducted in Western countries. University students, young adults, and

White/Caucasian people were overrepresented. Future research on gender differences in sex drive should focus more specifically on older adults, as well as people of non-White ethnicities and people from non-Western cultures. Additionally, restricting the search to articles written in English or German may have introduced cultural or language-based biases. We also note that the present analysis did not specifically address the sex drive of gender-non-binary and transgender people. This should also be addressed in future work.

## Conclusion

The key promise of meta-analyses is theoretical and empirical integration. The present work puts forth a coherent conceptualization of sex drive, grounded in trait theory, that directly translates into clear-cut indicators of the three postulated construct facets. Our meta-analysis documents that men's sex drive is stronger than women's, with a medium-to-large effect size ( $g = 0.69$ ). Men think and fantasize about sex more often, experience sexual affect such as desire more often, and more often engage in solitary sexual behavior (masturbation). Biased responding may have inflated these differences, but is unlikely to fully account for the effect. The conservative, response-bias-corrected effect estimate is still of moderate size ( $g = 0.54$ ). Natural language interpretations highlight that, despite the evidence for stronger sex drive in men on average, individual women exceeding individual men in sex drive is far from unusual.

## References

References marked by an asterisk are included in the meta-analysis.

- \* Aarøe, L., Osmundsen, M., & Petersen, M. B. (2016). Distrust as a disease avoidance strategy: Individual differences in disgust sensitivity regulate generalized social trust. *Frontiers in Psychology, 7*, Article 1038. <https://doi.org/10.3389/fpsyg.2016.01038>
- \* Aavik, T., & Raidam, G. (2012). *Personal values and sexual desire* [Paper presentation]. NACS 2012 conference in Helsinki, Finland.
- Ahlborg, T., Dahlöf, L., & Hallberg, L. R. (2005). Quality of the intimate and sexual relationship in first-time parents six months after delivery. *Journal of Sex Research, 42*(2), 167–174. <https://doi.org/10.1080/00224490509552270>
- Alexander, M. G., & Fisher, T. D. (2003). Truth and consequences: Using the bogus pipeline to examine sex differences in self-reported sexuality. *Journal of Sex Research, 40*(1), 27–35. <https://doi.org/10.1080/00224490309552164>
- \* Al-Shawaf, L., Lewis, D. M. G., & Buss, D. M. (2015). Disgust and mating strategy. *Evolution and Human Behavior, 36*(3), 199–205. <https://doi.org/10.1016/j.evolhumbehav.2014.11.003>
- \* Anslinger, J. (2019). *Measurement and evaluation of sexual objectification* [Doctoral dissertation, Bielefeld University]. PUB - Publications at Bielefeld University. doi:10.4119/unibi/2936501
- Arslan, R. C., Schilling, K. M., Gerlach, T. M., & Penke, L. (2018). Using 26,000 diary entries to show ovulatory changes in sexual desire and behavior. *Journal of Personality and Social Psychology*. Advance online publication. <https://doi.org/10.1037/pspp0000208>
- \* Asendorpf, J. B., Penke, L., & Back, M. D. (2011). From dating to mating and relating: Predictors of initial and long-term outcomes of speed-dating in a community sample. *European Journal of Personality, 25*(1), 16–30. <https://doi.org/10.1002/per.768>
- \* Ashdown, B., Hackathorn, J., & Clark, E. (2011). In and out of the bedroom: Sexual satisfaction in the marital relationship. *Journal of Integrated Social Sciences, 2*(1), 40–57.
- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543–554. <https://doi.org/10.1177/1745691612459060>
- \* Banai, B., & Pavela, I. (2015). Two-dimensional structure of the sociosexual orientation inventory and its personality correlates. *Evolutionary Psychology, 13*(3). <https://doi.org/10.1177/1474704915604541>
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.
- \* Barnett, M. D., Berry, K. E., Maciel, I. V., & Marsden III, A. D. (2017). The primal scene phenomenon: Witnessing parental sexual activity and sociosexual orientation. *Sexuality & Culture, 22*(1), 162–175. <https://doi.org/10.1007/s12119-017-9458-2>
- \* Barrada, J. R., Castro, Á., Correa, A. B., & Ruiz-Gómez, P. (2017). The tridimensional structure of sociosexuality: Spanish validation of the revised sociosexual orientation inventory. *Journal of Sex & Marital Therapy, 44*(2), 149–158. <https://doi.org/10.1080/0092623X.2017.1335665>
- \* Barriger, M., & Vélez-Blasini, C. J. (2013). Descriptive and injunctive social norm overestimation in hooking up and their role as predictors of hook-up activity in a college student sample. *Journal of Sex Research, 50*(1), 84–94. <https://doi.org/10.1080/00224499.2011.607928>
- Baumeister, R. F., Catanese, K. R., & Vohs, K. D. (2001). Is there a gender difference in strength of sex drive? Theoretical Views, conceptual distinctions, and a review of relevant evidence. *Personality and Social Psychology Review, 5*(3), 242–273. [https://doi.org/10.1207/S15327957PSPR0503\\_5](https://doi.org/10.1207/S15327957PSPR0503_5)
- Baumeister, R. F., Reynolds, T., Winegard, B., & Vohs, K. D. (2017). Competing for love: Applying sexual economics theory to mating contests. *Journal of Economic Psychology, 63*, 230–241. <https://doi.org/10.1016/j.joep.2017.07.009>
- Baumeister, R. F., & Vohs, K. D. (2004). Sexual economics: Sex as female resource for social exchange in heterosexual interactions. *Personality and Social Psychology Review, 8*(4), 339–363. [https://doi.org/10.1207/s15327957pspr0804\\_2](https://doi.org/10.1207/s15327957pspr0804_2)
- Beach, F. A. (1956). Characteristics of masculine „sex drive“. In M. R. Jones (Ed.), *Nebraska symposium on motivation* (pp. 1–32). University of Nebraska Press.
- \* Beall, A. T. (2012). *The attractiveness of emotion expression* [Master's thesis, University of British Columbia]. UBC Library Open Collections. <https://open.library.ubc.ca/media/download/pdf/24/1.0073005/1>
- \* Beall, A. T., & Schaller, M. (2014). Affective implications of the mating/parenting trade-off: Short-term mating motives and desirability as a short-term mate predict less intense tenderness responses to infants. *Personality and Individual Differences, 68*, 112–117. <https://doi.org/10.1016/j.paid.2014.03.049>
- \* Beall, A. T., & Schaller, M. (2017). Evolution, motivation, and the mating/parenting trade-off. *Self and Identity, 18*(1), 39–59. <https://doi.org/10.1080/15298868.2017.1356366>
- \* Beaulieu-Pelletier, G., Philippe, F. L., Lecours, S., & Couture, S. (2011). The role of attachment avoidance in extradyadic sex. *Attachment & Human Development, 13*(3), 293–313. <https://doi.org/10.1080/14616734.2011.562419>



- \* Bendixen, M., Asao, K., Wyckoff, J. P., Buss, D. M., & Kennair, L. E. O. (2017). Sexual regret in US and Norway: Effects of culture and individual differences in religiosity and mating strategy. *Personality and Individual Differences, 116*, 246–251. <https://doi.org/10.1016/j.paid.2017.04.054>
- \* Beutel, M. E., Stöbel-Richter, Y., & Brähler, E. (2007). Sexual desire and sexual activity of men and women across their lifespans: Results from a representative German community survey. *BJU International, 101*(1), 76–82. <https://doi.org/10.1111/j.1464-410X.2007.07204.x>
- \* Birnbaum, G. E., Mikulincer, M., & Gillath, O. (2011). In and out of a daydream: Attachment orientations, daily couple interactions, and sexual fantasies. *Personality and Social Psychology Bulletin, 37*(10), 1398–1410. <https://doi.org/10.1177/0146167211410986>
- \* Birnie-Porter, C., & Hunt, M. (2015). Does relationship status matter for sexual satisfaction? The roles of intimacy and attachment avoidance in sexual satisfaction across five types of ongoing sexual relationships. *The Canadian Journal of Human Sexuality, 24*(2), 174–183. <https://doi.org/10.3138/cjhs.242-A5>
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. <http://doi.wiley.com/10.1002/9780470743386>
- Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis:  $I^2$  is not an absolute measure of heterogeneity:  $I^2$  is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*(1), 5–18. <https://doi.org/10.1002/jrsm.1230>
- \* Bourdage, J. S., Lee, K., Ashton, M. C., & Perry, A. (2007). Big Five and HEXACO model personality correlates of sexuality. *Personality and Individual Differences, 43*(6), 1506–1516. <https://doi.org/10.1016/j.paid.2007.04.008>
- Breznysnyak, M., & Whisman, M. A. (2004). Sexual desire and relationship functioning: The effects of marital satisfaction and power. *Journal of Sex & Marital Therapy, 30*(3), 199–217. <https://doi.org/10.1080/00926230490262393>
- \* Brody, S. (2003). Alexithymia is inversely associated with women's frequency of vaginal intercourse. *Archives of Sexual Behavior, 32*(1), 73–77. <https://doi.org/10.1023/A:1021897530286>
- \* Brody, S. (2004). Slimness is associated with greater intercourse and lesser masturbation frequency. *Journal of Sex & Marital Therapy, 30*(4), 251–261. <https://doi.org/10.1080/00926230490422368>
- Brown, N. R., & Sinclair, R. C. (1999). Estimating number of lifetime sexual partners: Men and women do it differently. *Journal of Sex Research, 36*(3), 292–297. <https://doi.org/10.1080/00224499909551999>
- \* Buckels, E. E., Beall, A. T., Hofer, M. K., Lin, E. Y., Zhou, Z., & Schaller, M. (2015). Individual differences in activation of the parental care motivational system: Assessment, prediction, and implications. *Journal of Personality and Social Psychology, 108*(3), 497–514. <https://doi.org/10.1037/pspp0000023>
- Buss, D. M. (1998). Sexual strategies theory: Historical origins and current status. *Journal of Sex Research, 35*(1), 19–31. <https://doi.org/10.1080/00224499809551914>
- Buss, D. M. (2012). *Evolutionary psychology: The new science of the mind* (4th ed.). Allyn & Bacon.
- Buss, D. M., & Schmitt, D. P. (1993). Sexual Strategies Theory: An evolutionary perspective on human mating. *Psychological Review, 100*(2), 204–232. <https://doi.org/10.1037/0033-295X.100.2.204>
- Buss, D. M., & Schmitt, D. P. (2019). Mate preferences and their behavioral manifestations. *Annual Review of Psychology, 70*(1), 77–110. <https://doi.org/10.1146/annurev-psych-010418-103408>
- Bussey, K., & Bandura, A. (1999). Social cognitive theory of gender development and differentiation. *Psychological Review, 106*(4), 676–713. <https://doi.org/10.1037/0033-295X.106.4.676>
- Bussey, K., & Bandura, A. (2004). Social cognitive theory of gender development and functioning. In *The psychology of gender* (2nd ed., pp. 92–119). Guilford Press.
- \* Byers, E. S., Purdon, C., & Clark, D. A. (1998). Sexual intrusive thoughts of college students. *Journal of Sex Research, 35*(4), 359–369. <https://doi.org/10.1080/00224499809551954>
- \* Campbell, K. (2008). *The meaning of "I do": A mixed methods examination of newlyweds' marital expectations* [Doctoral dissertation, University of Georgia]. UGA Libraries. [https://getd.libs.uga.edu/pdfs/campbell\\_kelly\\_200805\\_phd.pdf](https://getd.libs.uga.edu/pdfs/campbell_kelly_200805_phd.pdf)
- \* Carpenter, C. J., & McEwan, B. (2016). The players of micro-dating: Individual and gender differences in goal orientations toward micro-dating apps. *First Monday, 21*(5). <https://doi.org/10.5210/fm.v21i5.6187>
- \* Carvalho, J., Santos, I., Soares, S., & Nobre, P. (2016). [Unpublished raw data on sexual desire in Portuguese men and women]. Lusófona University.
- \* Carvalho, J., Stulhofer, A., Vieira, A. L., & Jurin, T. (2015). Hypersexuality and high sexual desire: Exploring the structure of problematic sexuality. *The Journal of Sexual Medicine, 12*(6), 1356–1367. <https://doi.org/10.1111/jsm.12865>
- \* Cash, T. F. (2004). "Baring the body in the bedroom": Body image, sexual self-schemas, and sexual functioning among college women and men. *Electroni Journal of Human Sexuality, 7*. <http://mail.ejhs.org/volume7/bodyimage.html>
- \* Chadwick, S. B., Burke, S. M., Goldey, K. L., Bell, S. N., & van Anders, S. M. (2017). Sexual desire in sexual minority and majority women and men: The multifaceted sexual desire questionnaire. *Archives of Sexual Behavior, 46*(8), 2465–2484. <https://doi.org/10.1007/s10508-016-0895-z>
- \* Charles, N. E., & Alexander, G. M. (2011). The association between 2D:4D ratios and sociosexuality: A failure to replicate. *Archives of Sexual Behavior, 40*(3), 587–595. <https://doi.org/10.1007/s10508-010-9715-z>
- \* Chi, X., Yu, L., & Winter, S. (2012). Prevalence and correlates of sexual behaviors among university students: A study in Hefei, China. *BMC Public Health, 12*(1), Article 972. <https://doi.org/10.1186/1471-2458-12-972>
- \* Chiorri, C., Garofalo, C., & Velotti, P. (2017). Does the dark triad manifest similarly in men and women? Measurement invariance of the dirty dozen across sex. *Current Psychology, 38*(3), 659–675. <https://doi.org/10.1007/s12144-017-9641-5>
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.
- Conley, T. D., Moors, A. C., Matsick, J. L., Ziegler, A., & Valentine, B. A. (2011). Women, men, and the bedroom: Methodological and conceptual insights that narrow, reframe, and eliminate gender differences in sexuality. *Current Directions in Psychological Science, 20*(5), 296–300. <https://doi.org/10.1177/0963721411418467>
- \* Correa, A. B., Castro, Á., Barrada, J. R., & Ruiz-Gómez, P. (2017). Sociodemographic and psychosexual characteristics of students from a Spanish university who engage in casual sex. *Sexuality Research and Social Policy, 14*(4), 445–453. <https://doi.org/10.1007/s13178-017-0274-0>
- \* Couper, M. P., & Stinson, L. L. (1999). Completion of self-administered questionnaires in a sex survey. *The Journal of Sex Research, 36*(4), 321–330. <https://doi.org/10.1080/00224499909552004>
- Crawford, M., & Popp, D. (2003). Sexual double standards: A review and methodological critique of two decades of research. *The Journal of Sex Research, 40*(1), 13–26. <https://doi.org/10.1080/00224490309552163>
- \* Cross, C. P. (2010). Sex differences in same-sex direct aggression and sociosexuality: The role of risky impulsivity. *Evolutionary Psychology, 8*(4), 779–792. <https://doi.org/10.1177/147470491000800418>

- \* Daugherty, J. R., & Brase, G. (2010a). [Unpublished raw data on sociosexual orientation of men and women]. Kansas State University.
- \* Daugherty, J. R., & Brase, G. L. (2010b). Taking time to be healthy: Predicting health behaviors with delay discounting and time perspective. *Personality and Individual Differences*, *48*(2), 202–207. <https://doi.org/10.1016/j.paid.2009.10.007>
- Dawson, S. J., & Chivers, M. L. (2014). Gender differences and similarities in sexual desire. *Current Sexual Health Reports*, *6*(4), 211–219. <https://doi.org/10.1007/s11930-014-0027-5>
- Day, L. C., Muise, A., Joel, S., & Impett, E. A. (2015). To do it or not to do it? How communally motivated people navigate sexual interdependence dilemmas. *Personality and Social Psychology Bulletin*, *41*(6), 791–804. <https://doi.org/10.1177/0146167215580129>
- de Ridder, D. T. D., Lensvelt-Mulders, G., Finkenauer, C., Stok, F. M., & Baumeister, R. F. (2012). Taking stock of self-control: A meta-analysis of how Trait Self-Control relates to a wide range of behaviors. *Personality and Social Psychology Review*, *16*(1), 76–99. <https://doi.org/10.1177/1088868311418749>
- \* Del Giudice, M., Klimczuk, A. C. E., Traficonte, D. M., & Maestripieri, D. (2014). Autistic-like and schizotypal traits in a life history perspective: Diametrical associations with impulsivity, sensation seeking, and sociosexual behavior. *Evolution and Human Behavior*, *35*(5), 415–424. <https://doi.org/10.1016/j.evolhumbehav.2014.05.007>
- Diamond, L. M. (2004). Emerging perspectives on distinctions between romantic love and sexual desire. *Current Directions in Psychological Science*, *13*(3), 116–119. <https://doi.org/10.1111/j.0963-7214.2004.00287.x>
- \* Diaz-Loving, R., & Rodríguez, G. G. (2008). Sociosexual orientation and sexual behavior in Mexican adults. *Social and Personality Psychology Compass*, *2*(3), 1199–1217. <https://doi.org/10.1111/j.1751-9004.2008.00111.x>
- Dunne, M. P., Martin, N. G., Bailey, J. M., Heath, A. C., Bucholz, K. K., Madden, P. A., & Statham, D. J. (1997). Participation bias in a sexuality survey: Psychological and behavioural characteristics of responders and non-responders. *International Journal of Epidemiology*, *26*(4), 844–854. <https://doi.org/10.1093/ije/26.4.844>
- Eagly, A. H., & Wood, W. (1999). The origins of sex differences in human behavior: Evolved dispositions versus social roles. *American Psychologist*, *54*(6), 408. <https://doi.org/10.1037/0003-066X.54.6.408>
- \* Edelman, R. S., Chopik, W. J., & Kean, E. L. (2011). Sociosexuality moderates the association between testosterone and relationship status in men and women. *Hormones and Behavior*, *60*(3), 248–255. <https://doi.org/10.1016/j.yhbeh.2011.05.007>
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. <https://doi.org/10.1136/bmj.315.7109.629>
- Endendijk, J. J., van Baar, A. L., & Deković, M. (2020). He is a stud, she is a slut! A meta-analysis on the continued existence of sexual double standards. *Personality and Social Psychology Review*, *24*(2), 163–190. <https://doi.org/10.1177/1088868319891310>
- \* Eplov, L., Girdali, A., Davidsen, M., Garde, K., & Kamper-Jørgensen, F. (2007). Sexual desire in a nationally representative danish population. *The Journal of Sexual Medicine*, *4*(1), 47–56. <https://doi.org/10.1111/j.1743-6109.2006.00396.x>
- Fanelli, D. (2010). “Positive” results increase down the hierarchy of the sciences. *PLoS ONE*, *5*(4), Article e10068. <https://doi.org/10.1371/journal.pone.0010068>
- Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. <https://doi.org/10.1007/s11192-011-0494-7>
- Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120–128. <https://doi.org/10.1037/a0024445>
- \* Ferreira, L. C., Narciso, I., Novo, R. F., & Pereira, C. R. (2014). Predicting couple satisfaction: The role of differentiation of self, sexual desire and intimacy in heterosexual individuals. *Sexual and Relationship Therapy*, *29*(4), 390–404. <https://doi.org/10.1080/14681994.2014.957498>
- Fisher, T. D. (Ed.). (2011). *Handbook of sexuality-related measures* (3rd ed.). Routledge.
- Fisher, T. D., Moore, Z. T., & Pittenger, M.-J. (2012). Sex on the brain?: An examination of frequency of sexual cognitions as a function of gender, erotophilia, and social desirability. *Journal of Sex Research*, *49*(1), 69–77. <https://doi.org/10.1080/00224499.2011.565429>
- Fisher, Z., & Tipton, E. (2015). *robumeta: An R-package for robust variance estimation in meta-analysis*. arXiv. <https://arxiv.org/abs/1503.02220>
- Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*(6), 1011–1027. <https://doi.org/10.1037/0022-3514.80.6.1011>
- Fleeson, W. (2004). Moving personality beyond the person-situation debate: The challenge and the opportunity of within-person variability. *Current Directions in Psychological Science*, *13*(2), 83–87. <https://doi.org/10.1111/j.0963-7214.2004.00280.x>
- Fleeson, W., & Jayawickreme, E. (2015). Whole Trait Theory. *Journal of Research in Personality*, *56*, 82–92. <https://doi.org/10.1016/j.jrp.2014.10.009>
- \* Flynn, T.-J., & Gow, A. J. (2015). Examining associations between sexual behaviours and quality of life in older adults. *Age and Ageing*, *44*(5), 823–828. <https://doi.org/10.1093/ageing/afv083>
- \* Forbes, M. K., Baillie, A. J., & Schniering, C. A. (2014). Critical flaws in the Female Sexual Function Index and the International Index of Erectile Function. *The Journal of Sex Research*, *51*(5), 485–491. <https://doi.org/10.1080/00224499.2013.876607>
- \* Foulkes, L., Viding, E., McCrory, E., & Neumann, C. S. (2014). Social Reward Questionnaire (SRQ): Development and validation. *Frontiers in Psychology*, *5*, Article 201. <https://doi.org/10.3389/fpsyg.2014.00201>
- Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. <https://doi.org/10.1126/science.1255484>
- Friese, M., & Frankenbach, J. (2020). P-hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods*, *25*(4), 456–471. <https://doi.org/10.1037/met0000246>
- Friese, M., Frankenbach, J., Job, V., & Loschelder, D. D. (2017). Does self-control training improve self-control? A meta-analysis. *Perspectives on Psychological Science*, *12*(6), 1077–1099. <https://doi.org/10.1177/1745691617697076>
- Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. <https://doi.org/10.1177/2515245919847202>
- \* Gaither, G. A., & Sellbom, M. (2003). The Sexual Sensation Seeking Scale: Reliability and validity within a heterosexual college student sample. *Journal of Personality Assessment*, *81*(2), 157–167. [https://doi.org/10.1207/S15327752JPA8102\\_07](https://doi.org/10.1207/S15327752JPA8102_07)
- \* Gebauer, J. E., Baumeister, R. F., Sedikides, C., & Neberich, W. (2014a). Satisfaction–adaptation principles in sexual desire: Exploring gender differences across the life span. *Social Psychological and Personality Science*, *5*(2), 176–184. <https://doi.org/10.1177/1948550613490970>
- Gebauer, J. E., Baumeister, R. F., Sedikides, C., & Neberich, W. (2014b). Satisfaction–adaptation principles in sexual desire: Exploring gender differences across the life span. *Social Psychological and Personality Science*, *5*(2), 176–184. <https://doi.org/10.1177/1948550613490970>

- \* Gerressu, M., Mercer, C. H., Graham, C. A., Wellings, K., & Johnson, A. M. (2008). Prevalence of masturbation and associated factors in a British national probability survey. *Archives of Sexual Behavior, 37*(2), 266–278. <https://doi.org/10.1007/s10508-006-9123-6>
- \* Gewirtz-Meydan, A. (2017). Why do narcissistic individuals engage in sex? Exploring sexual motives as a mediator for sexual satisfaction and function. *Personality and Individual Differences, 105*, 7–13. <https://doi.org/10.1016/j.paid.2016.09.009>
- Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78. <https://doi.org/10.1016/j.paid.2016.06.069>
- \* Gray, P. B., Garcia, J. R., Crosier, B. S., & Fisher, H. E. (2015). Dating and sexual behavior among single parents of young children in the United States. *The Journal of Sex Research, 52*(2), 121–128. <https://doi.org/10.1080/00224499.2014.941454>
- \* Greengross, G., & Miller, G. (2011). Humor ability reveals intelligence, predicts mating success, and is higher in males. *Intelligence, 39*(4), 188–192. <https://doi.org/10.1016/j.intell.2011.03.006>
- \* Grøntvedt, T. V., Kennair, L. E. O., & Mehmetoglu, M. (2015). Factors predicting the probability of initiating sexual intercourse by context and sex. *Scandinavian Journal of Psychology, 56*(5), 516–526. <https://doi.org/10.1111/sjop.12215>
- Guest, O., & Martin, A. E. (2020). *How computational modeling can force theory building in psychological science*. PsyArXiv. <https://psyarxiv.com/rybh9/>
- \* Haddad, B., & Angman, M. (2016). Dark triad, sociosexual orientation and religious affiliation: An association and moderation study. *Clinical and Experimental Psychology, 2*(2), Article 1000124. <https://doi.org/10.4172/2471-2701.1000124>
- \* Hald, G. M. (2006). Gender differences in pornography consumption among young heterosexual Danish adults. *Archives of Sexual Behavior, 35*(5), 577–585. <https://doi.org/10.1007/s10508-006-9064-0>
- \* Hall, J. A., & Canterberry, M. (2011). Sexism and assertive courtship strategies. *Sex Roles, 65*(11–12), 840–853. <https://doi.org/10.1007/s11199-011-0045-y>
- Haus, K. R., & Thompson, A. E. (2020). An examination of the Sexual Double Standard pertaining to masturbation and the impact of assumed motives. *Sexuality & Culture, 24*(3), 809–834. <https://doi.org/10.1007/s12119-019-09666-8>
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107. <https://doi.org/10.2307/1164588>
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39–65. <https://doi.org/10.1002/jrsm.5>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature, 466*(7302), 29–29. <https://doi.org/10.1038/466029a>
- \* Hofer, J., Busch, H., Bond, M. H., Campos, D., Li, M., & Law, R. (2010). The implicit power motive and sociosexuality in men and women: Pancultural effects of responsibility. *Journal of Personality and Social Psychology, 99*(2), 380–394. <https://doi.org/10.1037/a0020053>
- \* Holmberg, D., & Blair, K. L. (2009). Sexual desire, communication, satisfaction, and preferences of men and women in same-sex versus mixed-sex relationships. *Journal of Sex Research, 46*(1), 57–66. <https://doi.org/10.1080/00224490802645294>
- \* Holtzman, N. S., & Strube, M. J. (2013). Above and beyond short-term mating, long-term mating is uniquely tied to human personality. *Evolutionary Psychology, 11*(5), 1101–1129. <https://doi.org/10.1177/147470491301100514>
- \* Hone, L. S. E. (n.d.-a). [Unpublished raw data on sociosexual orientation of undergraduate psychology students at University of Miami]. University of Miami.
- \* Hone, L. S. E. (n.d.-b). [Unpublished raw data on sociosexual orientation of 803 Amazon Mechanical Turk workers in the US]. University of Miami.
- \* Hone, L. S. E. (n.d.-c). [Unpublished raw data on sociosexual orientation of 737 Amazon Mechanical Turk workers in the US]. University of Miami.
- \* Hone, L. S. E., Carter, E. C., & McCullough, M. E. (2013). Drinking games as a venue for sexual competition. *Evolutionary Psychology, 11*(4), 889–906. <https://doi.org/10.1177/147470491301100413>
- \* Hone, L. S. E., & McCullough, M. (2015). Sexually selected sex differences in competitiveness explain sex differences in changes in drinking game participation. *Evolutionary Psychology, 13*(2), 397–410. <https://doi.org/10.1177/147470491501300206>
- Huang, M., Li, G., Liu, J., Li, Y., & Du, P. (2020). Is there an association between contraception and sexual dysfunction in women? A systematic review and meta-analysis based on Female Sexual Function Index. *The Journal of Sexual Medicine, 17*(10), 1942–1955. <https://doi.org/10.1016/j.jsxm.2020.06.008>
- Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60*(6), 581–592. <https://doi.org/10.1037/0003-066X.60.6.581>
- Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology, 65*(1), 373–398. <https://doi.org/10.1146/annurev-psych-010213-115057>
- \* Impett, E. A., Strachman, A., Finkel, E. J., & Gable, S. L. (2008). Maintaining sexual desire in intimate relationships: The importance of approach goals. *Journal of Personality and Social Psychology, 94*(5), 808–823. <https://doi.org/10.1037/0022-3514.94.5.808>
- Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science, 3*(1), 109–117. JSTOR.
- \* Jauk, E., Neubauer, A. C., Mairunteregger, T., Pemp, S., Sieber, K. P., & Rauthmann, J. F. (2016). How alluring are dark personalities? The dark triad and attractiveness in speed dating. *European Journal of Personality, 30*(2), 125–138. <https://doi.org/10.1002/per.2040>
- \* Jimenez, F. V. (2010). *The regulation of psychological distance in long-distance relationships* [Doctoral dissertation, Humboldt-Universität zu Berlin]. Deutsche Nationalbibliothek. <https://d-nb.info/1015169325/34>
- Joel, S., Eastwick, P. W., Allison, C. J., Arriaga, X. B., Baker, Z. G., Bar-Kalifa, E., Bergeron, S., Birnbaum, G. E., Brock, R. L., Brumbaugh, C. C., Carmichael, C. L., Chen, S., Clarke, J., Cobb, R. J., Coolsen, M. K., Davis, J., de Jong, D. C., Debrot, A., DeHaas, E. C., ... Wolf, S. (2020). Machine learning uncovers the most robust self-report predictors of relationship quality across 43 longitudinal couples studies. *Proceedings of the National Academy of Sciences, 117*(32), 19061–19071. <https://doi.org/10.1073/pnas.1917036117>
- Johnson, J. A. (1997). Units of analysis for the description and explanation of personality. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of Personality Psychology* (pp. 73–93). Academic Press. <https://linkinghub.elsevier.com/retrieve/pii/B9780121346454500044>
- Jonason, P. K. (2008). A mediation hypothesis to account for the sex difference in reported number of sexual partners: An intrasexual competition approach. *International Journal of Sexual Health, 19*(4), 41–49. [https://doi.org/10.1300/J514v19n04\\_05](https://doi.org/10.1300/J514v19n04_05)
- \* Jonason, P. K. (2013). Four functions for four relationships: Consensus definitions of university students. *Archives of Sexual Behavior, 42*(8), 1407–1414. <https://doi.org/10.1007/s10508-013-0189-7>
- \* Jonason, P. K., & Buss, D. M. (2012). Avoiding entangling commitments: Tactics for implementing a short-term mating strategy. *Personality and Individual Differences, 52*(5), 606–610. <https://doi.org/10.1016/j.paid.2011.12.015>
- \* Jonason, P. K., Foster, J. D., McCain, J., & Campbell, W. K. (2015). Where birds flock to get together: The who, what,

- where, and why of mate searching. *Personality and Individual Differences*, 80, 76–84.  
<https://doi.org/10.1016/j.paid.2015.02.018>
- \* Jonason, P. K., Garcia, J. R., Webster, G. D., Li, N. P., & Fisher, H. E. (2015). Relationship dealbreakers: Traits people avoid in potential mates. *Personality and Social Psychology Bulletin*, 41(12), 1697–1711.  
<https://doi.org/10.1177/0146167215609064>
- \* Jonason, P. K., Teicher, E. A., & Schmitt, D. P. (2011). The TIPI's validity confirmed: Associations with sociosexuality and self-esteem. *Individual Differences Research*, 9(1), 52–60.
- \* Jonason, P. K., Webster, G. D., & Gesselman, A. N. (2013). *The structure and content of long-term and short-term mate preferences*. PsychArchives.  
<http://dx.doi.org/10.23668/psycharchives.2173>
- \* Jones, B., & DeBruine, L. (n.d.). [Unpublished raw data on sexual desire collected via faceresearch.org]. University of Glasgow.
- Kafka, M. P. (2010). Hypersexual Disorder: A proposed diagnosis for DSM-V. *Archives of Sexual Behavior*, 39(2), 377–400. <https://doi.org/10.1007/s10508-009-9574-7>
- \* Kandrik, M., Jones, B. C., & DeBruine, L. M. (2015). Scarcity of female mates predicts regional variation in men's and women's sociosexual orientation across US states. *Evolution and Human Behavior*, 36(3), 206–210.  
<https://doi.org/10.1016/j.evolhumbehav.2014.11.004>
- Kaplan, H. S. (1977). Hypoactive sexual desire. *Journal of Sex & Marital Therapy*, 3(1), 3–9.  
<https://doi.org/10.1080/00926237708405343>
- \* Kar, N., & Koola, M. M. (2007). A pilot survey of sexual functioning and preferences in a sample of English-speaking adults from a small south Indian town. *The Journal of Sexual Medicine*, 4(5), 1254–1261.  
<https://doi.org/10.1111/j.1743-6109.2007.00543.x>
- \* Kardum, I., Hude-Knezevic, J., & Gracanin, A. (2006). Sociosexuality and mate retention in romantic couples. *Psychological Topics*, 15(2), 277–296.
- \* Kawamoto, T. (2015). Development of Japanese version of the long-term mating orientation scale (LTMO-J): Translation of LTMO into Japanese. *Japanese Psychological Research*, 57(4), 323–336. <https://doi.org/10.1111/jpr.12092>
- \* Kennair, L. E. O., & Bendixen, M. (2012). Sociosexuality as predictor of sexual harassment and coercion in female and male high school students. *Evolution and Human Behavior*, 33(5), 479–490.  
<https://doi.org/10.1016/j.evolhumbehav.2012.01.001>
- \* Kennair, L. E. O., Bendixen, M., & Buss, D. M. (2016). Sexual regret: Tests of competing explanations of sex differences. *Evolutionary Psychology*, 14(4).  
<https://doi.org/10.1177/1474704916682903>
- Kiliç Onar, D., Armstrong, H., & Graham, C. A. (2020). What Does Research Tell Us About Women's Experiences, Motives and Perceptions of Masturbation Within a Relationship Context?: A Systematic Review of Qualitative Studies. *Journal of Sex & Marital Therapy*, 46(7), 683–716.  
<https://doi.org/10.1080/0092623X.2020.1781722>
- Klusmann, D. (2002). Sexual motivation and the duration of partnership. *Archives of Sexual Behavior*, 31(3), 275–287.  
<https://doi.org/10.1023/a:1015205020769>
- \* Koban, K., Heid, A., & Ohler, P. (2016). *Love the way you lie. Sex-specific influence patterns of sociosexual mating orientation on acceptance of deception in human courtship* [Paper presentation]. 16th Annual Conference of the MVE list (Human Behavior in Evolutionary Perspective), Berlin, Germany.
- \* Koban, K., & Ohler, P. (2016). Ladies, know yourselves! Gentlemen, fool yourselves! Evolved self-promotion traits as predictors for promiscuous sexual behavior in both sexes. *Personality and Individual Differences*, 92, 11–15.  
<https://doi.org/10.1016/j.paid.2015.11.056>
- \* Kovacevic, K. (2017). *Investigating the masturbatory behaviours of Canadian midlife adults* [Master's thesis, University of Guelph]. The Atrium.  
[https://atrium.lib.uoguelph.ca/xmlui/bitstream/handle/10214/11607/Kovacevic\\_Katarina\\_201709\\_Msc.pdf?sequence=1&isAllowed=y](https://atrium.lib.uoguelph.ca/xmlui/bitstream/handle/10214/11607/Kovacevic_Katarina_201709_Msc.pdf?sequence=1&isAllowed=y)
- \* Kuhn, W., Koenig, J., Donoghue, A., Hillecke, T. K., & Warth, M. (2014). Psychometrische Eigenschaften einer deutschsprachigen Kurzversion des Sexual Desire Inventory (SDI-2). *Zeitschrift für Sexualforschung*, 27(2), 138–149.
- \* Kurzban, R., Dukes, A., & Weeden, J. (2010). Sex, drugs and moral goals: Reproductive strategies and views about recreational drugs. *Proceedings of the Royal Society B: Biological Sciences*, 277(1699), 3501–3508.  
<https://doi.org/10.1098/rspb.2010.0608>
- \* La Rocque, C., & Cioe, J. (2011). An evaluation of the relationship between body image and sexual avoidance. *Journal of Sex Research*, 48(4), 397–408.  
<https://doi.org/10.1080/00224499.2010.499522>
- Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, 4(1), 24.  
<https://doi.org/10.1186/s40359-016-0126-3>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159. <https://doi.org/10.2307/2529310>
- \* Långström, N., & Hanson, R. K. (2006). High rates of sexual behavior in the general population: Correlates and predictors. *Archives of Sexual Behavior*, 35(1), 37–52.  
<https://doi.org/10.1007/s10508-006-8993-y>
- \* Lawyer, S. R., & Schoepflin, F. J. (2013). Predicting domain-specific outcomes using delay and probability discounting for sexual versus monetary outcomes. *Behavioural Processes*, 96, 71–78.  
<https://doi.org/10.1016/j.beproc.2013.03.001>
- \* Leavitt, C. E., & Willoughby, B. J. (2015). Associations between attempts at physical intimacy and relational outcomes among cohabiting and married couples. *Journal of Social and Personal Relationships*, 32(2), 241–262.  
<https://doi.org/10.1177/0265407514529067>
- \* Lee, T., & Forbey, J. D. (2010). MMPI-2 correlates of sexual preoccupation as measured by the Sexuality Scale in a college setting. *Sexual Addiction & Compulsivity*, 17(3), 219–235. <https://doi.org/10.1080/10720162.2010.500500>
- \* Lehmann, R., Denissen, J. J. A., Allemand, M., & Penke, L. (2013). Age and gender differences in motivational manifestations of the Big Five from age 16 to 60. *Developmental Psychology*, 49(2), 365–383.  
<https://doi.org/10.1037/a0028277>
- Levine, S. B. (2003). The nature of sexual desire: A clinician's perspective. *Archives of Sexual Behavior*, 32(3), 279–285.  
<https://doi.org/10.1023/A:1023421819465>
- Lippa, R. A. (2006). Is high sex drive associated with increased sexual attraction to both sexes? It depends on whether you are male or female. *Psychological Science*, 17(1), 46–52.  
<https://doi.org/10.1111/j.1467-9280.2005.01663.x>
- Lippa, R. A. (2009). Sex differences in sex drive, sociosexuality, and height across 53 nations: Testing evolutionary and social structural theories. *Archives of Sexual Behavior*, 38(5), 631–651. <https://doi.org/10.1007/s10508-007-9242-8>
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE.
- \* Mark, K. P. (2012). The relative impact of individual sexual desire and couple desire discrepancy on satisfaction in heterosexual couples. *Sexual and Relationship Therapy*, 27(2), 133–146.  
<https://doi.org/10.1080/14681994.2012.678825>
- \* Mark, K. P. (2013a). *Good in bed surveys: Report #1* [Unpublished manuscript].  
[http://www.kristenmark.com/wp-content/uploads/2014/02/GIB\\_Relationship\\_Boredom.pdf](http://www.kristenmark.com/wp-content/uploads/2014/02/GIB_Relationship_Boredom.pdf)
- \* Mark, K. P. (2013b). *Good in bed surveys: Report #3* [Unpublished manuscript].  
[http://www.kristenmark.com/wp-content/uploads/2014/02/GIB\\_Orgasm.pdf](http://www.kristenmark.com/wp-content/uploads/2014/02/GIB_Orgasm.pdf)
- \* Mark, K. P. (2014). The impact of daily sexual desire and daily sexual desire discrepancy on the quality of the sexual experience in couples. *The Canadian Journal of Human*

- Sexuality*, 23(1), 27–33.  
<https://doi.org/10.3138/cjhs.23.1.A2>
- Mark, K. P. (2015). Sexual desire discrepancy. *Current Sexual Health Reports*, 7(3), 198–202.  
<https://doi.org/10.1007/s11930-015-0057-7>
- \* Marzec, M., & Lukasik, A. (2017). Love styles in the context of life history theory. *Polish Psychological Bulletin*, 48(2), 237–249. <https://doi.org/10.1515/ppb-2017-0027>
- Mastrich, Z., & Hernandez, I. (2021). Results everyone can understand: A review of common language effect size indicators to bridge the research-practice gap. *Health Psychology*, 40(10), 727–736.  
<https://doi.org/10.1037/hea0001112>
- \* Mattingly, B. A., Clark, E. M., Weidler, D. J., Bullock, M., Hackathorn, J., & Blankmeyer, K. (2011). Sociosexual orientation, commitment, and infidelity: A mediation analysis. *The Journal of Social Psychology*, 151(3), 222–226. <https://doi.org/10.1080/00224540903536162>
- \* Maxwell, J. A., Muise, A., MacDonald, G., Day, L. C., Rosen, N. O., & Impett, E. A. (2017). How implicit theories of sexuality shape sexual and relationship well-being. *Journal of Personality and Social Psychology*, 112(2), 238–279.  
<https://doi.org/10.1037/pspi0000078>
- McCrae, R. R., & Costa, P. T. (2003). *Personality in adulthood: A five-factor theory perspective*. Guilford Press.
- McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, 111(2), 361–365. <https://doi.org/10.1037/0033-2909.111.2.361>
- \* McIntyre, J. C., Barlow, F. K., & Hayward, L. E. (2015). Stronger sexual desires only predict bold romantic intentions and reported infidelity when self-control is low: Self-control, desire, and sexual behaviour. *Australian Journal of Psychology*, 67(3), 178–186.  
<https://doi.org/10.1111/ajpy.12073>
- McNulty, J. K., Maxwell, J. A., Meltzer, A. L., & Baumeister, R. F. (2019). Sex-differentiated changes in sexual desire predict marital dissatisfaction. *Archives of Sexual Behavior*, 48(8), 2473–2489. <https://doi.org/10.1007/s10508-019-01471-6>
- McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, 11(5), 730–749.  
<https://doi.org/10.1177/1745691616662243>
- \* Meana, M., & Nunnink, S. E. (2006). Gender differences in the content of cognitive distraction during sex. *Journal of Sex Research*, 43(1), 59–67.  
<https://doi.org/10.1080/00224490609552299>
- Mercer, C. H., Tanton, C., Prah, P., Erens, B., Sonnenberg, P., Clifton, S., Macdowall, W., Lewis, R., Field, N., Datta, J., Copas, A. J., Phelps, A., Wellings, K., & Johnson, A. M. (2013). Changes in sexual attitudes and lifestyles in Britain through the life course and over time: Findings from the National Surveys of Sexual Attitudes and Lifestyles (Natsal). *The Lancet*, 382(9907), 1781–1794.  
[https://doi.org/10.1016/S0140-6736\(13\)62035-8](https://doi.org/10.1016/S0140-6736(13)62035-8)
- \* Meskó, N., Láng, A., & Kocsor, F. (2014). *The Hungarian version of Sociosexual Orientation Inventory Revised (SOI-R): Sex and age differences*. PsychArchives.  
<http://dx.doi.org/10.23668/psycharchives.2185>
- Meston, C. M., & Buss, D. M. (2007). Why humans have sex. *Archives of Sexual Behavior*, 36(4), 477–507.  
<https://doi.org/10.1007/s10508-007-9175-2>
- Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, 56(2), 128–165. <https://doi.org/10.1037/0003-066X.56.2.128>
- \* Miller, J. D., Gentile, B., & Campbell, W. K. (2013). A test of the construct validity of the Five-Factor Narcissism Inventory. *Journal of Personality Assessment*, 95(4), 377–387. <https://doi.org/10.1080/00223891.2012.742903>
- Mitchell, K. R., Mercer, C. H., Prah, P., Clifton, S., Tanton, C., Wellings, K., & Copas, A. (2019). Why do men report more opposite-sex sexual partners than women? Analysis of the gender discrepancy in a British national probability survey. *The Journal of Sex Research*, 56(1), 1–8.  
<https://doi.org/10.1080/00224499.2018.1481193>
- \* Mogilski, J. K., Memering, S. L., Welling, L. L. M., & Shackelford, T. K. (2017). Monogamy versus consensual non-monogamy: Alternative approaches to pursuing a strategically pluralistic mating strategy. *Archives of Sexual Behavior*, 46(2), 407–417. <https://doi.org/10.1007/s10508-015-0658-2>
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, 6(7), Article e1000097.  
<https://doi.org/10.1371/journal.pmed.1000097>
- \* Moyano, N., Byers, E. S., & Sierra, J. C. (2016). Content and valence of sexual cognitions and their relationship with sexual functioning in Spanish men and women. *Archives of Sexual Behavior*, 45(8), 2069–2080.  
<https://doi.org/10.1007/s10508-015-0659-1>
- \* Moyano, N., Pérez, S., & Sierra, J. C. (2011). *Relationship between personality traits and sexual daydreaming* [Poster presentation]. 20th World Congress for Sexual Health, Glasgow, United Kingdom.
- \* Muise, A., Impett, E. A., Kogan, A., & Desmarais, S. (2012). Keeping the spark alive: Being motivated to meet a partner's sexual needs sustains sexual desire in long-term romantic relationships. *Social Psychological and Personality Science*, 4(3), 267–273.  
<https://doi.org/10.1177/1948550612457185>
- \* Muise, A., Stanton, S. C. E., Kim, J. J., & Impett, E. A. (2016). Not in the mood? Men under- (not over-) perceive their partner's sexual desire in established intimate relationships. *Journal of Personality and Social Psychology*, 110(5), 725–742. <https://doi.org/10.1037/pspi0000046>
- \* Nagoshi, J. L., Adams, K. A., Terrell, H. K., Hill, E. D., Brzuzy, S., & Nagoshi, C. T. (2008). Gender differences in correlates of homophobia and transphobia. *Sex Roles*, 59(7–8), 521–531. <https://doi.org/10.1007/s1199-008-9458-7>
- \* Nakamine, S. (2017). Does sociosexuality affect use and desirability of different types of opening lines among young Japanese? *Archives of Sexual Behavior*, 46(6), 1777–1783.  
<https://doi.org/10.1007/s10508-017-0940-6>
- Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6), 615–631.  
<https://doi.org/10.1177/1745691612459058>
- \* Oberlander, V. A., Ettinger, U., Banse, R., & Schmidt, A. F. (2016). Development of a cued pro- and antisaccade paradigm: An indirect measure to explore automatic components of sexual interest. *Archives of Sexual Behavior*, 46(8), 2377–2388. <https://doi.org/10.1007/s10508-016-0839-7>
- Ostovich, J. M., & Sabini, J. (2004). How are sociosexuality, sex drive, and lifetime number of sexual partners related? *Personality and Social Psychology Bulletin*, 30(10), 1255–1266. <https://doi.org/10.1177/0146167204264754>
- \* O'Sullivan, L. F., Byers, E. S., Brotto, L. A., Majerovich, J. A., & Fletcher, J. (2016). A longitudinal study of problems in sexual functioning and related sexual distress among middle to late adolescents. *Journal of Adolescent Health*, 59(3), 318–324.  
<https://doi.org/10.1016/j.jadohealth.2016.05.001>
- \* Patch, E. A., & Figueredo, A. J. (2017). Childhood stress, life history, psychopathy, and sociosexuality. *Personality and Individual Differences*, 115, 108–113.  
<https://doi.org/10.1016/j.paid.2016.04.023>
- \* Pedersen, W. (2014). Forbidden fruit? A longitudinal study of christianity, sex, and marriage. *The Journal of Sex Research*, 51(5), 542–550.  
<https://doi.org/10.1080/00224499.2012.753983>
- \* Peixoto, M., & Nobre, P. (2015). *Problematic sexual desire discrepancy impacts sexual satisfaction and dyadic adjustment in men and women with different sexual*

- orientation [Poster presentation]. 22nd Congress of the World Association for Sexual Health, Singapore. <http://f1000research.com/posters/4-1151>
- \* Penke, L., & Asendorpf, J. B. (2008). Beyond global sociosexual orientations: A more differentiated look at sociosexuality and its effects on courtship and romantic relationships. *Journal of Personality and Social Psychology, 95*(5), 1113–1135. <https://doi.org/10.1037/0022-3514.95.5.1113>
- Pervin, L. A. (1994). A critical analysis of current trait theory. *Psychological Inquiry, 5*(2), 103–113. [https://doi.org/10.1207/s15327965pli0502\\_1](https://doi.org/10.1207/s15327965pli0502_1)
- \* Peter, J., & Valkenburg, P. M. (2008). Adolescents' exposure to sexually explicit internet material and sexual preoccupancy: A three-wave panel study. *Media Psychology, 11*(2), 207–234. <https://doi.org/10.1080/15213260801994238>
- \* Peters, J. R., Eisenlohr-Moul, T. A., Pond, R. S., & DeWall, C. N. (2014). The downside of being sexually restricted: The effects of sociosexual orientation on relationships between jealousy, rejection, and anger. *Journal of Research in Personality, 51*, 18–22. <https://doi.org/10.1016/j.jrp.2014.04.002>
- Petersen, J. L., & Hyde, J. S. (2010). A meta-analytic review of research on gender differences in sexuality, 1993–2007. *Psychological Bulletin, 136*(1), 21–38. <https://doi.org/10.1037/a0017504>
- \* Peterson, A., Geher, G., & Kaufman, S. B. (2011). Predicting preferences for sex acts: Which traits matter most, and why? *Evolutionary Psychology, 9*(3), 371–389. <https://doi.org/10.1177/147470491100900308>
- \* Peterson, A. N. (2011). *Variability in mating strategies: Do individual differences in dispositional traits predict sexual preferences?* [Master's thesis, State University of New York at New Paltz]. Suny Digital Repository. <http://hdl.handle.net/1951/52592>
- \* Pinkerton, S., Cecil, H., Bogart, L., & Abramson, P. (2003). The pleasures of sex: An empirical investigation. *Cognition and Emotion, 17*(2), 341–353. <https://doi.org/10.1080/026999303022291>
- Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics, 47*(7), 702–709. <https://doi.org/10.1038/ng.3285>
- Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat. (2019). *2019 revision of World population prospects* [Data set]. [https://population.un.org/wpp/Download/Files/1\\_Indicators%20\(Standard\)/EXCEL\\_FILES/1\\_Population/WPP2019\\_POP\\_F10\\_1\\_SEX\\_RATIO\\_BY\\_BROAD\\_AGE\\_GROUP.xlsx](https://population.un.org/wpp/Download/Files/1_Indicators%20(Standard)/EXCEL_FILES/1_Population/WPP2019_POP_F10_1_SEX_RATIO_BY_BROAD_AGE_GROUP.xlsx)
- Pytlík Zillig, L. M., Hemenover, S. H., & Dienstbier, R. A. (2002). What do we assess when we assess a Big 5 trait? A content analysis of the affective, behavioral, and cognitive processes represented in Big 5 personality inventories. *Personality and Social Psychology Bulletin, 28*(6), 847–858. <https://doi.org/10.1177/0146167202289013>
- R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.0.1) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- \* Rammsayer, T. H., Bortner, N., & Troche, S. J. (2017). The effects of sex and gender-role characteristics on facets of sociosexuality in heterosexual young adults. *The Journal of Sex Research, 54*(2), 254–263. <https://doi.org/10.1080/00224499.2016.1236903>
- \* Rammsayer, T. H., & Troche, S. J. (2013). The relationship between sociosexuality and aspects of body image in men and women: A structural equation modeling approach. *Archives of Sexual Behavior, 42*(7), 1173–1179. <https://doi.org/10.1007/s10508-013-0114-0>
- \* Randler, C., Jankowski, K. S., Rahafar, A., & Díaz-Morales, J. F. (2016). Sociosexuality, morningness–eveningness, and sleep duration. *SAGE Open, 6*(1). <https://doi.org/10.1177/2158244015621958>
- \* Regan, P. (2000). The role of sexual desire and sexual activity in dating relationships. *Social Behavior and Personality: An International Journal, 28*(1), 51–59. <https://doi.org/10.2224/sbp.2000.28.1.51>
- \* Regan, P., & Anguiano, C. (2014). *Desire and romanticism among college students: Associations with age, gender, and ethnicity* [Paper presentation]. National Social Science Association conference, San Francisco, CA.
- \* Regan, P., & Atkins, L. (2006). Sex differences and similarities in frequency and intensity of sexual desire. *Social Behavior and Personality: An International Journal, 34*(1), 95–102. <https://doi.org/10.2224/sbp.2006.34.1.95>
- Reiser, B., & Faraggi, D. (1999). Confidence intervals for the overlapping coefficient: The normal equal variance case. *Journal of the Royal Statistical Society: Series D (The Statistician), 48*(3), 413–418. <https://doi.org/10.1111/1467-9884.00199>
- Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of Social Psychology quantitatively described. *Review of General Psychology, 7*(4), 331–363. <https://doi.org/10.1037/1089-2680.7.4.331>
- \* Richters, J., de Visser, R. O., Badcock, P. B., Smith, A. M. A., Rissel, C., Simpson, J. M., & Grulich, A. E. (2014). Masturbation, paying for sex, and other sexual activities: The Second Australian Study of Health and Relationships. *Sexual Health, 11*(5), 461–471. <https://doi.org/10.1071/SH14116>
- \* Richters, J., Grulich, A. E., de Visser, R. O., Smith, A. M. A., & Rissel, C. E. (2003). Sex in Australia: Autoerotic, esoteric and other sexual practices engaged in by a representative sample of adults. *Australian and New Zealand Journal of Public Health, 27*(2), 180–190. <https://doi.org/10.1111/j.1467-842X.2003.tb00806.x>
- Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *BMJ, 340*, Article c221. <https://doi.org/10.1136/bmj.c221>
- \* Robbins, C. L. (2011). Prevalence, frequency, and associations of masturbation with partnered sexual behaviors among US adolescents. *Archives of Pediatrics & Adolescent Medicine, 165*(12), 1087–1093. <https://doi.org/10.1001/archpediatrics.2011.142>
- Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. *Journal of Research in Personality, 43*(2), 137–145. <https://doi.org/10.1016/j.jrp.2008.12.015>
- Rodgers, M. A., & Pustejovsky, J. E. (2020). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000300>
- \* Rodrigues, D., & Lopes, D. (2017). Sociosexuality, commitment, and sexual desire for an attractive person. *Archives of Sexual Behavior, 46*(3), 775–788. <https://doi.org/10.1007/s10508-016-0814-3>
- \* Rodrigues, D., Lopes, D., & Pereira, M. (2016). “We agree and now everything goes my way”: Consensual sexual nonmonogamy, extradyadic sex, and relationship satisfaction. *Cyberpsychology, Behavior, and Social Networking, 19*(6), 373–379. <https://doi.org/10.1089/cyber.2016.0114>
- \* Rodrigues, D., Lopes, D., & Pereira, M. (2017). Sociosexuality, commitment, sexual infidelity, and perceptions of infidelity: Data from the second love web site. *The Journal of Sex Research, 54*(2), 241–253. <https://doi.org/10.1080/00224499.2016.1145182>
- \* Rodrigues, D., Lopes, D., & Smith, C. V. (2017). Caught in a “bad romance”? Reconsidering the negative association between sociosexuality and relationship functioning. *The Journal of Sex Research, 54*(9), 1118–1127. <https://doi.org/10.1080/00224499.2016.1252308>
- \* Rodriguez, C. G., & Ditto, P. (2020). *Respecting life or respecting women? Testing ideologically competing*

- explanations of abortion attitudes. PsyArXiv. <https://osf.io/s8kz7>
- \* Rowatt, W. C., & Schmitt, D. P. (2003). Associations between religious orientation and varieties of sexual experience. *Journal for the Scientific Study of Religion*, 42(3), 455–465.
- Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, 13(1), 19–30. <https://doi.org/10.1037/1082-989X.13.1.19>
- \* Sacco, D. F., Hugenberg, K., & Sefcek, J. A. (2009). Sociosexuality and face perception: Unrestricted sexual orientation facilitates sensitivity to female facial cues. *Personality and Individual Differences*, 47(7), 777–782. <https://doi.org/10.1016/j.paid.2009.06.021>
- \* Santos-Iglesias, P., Sierra, J. C., & Vallejo-Medina, P. (2013). Predictors of sexual assertiveness: The role of sexual desire, arousal, attitudes, and partner abuse. *Archives of Sexual Behavior*, 42(6), 1043–1052. <https://doi.org/10.1007/s10508-012-9998-3>
- \* Santtila, P., Wager, I., Witting, K., Harlaar, N., Jern, P., Johansson, A., Varjonen, M., & Sandnabba, N. K. (2007). Discrepancies between sexual desire and sexual activity: Gender differences and associations with relationship satisfaction. *Journal of Sex & Marital Therapy*, 34(1), 31–44. <https://doi.org/10.1080/00926230701620548>
- Schmitt, D. P. (2005). Sociosexuality from Argentina to Zimbabwe: A 48-nation study of sex, culture, and strategies of human mating. *Behavioral and Brain Sciences*, 28(2), 247–311. <https://doi.org/10.1017/s0140525x05000051>
- \* Schultheiss, O. C., Dargel, A., & Rohde, W. (2003). Implicit motives and sexual motivation and behavior. *Journal of Research in Personality*, 37(3), 224–230. [https://doi.org/10.1016/S0092-6566\(02\)00568-8](https://doi.org/10.1016/S0092-6566(02)00568-8)
- \* Schwarz, S. (2008). [Unpublished raw data on multidimensional sociosexuality in the US and Germany]. University of Wuppertal.
- \* Seehuus, M., & Rellini, A. H. (2013). Gender differences in the relationship between sexual satisfaction and propensity for risky sexual behavior. *Sexual and Relationship Therapy*, 28(3), 230–245. <https://doi.org/10.1080/14681994.2013.791748>
- \* Séguin, L. J. (2013). *Examining the relationships between men's and women's motives for pretending orgasm and levels of sexual desire, and relationship and sexual satisfaction* [Master's thesis, University of Guelph]. The Atrium. [https://atrium.lib.uoguelph.ca/xmlui/bitstream/handle/10214/7299/S%c3%a9guin\\_L%c3%a9a\\_201308\\_MSc.pdf?sequence=3&isAllowed=y](https://atrium.lib.uoguelph.ca/xmlui/bitstream/handle/10214/7299/S%c3%a9guin_L%c3%a9a_201308_MSc.pdf?sequence=3&isAllowed=y)
- \* Sevi, B., Aral, T., & Eskenazi, T. (2018). Exploring the hook-up app: Low sexual disgust and high sociosexuality predict motivation to use Tinder for casual sex. *Personality and Individual Differences*, 133, 17–20. <https://doi.org/10.1016/j.paid.2017.04.053>
- \* Shaughnessy, K., Byers, E. S., & Walsh, L. (2011). Online sexual activity experience of heterosexual students: Gender similarities and differences. *Archives of Sexual Behavior*, 40(2), 419–427. <https://doi.org/10.1007/s10508-010-9629-9>
- \* Shook, N. J., Terrizzi, J. A., Clay, R., & Oosterhoff, B. (2015). In defense of pathogen disgust and disease avoidance: A response to Tybur et al. (2015). *Evolution and Human Behavior*, 36(6), 498–502. <https://doi.org/10.1016/j.evolhumbehav.2015.06.003>
- Simpson, J. A., & Gangestad, S. W. (1991). Individual differences in sociosexuality: Evidence for convergent and discriminant validity. *Journal of Personality and Social Psychology*, 60(6), 870–883. <https://doi.org/10.1037/0022-3514.60.6.870>
- Singer, B., & Toates, F. M. (1987). Sexual motivation. *Journal of Sex Research*, 23(4), 481–501. <https://doi.org/10.1080/00224498709551386>
- \* Snowden, R. J., Curl, C., Jobbins, K., Lavington, C., & Gray, N. S. (2016). Automatic direction of spatial attention to male versus female stimuli: A comparison of heterosexual men and women. *Archives of Sexual Behavior*, 45(4), 843–853. <https://doi.org/10.1007/s10508-015-0678-y>
- \* Soyer, A. (2006). *An exploration of masculinity, femininity, sexual fantasy and masturbation as predictors of marital satisfaction* [Master's thesis, Middle East Technical University]. <http://etd.lib.metu.edu.tr/upload/12607652/index.pdf>
- Spector, I. P., Carey, M. P., & Steinberg, L. (1996). The Sexual Desire Inventory: Development, factor structure, and evidence of reliability. *Journal of Sex & Marital Therapy*, 22(3), 175–190. <https://doi.org/10.1080/00926239608414655>
- \* Stankovic, M., Miljkovic, S., Grbesa, G., & Visnjic, A. (2009). General characteristics of adolescent sexual behavior: National survey. *Srpski Arhiv Za Celokupno Lekarstvo*, 137(7–8), 409–415. <https://doi.org/10.2298/SARH0908409S>
- \* Stark, R., Kagerer, S., Walter, B., Vaitl, D., Klucken, T., & Wehrum-Osinski, S. (2015). Trait sexual motivation questionnaire: Concept and validation. *Journal of Sexual Medicine*, 12(4), 1080–1091. <https://doi.org/10.1111/jsm.12843>
- Sterling, T. 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. <https://doi.org/10.2307/2282137>
- Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, 49(1), 108–112. <https://doi.org/10.2307/2684823>
- Sterne, J. A., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99–110). Wiley.
- \* Stewart-Williams, S., Butler, C. A., & Thomas, A. G. (2017). Sexual history and present attractiveness: People want a mate with a bit of a past, but not too much. *The Journal of Sex Research*, 54(9), 1097–1105. <https://doi.org/10.1080/00224499.2016.1232690>
- Strassberg, D. S., & Lowe, K. (1995). Volunteer bias in sexuality research. *Archives of Sexual Behavior*, 24(4), 369–382. <https://doi.org/10.1007/BF01541853>
- \* Strouts, P. H., Brase, G. L., & Dillon, H. M. (2017). Personality and evolutionary strategies: The relationships between HEXACO traits, mate value, life history strategy, and sociosexuality. *Personality and Individual Differences*, 115, 128–132. <https://doi.org/10.1016/j.paid.2016.03.047>
- \* Sutherland, S. E., Rehman, U. S., Fallis, E. E., & Goodnight, J. A. (2015). Understanding the phenomenon of sexual desire discrepancy in couples. *The Canadian Journal of Human Sexuality*, 24(2), 141–150. <https://doi.org/10.3138/cjhs.242.A3>
- \* Thomas, A. G. (n.d.). [Unpublished raw data on sociosexuality]. Swansea University.
- \* Tidwell, N. D., & Eastwick, P. W. (2013). Sex differences in succumbing to sexual temptations: A function of impulse or control? *Personality and Social Psychology Bulletin*, 39(12), 1620–1633. <https://doi.org/10.1177/0146167213499614>
- Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, 20(3), 375–393. <https://doi.org/10.1037/met0000011>
- Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, 40(6), 604–634. <https://doi.org/10.3102/1076998615606099>
- \* Træen, B., Stigum, H., & Sørensen, D. (2002). Sexual diversity in urban Norwegians. *The Journal of Sex Research*, 39(4), 249–258. <https://doi.org/10.1080/00224490209552148>
- \* Trivedi, N., & Sabini, J. (1998). Volunteer bias, sexuality, and personality. *Archives of Sexual Behavior*, 27(2), 181–195.

- \* Trudel, G., Dargis, L., Villeneuve, L., Cadieux, J., Boyer, R., & Prévaille, M. (2014). Marital, sexual and psychological functioning of older couples living at home: The results of a national survey using longitudinal methodology (Part II). *Sexologies*, 23(2), e35–e48. <https://doi.org/10.1016/j.sexol.2013.03.007>
- United Nations Development Programme. (2019). *Beyond income, beyond averages, beyond today: Inequalities in human development in the 21st century* (Human Development Reports). <http://hdr.undp.org/sites/default/files/hdr2019.pdf>
- \* Vaillancourt-Morel, M.-P., Blais-Lecours, S., Labadie, C., Bergeron, S., Sabourin, S., & Godbout, N. (2017a). [Unpublished raw data on sexual cognition]. Université Laval.
- \* Vaillancourt-Morel, M.-P., Blais-Lecours, S., Labadie, C., Bergeron, S., Sabourin, S., & Godbout, N. (2017b). Profiles of cyberpornography use and sexual well-being in adults. *The Journal of Sexual Medicine*, 14(1), 78–85. <https://doi.org/10.1016/j.jsxm.2016.10.016>
- \* Vallejo-Medina, P., Marchal-Bertrand, L., Gómez-Lugo, M., Espada, J. P., Sierra, J. C., Soler, F., & Morales, A. (2016). Adaptation and validation of the Brief Sexual Opinion Survey (SOS) in a Colombian sample and factorial equivalence with the Spanish version. *PLOS ONE*, 11(9), Article e0162531. <https://doi.org/10.1371/journal.pone.0162531>
- \* van Anders, S. M., & Dunn, E. J. (2009). Are gonadal steroids linked with orgasm perceptions and sexual assertiveness in women and men? *Hormones and Behavior*, 56(2), 206–213. <https://doi.org/10.1016/j.yhbeh.2009.04.007>
- \* van Anders, S. M., & Goldey, K. L. (2010). Testosterone and partnering are linked via relationship status for women and ‘relationship orientation’ for men. *Hormones and Behavior*, 58(5), 820–826. <https://doi.org/10.1016/j.yhbeh.2010.08.005>
- \* van Anders, S. M., Hamilton, L. D., & Watson, N. V. (2007). Multiple partners are associated with higher testosterone in North American men and women. *Hormones and Behavior*, 51(3), 454–459. <https://doi.org/10.1016/j.yhbeh.2007.01.002>
- \* van Anders, S. M., Hipp, L. E., & Kane Low, L. (2013). Exploring co-parent experiences of sexuality in the first 3 months after birth. *The Journal of Sexual Medicine*, 10(8), 1988–1999. <https://doi.org/10.1111/jsm.12194>
- Van Erp, S., Verhagen, J., Grasman, R. P. P., & Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in *Psychological Bulletin* from 1990–2013. *Journal of Open Psychology Data*, 5(1), Article 4. <https://doi.org/10.5334/jopd.33>
- \* Van Slyke, J. A., & Wasemiller, A. (2017). Short-term mating strategies are negatively correlated with religious commitment: Exploring evolutionary variables for religiosity at a small christian liberal arts college. *Evolutionary Psychological Science*, 3(3), 253–260. <https://doi.org/10.1007/s40806-017-0093-9>
- \* Velten, J., & Margraf, J. (2017). Satisfaction guaranteed? How individual, partner, and relationship factors impact sexual satisfaction within partnerships. *PLOS ONE*, 12(2), Article e0172855. <https://doi.org/10.1371/journal.pone.0172855>
- Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, 60(3), 419–435. <https://doi.org/10.1007/BF02294384>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48. <https://doi.org/10.18637/jss.v036.i03>
- \* von Borell, C. J., Kordsmeyer, T. L., Gerlach, T. M., & Penke, L. (2019). An integrative study of facultative personality calibration. *Evolution and Human Behavior*, 40(2), 235–248. <https://doi.org/10.1016/j.evolhumbehav.2019.01.002>
- \* Wagstaff, D. (n.d.). [Unpublished raw data on sociosexuality]. Federation University Australia.
- \* Waite, L. J., Iveniuk, J., Laumann, E. O., & McClintock, M. K. (2017). Sexuality in older couples: Individual and dyadic characteristics. *Archives of Sexual Behavior*, 46(2), 605–618. <https://doi.org/10.1007/s10508-015-0651-9>
- Walter, K. V., Conroy-Beam, D., Buss, D. M., Asao, K., Sorokowska, A., Sorokowski, P., Aavik, T., Akello, G., Alhabahba, M. M., Alm, C., Amjad, N., Anjum, A., Atama, C. S., Atamtürk Duyar, D., Ayebare, R., Batres, C., Bendixen, M., Bensafia, A., Bizumic, B., ... Zupančić, M. (2020). Sex differences in mate preferences across 45 countries: A large-scale replication. *Psychological Science*, 31(4), 408–423. <https://doi.org/10.1177/0956797620904154>
- \* Walton, M. T., Lykins, A. D., & Bhullar, N. (2016). Sexual arousal and sexual activity frequency: Implications for understanding hypersexuality. *Archives of Sexual Behavior*, 45(4), 777–782. <https://doi.org/10.1007/s10508-016-0727-1>
- \* Waterink, W. (2014). In steady heterosexual relationships men masturbate more than women because of gender differences in sex drive. *New Voices in Psychology*, 10(1), 96–108. <https://doi.org/10.25159/1812-6371/3419>
- \* Weber, M. (2021). [Unpublished raw data on sex drive]. Department of Psychology, Saarland University.
- Weber, M., Frankenbach, J., Hofmann, W., & Friese, M. (2022a). *Sex drive in everyday life: Characteristics, antecedents, and consequences* [Unpublished manuscript]. Department of Psychology, Saarland University.
- \* Weber, M., Frankenbach, J., Hofmann, W., & Friese, M. (2022b). [Unpublished raw experience-sampling data on sex drive in everyday life]. Department of Psychology, Saarland University.
- \* Webster, G. D., & Bryan, A. (2007). Sociosexual attitudes and behaviors: Why two factors are better than one. *Journal of Research in Personality*, 41(4), 917–922. <https://doi.org/10.1016/j.jrp.2006.08.007>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), Article 1686. <https://doi.org/10.21105/joss.01686>
- Wiederman, M. W. (1997). The truth must be in here somewhere: Examining the gender discrepancy in self-reported lifetime number of sex partners. *Journal of Sex Research*, 34(4), 375–386. <https://doi.org/10.1080/00224499709551905>
- \* Winters, J., Christoff, K., & Gorzalka, B. B. (2010). Dysregulated sexuality and high sexual desire: Distinct constructs? *Archives of Sexual Behavior*, 39(5), 1029–1043. <https://doi.org/10.1007/s10508-009-9591-6>
- \* Wlodarski, R. (2015). The relationship between cognitive and affective empathy and human mating strategies. *Evolutionary Psychological Science*, 1(4), 232–240. <https://doi.org/10.1007/s40806-015-0027-3>
- \* Wlodarski, R., Manning, J., & Dunbar, R. I. M. (2015). Stay or stray? Evidence for alternative mating strategy phenotypes in both men and women. *Biology Letters*, 11(2), Article 20140977. <https://doi.org/10.1098/rsbl.2014.0977>
- Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. In J. M. Olson & M. P. Zanna (Eds.), *Advances in Experimental Social Psychology* (Vol. 46, pp. 55–123). Academic Press. <https://doi.org/10.1016/B978-0-12-394281-4.00002-7>
- Zhang, L., Lee, A. J., DeBruine, L. M., & Jones, B. C. (2019). Are sex differences in preferences for physical attractiveness and good earning capacity in potential mates smaller in countries with greater gender equality? *Evolutionary Psychology*, 17(2), Article 1474704919852921. <https://doi.org/10.1177/1474704919852921>
- \* Zheng, L., & Zheng, Y. (2014). Online sexual activity in Mainland China: Relationship to sexual sensation seeking and sociosexuality. *Computers in Human Behavior*, 36, 323–329. <https://doi.org/10.1016/j.chb.2014.03.062>



**Table 1***Overview of Included Items*

Item	Abbr.	Role	Example
Frequency of sexual cognitions	CF	Sex drive manifestation	During the last month, how often have you had sexual thoughts?
Frequency of sexual affect	AF	Sex drive manifestation	How frequently do you feel sexual desire?
Frequency of sexual behavior	BF	Sex drive manifestation	How many times did you masturbate during the last week?
Intensity of sexual affect	AI	Indicator of latent sex drive	My desire for sex with my partner is strong.
Self-rated sex drive	SRSD	Indicator of latent sex drive	I have a strong sex drive.
Sexual intercourse frequency	SIF	Indicator of potentially biased responding	On average, how many times per month do you and your partner have sex?
Total one night stands	ONS	Indicator of potentially biased responding	With how many partners have you had intercourse on one and only one occasion?
Total sex partners	TSP	Indicator of potentially biased responding	With how many partners have you had intercourse in your lifetime?
Total sex partners in last year	TSPY	Indicator of potentially biased responding	How many people have you had sex with in the last year?



Table 2

Main Results

Role	Indicator	Summary Effect									k	m	Test of Moderation			$\tau$	$\rho^2$	
		g	SE	t	df	p	CI <sub>95</sub>	U <sub>3</sub>	OVL	CL			AHZ	df	p			
Sex Drive Manifestations (Global Summary Effect)																	0.09	87.59
	Sex Drive Manifestations	0.69 (0.55)	0.06	12.10		< .001	[0.58, 0.81]	0.76 (0.71)	0.73 (0.78)	0.69 (0.65)	195	439						
Bias Indicators (Global Summary Effect)																	0.05	77.64
	Bias Indicators	0.15	0.04	4.07		< .001	[0.08, 0.22]	0.56	0.94	0.54	123	244						
Sex Drive Manifestations														7.26	68.53	.001	0.21	91.03
	Affect Frequency	0.58 (0.43)	0.04	13.24	42.09	< .001	[0.49, 0.66]	0.72 (0.67)	0.77 (0.83)	0.66 (0.62)	57	94						
	Behavior Frequency	0.75 (0.60)	0.04	17.99	30.17	< .001	[0.66, 0.84]	0.77 (0.73)	0.71 (0.76)	0.70 (0.67)	44	63						
	Cognition Frequency	0.76 (0.61)	0.02	35.69	138.72	< .001	[0.71, 0.80]	0.78 (0.73)	0.71 (0.76)	0.70 (0.67)	161	282						
Indicators of Latent Sex Drive														5.75	4.00	.074	0.15	90.45
	Affect Intensity	0.40 (0.25)	0.03	15.59	42.34	< .001	[0.35, 0.45]	0.66 (0.60)	0.84 (0.90)	0.61 (0.57)	50	166						
	Self Rated Sex Drive	0.63 (0.49)	0.09	6.74	3.28	.005	[0.35, 0.92]	0.74 (0.69)	0.75 (0.81)	0.67 (0.63)	7	7						
Bias Indicators														7.49	27.75	< .001	0.16	80.41
	Intercourse Frequency	0.04	0.06	0.63	14.05	.541	[-0.09, 0.17]	0.52	0.98	0.51	18	20						
	Sex Partners in Last Year	0.15	0.02	8.09	96.82	< .001	[0.11, 0.19]	0.56	0.94	0.54	106	106						
	Total One Night Stand	0.21	0.02	11.93	94.98	< .001	[0.18, 0.25]	0.58	0.92	0.56	106	106						
	Total Sex Partners	0.19	0.07	2.61	7.28	.034	[0.02, 0.36]	0.57	0.93	0.55	12	12						

Note. Global and group-wise summary results for gender differences in sex drive manifestations, indicators of latent sex drive, and bias indicators. *g* = Hedges' *g* effect size (positive favors males). *SE* = standard error associated with the *g*-value in the same row. *t* = *t*-value associated with the *g*-value in the same row. *df* = degrees-of-freedom associated with the *g*-value in the same row. *p* = *p*-value associated with the *g*-value in the same row. *CI95* = 95% confidence interval. *U*<sub>3</sub> = Cohen's *U*<sub>3</sub> effect size of non-overlap. *OVL* = overlap effect size. *CL* = Common-language effect size, or probability of superiority. *k* = number of studies per subgroup/total. *m* = number of effect sizes per subgroup/total. *AHZ* = Hotelling- *T*-approximated test statistic. *df* = small sample corrected degrees of freedom. *p* = *p*-value associated with the test statistic and *df* in the same row.  $\rho^2$  = proportion of the variation in observed effects that is due to variation in true effects.  $\tau$  = estimated standard deviation of the true effects. Values in parentheses have been bias-corrected. For the correction, the global summary effect of the bias indicators has been subtracted from the respective summary effect.

Table 3

Tests for Moderation (Sex Drive Manifestations)

Moderator	Cognition Frequency					Affect Frequency					Behavior Frequency				
	AHZ	df	p	F <sup>2</sup>	τ	AHZ	df	p	F <sup>2</sup>	τ	AHZ	df	p	F <sup>2</sup>	τ
<b>Outcome-level Moderators</b>															
Aggregation Span	8.46	5.57	.029	79.18	0.12	4.50	9.81	.060	93.43	0.25	1.05	14.67	.323	92.63	0.18
Item Content						4.24	39.15	.046	90.47	0.21					
Item Context											0.18	5.99	.685	94.88	0.20
Type of Response Scale	2.43	7.03	.163	86.40	0.19	1.70	4.95	.250	92.20	0.22	1.46	31.73	.235	94.90	0.19
Scale Range	1.39	50.95	.243	86.57	0.19	1.71	22.19	.204	92.20	0.23	0.51	6.62	.500	95.07	0.17
Item Wording	0.84	5.52	.524	86.40	0.20	1.67	49.28	.202	92.23	0.23	2.09	1.38	N/A	94.98	0.19
<b>Publication-level Moderators</b>															
Aim to Find Gender Differences in Sex Drive	1.97	14.82	.181	86.63	0.20	0.17	19.72	.684	93.08	0.24	0.34	6.76	.580	95.57	0.19
Focus on Anonymity	0.11	120.36	.739	86.44	0.20	0.31	45.75	.583	92.33	0.24	0.03	31.41	.858	94.68	0.19
Focus on Gender Differences in Sex Drive	0.48	27.72	.496	86.03	0.19	0.40	29.13	.531	93.06	0.24	0.19	33.31	.668	95.55	0.20
Focus on Gender Differences	0.94	101.22	.336	85.74	0.19	1.35	19.65	.259	92.73	0.23	2.21	10.40	.167	95.51	0.19
Gender of First Author	1.81	77.35	.182	86.24	0.20	4.36	45.05	.043	91.98	0.23	0.04	32.50	.846	95.03	0.20
Mean Author Gender	0.04	55.58	.848	86.50	0.20	9.22	23.18	.006	91.36	0.21	0.04	18.08	.845	95.02	0.20
Publication Status	2.19	37.81	.147	85.77	0.19	2.05	21.03	.167	92.26	0.23	1.08	13.78	.316	94.46	0.18
Sexuality Journal	1.89	68.08	.174	86.49	0.20	0.71	33.56	.405	92.29	0.23	0.04	29.86	.851	95.01	0.20
<b>Sample-level Moderators</b>															
Mean Age	0.19	31.05	.664	85.07	0.18	0.37	5.80	.566	91.77	0.23	0.43	6.84	.533	93.32	0.21
Percent White	0.10	11.17	.759	88.18	0.21	0.29	4.26	.616	91.78	0.28	0.49	4.89	.515	88.91	0.20
Country-Level Gender Development	2.74	40.36	.105	85.61	0.19	0.25	17.01	.623	92.06	0.24	0.82	19.32	.377	94.74	0.19
Country-Level Gender Inequality	0.64	19.31	.435	85.07	0.19	0.01	35.59	.925	92.02	0.24	1.92	7.46	.206	95.00	0.19
Percent Heterosexual	0.13	10.89	.728	88.61	0.19	1.43	12.60	.254	93.97	0.26	0.42	3.45	N/A	95.34	0.26
Average Partnership Duration in Weeks	0.36	7.48	.566	74.72	0.18	4.16	2.03	N/A	83.31	0.26	2.33	1.74	N/A	60.95	0.16
Percent Parents	0.76	6.18	.417	83.57	0.24	7.58	2.59	N/A	89.12	0.31	2.27	2.38	N/A	96.29	0.33
Country-Level Sex Ratio	1.58	21.73	.221	85.79	0.20	0.51	12.37	.489	92.05	0.24	0.95	7.13	.362	95.19	0.20
Study Restricted to Sexually Active	4.99	20.04	.037	80.26	0.16	0.15	8.60	.707	93.17	0.25					
Percent Single	7.21	26.18	.012	85.80	0.20	5.75	12.72	.033	90.26	0.21	0.91	10.19	.361	94.59	0.24
Percent University Students	0.33	11.63	.576	82.54	0.20	0.27	7.53	.616	87.00	0.22	9.54	7.80	.015	81.73	0.18
<b>Study-level Moderators</b>															
Anonymity Reassurance	0.53	85.80	.468	86.39	0.20	3.18	36.86	.083	92.42	0.24	0.97	33.62	.333	94.99	0.20
Participant Compensation	0.63	48.85	.600	86.78	0.22	0.62	8.42	.618	90.77	0.27	1.94	2.98	N/A	93.68	0.26
Electronic Data Collection	0.10	54.90	.756	87.39	0.20	0.19	4.68	.830	92.10	0.22	0.00	35.56	.949	94.89	0.19
Group Assessment	3.86	20.13	.038	87.13	0.19	0.96	3.52	N/A	93.78	0.23	0.48	5.23	.519	95.63	0.19
Personal Contact	0.32	11.30	.730	86.08	0.19	0.06	8.22	.947	92.80	0.24	2.06	28.66	.162	94.88	0.19
Sexuality Study	4.37	37.96	.043	82.05	0.16	0.00	20.29	.986	93.31	0.27	0.01	17.52	.944	91.77	0.16
Year of Study	0.73	46.47	.397	86.44	0.20	0.10	15.91	.755	91.73	0.24	0.16	14.27	.698	94.99	0.20

Note. Tests for moderation of the sex drive manifestations. The tests indicate significance of the slope for continuous moderators or differences between subgroups for categorical moderators. For cognition frequency, the results are statistically controlled for item content (extra-pair partner vs. any partner/no target). Results for the control variable are not reported. Some models could not be fitted because the number of available codings was insufficient. These are left blank. AHZ = Hotelling-T-approximated test statistic. df = small-sample-corrected degrees of freedom. p = p-value associated with the test statistic and df in the same row. F<sup>2</sup> = proportion of the variation in observed effects that is due to variation in true effects. τ = estimated standard deviation of the true effects. Note that if degrees of freedom fall below 4, significance tests are unreliable. p-values for unreliable tests are not reported (N/A).

Table 4

Regression Tables for Moderation Analyses (Sex Drive Manifestations)

Moderator	Cognition Frequency							Affect Frequency							Behavior Frequency						
	<i>g</i>	<i>SE</i>	<i>k</i>	<i>m</i>	<i>t</i>	<i>df</i>	<i>p</i>	<i>g</i>	<i>SE</i>	<i>k</i>	<i>m</i>	<i>t</i>	<i>df</i>	<i>p</i>	<i>g</i>	<i>SE</i>	<i>k</i>	<i>m</i>	<i>t</i>	<i>df</i>	<i>p</i>
<b>Outcome-level Moderators</b>																					
Aggregation Span																					
Intercept	0.31	0.08	30	46	3.84	3.96	N/A	0.36	0.09	36	70	4.08	6.56	.005	0.66	0.12	28	42	5.33	8.12	<.001
Slope	0.01	0.00			2.91	5.57	.029	0.01	0.00			2.12	9.81	.060	0.01	0.01			1.02	14.67	.323
Item Content																					
Unspecified partner								0.40	0.06	20	20	6.87	18.08	<.001							
No target								0.57	0.06	36	54	10.29	31.59	<.001							
Item Context																					
Alone															0.74	0.12	6	6	6.24	4.45	.002
Not specified															0.79	0.04	39	57	18.47	34.63	<.001
Type of Response Scale																					
No	0.59	0.03	157	268	18.51	47.43	<.001	0.56	0.04	55	82	14.26	48.84	<.001	0.82	0.05	29	32	15.20	24.27	<.001
Yes	0.43	0.10	8	14	4.29	5.89	.005	0.41	0.11	6	12	3.62	4.20	.021	0.72	0.06	19	31	12.45	15.91	<.001
Scale Range																					
Intercept	0.51	0.07	155	265	7.02	40.78	<.001	0.80	0.19	54	81	4.17	18.38	<.001	0.67	0.21	28	31	3.23	7.47	.013
Slope	0.01	0.01			1.18	50.95	.243	-0.04	0.03			-1.31	22.19	.204	0.03	0.04			0.71	6.62	.500
Item Wording																					
Daydreams	0.39	0.12	6	10	3.14	4.02	.034														
Fantasies	0.66	0.10	120	189	6.42	11.49	<.001														
Other	0.38	0.24	4	5	1.56	1.97	N/A	0.50	0.05	27	49	9.76	23.37	<.001	0.47	0.12	4	7	3.81	2.79	N/A
Thoughts	0.57	0.03	46	74	20.81	38.89	<.001														
Desire								0.59	0.05	35	45	11.23	30.16	<.001							
Masturbation															0.80	0.04	42	49	19.20	36.25	<.001
Self-stimulation															0.71	0.08	2	6	8.45	1.00	N/A
<b>Publication-level Moderators</b>																					
Aim to Find Gender Differences in Sex Drive																					
No	0.60	0.04	127	215	15.68	37.61	<.001	0.58	0.05	35	53	10.82	32.09	<.001	0.79	0.05	32	44	16.38	28.62	<.001
Yes	0.49	0.07	11	28	7.41	9.69	<.001	0.54	0.07	12	25	7.46	10.84	<.001	0.71	0.12	6	13	5.68	4.83	.003
Focus on Anonymity																					
No	0.58	0.04	74	136	14.33	47.69	<.001	0.59	0.06	23	32	9.31	21.12	<.001	0.79	0.07	17	25	11.57	14.94	<.001
Yes	0.56	0.04	70	118	15.13	48.88	<.001	0.54	0.05	29	51	10.75	26.14	<.001	0.77	0.05	27	38	15.09	23.52	<.001
Focus on Gender Differences in Sex Drive																					
No	0.59	0.04	118	201	15.72	38.04	<.001	0.58	0.06	32	48	9.80	29.36	<.001	0.75	0.06	21	32	11.80	18.10	<.001
Yes	0.54	0.06	20	42	9.16	19.15	<.001	0.53	0.06	15	30	9.50	13.79	<.001	0.79	0.06	17	25	12.72	15.55	<.001
Focus on Gender Differences																					
No	0.60	0.04	55	93	13.99	45.78	<.001	0.65	0.09	13	21	7.46	11.39	<.001	0.68	0.06	8	11	11.68	6.48	<.001
Yes	0.56	0.04	83	150	15.38	47.88	<.001	0.54	0.05	34	57	11.04	31.35	<.001	0.80	0.05	30	46	14.97	27.01	<.001
Gender of First Author																					
Female	0.60	0.03	57	102	17.75	49.38	<.001	0.61	0.05	33	47	11.48	30.39	<.001	0.79	0.06	16	20	13.43	14.50	<.001
Male	0.55	0.04	100	173	13.81	43.76	<.001	0.46	0.05	24	47	10.08	21.04	<.001	0.77	0.06	28	43	13.73	24.36	<.001
Mean Author Gender																					
Intercept	0.58	0.04	157	275	15.88	46.23	<.001	0.67	0.06	57	94	11.13	25.30	<.001	0.77	0.08	44	63	9.07	12.75	<.001
Slope	-0.01	0.06			-0.19	55.58	.848	-0.26	0.08			-3.04	23.18	.006	0.02	0.12			0.20	18.08	.845
Publication Status																					
Published	0.59	0.03	132	220	17.97	52.77	<.001	0.58	0.04	42	61	12.91	38.90	<.001	0.81	0.04	33	42	18.69	29.57	<.001
Unpublished	0.52	0.04	29	62	11.79	29.43	<.001	0.46	0.07	15	33	6.87	12.74	<.001	0.69	0.10	11	21	6.75	8.60	<.001
Sexuality Journal																					
No	0.60	0.04	117	196	15.92	48.10	<.001	0.53	0.05	39	65	11.28	35.34	<.001	0.79	0.06	28	43	14.07	25.16	<.001
Yes	0.55	0.03	44	86	15.66	41.51	<.001	0.59	0.07	18	29	8.95	16.31	<.001	0.77	0.06	16	20	13.32	13.68	<.001
<b>Sample-level Moderators</b>																					
Mean Age																					
Intercept	0.61	0.10	137	234	6.27	41.80	<.001	0.42	0.20	52	83	2.11	8.45	.066	0.90	0.14	31	42	6.18	10.17	<.001
Slope	-0.00	0.00			-0.44	31.05	.664	0.00	0.01			0.61	5.80	.566	-0.00	0.00			-0.66	6.84	.533
Percent White																					
Intercept	0.71	0.12	47	79	6.03	8.17	<.001	0.52	0.10	19	31	5.30	3.08	N/A	0.68	0.18	10	16	3.73	3.61	N/A
Slope	-0.00	0.00			-0.31	11.17	.759	0.00	0.00			0.54	4.26	.616	0.00	0.00			0.70	4.89	.515
Country-Level Gender Development																					
Intercept	-2.26	1.73	143	253	-1.31	39.64	.198	-0.74	2.60	55	91	-0.29	16.96	.778	3.50	3.00	41	60	1.16	19.07	.259
Slope	2.89	1.74			1.66	40.36	.105	1.32	2.63			0.50	17.01	.623	-2.77	3.07			-0.90	19.32	.377
Country-Level Gender Inequality																					
Intercept	0.52	0.08	146	256	6.84	35.87	<.001	0.56	0.10	55	91	5.36	34.77	<.001	0.90	0.09	42	61	10.24	15.13	<.001

Moderator	Cognition Frequency							Affect Frequency							Behavior Frequency							
	<i>g</i>	<i>SE</i>	<i>k</i>	<i>m</i>	<i>t</i>	<i>df</i>	<i>p</i>	<i>g</i>	<i>SE</i>	<i>k</i>	<i>m</i>	<i>t</i>	<i>df</i>	<i>p</i>	<i>g</i>	<i>SE</i>	<i>k</i>	<i>m</i>	<i>t</i>	<i>df</i>	<i>p</i>	
Slope	0.30	0.37			0.80	19.31	.435	-0.05	0.55			-0.09	35.59	.925	-0.65	0.47			-1.38	7.46	.206	
Percent Heterosexual																						
Intercept	0.62	0.14	81	150	4.31	9.31	.002	0.14	0.37	33	63	0.38	10.19	.712	0.61	0.19	24	37	3.19	2.94	N/A	
Slope	-0.00	0.00			-0.36	10.89	.728	0.00	0.00			1.19	12.60	.254	0.00	0.00			0.65	3.45	N/A	
Average Partnership Duration in Weeks																						
Intercept	0.67	0.09	31	46	7.17	9.85	<.001	0.51	0.10	17	21	5.23	9.31	<.001	0.96	0.10	9	9	9.65	4.94	<.001	
Slope	-0.00	0.00			-0.60	7.48	.566	0.00	0.00			2.04	2.03	N/A	-0.00	0.00			-1.53	1.74	N/A	
Percent Parents																						
Intercept	0.38	0.19	21	37	1.99	4.30	.112	0.36	0.13	12	17	2.75	8.91	.023	0.97	0.14	11	12	7.14	5.59	<.001	
Slope	0.00	0.00			0.87	6.18	.417	0.00	0.00			2.75	2.59	N/A	-0.01	0.01			-1.51	2.38	N/A	
Country-Level Sex Ratio																						
Intercept	-0.56	0.91	146	256	-0.62	21.58	.544	2.04	2.09	55	91	0.97	12.17	.349	-1.44	2.29	42	61	-0.63	7.01	.548	
Slope	0.01	0.01			1.26	21.73	.221	-0.01	0.02			-0.71	12.37	.489	0.02	0.02			0.97	7.13	.362	
Study Restricted to Sexually Active																						
No	0.62	0.04	97	172	15.70	35.13	<.001	0.56	0.05	36	64	11.24	33.13	<.001								
Yes	0.49	0.06	16	29	8.41	14.93	<.001	0.60	0.09	7	12	6.53	5.84	<.001								
Percent Single																						
Intercept	0.51	0.04	88	158	12.93	32.26	<.001	0.62	0.06	42	71	11.00	21.85	<.001	0.72	0.10	33	47	7.05	12.45	<.001	
Slope	0.00	0.00			2.69	26.18	.012	-0.00	0.00			-2.40	12.72	.033	0.00	0.00			0.96	10.19	.361	
Percent University Students																						
Intercept	0.54	0.09	80	132	5.78	9.34	<.001	0.47	0.10	25	39	4.62	4.81	.006	0.62	0.07	20	30	8.71	4.53	<.001	
Slope	0.00	0.00			0.58	11.63	.576	0.00	0.00			0.52	7.53	.616	0.00	0.00			3.09	7.80	.015	
<b>Study-level Moderators</b>																						
Anonymity Reassurance																						
No	0.58	0.04	92	164	15.43	46.42	<.001	0.61	0.05	30	43	11.37	28.01	<.001	0.75	0.05	25	33	14.51	22.88	<.001	
Yes	0.55	0.04	49	86	12.40	41.64	<.001	0.47	0.06	20	37	8.51	17.54	<.001	0.83	0.07	19	30	12.57	15.92	<.001	
Participant Compensation																						
Course-credit	0.64	0.06	29	53	10.12	27.41	<.001	0.68	0.14	5	7	5.00	3.74	N/A	1.02	0.12	4	5	8.43	2.74	N/A	
Material	0.54	0.05	35	64	10.62	29.14	<.001	0.57	0.10	12	26	5.63	10.66	<.001	0.60	0.09	7	13	6.54	5.88	<.001	
Mixed	0.60	0.06	18	29	10.97	21.59	<.001	0.60	0.06	4	8	10.64	2.97	N/A	0.71	0.08	2	6	8.42	1.00	N/A	
None	0.62	0.07	18	35	9.29	22.53	<.001	0.46	0.09	6	11	5.05	4.68	.005	0.87	0.08	7	10	10.30	5.81	<.001	
Electronic Data Collection																						
No	0.59	0.05	36	53	10.93	35.42	<.001	0.57	0.07	13	18	7.99	10.89	<.001	0.79	0.06	20	27	13.32	17.38	<.001	
Yes	0.57	0.03	96	181	17.12	41.31	<.001	0.55	0.05	34	58	11.06	31.43	<.001	0.80	0.06	21	32	13.46	18.36	<.001	
Mixed								0.68	0.18	3	5	3.73	1.85	N/A								
Group Assessment																						
Mixed	0.38	0.07	10	13	5.51	9.70	<.001	0.52	0.05	4	8	10.67	2.84	N/A								
No	0.58	0.03	103	180	16.61	41.61	<.001	0.58	0.05	39	63	12.29	35.81	<.001	0.78	0.04	33	45	18.11	29.25	<.001	
Yes	0.63	0.08	19	30	7.86	22.64	<.001								0.90	0.16	5	5	5.46	3.91	N/A	
Personal Contact																						
Mixed	0.56	0.06	5	8	8.99	3.82	N/A	0.58	0.07	4	8	7.82	2.87	N/A								
No	0.56	0.03	84	151	16.55	44.14	<.001	0.57	0.06	29	48	10.00	26.82	<.001	0.75	0.05	26	34	15.18	23.42	<.001	
Yes	0.60	0.04	57	97	13.53	48.20	<.001	0.55	0.06	19	28	8.67	16.78	<.001	0.87	0.06	17	27	13.75	14.36	<.001	
Sexuality Study																						
No	0.66	0.05	25	42	12.44	21.37	<.001	0.61	0.05	12	18	11.46	10.30	<.001	0.79	0.09	13	20	9.11	10.01	<.001	
Yes	0.55	0.04	44	87	14.96	26.67	<.001	0.61	0.07	21	39	8.60	19.45	<.001	0.79	0.05	23	31	16.59	20.62	<.001	
Year of Study																						
Intercept	9.65	10.64	152	266	0.91	46.36	.369	-5.25	18.27	56	92	-0.29	15.90	.778	-4.04	12.16	44	63	-0.33	14.24	.745	
Slope	-0.00	0.01			-0.85	46.47	.397	0.00	0.01			0.32	15.91	.755	0.00	0.01			0.40	14.27	.698	

Note. Meta-regression tables for moderation of the sex drive manifestations. For categorical moderators, point estimates for subgroups and corresponding significance tests are presented. For continuous moderators, values are presented for the intercept and slope. For cognition frequency, results are statistically controlled for item content (extra-pair partner vs. any partner/no target). Results for the control variable are not reported. Some models could not be fitted because the number of available codings was insufficient. These are left blank. *g* = Hedges' *g* effect size (positive favors males). *SE* = Standard error for Hedges' *g* effect size. *k* = number of studies per subgroup. *m* = number of effect sizes per subgroup. *t*-value from *t*-test testing the parameter against zero. *df* = small sample corrected degrees of freedom. *p* = *p*-value associated with the *t*-value and *df* in the same row. Note that if degrees-of-freedom fall below 4, significance tests are unreliable. *p*-values for unreliable tests are not reported (N/A).

Table 5

Moderator Overview

Moderator	Total		Cognition Frequency			Affect Frequency			Behavior Frequency		
	<i>m</i>	Compl.	<i>m</i>	Compl.	Distribution	<i>m</i>	Compl.	Distribution	<i>m</i>	Compl.	Distribution
<b>Outcome-level Moderators</b>											
Item Content	558	62%	282	100%	unspecified partner (m = 32), extra-pair partner (m = 168), no target (m = 82)	76	81%	unspecified partner (m = 20), masturbation (m = 1), no target (m = 54), own partner (m = 1), NA (m = 18)	0	0%	NA (m = 63)
Item Context	639	71%	282	100%	asleep (m = 1), at work (m = 4), before sleep (m = 1), being bored (m = 2), everyday life (m = 1), in everyday life (m = 1), not specified (m = 266), on the way to work (m = 2), traveling on a train or bus (m = 2), while reading (m = 1), while working at a job (m = 1)	94	100%	in contact with an extrapair person (m = 1), not specified (m = 93)	63	100%	alone (m = 6), not specified (m = 57)
Type of Response Scale	897	100%	282	100%	no (m = 268), yes (m = 14)	94	100%	no (m = 82), yes (m = 12)	63	100%	no (m = 32), yes (m = 31)
Scale Range	732	82%	265	94%	Q = [2.00, 4.00, 7.00, 8.00, 9.00], M = 6.37, SD = 1.72	81	86%	Q = [3.00, 6.00, 7.00, 7.00, 9.00], M = 6.14, SD = 1.34	31	49%	Q = [2.00, 4.00, 5.00, 5.50, 8.00], M = 4.84, SD = 1.24
Item Wording	639	71%	282	100%	daydreams (m = 10), dreams (m = 1), fantasies (m = 189), other (m = 5), pre-occupation (m = 3), thoughts (m = 74)	94	100%	desire (m = 45), other (m = 49)	63	100%	masturbation (m = 49), other (m = 7), self-stimulation (m = 6), touch and explore body (m = 1)
Aggregation Span	185	21%	48	17%	Q = [1.00, 1.00, 30.00, 30.00, 183.00], M = 23.67, SD = 27.71	71	76%	Q = [0.00, 5.00, 30.00, 30.00, 365.00], M = 26.17, SD = 42.75	49	78%	Q = [7.00, 7.00, 28.00, 30.00, 365.00], M = 48.04, SD = 89.43
<b>Publication-level Moderators</b>											
Aim to Find Gender Differences in Sex Drive	779	87%	243	86%	no (m = 215), yes (m = 28), NA (m = 39)	78	83%	no (m = 53), yes (m = 25), NA (m = 16)	57	90%	no (m = 44), yes (m = 13), NA (m = 6)
Focus on Gender Differences in Sex Drive	780	87%	243	86%	no (m = 201), yes (m = 42), NA (m = 39)	78	83%	no (m = 48), yes (m = 30), NA (m = 16)	57	90%	no (m = 32), yes (m = 25), NA (m = 6)
Focus on Gender Differences	780	87%	243	86%	no (m = 93), yes (m = 150), NA (m = 39)	78	83%	no (m = 21), yes (m = 57), NA (m = 16)	57	90%	no (m = 11), yes (m = 46), NA (m = 6)
Gender of First Author	884	99%	275	98%	female (m = 102), male (m = 173), NA (m = 7)	94	100%	female (m = 47), male (m = 47)	63	100%	female (m = 20), male (m = 43)
Publication Status	897	100%	282	100%	published (m = 220), unpublished (m = 62)	94	100%	published (m = 61), unpublished (m = 33)	63	100%	published (m = 42), unpublished (m = 21)
Sexuality Journal	897	100%	282	100%	No (m = 196), Yes (m = 86)	94	100%	No (m = 65), Yes (m = 29)	63	100%	No (m = 43), Yes (m = 20)
Focus on Anonymity	815	91%	254	90%	no (m = 136), yes (m = 118), NA (m = 28)	83	88%	no (m = 32), yes (m = 51), NA (m = 11)	63	100%	no (m = 25), yes (m = 38)
Mean Author Gender	884	99%	275	98%	Q = [0.00, 0.50, 0.67, 1.00, 1.33], M = 0.65, SD = 0.35	94	100%	Q = [0.00, 0.21, 0.50, 1.00, 1.33], M = 0.56, SD = 0.43	63	100%	Q = [0.00, 0.50, 1.00, 1.00, 1.33], M = 0.73, SD = 0.43

Moderator	Total		Cognition Frequency			Affect Frequency			Behavior Frequency		
	<i>m</i>	Compl.	<i>m</i>	Compl.	Distribution	<i>m</i>	Compl.	Distribution	<i>m</i>	Compl.	Distribution
<b>Sample-level Moderators</b>											
Mean Age	775	86%	235	83%	Q = [16.78, 20.85, 23.96, 29.89, 70.55], M = 25.89, SD = 6.87	83	88%	Q = [18.60, 24.05, 26.47, 30.45, 51.28], M = 27.77, SD = 5.89	43	68%	Q = [15.55, 22.82, 25.95, 30.75, 74.00], M = 29.22, SD = 10.29
Percent White	323	36%	79	28%	Q = [15.50, 60.00, 78.60, 90.50, 100.00], M = 73.92, SD = 19.73	31	33%	Q = [15.50, 62.42, 78.60, 88.20, 98.00], M = 70.91, SD = 24.69	16	25%	Q = [32.20, 50.88, 58.98, 86.25, 100.00], M = 65.90, SD = 24.93
Country-Level Gender Inequality	828	92%	256	91%	Q = [0.05, 0.10, 0.18, 0.25, 0.63], M = 0.18, SD = 0.09	91	97%	Q = [0.08, 0.09, 0.13, 0.24, 0.26], M = 0.16, SD = 0.07	61	97%	Q = [0.05, 0.10, 0.16, 0.25, 0.56], M = 0.18, SD = 0.10
Country-Level Gender Development	828	92%	256	91%	Q = [0.76, 0.97, 0.99, 0.99, 1.03], M = 0.98, SD = 0.02	91	97%	Q = [0.94, 0.97, 0.99, 0.99, 1.03], M = 0.98, SD = 0.01	61	97%	Q = [0.85, 0.96, 0.98, 0.99, 1.00], M = 0.97, SD = 0.02
Percent Heterosexual	516	58%	150	53%	Q = [31.66, 85.48, 100.00, 100.00, 100.00], M = 92.31, SD = 12.19	63	67%	Q = [47.41, 77.75, 95.10, 100.00, 100.00], M = 88.13, SD = 14.49	37	59%	Q = [31.66, 77.02, 96.50, 100.00, 100.00], M = 87.93, SD = 15.65
Percent Single	548	61%	158	56%	Q = [0.00, 0.00, 37.51, 50.00, 100.00], M = 33.85, SD = 27.67	71	76%	Q = [0.00, 0.00, 35.10, 43.78, 100.00], M = 27.30, SD = 26.11	47	75%	Q = [0.00, 30.27, 45.98, 50.67, 100.00], M = 40.08, SD = 25.04
Percent University Students	424	47%	132	47%	Q = [0.00, 89.18, 100.00, 100.00, 100.00], M = 86.02, SD = 26.36	39	41%	Q = [14.20, 58.99, 92.00, 100.00, 100.00], M = 78.59, SD = 28.14	30	48%	Q = [0.00, 53.57, 86.38, 100.00, 100.00], M = 71.67, SD = 33.70
Average Partnership Duration in Weeks	191	21%	46	16%	Q = [1.80, 25.83, 52.52, 58.05, 158.40], M = 56.14, SD = 43.73	21	22%	Q = [3.91, 35.66, 54.94, 78.89, 287.76], M = 70.78, SD = 63.41	9	14%	Q = [2.74, 54.42, 54.94, 55.80, 287.76], M = 82.38, SD = 87.74
Percent Parents	147	16%	37	13%	Q = [0.00, 14.55, 20.53, 36.00, 100.00], M = 30.77, SD = 28.51	17	18%	Q = [14.55, 17.30, 25.00, 64.00, 100.00], M = 44.44, SD = 35.03	12	19%	Q = [0.00, 16.37, 19.26, 30.00, 59.70], M = 24.95, SD = 16.81
Study Restricted to Sexually Active	662	74%	201	71%	no (m = 172), yes (m = 29), NA (m = 81)	76	81%	no (m = 64), yes (m = 12), NA (m = 18)	52	83%	no (m = 49), yes (m = 3), NA (m = 11)
Country-Level Sex Ratio	828	92%	256	91%	Q = [92.17, 100.88, 101.19, 103.00, 108.35], M = 101.41, SD = 2.46	91	97%	Q = [93.92, 100.88, 101.23, 103.24, 104.88], M = 101.69, SD = 2.47	61	97%	Q = [92.17, 100.88, 101.36, 103.85, 104.92], M = 101.80, SD = 2.72
<b>Study-level Moderators</b>											
Anonymity Reassurance	798	89%	250	89%	no (m = 164), yes (m = 86), NA (m = 32)	80	85%	no (m = 43), yes (m = 37), NA (m = 14)	63	100%	no (m = 33), yes (m = 30)
Participant Compensation	571	64%	181	64%	course-credit (m = 53), material (m = 64), mixed (m = 29), none (m = 35), NA (m = 101)	52	55%	course-credit (m = 7), material (m = 26), mixed (m = 8), none (m = 11), NA (m = 42)	34	54%	course-credit (m = 5), material (m = 13), mixed (m = 6), none (m = 10), NA (m = 29)
Sexuality Study	454	51%	129	46%	no (m = 42), yes (m = 87), NA (m = 153)	57	61%	no (m = 18), yes (m = 39), NA (m = 37)	51	81%	no (m = 20), yes (m = 31), NA (m = 12)
Year of Study	852	95%	266	94%	Q = [1996.00, 2008.00, 2012.00, 2015.00, 2019.00], M = 2011.38, SD = 4.58	92	98%	Q = [2000.00, 2008.75, 2012.00, 2015.63, 2019.00], M = 2012.03, SD = 4.71	63	100%	Q = [1992.00, 2004.50, 2008.00, 2014.50, 2019.00], M = 2008.70, SD = 7.01
Face-to-Face Interview	848	95%	266	94%	no (m = 265), yes (m = 1), NA (m = 16)	83	88%	no (m = 82), yes (m = 1), NA (m = 11)	62	98%	no (m = 60), yes (m = 2), NA (m = 1)
Electronic Data Collection	786	88%	238	84%	mixed (m = 4), no (m = 53), yes (m = 181), NA (m = 44)	81	86%	mixed (m = 5), no (m = 18), yes (m = 58), NA (m = 13)	60	95%	mixed (m = 1), no (m = 27), yes (m = 32), NA (m = 3)
Group Assessment	732	82%	223	79%	mixed (m = 13), no (m = 180), yes (m = 30), NA (m = 59)	75	80%	mixed (m = 8), no (m = 63), yes (m = 4), NA (m = 19)	50	79%	no (m = 45), yes (m = 5), NA (m = 13)
Personal Contact	829	92%	256	91%	mixed (m = 8), no (m = 151), yes (m = 97), NA (m = 26)	84	89%	mixed (m = 8), no (m = 48), yes (m = 28), NA (m = 10)	61	97%	no (m = 34), yes (m = 27), NA (m = 2)

Note. *m*: Absolute number of effect sizes for which the corresponding characteristic could be coded. Compl.: Percentage of effect sizes for which the corresponding characteristic could be coded. Distribution: Information about the distribution of the coded characteristics. For categorical characteristics, the number of effect sizes per subgroup is reported. For continuous characteristics, Q are quartiles (minimum, 25% quartile, median, 75% quartile, maximum), M is the mean, and SD is the standard deviation.

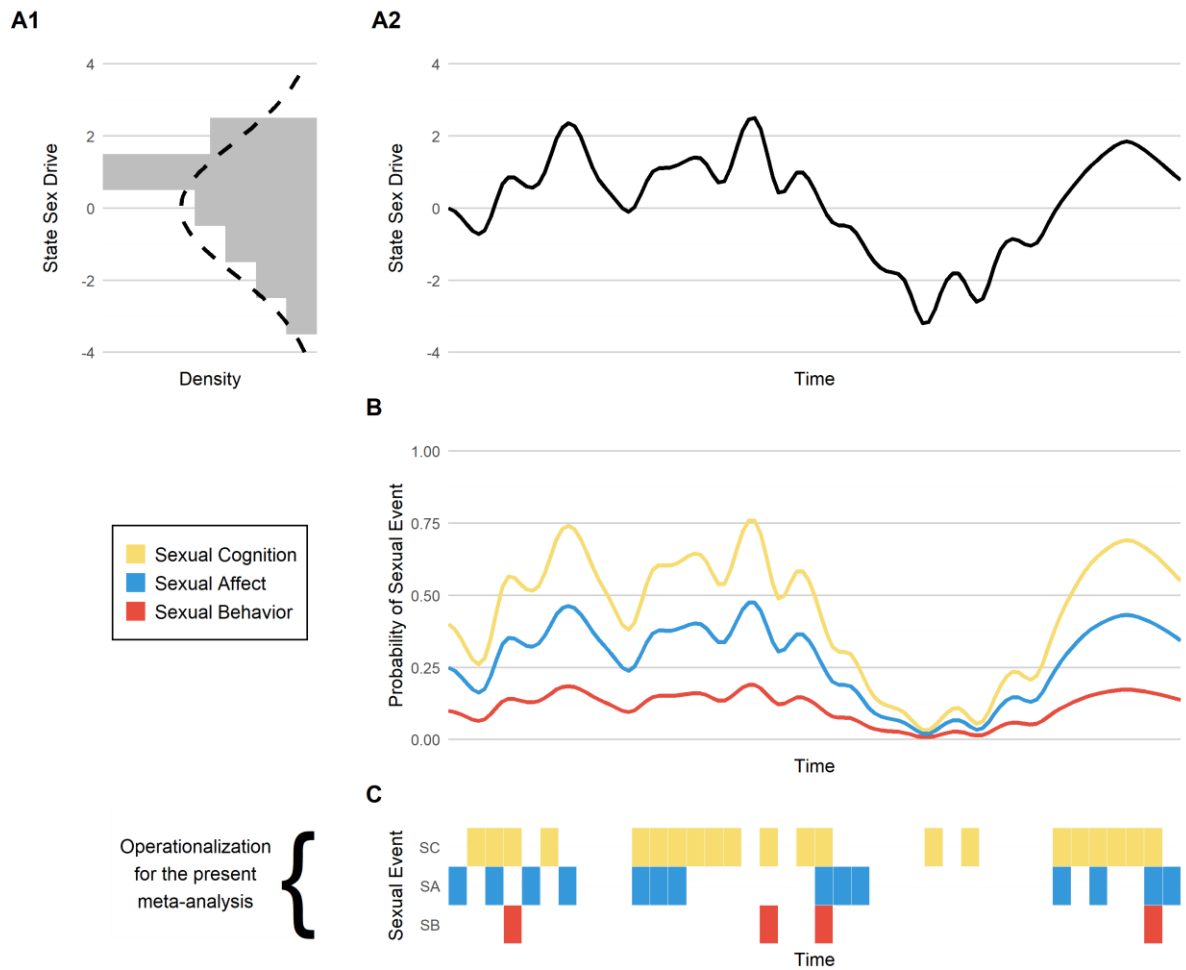


Moderator	Total		Cognition Frequency			Affect Frequency			Behavior Frequency		
	<i>m</i>	Compl.	<i>m</i>	Compl.	Distribution	<i>m</i>	Compl.	Distribution	<i>m</i>	Compl.	Distribution

Note that summaries for continuous moderators are computed on the effect size level for this table. In the results section, some of this information was presented on the level of individual participants (i.e., as summaries weighted by sample size). Some values may therefore differ.

**Figure 1**

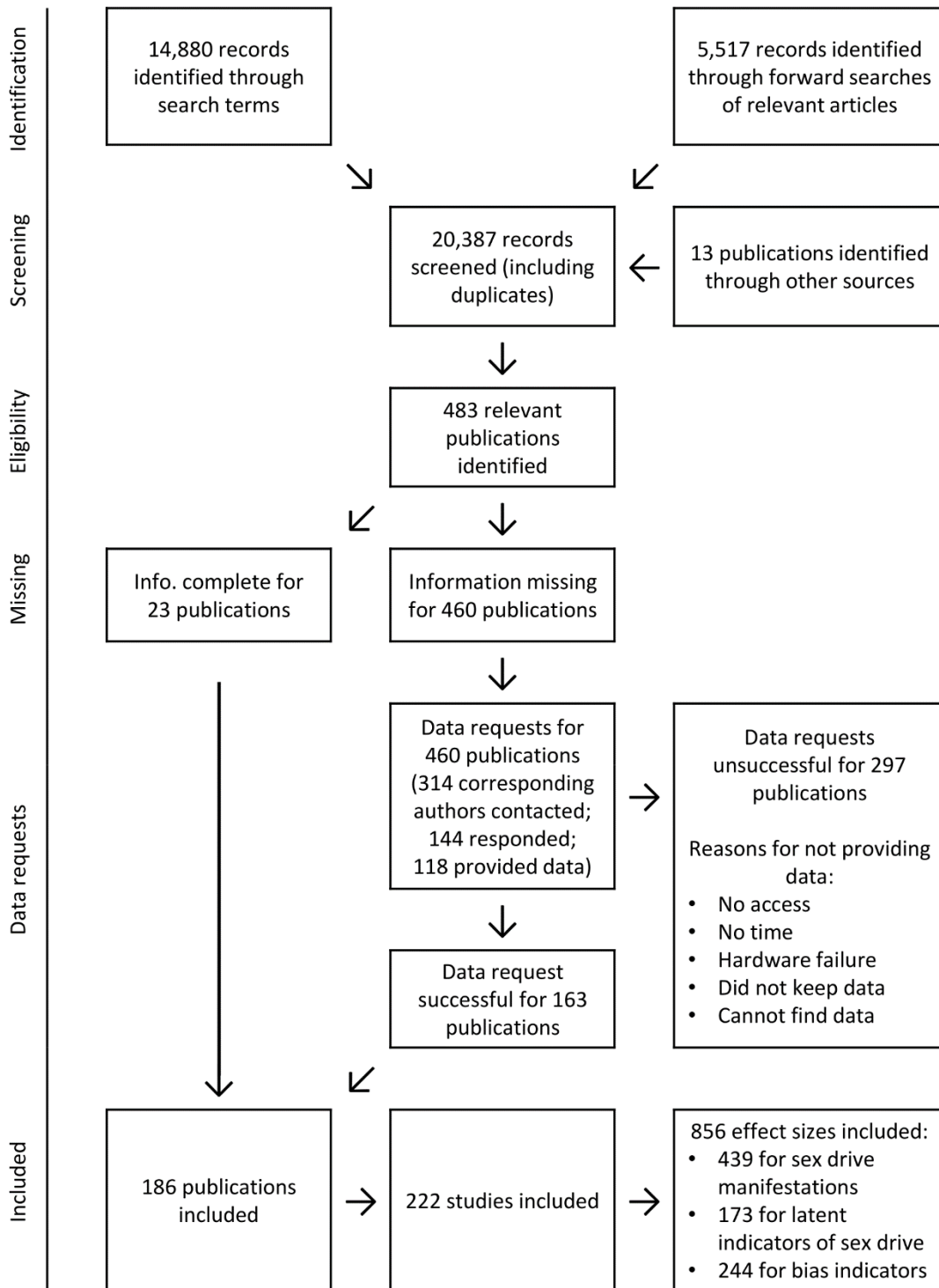
*Conceptualization of Sex Drive*



*Note.* This figure visualizes random data for one hypothetical person generated under our theoretical conceptualization of sex drive. Panel A1 depicts a histogram of the observed density (grey rectangles) and the curve of the expected density (dashed line) of the distribution of state sex drive depicted in Panel A2. In Panel A2, the black line displays the fluctuation of state sex drive over time. Panel B depicts the probability of sexual events over time. Panel C depicts the occurrence of sexual events over time. A colored rectangle indicates that the respective sexual event occurred at a given point in time. The depicted occurrences are the result of random sampling according to the probabilities depicted in Panel B. In Panels B and C, yellow denotes sexual cognition (SC), blue denotes sexual affect (SA), and red denotes sexual behavior (SB). Gender differences in this quantity will be meta-analyzed in the present study.

**Figure 2**

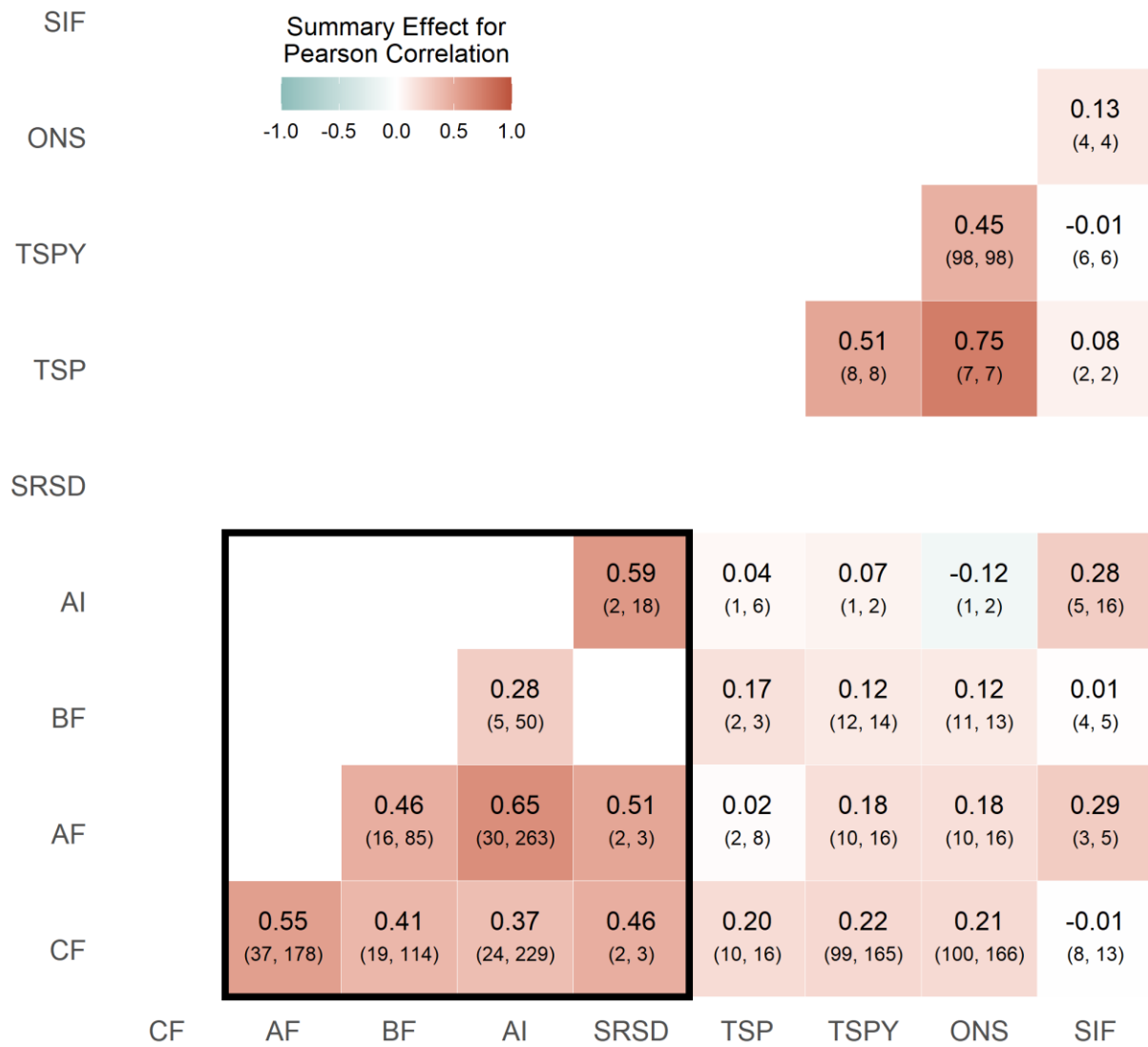
*Flow of Data into the Research Synthesis*



*Note.* Flowchart of the literature search and study coding. Note that some corresponding authors contributed more than one publication. Of the 20,387 publications identified during the identification stage, 19,904 did not meet the inclusion criteria (reported at the beginning of the Method section) as became evident either when screening the abstract or the full text.

**Figure 3**

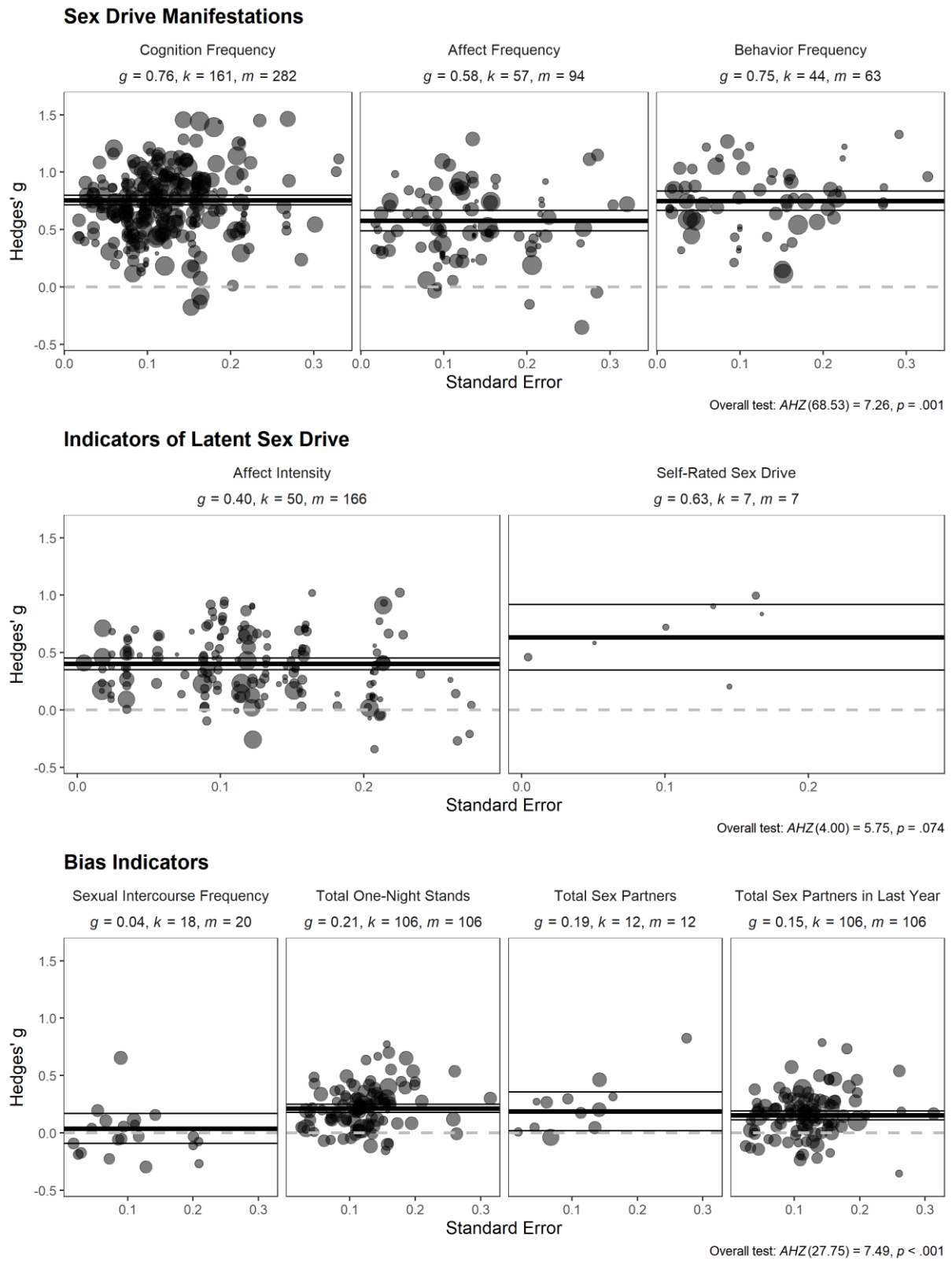
*Meta-Analytic Correlation Table*



*Note.* Meta-analytic correlation table displaying convergent validity for sex drive indicators. Values in the table without parentheses are summary effects for Pearson correlations for pairwise complete observations. The first value in parentheses denotes the number of studies that contributed to the summary effect (*k*) and the second value denotes the number of effect sizes (*m*). The solid box contains correlations among the sex drive manifestations and indicators of latent sex drive (convergent validity). CF = Cognition Frequency; AF = Affect Frequency; BF = Behavior Frequency; AI = Affect Intensity; SRSD = Self-Rated Sex Drive; TSP = Total Sexual Partners; TSPY = Total Sexual Partners in Last Year; ONS = Total One-night Stands; SIF = Sexual Intercourse Frequency.

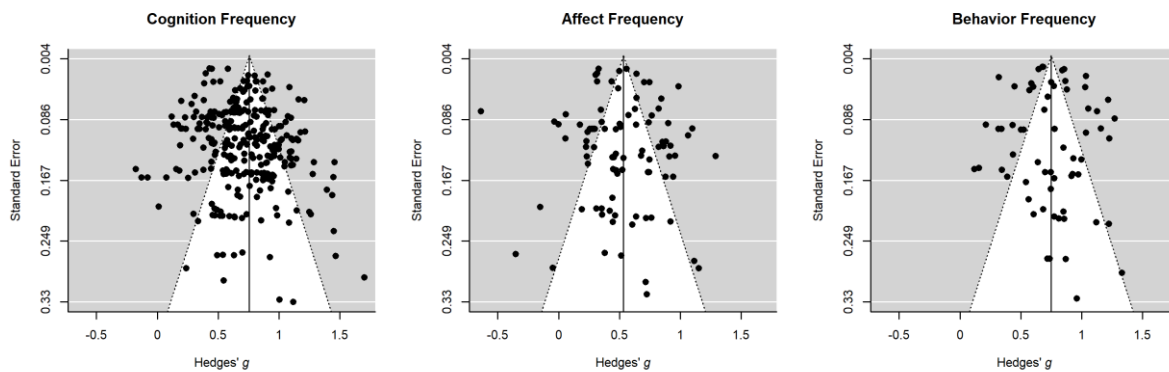
**Figure 4**

*Main Results*



*Note.* Main summary effects and confidence intervals for gender differences in sex drive manifestations (i.e., sex drive facets, top panel), indicators of latent sex drive (middle panel), and bias indicators (bottom panel).  $g =$

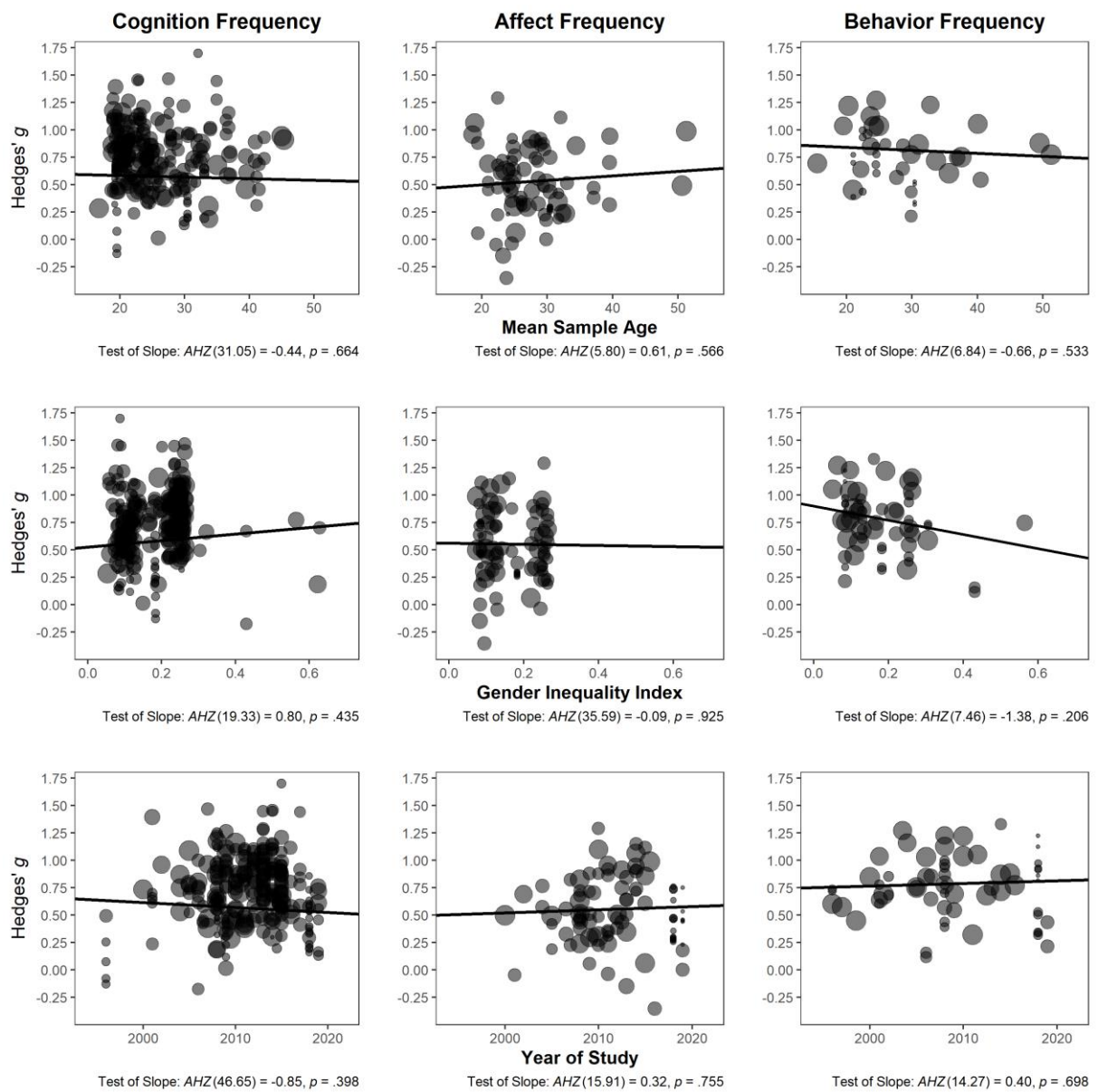
Hedges'  $g$  summary effect within the respective subgroup (positive values indicate larger values in men);  $AHZ = AHZ$  value for the test of group differences;  $p = p$ -value for the test of group differences;  $k =$  number of studies per subgroup;  $m =$  number of effect sizes per subgroup. Black dots represent individual effect sizes. The thick black horizontal lines represent the meta-analytic summary effects within the subgroups. The thin black horizontal lines represent the borders of the 95% confidence interval. The dashed grey horizontal line represents the null effect at  $g = 0$ . Standard error for each effect is depicted on the  $x$ -axis. Circle size represents the weight of the respective effect size in the RVE meta-regression model. Darker circles are due to multiple, overlapping effect sizes.

**Figure 5***Funnel Plots for Sex Drive Manifestations*

*Note.* The solid vertical lines represent the within-subgroup summary effects. x-axis: Hedges'  $g$  effect sizes, positive values indicate larger values in men. y-axis: Standard error of effect sizes. The dotted lines denote the area in which 95% of effect sizes are expected to fall in the absence of heterogeneity. Leave-one-out analyses identified one outlier in the center plot at  $g < -0.5$  and Standard Error  $< 0.086$ . This effect was removed for all other analyses. Summary effects displayed in the figure were computed after removing the outlier.

**Figure 6**

*Moderation Results for Mean Sample Age, Gender Inequality, and Year of Study.*



*Note.* Depicted are scatterplots for meta-regression analyses.  $g$  = Hedges'  $g$  effect size (positive favors males);  $AHZ$  =  $AHZ$  value for the test of the slope;  $p$  =  $p$ -value for the test of the slope. The solid black lines represent the slopes of the meta-regressions. Circle size represents the weight of the respective effect size in the RVE meta-regression model. Darker circles are due to multiple, overlapping effect sizes. Higher values for the Gender Inequality Index denote higher inequality.



Supplemental Materials

Sex Drive: Theoretical Conceptualization and Meta-Analytic Review of Gender Differences

## Introduction

These supplemental materials contain additional figures and tables to complement the analyses reported in the main article. Table S1 summarizes the terms for the electronic literature search. Table S2 gives an overview of the psychometric inventories from which we drew individual items to included them in the meta-analysis. Tables S3 to S5 summarize results for meta-moderation analyses for the indicators of latent sex drive. Table S6 gives references for all psychometric inventories from which we drew questionnaire items. Figure 1 shows the distribution of effect sizes for the sex drive manifestations and indicators of latent sex drive, as well as for the bias indicators. Figure S2 depicts results of the leave-one-out analyses for outlier detection. Figures S3 and S4 show funnel plots for the indicators of latent sex drive and bias indicators, respectively. Figure S5 shows how the distributions of men's and women's sex drive overlap under assumptions of normality.

## A Formalized Definition of Sex Drive

In line with recent calls to incorporate more formal modelling into psychology (Guest & Martin, 2020), we now present a mathematical definition of our conceptualization of sex drive. The illustrative data in Figure 1 in the main article were generated according to these equations. We describe the fluctuation of state sex drive in an individual over time  $X$  as a simple auto-regressive process (AR(1) process; Mills, 1991), assuming that state sex drive at a given point in time  $X_t$  depends by some factor  $\beta$  on state sex drive at the previous point in time  $X_{t-1}$ :

$$X_t = c + \beta X_{t-1} + \epsilon_t \quad (1)$$

where  $\epsilon_t$  is Gaussian white noise with zero mean and constant variance  $\sigma_\epsilon^2$  and  $c$  is a constant.

The relationship between state sex drive  $X$  and the probability  $p$  that a particular sexual event  $s$  occurs at a given point in time  $t$ , may be described by some function  $\Phi$ :

$$p(s_t) = \Phi(X_t) \quad (2)$$

While the exact nature of the relationship between state sex drive and the probability of sexual events, that is,  $\Phi$ , is beyond the scope of the current work, we may engage in some speculation. For instance, it appears reasonable to assume that the relationship follows a monotonic function, such that higher sex drive implies a higher probability of sexual events. It may also be reasonable to assume that the relationship is not identical for the events of sexual cognition, affect, and behavior. For example, sexual cognition may be more likely than sexual behavior at any given point in time. There are many potential functions that satisfy these requirements. A natural contender could for example be the logistic function multiplied with some constant  $b$ :

$$\Phi(z) = b \left( \frac{1}{1 + e^{-z}} \right) \quad (3)$$

where  $b$  is different for each type of sexual event, that is, sexual cognition, affect, and behavior. We wish to emphasize again that this is speculative, and that finding the true relationship is beyond the scope of this work. For full transparency: Our choice of function was in part guided by aesthetic considerations for generating the data in Figure 1. The same is true for our choice of parametrizations when generating the data. We assumed that  $\sigma_\epsilon^2 = 0.64$ ,  $\beta = 0.9$ ,  $c = 0$ ,  $b_{SexCog} = 0.8$ ,  $b_{SexAff} = 0.5$ , and  $b_{SexBeh} = 0.2$ . We note that some obvious shortcomings of this model come immediately to mind. The fact that  $b$  effectively caps the probability of sexual events, such that for example the probability of sexual behavior never exceeds 20%, is unrealistic. It is also unrealistic that the probabilities for each type of sexual event always move in tandem, that is, that the lines never cross. Furthermore, we have omitted "hard" situational constraints on sexual behavior from the model (i.e., in most situations, behavioral acts to attain sexual gratification are not an option). However, addressing these limitations would severely complicate the model, and a more complicated model would not serve our purpose of illustration. For the moment, we view the model as oversimplified, yet useful.

### References

- Guest, O., & Martin, A. E. (2020). *How computational modeling can force theory building in psychological science*. PsyArXiv. <https://psyarxiv.com/rybh9/>
- Mills, T. C. (1991). *Time series techniques for economists*. Cambridge University Press.

**Table S1***Search Terms for Literature Search*

Domain	Terms	a	b	c	d	e	f	g
Cognition	Terms 1	sex*	for sex	erotic				
	Terms 2	thought*	cogniti*	reverie*	fantas*	daydream*	think*	ruminat*
	Terms 3	disorder	disfunction					
Affect	Terms 1	sex*	for sex	erotic				
	Terms 2	desire*	urge*	impuls*	crav*	drive*	motiv*	
	Terms 3	disorder	disfunction					
Behavior	Terms 1	masturbate*						
	Terms 2							
	Terms 3	disorder	disfunction					

*Note.* Search terms for each domain were composed as "One of term 1 AND one of term 2 NOT one of term 3".

**Table S2***Overview of Psychometric Inventories*

Indicator	Inventories
Cognition Frequency	Sociosexual Orientation Inventory Revised, Sociosexual Orientation Inventory, Sexual Desire Inventory, Derogatis Sexual Functioning Inventory, Sexual Compulsivity Scale, Sexuality Scale, Imaginal Process Inventory, Multidimensional Sociosexual Orientation Inventory, Multidimensional Sexuality Questionnaire, Hurlbert Index of Sexual Desire, Sexual Behavior Inventory, Sex Knowledge and Attitudes Test Adolescents, Changes in Sexual Functioning Inventory, Trait Sex Drive Scale
Affect Frequency	Sex Drive Questionnaire, Sexual Desire Inventory, International Index of Erectile Function / Female Sexual Functioning Index, Sociosexual Orientation Inventory Revised, Israeli Sexual Behaviour Inventory, Hurlbert Index of Sexual Desire, Changes in Sexual Functioning Inventory, Trait Sex Drive Scale
Behavior Frequency	Sex Drive Questionnaire, Derogatis Sexual Functioning Inventory, Sexual Behavior Inventory, Sex Knowledge and Attitudes Test Adolescents, Trait Sex Drive Scale
Affect Intensity	Sexual Desire Inventory, International Index of Erectile Function / Female Sexual Functioning Index, Multidimensional Sexuality Questionnaire, Hurlbert Index of Sexual Desire
Self-Rated Sex Drive	Hurlbert Index of Sexual Desire
Sexual Intercourse Frequency	Sociosexual Orientation Inventory, Derogatis Sexual Functioning Inventory, Israeli Sexual Behaviour Inventory, Sex Knowledge and Attitudes Test Adolescents
Total One-Night Stands	Sociosexual Orientation Inventory Revised, Sociosexual Orientation Inventory, Multidimensional Sociosexual Orientation Inventory
Total Sex Partners	Multidimensional Sociosexual Orientation Inventory, Sociosexual Orientation Inventory, Sexuality Scale
Total Sex Partners in Last Year	Sociosexual Orientation Inventory Revised, Sociosexual Orientation Inventory, Multidimensional Sociosexual Orientation Inventory

*Note.* See the supplementary online materials for a complete list of the inventories with references.

**Table S3***List of Included Inventories with References*

Inventory	Abbreviation	Reference
Changes in Sexual Functioning Inventory	CSFQ	Clayton, A. H., McGarvey, E. L., & Clavet, G. J. (1997). The Changes in Sexual Functioning Questionnaire (CSFQ): Development, reliability, and validity. <i>Psychopharmacology Bulletin</i> , 33(4), 731–745.
Derogatis Sexual Functioning Inventory	DSFI	Derogatis, L. R., & Melisaratos, N. (1979). The DSFI: A multidimensional measure of sexual functioning. <i>Journal of Sex &amp; Marital Therapy</i> , 5(3), 244–281. <a href="https://doi.org/10.1080/00926237908403732">https://doi.org/10.1080/00926237908403732</a>
Hurlbert Index of Sexual Desire	HISD	Apt, C. V., & Hurlbert, D. F. (1992). Motherhood and female sexuality beyond one year postpartum: A study of military wives. <i>Journal of Sex Education and Therapy</i> , 18(2), 104–114. <a href="https://doi.org/10.1080/01614576.1992.11074044">https://doi.org/10.1080/01614576.1992.11074044</a>
Imaginal Process Inventory	ISI	Giambra, L. M. (1980). A factor analysis of the items of the Imaginal Processes Inventory. <i>Journal of Clinical Psychology</i> , 36(2), 383–409. <a href="https://doi.org/10.1002/jclp.6120360203">https://doi.org/10.1002/jclp.6120360203</a>
International Index of Erectile Function / Female Sexual Functioning Index	IIEF/FSFI	Rosen, R., Brown, C., Heiman, J., Meston, C., Shabsigh, R., Ferguson, D., & D'Agostino, R. (2000). The Female Sexual Function Index (FSFI): A multidimensional self-report instrument for the assessment of female sexual function. <i>Journal of Sex &amp; Marital Therapy</i> , 26(2), 191–208. <a href="https://doi.org/10.1080/009262300278597">https://doi.org/10.1080/009262300278597</a> Rosen, R., Riley, A., Wagner, G., Osterloh, I. H., Kirkpatrick, J., & Mishra, A. (1997). The international index of erectile function (IIEF): A multidimensional scale for assessment of erectile dysfunction. <i>Urology</i> , 49(6), 822–830. <a href="https://doi.org/10.1016/S0090-4295(97)00238-0">https://doi.org/10.1016/S0090-4295(97)00238-0</a>
Israeli Sexual Behaviour Inventory	ISBI	Kravetz, S. (1999). The Israeli Sexual Behavior Inventory (ISBI): Scale construction and preliminary validation. <i>Sexuality and Disability</i> , 17(2), 115–128. <a href="https://doi.org/10.1023/A:1021420300693">https://doi.org/10.1023/A:1021420300693</a>
Multidimensional Sexuality Questionnaire	MSQ	Snell, W. E., Fisher, T. D., & Walters, A. S. (1993). The Multidimensional Sexuality Questionnaire: An objective self-report measure of psychological tendencies associated with human sexuality. <i>Annals of Sex Research</i> , 6, 27–55. <a href="https://doi.org/10.1007/BF00849744">https://doi.org/10.1007/BF00849744</a>
Multidimensional Sociosexual Orientation Inventory	M-SOI	Jackson, J. J., & Kirkpatrick, L. A. (2007). The structure and measurement of human mating strategies: Toward a multidimensional model of sociosexuality. <i>Evolution and Human Behavior</i> , 28(6), 382–391. <a href="https://doi.org/10.1016/j.evolhumbehav.2007.04.005">https://doi.org/10.1016/j.evolhumbehav.2007.04.005</a>
Sex Drive Questionnaire	SDQ	Ostovich, J. M., & Sabini, J. (2004). How are sociosexuality, sex drive, and lifetime number of sexual partners related? <i>Personality and Social Psychology Bulletin</i> , 30(10), 1255–1266. <a href="https://doi.org/10.1177/0146167204264754">https://doi.org/10.1177/0146167204264754</a>
Sex Knowledge and Attitudes Test Adolescents	SKAT-A	Lief, H. I., Fullard, W., & Devlin, S. J. (1990). A new measure of adolescent sexuality: SKAT-A. <i>Journal of Sex Education and Therapy</i> , 16(2), 79–91. <a href="https://doi.org/10.1080/01614576.1990.11074980">https://doi.org/10.1080/01614576.1990.11074980</a>
Sexual Behavior Inventory	SBI	Díaz-Loving, R., & Rodríguez, G. G. (2008). Sociosexual orientation and sexual behavior in Mexican adults. <i>Social and Personality Psychology Compass</i> , 2(3), 1199–1217. <a href="https://doi.org/10.1111/j.1751-9004.2008.00111.x">https://doi.org/10.1111/j.1751-9004.2008.00111.x</a>
Sexual Compulsivity Scale	SCS	Kalichman, S. C., & Rompa, D. (1995). Sexual sensation seeking and sexual compulsivity scales: Reliability, validity, and predicting HIV risk behavior. <i>Journal of Personality Assessment</i> , 65(3), 586–601. <a href="https://doi.org/10.1207/s15327752jpa6503_16">https://doi.org/10.1207/s15327752jpa6503_16</a>
Sexual Desire Inventory	SDI	Spector, I. P., Carey, M. P., & Steinberg, L. (1996). The Sexual Desire Inventory: Development, factor structure, and evidence of reliability. <i>Journal of Sex &amp; Marital Therapy</i> , 22(3), 175–190. <a href="https://doi.org/10.1080/00926239608414655">https://doi.org/10.1080/00926239608414655</a>

Inventory	Abbreviation	Reference
Sexuality Scale	SS	Snell, W. E., & Papini, D. R. (1989). The Sexuality Scale: An instrument to measure sexual-esteem, sexual-depression, and sexual-preoccupation. <i>Journal of Sex Research</i> , 26(2), 256–263. <a href="https://doi.org/10.1080/00224498909551510">https://doi.org/10.1080/00224498909551510</a>
Sociosexual Orientation Inventory	SOI	Simpson, J. A., & Gangestad, S. W. (1991). Individual differences in sociosexuality: Evidence for convergent and discriminant validity. <i>Journal of Personality and Social Psychology</i> , 60(6), 870–883. <a href="https://doi.org/10.1037//0022-3514.60.6.870">https://doi.org/10.1037//0022-3514.60.6.870</a>
Sociosexual Orientation Inventory Revised	SOI-R	Penke, L., & Asendorpf, J. B. (2008). Beyond global sociosexual orientations: A more differentiated look at sociosexuality and its effects on courtship and romantic relationships. <i>Journal of Personality and Social Psychology</i> , 95(5), 1113–1135. <a href="https://doi.org/10.1037/0022-3514.95.5.1113">https://doi.org/10.1037/0022-3514.95.5.1113</a>

*Note.* The inventories listed here are not analyzed in full in the present meta-analysis. Rather, we retrieved one or more items that matched our conceptualization of sex drive from these inventories and meta-analysed gender differences on the level of individual items.

Table S4

Tests for Moderation (Indicators of Latent Sex Drive)

Moderator	Cognition Frequency (uncontrolled)					Affect Intensity					Self-Rated Sex Drive				
	AHZ	df	p	f <sup>2</sup>	τ	AHZ	df	p	f <sup>2</sup>	τ	AHZ	df	p	f <sup>2</sup>	τ
<b>Outcome-level Moderators</b>															
Aggregation Span	4.05	7.26	.083	79.81	0.12	7.07	11.37	.022	81.97	0.20					
Item Content	21.00	49.54	< .001	85.78	0.19	15.76	23.05	< .001	88.92	0.15					
Item Context						21.41	14.16	< .001	81.58	0.11					
Type of Response Scale	9.63	6.44	.019	89.93	0.23										
Scale Range	12.90	55.53	< .001	90.00	0.23	1.42	28.19	.243	90.30	0.16	0.64	2.75	N/A	86.70	0.16
Item Wording	12.60	5.56	.007	87.66	0.20	0.04	2.26	N/A	90.82	0.15					
<b>Publication-level Moderators</b>															
Aim to Find Gender Differences in Sex Drive	14.71	11.62	.003	89.28	0.22	0.53	17.25	.478	89.68	0.19					
Focus on Anonymity	1.62	135.29	.205	89.97	0.23	0.38	38.01	.544	89.64	0.16					
Focus on Gender Differences in Sex Drive	6.85	25.63	.015	87.62	0.21	0.07	29.19	.795	89.49	0.16					
Focus on Gender Differences	2.39	108.91	.125	88.86	0.22	0.02	23.01	.890	89.71	0.15					
Gender of First Author	1.10	112.97	.297	89.72	0.23	1.23	16.39	.283	90.57	0.16					
Mean Author Gender	4.20	67.90	.044	89.80	0.23	1.47	15.68	.244	90.68	0.17	2.53	3.25	N/A	79.86	0.14
Publication Status	2.68	36.44	.110	89.11	0.22	0.08	3.45	N/A	90.60	0.16					
Sexuality Journal	9.87	76.50	.002	89.65	0.23	0.32	36.19	.577	90.61	0.18					
<b>Sample-level Moderators</b>															
Mean Age	0.88	33.02	.356	89.79	0.23	2.70	6.36	.148	89.94	0.16	0.00	2.47	N/A	87.67	0.20
Percent White	0.18	11.91	.683	89.40	0.22	0.77	9.06	.402	89.25	0.27	0.86	2.02	N/A	78.16	0.25
Country-Level Gender Development	3.60	40.58	.065	88.88	0.22	3.17	7.05	.118	88.73	0.21	0.45	2.30	N/A	79.23	0.21
Country-Level Gender Inequality	0.52	18.89	.479	88.58	0.22	0.22	35.90	.644	89.13	0.22	1.54	2.05	N/A	75.75	0.21
Percent Heterosexual	1.71	10.29	.220	90.65	0.21	2.54	4.78	.175	92.55	0.18	0.12	1.18	N/A	71.25	0.15
Average Partnership Duration in Weeks	1.19	7.58	.308	75.39	0.18	0.01	1.66	N/A	84.36	0.27					
Percent Parents	0.18	5.17	.685	89.56	0.28	0.73	3.05	N/A	89.99	0.37					
Country-Level Sex Ratio	0.48	21.38	.494	88.72	0.22	0.01	8.64	.937	89.42	0.22	0.09	1.54	N/A	79.59	0.24
Study Restricted to Sexually Active	9.97	19.41	.005	83.79	0.18	0.04	14.61	.841	90.43	0.19					
Percent Single	12.92	31.11	.001	88.86	0.22	0.00	7.74	.974	90.13	0.23	7.47	1.00	N/A	51.57	0.14
Percent University Students	0.72	11.58	.414	85.20	0.22	0.66	4.44	.459	92.16	0.26					
<b>Study-level Moderators</b>															
Anonymity Reassurance	2.32	90.82	.131	90.14	0.23	3.41	11.55	.090	89.62	0.16					
Participant Compensation	1.68	51.20	.182	89.87	0.25	0.90	2.85	N/A	91.36	0.38					
Electronic Data Collection	0.21	57.29	.647	90.55	0.23	0.36	4.13	.720	91.15	0.16					
Group Assessment	6.18	20.36	.008	90.35	0.22	1.47	3.39	N/A	92.21	0.16					
Personal Contact	2.66	10.96	.115	90.18	0.23	0.51	7.84	.619	91.51	0.16					
Sexuality Study	6.57	41.89	.014	86.18	0.19	0.41	5.63	.547	92.22	0.23					
Year of Study	0.18	42.54	.676	89.52	0.23	0.11	13.56	.747	89.07	0.16	0.92	4.18	.388	83.64	0.15

Note. Tests for moderation of the indicators of latent sex drive and cognition frequency (not controlled for item content). The tests indicate the significance of the slope for continuous moderator or differences between subgroups for categorical moderators. Some models could not be fitted because the number of available codings was insufficient. These are left blank. AHZ = Hotelling-T-approximated test statistic. df = small-sample-corrected degrees of freedom. p = p-value associated with the test statistic and df in the same row. f<sup>2</sup> = proportion of the variation in observed effects that is due to variation in true effects. τ = estimated standard deviation of the true effects. Note that if degrees of freedom fall below 4, significance tests are unreliable. p-values for unreliable tests are not reported (N/A).



Table S5

Regression Tables for Moderation Analyses (Indicators of Latent Sex Drive)

Moderator	Cognition Frequency (uncontrolled)							Affect Intensity							Self-Rated Sex Drive							
	<i>g</i>	<i>SE</i>	<i>k</i>	<i>m</i>	<i>t</i>	<i>df</i>	<i>p</i>	<i>g</i>	<i>SE</i>	<i>k</i>	<i>m</i>	<i>t</i>	<i>df</i>	<i>p</i>	<i>g</i>	<i>SE</i>	<i>k</i>	<i>m</i>	<i>t</i>	<i>df</i>	<i>p</i>	
<b>Outcome-level Moderators</b>																						
Aggregation Span																						
Intercept	0.37	0.09	30	46	3.87	4.91	.012	0.18	0.07	16	17	2.72	5.72	.036								
Slope	0.01	0.00			2.01	7.26	.083	0.01	0.00			2.66	11.37	.022								
Item Content																						
Extra pair partner	0.82	0.02	106	168	36.77	98.63	< .001															
No target	0.57	0.04	43	82	14.85	36.36	< .001	0.45	0.03	39	88	13.34	29.95	< .001								
Unspecified partner	0.58	0.05	24	32	11.74	19.20	< .001	0.27	0.03	20	39	8.51	16.49	< .001								
Masturbation								0.49	0.04	19	20	11.94	16.22	< .001								
Own partner								0.27	0.04	13	19	6.81	10.01	< .001								
Item Context																						
First seeing an attractive person								0.67	0.04	19	19	15.21	14.92	< .001								
Not specified								0.43	0.03	47	81	12.80	31.17	< .001								
Romantic situation								0.09	0.04	20	20	2.51	15.43	.024								
While having sexual thoughts								0.23	0.04	20	20	6.25	15.43	< .001								
While spending time with an attractive								0.50	0.04	20	20	13.46	15.43	< .001								
Type of Response Scale																						
No	0.74	0.02	157	268	36.61	149.2	< .001															
Yes	0.43	0.10	8	14	4.23	6.03	.005															
Scale Range																						
Intercept	0.46	0.08	155	265	5.52	41.91	< .001	0.52	0.12	49	165	4.36	20.88	< .001	1.05	0.45	7	7	2.34	2.60	N/A	
Slope	0.04	0.01			3.59	55.53	< .001	-0.02	0.02			-1.19	28.19	.243	-0.07	0.08			-0.80	2.75	N/A	
Item Wording																						
Daydreams	0.39	0.12	6	10	3.11	4.00	.036															
Fantasies	0.81	0.02	120	189	35.24	111.3	< .001															
Other	0.37	0.24	4	5	1.55	1.98	N/A	0.39	0.09	4	5	4.44	2.14	N/A								
Thoughts	0.57	0.03	46	74	20.89	39.17	< .001															
Desire								0.40	0.03	48	156	15.27	40.42	< .001								
<b>Publication-level Moderators</b>																						
Aim to Find Gender Differences in Sex Drive																						
No	0.76	0.02	127	215	33.40	120.9	< .001	0.39	0.03	35	124	12.06	31.34	< .001								
Yes	0.50	0.06	11	28	8.09	9.75	< .001	0.43	0.04	11	33	9.93	9.63	< .001								
Focus on Anonymity																						
No	0.76	0.03	74	136	26.37	70.99	< .001	0.42	0.04	26	80	11.13	22.31	< .001								
Yes	0.71	0.03	70	118	21.27	65.60	< .001	0.39	0.04	20	77	10.84	17.36	< .001								
Focus on Gender Differences in Sex Drive																						
No	0.76	0.02	118	201	32.50	111.7	< .001	0.41	0.04	31	110	11.36	26.63	< .001								
Yes	0.60	0.06	20	42	10.72	18.24	< .001	0.40	0.04	15	47	10.84	13.14	< .001								
Focus on Gender Differences																						
No	0.78	0.04	55	93	21.07	51.79	< .001	0.40	0.06	15	45	7.31	12.36	< .001								
Yes	0.71	0.03	83	150	25.92	78.77	< .001	0.41	0.03	31	112	13.77	26.90	< .001								
Gender of First Author																						
Female	0.71	0.03	57	102	22.08	54.00	< .001	0.41	0.03	39	122	12.53	33.86	< .001								
Male	0.75	0.03	100	173	27.91	95.14	< .001	0.36	0.03	11	44	11.51	9.13	< .001								
Mean Author Gender																						
Intercept	0.66	0.04	157	275	18.33	51.38	< .001	0.43	0.04	50	166	11.00	25.32	< .001	0.82	0.16	7	7	5.18	2.21	N/A	
Slope	0.11	0.06			2.05	67.90	.044	-0.09	0.07			-1.21	15.68	.244	-0.27	0.17			-1.59	3.25	N/A	
Publication Status																						
Published	0.75	0.02	133	222	33.24	126.9	< .001	0.40	0.03	45	151	15.32	39.56	< .001								
Unpublished	0.66	0.05	28	60	14.57	25.58	< .001	0.37	0.12	5	15	2.94	2.95	N/A								
Sexuality Journal																						
No	0.77	0.02	117	196	30.82	111.5	< .001	0.38	0.03	32	84	11.02	27.55	< .001								
Yes	0.64	0.03	44	86	20.63	41.50	< .001	0.41	0.04	18	82	10.25	16.19	< .001								
<b>Sample-level Moderators</b>																						
Mean Age																						
Intercept	0.83	0.09	137	234	9.22	48.29	< .001	0.24	0.10	48	164	2.35	10.10	.041	0.64	0.50	6	6	1.30	1.80	N/A	
Slope	-0.00	0.00			-0.94	33.02	.356	0.01	0.00			1.64	6.36	.148	-0.00	0.02			-0.03	2.47	N/A	
Percent White																						
Intercept	0.84	0.13	47	79	6.43	9.92	< .001	0.29	0.11	27	98	2.56	6.26	.041	0.51	0.18	6	6	2.77	1.76	N/A	
Slope	-0.00	0.00			-0.42	11.91	.683	0.00	0.00			0.88	9.06	.402	0.00	0.00			0.92	2.02	N/A	
Country-Level Gender Development																						
Intercept	-2.39	1.66	143	253	-1.44	39.93	.157	3.52	1.76	48	159	2.00	7.01	.085	36.92	53.99	6	6	0.68	2.31	N/A	
Slope	3.19	1.68			1.90	40.58	.065	-3.16	1.78			-1.78	7.05	.118	-36.52	54.52			-0.67	2.30	N/A	
Country-Level Gender Inequality																						
Intercept	0.70	0.07	146	256	10.48	32.25	< .001	0.43	0.08	48	159	5.05	27.57	< .001	1.04	0.25	6	6	4.15	1.24	N/A	
Slope	0.25	0.35			0.72	18.89	.479	-0.19	0.41			-0.47	35.90	.644	-1.60	1.29			-1.24	2.05	N/A	
Percent Heterosexual																						

Moderator	Cognition Frequency (uncontrolled)							Affect Intensity							Self-Rated Sex Drive						
	<i>g</i>	<i>SE</i>	<i>k</i>	<i>m</i>	<i>t</i>	<i>df</i>	<i>p</i>	<i>g</i>	<i>SE</i>	<i>k</i>	<i>m</i>	<i>t</i>	<i>df</i>	<i>p</i>	<i>g</i>	<i>SE</i>	<i>k</i>	<i>m</i>	<i>t</i>	<i>df</i>	<i>p</i>
Intercept	0.48	0.17	81	150	2.86	9.18	.018	0.19	0.13	35	129	1.44	4.21	.219	0.83	0.27	4	4	3.09	1.00	N/A
Slope	0.00	0.00			1.31	10.29	.220	0.00	0.00			1.59	4.78	.175	-0.00	0.00			-0.35	1.18	N/A
Average Partnership Duration in Weeks																					
Intercept	0.76	0.06	31	46	12.50	19.13	<.001	0.35	0.06	23	57	6.09	18.30	<.001							
Slope	-0.00	0.00			-1.09	7.58	.308	0.00	0.00			0.08	1.66	N/A							
Percent Parents																					
Intercept	0.80	0.12	21	37	6.49	13.45	<.001	0.30	0.26	7	34	1.17	3.14	N/A							
Slope	-0.00	0.00			-0.43	5.17	.685	0.00	0.00			0.85	3.05	N/A							
Country-Level Sex Ratio																					
Intercept	0.10	0.92	146	256	0.11	21.06	.915	0.55	1.89	48	159	0.29	8.52	.779	2.36	5.66	6	6	0.42	1.49	N/A
Slope	0.01	0.01			0.70	21.38	.494	-0.00	0.02			-0.08	8.64	.937	-0.02	0.06			-0.29	1.54	N/A
Study Restricted to Sexually Active																					
No	0.78	0.02	97	172	32.40	89.47	<.001	0.40	0.04	29	94	11.41	25.83	<.001							
Yes	0.59	0.06	16	29	10.23	14.16	<.001	0.39	0.06	10	36	6.59	8.25	<.001							
Percent Single																					
Intercept	0.61	0.04	88	158	17.24	40.77	<.001	0.36	0.04	39	130	9.43	27.16	<.001	0.83	0.13	3	3	6.16	1.00	N/A
Slope	0.00	0.00			3.59	31.11	.001	0.00	0.00			0.03	7.74	.974	-0.01	0.00			-2.73	1.00	N/A
Percent University Students																					
Intercept	0.70	0.10	80	132	7.05	8.84	<.001	0.30	0.09	20	75	3.49	2.87	N/A							
Slope	0.00	0.00			0.85	11.58	.414	0.00	0.00			0.81	4.44	.459							
<b>Study-level Moderators</b>																					
Anonymity Reassurance																					
No	0.76	0.03	92	164	28.82	88.65	<.001	0.42	0.03	36	116	12.73	31.20	<.001							
Yes	0.69	0.04	49	86	17.20	45.39	<.001	0.35	0.03	9	35	13.47	7.19	<.001							
Participant Compensation																					
Coursecredit	0.80	0.05	29	53	15.68	27.46	<.001	0.43	0.04	4	14	11.20	2.91	N/A							
Material	0.66	0.05	35	64	13.13	32.95	<.001	0.35	0.05	18	58	7.20	16.83	<.001							
Mixed	0.77	0.04	18	29	19.79	16.61	<.001	0.30	0.06	7	29	5.51	5.93	.002							
None	0.79	0.06	18	35	12.92	16.36	<.001	0.46	0.13	2	18	3.44	1.00	N/A							
Electronic Data Collection																					
No	0.71	0.05	36	53	14.06	33.40	<.001	0.43	0.05	10	39	8.37	8.24	<.001							
Yes	0.73	0.03	96	181	29.21	91.98	<.001	0.39	0.03	33	103	12.86	28.82	<.001							
Mixed								0.49	0.14	3	15	3.61	1.69	N/A							
Group Assessment																					
Mixed	0.52	0.06	10	13	8.78	8.66	<.001	0.35	0.03	4	26	11.70	2.65	N/A							
No	0.74	0.02	103	180	30.09	97.61	<.001	0.40	0.03	36	112	12.85	31.30	<.001							
Yes	0.79	0.07	19	30	11.39	17.45	<.001														
Personal Contact																					
Mixed	0.62	0.05	5	8	13.17	3.84	N/A	0.39	0.04	4	26	8.97	2.69	N/A							
No	0.73	0.03	84	151	27.45	80.36	<.001	0.43	0.04	24	83	11.96	21.29	<.001							
Yes	0.75	0.04	57	97	19.63	53.81	<.001	0.37	0.04	18	53	8.42	15.09	<.001							
Sexuality Study																					
No	0.80	0.04	25	42	18.31	21.98	<.001	0.35	0.10	5	25	3.52	3.94	N/A							
Yes	0.65	0.03	44	87	18.69	40.93	<.001	0.42	0.04	25	68	10.69	22.82	<.001							
Year of Study																					
Intercept	-4.29	11.95	152	266	-0.36	42.48	.722	5.29	14.82	49	160	0.36	13.53	.727	-39.65	41.85	7	7	-0.95	4.18	.395
Slope	0.00	0.01			0.42	42.54	.676	-0.00	0.01			-0.33	13.56	.747	0.02	0.02			0.96	4.18	.388

Note. Meta-regression tables for moderation of the indicators of latent sex drive and cognition frequency (not controlled for item content). For categorical moderators, point estimates for subgroups and corresponding significance tests are presented. For continuous moderators, values are presented for the intercept and slope. Some models could not be fitted because the number of available codings was insufficient. These are left blank. *g* = Hedges' *g* effect size (positive favors males). *SE* = Standard Error for Hedges' *g* effect size. *k* = number of studies per subgroup. *m* = number of effect sizes per subgroup. *t*-value from *t*-test testing the parameter against zero. *df* = small-sample-corrected degrees of freedom. *p* = *p*-value associated with the *t*-value and *df* in the same row. Note that if degrees of freedom fall below 4, significance tests are unreliable. *p*-values for unreliable tests are not reported (N/A).

Table S6

Moderator Overview for Secondary Indicators

Moderator	Total		Affect Intensity			Self-Rated Sex Drive		
	<i>m</i>	Compl.	<i>m</i>	Compl.	Distribution	<i>m</i>	Compl.	Distribution
<b>Outcome-level Moderators</b>								
Item Content	524	61%	166	100%	masturbation ( <i>m</i> = 20), no target ( <i>m</i> = 88), own partner ( <i>m</i> = 19), unspecified partner ( <i>m</i> = 39)	0	0%	NA ( <i>m</i> = 7)
Item Context	605	71%	166	100%	being attracted to someone wrong to pursue ( <i>m</i> = 3), during sex ( <i>m</i> = 1), first seeing an attractive person ( <i>m</i> = 19), not specified ( <i>m</i> = 81), prior to sex ( <i>m</i> = 1), romantic situation ( <i>m</i> = 20), seeing an attractive person ( <i>m</i> = 1), while having sexual thoughts ( <i>m</i> = 20), while spending time with an attractive person ( <i>m</i> = 20)	0	0%	NA ( <i>m</i> = 7)
Type of Response Scale	856	100%	166	100%	no ( <i>m</i> = 166)	7	100%	no ( <i>m</i> = 7)
Scale Range	691	81%	165	99%	Q = [4.00, 7.00, 8.00, 8.00, 9.00], M = 7.19, SD = 1.50	7	100%	Q = [4.00, 5.00, 6.00, 7.00, 8.00], M = 6.00, SD = 1.63
Item Wording	605	71%	166	100%	appetite ( <i>m</i> = 2), desire ( <i>m</i> = 156), libido ( <i>m</i> = 1), motivation ( <i>m</i> = 2), other ( <i>m</i> = 5)	0	0%	NA ( <i>m</i> = 7)
Aggregation Span	185	22%	17	10%	Q = [1.00, 1.00, 28.00, 28.00, 28.00], M = 15.65, SD = 13.58	0	0%	Q = [NA, NA, NA, NA, NA], M = NaN, SD = NA
<b>Publication-level Moderators</b>								
Aim to Find Gender Differences in Sex Drive	745	87%	157	95%	no ( <i>m</i> = 124), yes ( <i>m</i> = 33), NA ( <i>m</i> = 9)	6	86%	no ( <i>m</i> = 1), yes ( <i>m</i> = 5), NA ( <i>m</i> = 1)
Focus on Gender Differences in Sex Drive	746	87%	157	95%	no ( <i>m</i> = 110), yes ( <i>m</i> = 47), NA ( <i>m</i> = 9)	7	100%	no ( <i>m</i> = 1), yes ( <i>m</i> = 6)
Focus on Gender Differences	746	87%	157	95%	no ( <i>m</i> = 45), yes ( <i>m</i> = 112), NA ( <i>m</i> = 9)	7	100%	yes ( <i>m</i> = 7)
Gender of First Author	843	98%	166	100%	female ( <i>m</i> = 122), male ( <i>m</i> = 44)	7	100%	female ( <i>m</i> = 4), male ( <i>m</i> = 3)
Publication Status	856	100%	166	100%	published ( <i>m</i> = 151), unpublished ( <i>m</i> = 15)	7	100%	published ( <i>m</i> = 6), unpublished ( <i>m</i> = 1)
Sexuality Journal	856	100%	166	100%	No ( <i>m</i> = 84), Yes ( <i>m</i> = 82)	7	100%	No ( <i>m</i> = 5), Yes ( <i>m</i> = 2)
Focus on Anonymity	776	91%	157	95%	no ( <i>m</i> = 80), yes ( <i>m</i> = 77), NA ( <i>m</i> = 9)	7	100%	no ( <i>m</i> = 4), yes ( <i>m</i> = 3)
Mean Author Gender	843	98%	166	100%	Q = [0.00, 0.00, 0.25, 0.50, 1.00], M = 0.35, SD = 0.32	7	100%	Q = [0.00, 0.38, 0.50, 1.00, 1.00], M = 0.61, SD = 0.40
<b>Sample-level Moderators</b>								
Mean Age	735	86%	165	99%	Q = [18.04, 24.21, 27.57, 31.70, 74.59], M = 28.51, SD = 7.22	6	86%	Q = [18.60, 24.44, 32.02, 35.90, 39.56], M = 30.22, SD = 8.29
Percent White	302	35%	98	59%	Q = [0.00, 65.83, 84.75, 91.00, 98.00], M = 74.00, SD = 24.18	6	86%	Q = [42.00, 63.90, 73.30, 88.25, 98.00], M = 73.43, SD = 20.47
Country-Level Gender Inequality	790	92%	159	96%	Q = [0.08, 0.12, 0.22, 0.25, 0.31], M = 0.18, SD = 0.07	6	86%	Q = [0.11, 0.16, 0.25, 0.26, 0.26], M = 0.21, SD = 0.07
Country-Level Gender Development	790	92%	159	96%	Q = [0.96, 0.99, 0.99, 1.00, 1.03], M = 0.99, SD = 0.01	6	86%	Q = [0.99, 0.99, 0.99, 0.99, 1.00], M = 0.99, SD = 0.00
Percent Heterosexual	487	57%	129	78%	Q = [0.00, 76.30, 90.00, 100.00, 100.00], M = 84.04, SD = 18.53	4	57%	Q = [0.00, 72.20, 98.14, 100.00, 100.00], M = 74.07, SD = 49.41
Percent Single	517	60%	130	78%	Q = [0.00, 0.00, 0.00, 35.10, 95.00], M = 16.28, SD = 22.42	3	43%	Q = [0.00, 0.00, 0.00, 14.50, 29.00], M = 9.67, SD = 16.74
Percent University Students	398	46%	75	45%	Q = [0.00, 54.80, 91.53, 100.00, 100.00], M = 75.17, SD = 30.67	2	29%	Q = [100.00, 100.00, 100.00, 100.00, 100.00], M = 100.00, SD = 0.00
Average Partnership Duration in Weeks	179	21%	57	34%	Q = [1.80, 14.00, 53.40, 109.20, 554.40], M = 74.15, SD = 87.93	2	29%	Q = [109.20, 121.50, 133.80, 146.10, 158.40], M = 133.80, SD = 34.79

Moderator	Total		Affect Intensity			Self-Rated Sex Drive		
	<i>m</i>	Compl.	<i>m</i>	Compl.	Distribution	<i>m</i>	Compl.	Distribution
Percent Parents	134	16%	34	20%	Q = [17.30, 30.75, 64.00, 100.00, 100.00], M = 64.32, SD = 32.15	1	14%	Q = [64.00, 64.00, 64.00, 64.00, 64.00], M = 64.00, SD = NA
Study Restricted to Sexually Active	629	73%	130	78%	no (m = 94), yes (m = 36), NA (m = 36)	6	86%	no (m = 6), NA (m = 1)
Country-Level Sex Ratio	790	92%	159	96%	Q = [93.92, 100.88, 101.03, 101.23, 104.69], M = 100.90, SD = 2.19	6	86%	Q = [95.65, 100.10, 100.88, 100.99, 101.03], M = 99.88, SD = 2.12
<b>Study-level Moderators</b>								
Anonymity Reassurance	761	89%	151	91%	no (m = 116), yes (m = 35), NA (m = 15)	7	100%	no (m = 7)
Participant Compensation	545	64%	119	72%	coursecredit (m = 14), material (m = 58), mixed (m = 29), none (m = 18), NA (m = 47)	5	71%	coursecredit (m = 1), material (m = 2), none (m = 2), NA (m = 2)
Sexuality Study	428	50%	93	56%	no (m = 25), yes (m = 68), NA (m = 73)	2	29%	no (m = 1), yes (m = 1), NA (m = 5)
Year of Study	813	95%	160	96%	Q = [1998.00, 2008.00, 2011.00, 2013.00, 2016.00], M = 2010.56, SD = 3.10	7	100%	Q = [2004.00, 2004.50, 2011.00, 2011.50, 2014.00], M = 2008.71, SD = 4.23
Face-to-Face Interview	807	94%	163	98%	no (m = 161), yes (m = 2), NA (m = 3)	7	100%	no (m = 7)
Electronic Data Collection	745	87%	157	95%	mixed (m = 15), no (m = 39), yes (m = 103), NA (m = 9)	5	71%	no (m = 1), yes (m = 4), NA (m = 2)
Group Assessment	695	81%	141	85%	mixed (m = 26), no (m = 112), yes (m = 3), NA (m = 25)	4	57%	no (m = 4), NA (m = 3)
Personal Contact	788	92%	162	98%	mixed (m = 26), no (m = 83), yes (m = 53), NA (m = 4)	5	71%	no (m = 3), yes (m = 2), NA (m = 2)

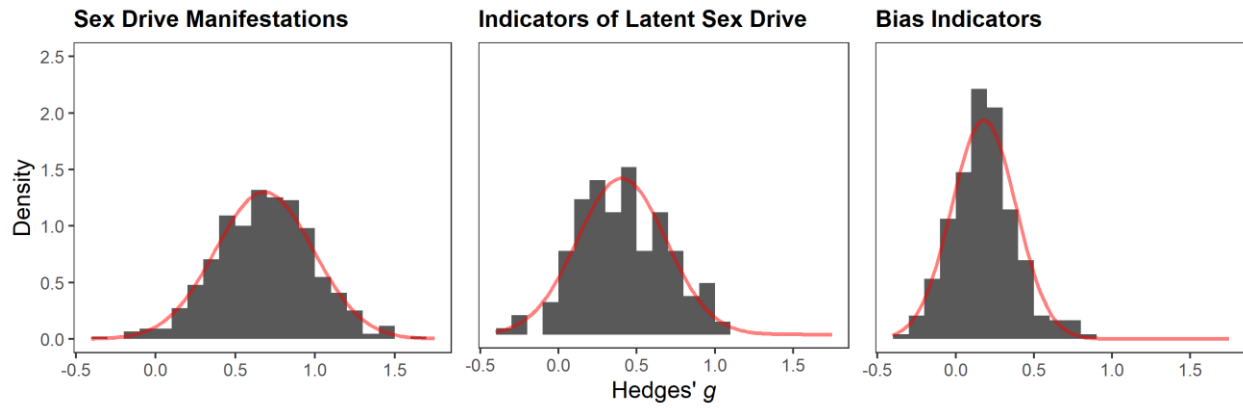
*Note.* *m*: Absolute number of effect sizes for which the corresponding characteristic could be coded. Compl.: Percentage of effect sizes for which the corresponding characteristic could be coded. Distribution: Information about the distribution of the coded characteristics. For categorical characteristics, the number of effect sizes per subgroup is reported. For continuous characteristics, Q are quartiles (minimum, 25% quartile, median, 75% quartile, maximum), M is the mean, and SD is the standard deviation. Note that summaries for continuous moderators are computed on the effect size level for this table. In the results section, some of this information was presented on the level of individual participants (i.e. as summaries weighted by sample size). Some values may therefore differ.

Table S7

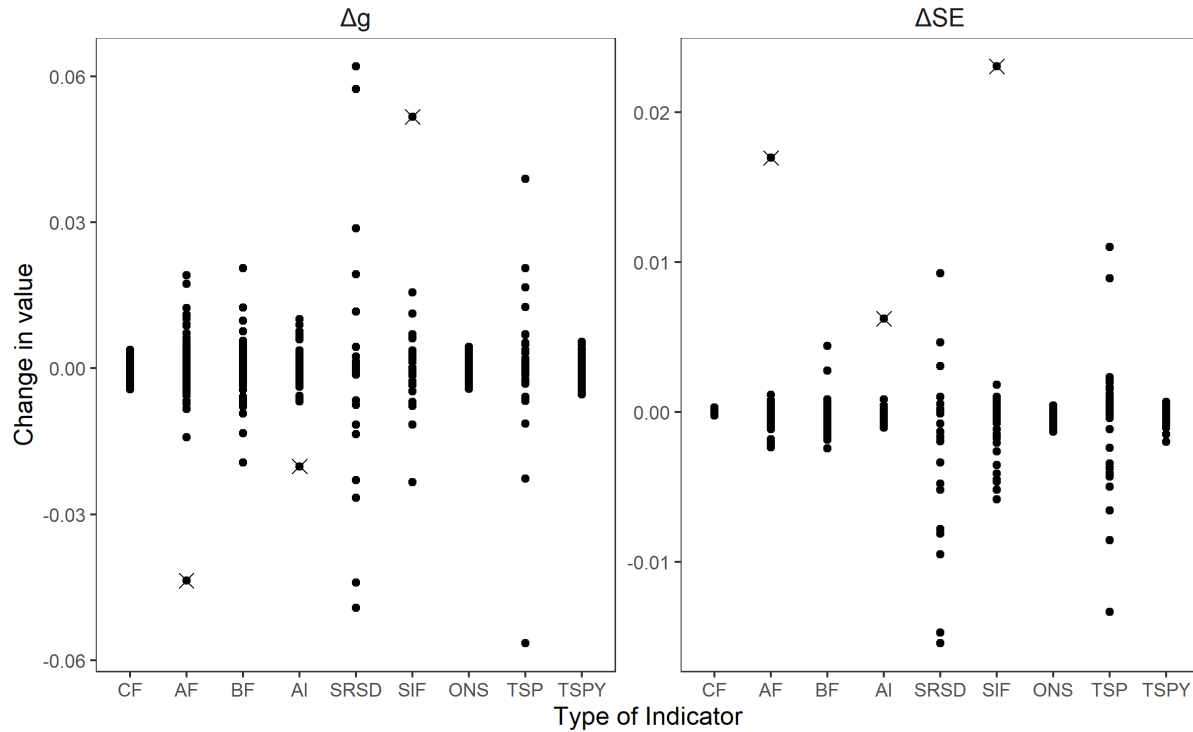
## Interrater Reliability

Moderator	Reliability	Statistic	No. of Categories	Type of Moderator	NA Match	NA Mismatch
Gender of First Author	0.86	Cohen's $\kappa$	3			
Focus on Gender Differences	0.79	Cohen's $\kappa$	3			
Focus on Gender Differences in Sex Drive	0.79	Cohen's $\kappa$	3			
Aim to Find Gender Differences in Sex Drive	0.57	Cohen's $\kappa$	3			
Face-to-Face Interview	0.93	Cohen's $\kappa$	3			
Personal Contact	1.00	Cohen's $\kappa$	3			
Group Assessment	0.93	Cohen's $\kappa$	3			
Electronic Data Collection	1.00	Cohen's $\kappa$	3			
Focus on Anonymity	1.00	Cohen's $\kappa$	3			
Anonymity Reassurance	1.00	Cohen's $\kappa$	3			
Participant Compensation	0.94	Cohen's $\kappa$	4			
Sexuality Study	0.79	Cohen's $\kappa$	3			
Study Restricted to Sexually Active	0.71	Cohen's $\kappa$	3			
Journal (open)	1.00	Percent Agreement		Unknown no. of categories		
Nation (open)	0.90	Percent Agreement		Unknown no. of categories		
Mean Author Gender	0.94	Pearson Correlation		Numeric moderator	0	0
Percent University Students	0.71	Pearson Correlation		Numeric moderator	4	9
Percent Single	0.94	Pearson Correlation		Numeric moderator	1	10
Average Partnership Duration in Weeks	1.00	Pearson Correlation		Numeric moderator	1	17
Percent Parents		Pearson Correlation		Numeric moderator	4	16
Percent White	1.00	Pearson Correlation		Numeric moderator	0	14
Percent Heterosexual	1.00	Pearson Correlation		Numeric moderator	0	7
Mean Age	1.00	Pearson Correlation		Numeric moderator	4	11
Year the Study was Published	0.99	Pearson Correlation		Numeric moderator	0	0
Year the Study was Conducted	1.00	Pearson Correlation		Numeric moderator	2	16
Year the Study was Submitted	1.00	Pearson Correlation		Numeric moderator	1	11

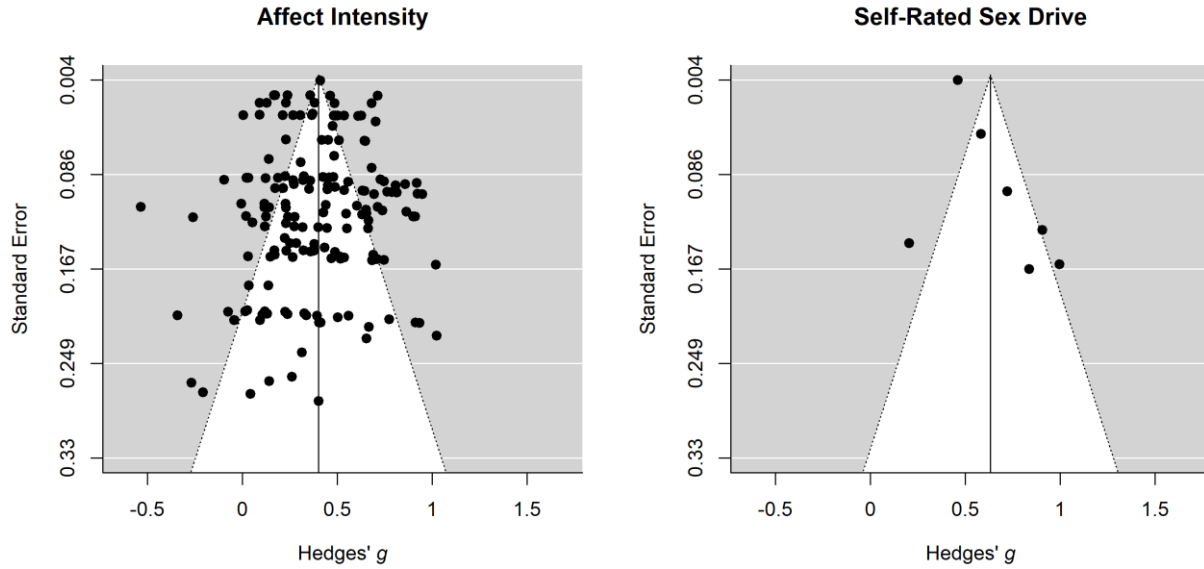
*Note.* Interrater reliability for study and publication level moderators. The results in this table were derived from 21 studies that were coded by two coders. Some moderators reported in the main manuscript were derived from other raw codings and hence do not appear in this table. The codings 'Publication Status' and 'Sexuality Journal' were derived from 'Journal (open)'. The codings 'Country-Level Gender Inequality', 'Country-Level Gender Development', and 'Country-Level Sex Ratio' were derived from 'Nation'. The coding 'Year' was derived from 'Year the study was conducted', 'Year the study was published', and 'Year the study was submitted'. We report different reliability indicators for different moderators. For categorical codings, we report Cohen's  $\kappa$  along with the number of possible categories. For categorical codings with an unknown number of categories, we report percent agreement. For numerical codings, we report Pearson's correlation coefficient along with the number of cases where coders agreed that the information was missing (NA Match) and the number of cases where only one coder found information (NA Mismatch). For categorical codings, 'information missing' was treated as a normal category.

**Figure S1***Distribution of Effect Sizes*

*Note.* The red curve denotes the fitted normal density curve for the unweighted effect sizes.

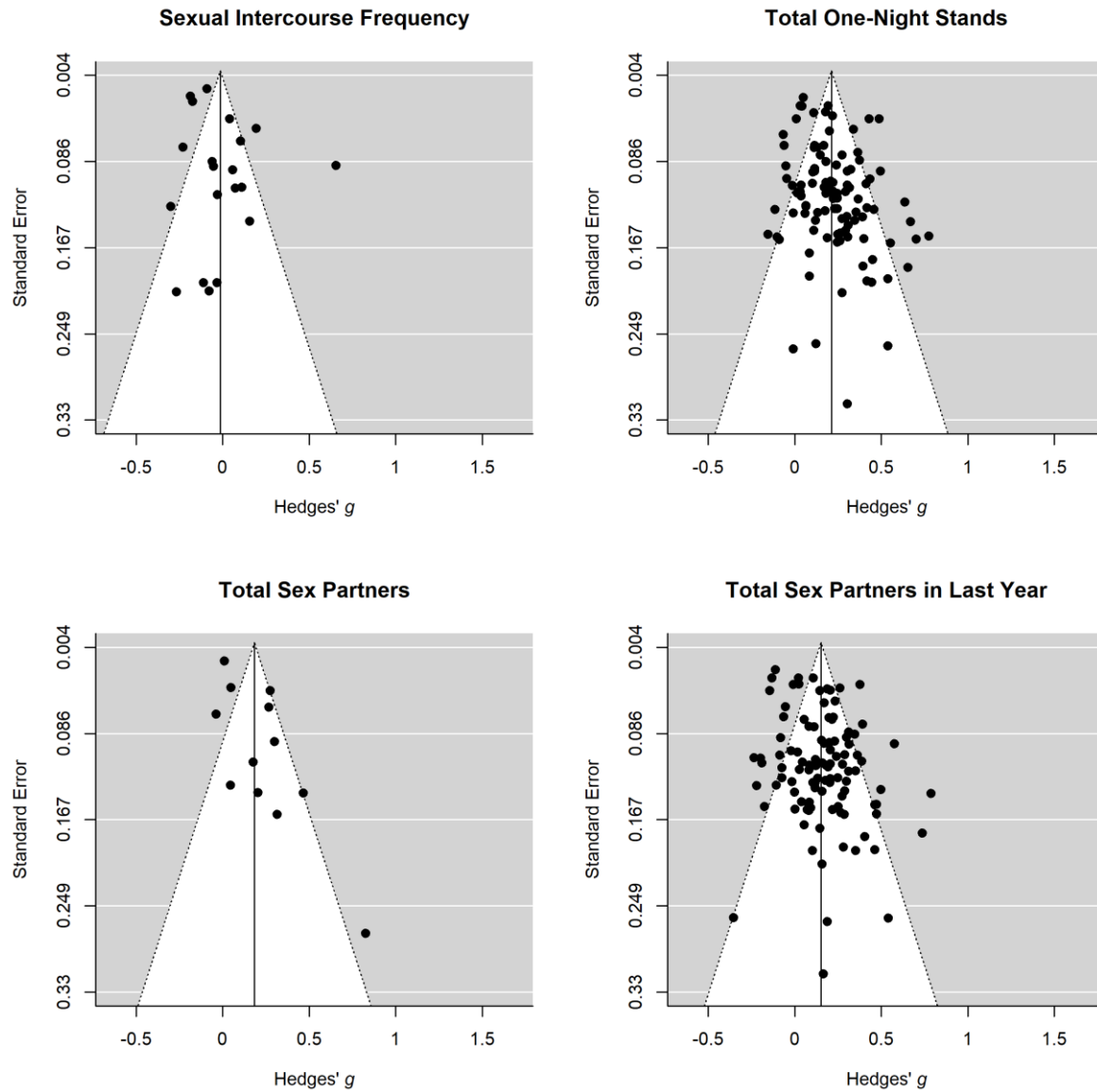
**Figure S2***Results for Leave-One-Out Analyses*

*Note.* This figure illustrates results from a leave-one-out analysis to detect outliers for the sex drive manifestations and latent sex drive indicators and the bias indicators. The left pane depicts changes in Hedges'  $g$  summary effects when effect sizes are removed iteratively. The right pane depicts changes in the standard error of Hedges'  $g$ . CF = Cognition Frequency; AF = Affect Frequency; BF = Behavior Frequency; AI = Affect Intensity; SRSD = Self-Rated Sex Drive; TSP = Total Sexual Partners; TSPY = Total Sexual Partners in Last Year; ONS = Total One-night Stands; SIF = Sexual Intercourse Frequency. Crossed out effect sizes were removed from all subsequent analyses.

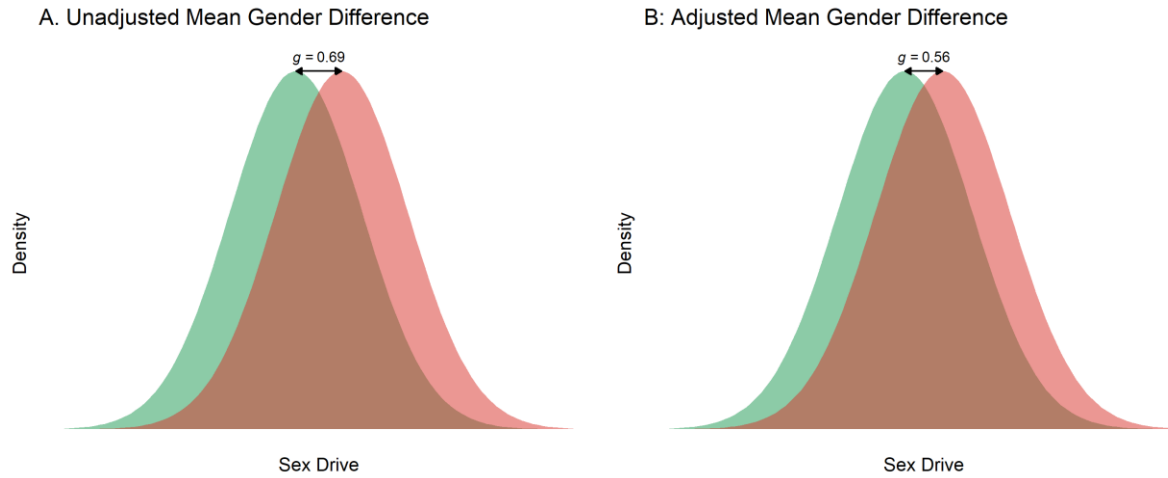
**Figure S3***Funnel Plots for Indicators of Latent Sex Drive*

*Note.* The solid vertical lines represent the within-subgroup summary effects. The dotted lines denote the area in which 95% of effect sizes are expected to fall in the absence of heterogeneity.  $x$ -axis: Hedges'  $g$  effect sizes, positive values indicate larger values in men.  $y$ -axis: Standard error of effect sizes. Leave-one-out analyses identified one outlier in the left plot at  $g < -0.5$  and Standard Error  $< 0.167$ . This effect was removed for all other analyses. Summary effects displayed in the figure were computed after removing the outlier.



**Figure S4***Funnel Plots for Bias Indicators*

*Note.* The solid vertical lines represent the within-subgroup summary effects. The dotted lines denote the area in which 95% of effect sizes are expected to fall in the absence of heterogeneity. *x*-axis: Hedges' *g* effect sizes, positive values indicate larger values in men. *y*-axis: Standard error of effect sizes. Leave-one-out analyses identified one outlier in the upper left plot at  $g > 0.5$  and Standard Error  $> 0.249$ . This effect was removed for all other analyses. Summary effects displayed in the figure were computed after removing the outlier.

**Figure S5***Distribution Overlap for Adjusted and Unadjusted Gender Difference in Sex Drive*

*Note.* This figure displays the overlap of the density distributions for female (green) and male sex drive (red) under normality assumptions. The gender difference displayed in the left panel (A) is  $g = 0.69$ , which is the unadjusted global summary effect for sex drive manifestation indicators (i.e., averaged across cognition, affect, and behavior frequency). The gender difference displayed in the right panel (B) is  $g = 0.56$ , which is the global summary effect for sex drive manifestation indicators ( $g = 0.69$ ) adjusted for response bias tendencies, that is, the global summary effect for bias indicators (indicating potentially biased responding;  $g = 0.13$ ).

## General Discussion

Part I of the dissertation project highlighted challenges to meta-analytic research in social psychology and personality research. Paper 1 attempted to find summary evidence for the effectiveness of self-control training, but the summary effect was small, likely inflated by publication bias, and could not be attributed beyond doubt to a theoretical mechanism. Paper 2 reported on a simulation study that showed how multiple sources of bias (publication bias, p-hacking) can interact with contextual factors and each other to create significant meta-analytic evidence from very small or even zero true effects. Together, the findings of these two papers generated the guiding question of this dissertation project: How can meta-scientific work advance social-psychological and personality theory despite an unknowable risk of bias in the literature? Part II of the dissertation project is an attempt at an answer. Both papers of part II make use of one key idea: Re-using existing raw research data to test novel theoretical ideas in a meta-analysis. As it turns out, this idea helps towards both goals of the dissertation project, that is, building theoretical coherence and reducing risk-of-bias. In the subsequent discussion, I will focus on how and why such secondary data analyses can reduce bias and increase theoretical coherence, but first, I will discuss the idea in more detail.

### **Key Idea: Secondary Data Analyses to Test Novel Research Questions**

Many research studies in psychology do not only measure the primary variables relevant to the hypotheses, but also additional, secondary variables. In paper 1 of part II, experiments on nostalgia inductions also routinely included measurements of the Big Five personality traits. In paper 2 of part II, studies recorded participant gender without exception. My coauthors and I correlated these secondary measurements with other measurements that were more focal. More specifically, we correlated trait neuroticism with the effect of the nostalgia induction in paper 1

of part II, and we correlated gender with sex drive in paper 2 of part II. These two secondary measures, gender and personality, are probably prime examples for general variables that are routinely measured in psychological studies. It will always be informative to know how they relate to other focal variables in a research field. This is, of course, why they are routinely measured. With a meta-analytic perspective, we can examine these correlations across multiple studies, not just within one primary study. Consider this: Since gender and personality are routinely measured, there is sufficient data out there to examine how they relate to almost every other psychological construct. If we imagine a grand correlation table of key psychological constructs, be they personality traits, experimental effects, longitudinal trajectories, or even life outcomes, we will soon realize that there are many blank spots. Some of these spots are blank because there simply is no data. However, I argue that for many correlations, there is sufficient data in the world to get good estimates. It seems just a matter of going back, retrieving the raw data from the original authors, and aggregating them.

### **The Role of Theory in Secondary Data Analyses**

In the way I have described my perspective on secondary data analyses in the previous paragraph, it may seem purely as an exercise in data acquisition and organization: Construct a correlation table of psychological variables and find existing data to fill the spots. I think that this is a valid perspective. In my view, there is value in the work of just organizing and structuring the data, even when leaving theoretical work for later. However, doing this work does become more interesting and stimulating if theory is involved. In the two papers of the dissertation project that made use of secondary data analysis, theory played different roles. In paper 1 of part II, the meta-analysis of the nostalgia-neuroticism interaction, our hypothesis was derived from theory. In essence, we wanted to know if nostalgia is good for everyone. Given that nostalgia,

due to its bittersweetness, also entails negative emotions, people high in neuroticism may be more susceptible to the “bitterness”, and hence benefit less. That made neuroticism a plausible candidate to moderate the nostalgia main effect. In paper 2 of part II, we correlated sex drive with gender. Sex drive is a widely used concept, but there was no consensus in the literature on its definition or even a theoretical basis for a definition. Hence, we needed to do some theoretical work ourselves in order to make the secondary data analysis possible in the first place. This theoretical work along with secondary data analysis proved a perfect fit, because we could select precisely the kinds of data that matched our theoretical definition. As a result, we were able to do a meta-analysis (with all typical advantages resulting from the large data base, including precise estimates, high generalizability, and consideration of contextual factors), while also achieving high theoretical coherence. As pointed out in the introduction, this typically a challenge for meta-analytic work.

### **How Secondary Data Analyses Can Reduce Risk of Bias**

In the previous paragraph, I pointed out how secondary data analyses can increase theoretical coherence, addressing the first part of the guiding question of the dissertation project. The second part concerns the reduction of risk-of-bias in meta-scientific work. My colleagues and I found that secondary data analyses based on raw data can indeed reduce risk-of-bias. I will first discuss the most pressing concern: publication bias. Typically, there is “publication pressure” on quite specific parts of a statistical analysis. More often than not, this is the  $p$ -value for the null hypothesis significance test (NHST) testing the main hypothesis. With secondary data analysis as I define it, the meta-analyst aggregates correlations that were usually not focal in the original study. For example, in paper 2 of part II most studies were not concerned with sex drive directly, but rather relationship dynamics in romantic dyads, the role of testosterone in

sexuality, evolutionary strategies, or something else entirely. Thus, whether or not gender and sex drive were correlated likely had no bearing on whether or not the study would be published. Thus, with regard to our goal, there was no reason to expect any direct publication bias at all. To be sure, there was publication bias on other values in the manuscripts. I do not mean to argue that the data base was entirely unbiased and representative. Yet, I would expect these distorting influences that are unrelated to the focal question to cancel out with regard to the correlation we were interested in. This was quite a revelation, to have a large data base and no pressing reason to question its validity. In many ways, the biasing processes for publication bias and *p*-hacking, the second important class of bias we considered in paper 2 of part I, are quite similar. Publication bias is goal directed selection of studies and manuscript, while *p*-hacking often involves goal directed selection of data points, variables, data preprocessing steps, or analytical approaches. Consequently, the same advantages of the secondary data analysis approach apply. Again, for the example of paper 2 of part II, the original researchers were not interested in the correlation between sex drive and gender, and thus likely did not (consciously or unconsciously) manipulate their analyses to minimize or maximize this correlation. Beyond publication bias and *p*-hacking, there were also some other bias-reducing advantages of the secondary data analysis approach. For one, having access to the raw data makes it possible to run meta-analytic psychometric analyses. In paper 1 of part II, we examined the data for range restriction and unreliability. In paper 2 of part II, we were able to conduct a meta-analytic test of convergent validity for the measures we included.

In summary, we found that secondary data analysis can yield meta-analytic analyses that have high theoretical coherence and low risk of bias. In the next section of the discussion, I

would like to broaden the scope and attempt to situate the approach of secondary data analysis as described above in the grander scheme of psychological research methodology.

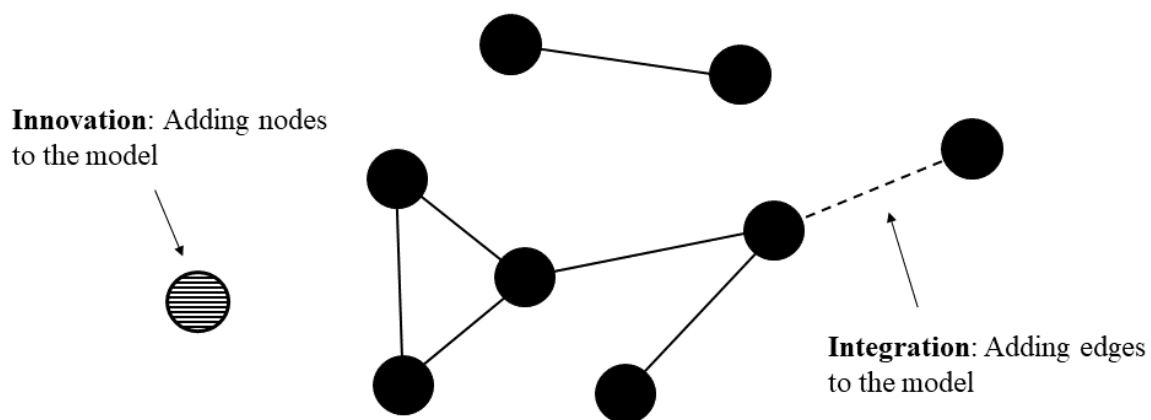
### **Integration versus Innovation in Psychological Research**

When I was first introduced to psychological research, my impression was that the field proceeds in a somewhat linear, cumulative fashion. There are solid foundations, and new research constitutes more and more intricate additions that extend and ornament what is already there. Today, I sometimes feel like everyone is just tossing new rocks on a huge pile of rubble. The new stuff buries the old stuff and walking up the pile is unsafe and strenuous, leaving one wishing for stairs. Of course, this is an outrageous oversimplification and overgeneralization, and does not do justice to the many subfields that do solid, cumulative work. But I do think that the underlying question of how the field prioritizes innovation versus integration is valid and important. To quickly define the terms, I would draw on a view of psychological science in terms of a graph model (Koller & Friedman, 2009), where psychological variables and theories are nodes and connections between theories and variables are edges (see Figure 1). I would consider innovation to mean adding nodes to the model. Conversely, integration is A) adding edges, that is, (empirical or theoretical) connections between the nodes to the model, or B) reshaping nodes so that connections can even be made in the first place. This is what my colleagues and I aimed for in part II of the dissertation project. In paper 2 of part II, we connected gender and sex drive. In order to do that, we had to develop a new theory of sex drive because there was no agreement in the literature (thus reshaping the ‘sex drive’ node; we obviously did not invent the concept). The theory added further connections, tying sex drive to personality theory (Fleeson & Jayawickreme, 2015; McCrae & Costa, 2003) and foundational psychological concepts like cognition, emotion, and behavior. In the competitive space that is the academic job market,

innovation is highly prized and is essentially a requirement for career advancement. Researchers need their own theories, and as the saying goes, theories are like toothbrushes—no self-respecting person wants to use anyone else’s (Mischel, 2008). In this way, the psyche is carved up into ever smaller bits that generate a research niche, and research proceeds in parallel lines, carefully avoiding to trespass someone else’s territory. This is understandable, but unsatisfying. Integration and cooperation are needed if the field is to make progress. Working on integration also does not mean foregoing theory. On the contrary, theoretical work is often required to enable new connections (i.e., theories and variables need to be made ‘connectable’). New connections themselves can be of theoretical nature, and new empirical connections can be inspired by theory. I would appreciate to see more emphasis on such integrative work in the future.

### Figure 1

#### *A Graph-based Model of Psychological Science*





*Note:* Circles (nodes) represent psychological theories or variables. Lines between the circles (edges) represent connections between variables or theories.

### **To Reduce Risk-of-Bias, What About Open Science?**

I have defined the reduction of risk-of-bias, especially publication bias, as one of the main goals of the dissertation project. Of course, this has been a key focus of the wider open science movement to solve what has often been coined a replication crisis or credibility crisis in psychological science (Nelson et al., 2018). So why have I thus far given little attention to the innovations developed by the open science movement, such as preregistration, open data and materials, registered reports, etc. (Munafò et al., 2017; Nosek et al., 2015, 2018; Nosek & Lakens, 2014). Mainly, because these measures are prospective: they increase the credibility of future research. My objective was to increase the credibility of summary research that aggregates previous work. In other words, I wanted to salvage what was there already. That is not meant to say that they are not important or beneficial for meta-analytic work. First, the principles also improve the quality of meta-analytic work directly. All papers included in this dissertation project have been preregistered (except for the simulation study, paper 2 of part I) and data for all projects is publicly available on the open science framework. Second, I suspect that the discussion around open science has made original authors more open to sharing their data than they would have been otherwise. Third, some (but not many) of the raw datasets we collected to conduct secondary data analyses were already publicly available, which simplified discovery and access. This leads me to the next point I would like to discuss, which is how new technological developments may make secondary data analyses more feasible in the future.

## **How Technological Advancements May Enable More Secondary Data Analyses**

Secondary data analyses as described here involve considerable effort on the part of the analyst. Relevant data need to be identified, acquired, validated, preprocessed, and finally jointly analyzed. In this section, I will briefly reflect on ideas from statistics and computer science that may facilitate such analyses in the future.

Starting with the statistical aspect, it is worth noting that the statistical models we used were strictly rooted in the traditional meta-analysis perspective (e.g., Borenstein et al., 2009). That is, we computed effect sizes for the original outcomes and modelled these in meta-analytic models that are essentially multi-level models with a hidden level (i.e., the level of raw data is assumed by the model). This conventional way of doing meta-analysis has some advantages. For one, casting studies or outcomes into effect sizes is quite descriptive and allows for some useful visualizations, such as funnel plots (Sterne & Egger, 2005). It can also be easier to ask original authors for effect sizes rather than a full data set. However, using effect sizes entails a loss of data when they are computed from raw data. Instead, summary estimates will be more precise when the raw data from multiple studies are modelled directly in a multi-level model (Riley et al., 2010). Going forward, this should become standard practice for meta-analyses, since there is little reason not to do it. In many ways, the effect size-based way is a workaround stemming from times when sharing data was cumbersome, which is now much easier with the internet and open repositories.

Accessing and reanalyzing existing data is indeed much easier than it probably was 20 years ago, but, as I learned, still quite effortful. Datasets shared by the original authors are often sparsely documented, leading to a back-and-forth on what certain variables mean. Sometimes the data are incomplete or even faulty. We discovered quite a few errors in the original data files.

For example, in one case an author misinterpreted their own gender coding in the analyses. Beyond these properties of the data themselves, the form in which they are transmitted can also be a challenge. Many authors rely on various proprietary file formats. Versioning and country locales can lead to problems. Authors often adjust psychological measurement in order to suit their purpose, for example by changing the response scale, or by removing, rewording, or adding scale items. I will discuss potential solutions to these issues in turn.

Proprietary datafile types is becoming less of an issue with the advent of open-source statistical packages such as *R* (R Core Team, 2022). With *R*, most proprietary datafiles can be read without issue, but problems may still prevail for older versions.

Authors changing psychological scales, sometimes without explicitly reporting this in the manuscript, and thus potentially distorting the measurement is a known problem. This has sometimes to do with the fact that there are multiple references for a psychometric scale that report different versions. Sometimes there is only one reference, but the reference does not include the full text items. A solution could be to create a central registry for psychological measurements that supports versioning and forking (as in common version control systems such as git; Junio Hamano & others, 2022). This registry would assign stable links to psychological measures (including all full-text items and the measurement scales). Reviewers could then ask authors to link to the specific version they have used and confirm that they did not alter the scale. If they did want to alter the scale, they would need to create a new fork in the registry and link to that. Recently, a proposal has been made for a central register for studies to make negative findings more discoverable and thus curb publication bias (Laitin et al., 2021). Similarly, widespread use of a registry for psychometric scales would also facilitate the discoverability of data for secondary analyses.

Unclear documentation of data and code is also a common problem that may even worsen as psychological researchers perform more and more coding and data processing without formal training for such tasks. Beyond better training, one solution could be to incorporate datasets and code more explicitly into the review process, but this would place even higher burdens on voluntary reviewers. Perhaps there could be a similar solution like the registry for psychometric scales I proposed previously. Repositories for open data such as the OSF could offer features for uploading datasets that enforce certain quality standards for the data. Such a system could implement concepts from data base theory (Codd, 1970), such as data consistency tests, check the data for missings, and force users to label and explain variables and codings. Of course, such a system could only work on a voluntary basis, as there will always be exotic data structures or types that cannot be foreseen. However, many psychological studies have relatively standard formats, such as experiments or cross-sectional questionnaire studies, that could be implemented as templates in the system. If it were clear that a dataset conforms to a standard format, it would be much easier to process for secondary users, or perhaps even fully programmatically.

Taking these thoughts on standardizing data and study formats as well as psychometric measurement further, one could envision a future of machine-readable psychological research. Currently, to understand a psychological study, one needs to retrieve the manuscript and read the plain-text methods section. These sections are still far from standardized, and the information conveyed can vary significantly between manuscripts. In many cases, one even requires insider knowledge of the respective field of research to comprehend the study. Imagine that instead, psychological studies were sufficiently standardized to be processed computationally. One data file would include full information of the study design as well as any measurements, manipulations, and so forth. If these data were freely discoverable and accessible, secondary data

analyses could be more and more automated utilizing innovations in big data processing and cloud computing, such that analysts can search and retrieve data on specific manipulations or measures. Considering how diverse psychological studies are, one may find this idea farfetched. But consider how far computational communication through standardizing protocols has progressed. Servers all over the world exchange any kinds of data in real time over the internet, connecting warehouses, financial markets, factory components, or even cars. Finding exchange formats and protocols for psychological research will be challenging but could ultimately pave the way to truly connected research with a cumulative, accessible, and comprehensible evidence base.

### **Limitations**

The vision for psychological research that I tried to develop in this dissertation perhaps deviates from the mainstream view. Instead of expanding the concept space in (social and personality) psychology in an ever-faster pace and aiming for breakthrough findings, I argue for a slower, perhaps more boring and bureaucratic way of doing science. In this boring science, there would be higher thresholds for introducing new ideas. More time would be spent on integrative work that standardizes and connects theories and data. The work of organizing and curating previous evidence would be cherished, rather than dismissed as grunt work for the big storytellers. Progress would seem slower but would be more stable and incremental. However, I am uncertain how compatible such a view of psychological science is with the current social architecture of academia, where jobs are sparse and visibility is key.

A related but unresolved issue is how to give credit to the original authors when doing extensive secondary data analyses. The analyses in this dissertation depended heavily on others' previous work and often ad hoc support with making data accessible and understandable, yet the

only credit these authors received was a citation, often for unpublished work. An alternative could be to involve authors as co-authors with pre-determined responsibilities limited to support with curating the data. The final article would then include a note on the contributions of all authors. Such article types with a large number of authors are becoming more common (for example, the multilab replications of ego depletion; Vohs et al., 2021).

Another potential limitation is that the proposed approach of re-using existing research data may be more feasible for some types of research than for others. The approach seems especially suited for correlational research, which was the primary approach of paper 2 of part II (Frankenbach et al., 2022). In paper 1 of part II, we relied on experimental data, but the approach was in essence also correlational, since we “correlated” the experimental effect with a measurement of personality. This point is self-evident, since secondary research can only use measures and manipulations that were in the original research, and incidental measurements are much more common than incidental manipulations (if they exist at all).

One issue that I have not definitively discussed is whether it is under all circumstances valid to detach measurements from the theoretical context in which they were conceived. For example, in paper 2 of part II, we retrieved the item “During the last month, how often have you had sexual thoughts involving a partner?” from the Sexual Desire Inventory (Spector et al., 1996) and classified it according to its literal, “atomic” meaning as a measure of the frequency of sexual thoughts. In the inventory, however, the item is thought to reflect a construct called “dyadic sexual desire”. This also raises the question if questionnaire items only reflect their verbatim meaning, or if the item context like instructions, previous items, or the overall study context are also reflected in the item response.

## **Conclusion**

Part I of the present dissertation project illustrated how and why dysfunctional research processes in psychological science can bias entire research literatures. Part II demonstrated a potentially viable solution: secondary data analyses. Two papers of part II showed how secondary data analyses can be used to test novel research questions in a theoretically coherent way with low risk-of-bias. These demonstrations could serve as a starting point to further facilitate secondary data analyses in the future, perhaps even making progress toward a vision of machine-readable psychological research. Potential limitations of such a vision include problems with assigning credit to contributors, psychometric concerns about context dependency of measurements, and a limited set of research designs that support secondary data analyses. Such developments toward a psychological science that is truly cumulative on the data-level may require alterations to the social architecture of academia, away from a system that encourages distinguishable, individual contributions toward a more collective approach.

### References

- American Psychological Association (Ed.). (2010). *Publication manual of the American Psychological Association* (6th ed). American Psychological Association.
- Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*(5), 1252–1265. <https://doi.org/10.1037/0022-3514.74.5.1252>
- Baumeister, R. F., Tice, D. M., & Vohs, K. D. (2018). The strength model of self-regulation: Conclusions from the second decade of willpower research. *Perspectives on Psychological Science*, *13*(2), 141–145. <https://doi.org/10.1177/1745691617716946>
- Baumeister, R. F., Vohs, K. D., & Tice, D. M. (2007). The strength model of self-control. *Current Directions in Psychological Science*, *16*(6), 351–355. <https://doi.org/10.1111/j.1467-8721.2007.00534.x>
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. <http://doi.wiley.com/10.1002/9780470743386>
- Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, *144*(4), 796–815. <https://doi.org/10.1037/xge0000083>
- Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: Has the evidence for ego depletion been overestimated? *Frontiers in Psychology*, *5*. <https://doi.org/10.3389/fpsyg.2014.00823>
- Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, *13*(6), 377–387. <https://doi.org/10.1145/362384.362685>



- Dang, J., Barker, P., Baumert, A., Bentvelzen, M., Berkman, E., Buchholz, N., Buczny, J., Chen, Z., De Cristofaro, V., de Vries, L., Dewitte, S., Giacomantonio, M., Gong, R., Homan, M., Imhoff, R., Ismail, I., Jia, L., Kubiak, T., Lange, F., ... Zinkernagel, A. (2021). A multilab replication of the ego depletion effect. *Social Psychological and Personality Science*, *12*(1), 14–24. <https://doi.org/10.1177/1948550619887702>
- Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120–128. <https://doi.org/10.1037/a0024445>
- Fleeson, W., & Jayawickreme, E. (2015). Whole Trait Theory. *Journal of Research in Personality*, *56*, 82–92. <https://doi.org/10.1016/j.jrp.2014.10.009>
- Frankenbach, J. (2015). *Versatile means of overcoming ego depletion*. Saarland University.
- Frankenbach, J., Weber, M., Loschelder, D. D., Kilger, H., & Friese, M. (2022). *Sex drive: Theoretical conceptualization and meta-analytic review of gender differences* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/9yk8e>
- Frankenbach, J., Wildschut, T., Juhl, J., & Sedikides, C. (2020). Does Neuroticism Disrupt the Psychological Benefits of Nostalgia? A Meta-analytic Test. *European Journal of Personality*, per.2276. <https://doi.org/10.1002/per.2276>
- Friese, M., & Frankenbach, J. (2020). P-hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods*, *25*(4), 456–471. <https://doi.org/10.1037/met0000246>
- Friese, M., Frankenbach, J., Job, V., & Loschelder, D. D. (2017). Does self-control training improve self-control? A meta-analysis. *Perspectives on Psychological Science*, *12*(6), 1077–1099. <https://doi.org/10.1177/1745691617697076>

- Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., ... Zwieneberg, M. (2016). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science, 11*(4), 546–573.  
<https://doi.org/10.1177/1745691616652873>
- Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin, 136*(4), 495–525.  
<https://doi.org/10.1037/a0019486>
- Inzlicht, M., & Schmeichel, B. J. (2012). What is Ego Depletion? Toward a mechanistic revision of the resource model of self-control. *Perspectives on Psychological Science, 7*(5), 450–463. <https://doi.org/10.1177/1745691612454134>
- Junio Hamano & others. (2022). *Git* (2.37.1) [Computer software].
- Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT Press.
- Laitin, D. D., Miguel, E., Alrababa'h, A., Bogdanoski, A., Grant, S., Hoerberling, K., Hyunjung Mo, C., Moore, D. A., Vazire, S., Weinstein, J., & Williamson, S. (2021). Reporting all results efficiently: A RARE proposal to open up the file drawer. *Proceedings of the National Academy of Sciences, 118*(52), e2106178118.  
<https://doi.org/10.1073/pnas.2106178118>
- McCrae, R. R., & Costa, P. T. (2003). *Personality in adulthood: A five-factor theory perspective*. Guilford Press.

- Mischel, W. (2008, December 1). The Toothbrush Problem. *APS Observer*.  
<https://www.psychologicalscience.org/observer/the-toothbrush-problem>
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021.  
<https://doi.org/10.1038/s41562-016-0021>
- Muraven, M., Baumeister, R. F., & Tice, D. M. (1999). Longitudinal improvement of self-regulation through practice: Building self-control strength through repeated exercise. *The Journal of Social Psychology*, *139*(4), 446–457.  
<https://doi.org/10.1080/00224549909598404>
- Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*(1), 511–534. <https://doi.org/10.1146/annurev-psych-122216-011836>
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425.  
<https://doi.org/10.1126/science.aab2374>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606.  
<https://doi.org/10.1073/pnas.1708274114>
- Nosek, B. A., & Lakens, D. (2014). Registered Reports: A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141. <https://doi.org/10.1027/1864-9335/a000192>

- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. <https://doi.org/10.1126/science.aac4716>
- R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.2.1) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *BMJ*, *340*, Article c221. <https://doi.org/10.1136/bmj.c221>
- Sedikides, C., Wildschut, T., Routledge, C., Arndt, J., Hepper, E. G., & Zhou, X. (2015). To nostalgize: Mixing memory with affect and desire. In *Advances in Experimental Social Psychology* (Vol. 51, pp. 189–273). Elsevier. <https://doi.org/10.1016/bs.aesp.2014.10.001>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Spector, I. P., Carey, M. P., & Steinberg, L. (1996). The Sexual Desire Inventory: Development, factor structure, and evidence of reliability. *Journal of Sex & Marital Therapy*, *22*(3), 175–190. <https://doi.org/10.1080/00926239608414655>
- Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78. <https://doi.org/10.1002/jrsm.1095>
- Sterne, J. A., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99–110). Wiley.

- Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality, 72*(2), 271–324. <https://doi.org/10.1111/j.0022-3506.2004.00263.x>
- Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A. J., Ainsworth, S. E., Alquist, J. L., Baker, M. D., Brizi, A., Bunyi, A., Butschek, G. J., Campbell, C., Capaldi, J., Cau, C., Chambers, H., Chatzisarantis, N. L. D., Christensen, W. J., Clay, S. L., Curtis, J., ... Albarracín, D. (2021). A multisite preregistered paradigmatic test of the ego-depletion effect. *Psychological Science, 32*(10), 1566–1581. <https://doi.org/10.1177/0956797621989733>