ANNUAL REVIEWS

Annual Review of Clinical Psychology

Psychology's Replication Crisis and Clinical Psychological Science

Jennifer L. Tackett,[1] Cassandra M. Brandes,[1] Kevin M. King,[2] and Kristian E. Markon[3]

[1]Department of Psychology, Northwestern University, Evanston, Illinois 60208, USA; email: jennifer.tackett@northwestern.edu

[2]Department of Psychology, University of Washington, Seattle, Washington 98195, USA

[3]Department of Psychological and Brain Sciences, University of Iowa, Iowa City, Iowa 52242, USA

Annu. Rev. Clin. Psychol. 2019. 15:579–604

First published as a Review in Advance on January 23, 2019

The *Annual Review of Clinical Psychology* is online at clinpsy.annualreviews.org

https://doi.org/10.1146/annurev-clinpsy-050718-095710

Keywords

clinical psychology, clinical science, metapsychology, open science, replication crisis, reproducibility

Abstract

Despite psychological scientists' increasing interest in replicability, open science, research transparency, and the improvement of methods and practices, the clinical psychology community has been slow to engage. This has been shifting more recently, and with this review, we hope to facilitate this emerging dialogue. We begin by examining some potential areas of weakness in clinical psychology in terms of methods, practices, and evidentiary base. We then discuss a select overview of solutions, tools, and current concerns of the reform movement from a clinical psychological science perspective. We examine areas of clinical science expertise (e.g., implementation science) that should be leveraged to inform open science and reform efforts. Finally, we reiterate the call to clinical psychologists to increase their efforts toward reform that can further improve the credibility of clinical psychological science.

Contents

INTRODUCTION

In recent years, replicability, open science, research transparency, and the improvement of methods and practices have been at the forefront of scientific discussions in the psychological community, with a renewed focus on these topics emerging around 2011 (e.g., Simmons et al. 2011). Although clinical psychology has been slow to engage in these ongoing efforts, a number of clinical scientists produced a paper calling clinical psychologists to the replicability conversation (Tackett et al. 2017), and movement within clinical psychology and related disciplines has been growing. For example, a recent call to establish an initial database of clinical psychologists interested in open science and replicability topics received an overwhelming response, and more than 200 individuals identifying as clinical scientists signed up within a couple of weeks (see **https://goo.gl/forms/tvAijds0as7peCdG2**). These individuals indicated a wide range of topical interests in open science, methodology, and replicability, and they reflected a broad range of training levels, including graduate students and postdoctoral fellows (38.4%), assistant professors (26.4%), and associate/full professors (25.1%). Thus, despite its slow emergence thus far, substantial interest exists in the field and is motivating researchers to develop collaborative and constructive ways to move forward.

In the current review, we aim to contribute to this emerging conversation, hoping that clinical psychologists will continue to increase their involvement and the potential solutions generated from it. We examine a few areas in clinical psychological science in which replicability and credibility concerns are likely. First, we call for a more critical look at those portions of our existing research base (and the associated methods, practices, and dissemination culture) where findings may not be replicable or robust. Then, we briefly review some of the current open science tools, proposals, and solutions that clinical psychological scientists could begin integrating into their

research and workflow. We next discuss some of the challenges in defining and operationalizing replication in clinical psychological research, at both the theoretical and methodological level. We then turn to a consideration of how the broader principles and framework of implementation science might be leveraged to increase uptake and retention of better methods and practices across the field of clinical psychology. We end with some basic conclusions and next steps, including a plea for other clinical psychological scientists to join us in this work.

IDENTIFYING THE WEAK SPOTS IN THE CLINICAL SCIENCE RESEARCH BASE

The fact that clinical psychology has been less involved in the replicability conversation to date should not be taken to suggest that the field does not have replicability concerns. Early efforts from social, personality, and cognitive psychology have begun highlighting areas in which replicability problems may exist (e.g., Open Sci. Collab. 2015). Differences in methodology and statistical approaches [e.g., an emphasis on effect size in clinical research over focal hypotheses examined via null hypothesis significance testing (NHST) in experimental social and cognitive psychology] may point to potential differences in overall replication of findings, but until this issue is explicitly and comprehensively examined, it remains an empirical question (Tackett et al. 2017). One recent effort to replicate comprehensive associations between personality traits and life outcomes demonstrated remarkably high replication rates (Soto 2018), perhaps in part because of the extensive body of research underlying the conceptualization and measurement of personality traits. To the extent that certain areas of clinical psychology, such as personality psychology, similarly rely heavily on correlational (or observational) estimates of descriptive associations, it is certainly possible that these areas will prove more robust.

Reardon and colleagues (K.W. Reardon, A.J. Smack, K. Herzhoff & J.L. Tackett, manuscript in review) coded median sample sizes (which are but one aspect informing the statistical power in a study) in studies published in the two top clinical journals of the American Psychological Association—the *Journal of Abnormal Psychology* and the *Journal of Consulting and Clinical Psychology*—in 2000, 2005, 2010, and 2015. This study found that sample sizes in studies published in top clinical psychological outlets were larger than those reported in top social psychological outlets as well as the field-wide flagship journal *Psychological Science* (Fraley & Vazire 2014; K.W. Reardon, A.J. Smack, K. Herzhoff & J.L. Tackett, manuscript in review). Clinical sample sizes were more comparable to those reported in personality journals, although the clinical journals had higher associated impact factors—that is, correlations between journal impact factor and the clinical N-factor show that clinical outlets may reward higher sample sizes. Nonetheless, most samples were not large enough to detect small effect sizes (e.g., correlations in the 0.10–0.20 range, which are likely to be most common for psychological research). Furthermore, from 2000 to 2015, a change in sample size was not evidenced for the *Journal of Consulting and Clinical Psychology* and was only beginning to trend up toward higher sample sizes for the *Journal of Abnormal Psychology*. Reardon et al. (K.W. Reardon, A.J. Smack, K. Herzhoff & J.L. Tackett, manuscript in review) reveal some mixed news for clinical psychology: Although sample sizes in clinical psychology studies have been higher than in some psychological subfields (e.g., social psychology), and clinical journals evidence a higher correlation between median sample size and impact factor, an improvement in sample sizes over the last two decades is not clearly evident across these prominent journals. Moreover, our samples are often too small to detect small effect sizes, which may be a common magnitude for many psychological effects.

Historically, statistical power has been poor in psychological research (Cohen 1962, Fraley & Vazire 2014), and this is especially true for resource-intensive areas. Two such areas in clinical

psychology include clinical neuroscience and treatment/intervention research. Research on the neuroscientific mechanisms underlying psychopathology has been increasingly prioritized by funding bodies, including the National Institutes of Health (NIH), creating a heightened market for this type of research. These studies are costly in time and money, and thus it is unsurprising that they often rely on small samples to examine small- to medium-sized effects, resulting in poor statistical power to detect true effects (Button et al. 2013). In addition to the frequent use of small samples, another factor that can minimize the statistical power of this research is that often researchers simultaneously use a large number of dependent variables and thus need to correct for multiple comparisons (Cremers et al. 2017). Of course, these problems also generalize to clinical neuroscience research and present a number of methodological implications, all of which can affect replicability (Abram & DeYoung 2017). Studies of statistical power in treatment research have been similarly portentous. For example, upon review of comparative outcome trials for depression, researchers found that the reviewed studies included a maximum of 40% of the participants needed to detect a minimally clinically important effect size ($d = 0.24$) (Cuijpers 2016). These initial investigations indicate that clinical neuroscience and treatment/intervention studies are likely areas for concern when examining clinical psychology's replicability weak spots.

Another potential area for concern in clinical psychology stems from diagnostic unreliability. Interrater reliability for diagnoses has historically been mixed at best (Frances & Widiger 2012), and with the transition from DSM-IV to DSM-5, reliability has improved only for some diagnoses [e.g., posttraumatic stress disorder (PTSD)], whereas major problems remain for many common disorders (e.g., generalized anxiety and major depressive disorders) (Freedman et al. 2013, Regier et al. 2013). In addition, categorical diagnoses have poor predictive power for future diagnoses in comparison to dimensional alternatives (Kim & Eaton 2015), which is likely related in part to the loss of power that accompanies dichotomization (MacCallum et al. 2002). Given the primacy of DSM-defined diagnoses in the vast majority of clinical psychological research, such issues pose clear barriers to establishing a replicable and robust literature.

Yet another consideration is the impact of publication bias in clinical science, which has been discussed for decades (e.g., Meehl 1990). Across psychology, significant results are disproportionately published relative to null results, with 96% of published studies showing significant effects (Bakker et al. 2012); this figure is elevated even in relation to the 80% average across the sciences (Fanelli 2010). In clinical psychology specifically, researchers have documented a consistent bias toward significant results in intervention studies, resulting in an overestimation of treatment effects for disorders such as depression, generalized anxiety disorder, and attention-deficit/hyperactivity disorder (Cuijpers et al. 2010, 2014; Rapport et al. 2013). This is consistent with evidence that positive (i.e., statistically significant) findings from clinical trials are published more often, and more quickly, than negative/null findings (Hopewell et al. 2009). Descriptive psychopathology research suffers from publication bias as well; one high-profile example suggested that the association between the 5-HTTLPR serotonin transporter gene and unipolar depression was likely a false positive finding initially stoked by publication bias (Culverhouse et al. 2018, Duncan & Keller 2011). Similarly, one recent investigation found that substantial publication bias favoring large positive effects had drastically inflated reported associations between depression and mortality (Miloyan & Fried 2017). Thus, publication bias is likely a problem across various methodological approaches and research topics in clinical psychology.

A related concern is the use of questionable research practices (QRPs), or researcher behaviors that increase the likelihood of achieving preferred (typically, statistically significant) results (Simmons et al. 2011). These practices include the selective reporting of desirable results, flexibility in data analytic strategies, and hypothesizing after results are known (HARKing), to name a few. In a study of QRPs among psychologists, John and colleagues (2012) found that an estimated

91% to 94% of researchers across subfields reported engaging in at least one QRP, with this rate being lower among clinical researchers compared to researchers in other subfields of psychology. Nonetheless, there is evidence of QRPs from within the clinical literature, as well; for example, the selective reporting of results in clinical treatment studies is an established phenomenon often discussed as allegiance effects (Leichsenring et al. 2017). Allegiance effects, or the tendency of treatment studies to support the theoretical orientation of the author (e.g., cognitive behavioral or psychodynamic therapies), have been documented for several decades (Munder et al. 2013). This selective reporting may contribute to null findings for comparative outcome studies or to the so-called Dodo Bird Verdict in treatment research (Cuijpers 2016, Wampold et al. 1997). In a recent critical examination, Leichsenring and colleagues (2017) identified 13 biases they believed to affect the replicability of psychotherapy research and offered suggestions on how to counter each of them. This is a model future areas of clinical psychological research can follow to begin evaluating their own practices and biases.

To combat biased reporting, QRPs, and other problematic practices, one ameliorative effort has been to create guidelines and norms for the registration of intervention trials in psychology (e.g., APA Publ. Commun. Board Work. Group J. Article Rep. Stand. 2008). However, adherence to these guidelines is not required by most clinical psychology journals. Therefore, it is unsurprising that research shows that the vast majority of randomized controlled trials (RCTs) in clinical psychology are not preregistered according to best practice recommendations. Cybulski and colleagues (2016) showed that among RCTs published in 2013 in the 10 highest-impact clinical psychology journals ($k = 165$), only 25% of trials were registered prospectively (i.e., before data collection) and a mere 1% ($k = 2$) were registered prospectively and defined all primary outcomes in their registration. A systematic comparison of registrations to publications has not been conducted in clinical psychology, but evidence from health intervention research suggests that anywhere between 40% and 62% of RCTs contain at least one unreported change to the research design following registration (Dwan et al. 2013). It is also worth noting that it is common to not report results of clinical trials; one estimate suggests that just 63% of all clinical trials report their findings (see **http://fdaaa.trialstracker.net/**). These results collectively suggest that clinical psychologists, like researchers in other areas, likely engage in practices that increase the likelihood of replicability problems in the long term. Further, this evidence indicates a need to establish better norms and to strengthen tools that combat bias and increase openness in clinical science.

Meta-Analysis and Its Relevant History in Clinical Psychology

Historically, the field of clinical psychology has been central to the study of scientific replicability, not only within psychology and the behavioral sciences but also in the sciences more broadly. By pooling published results and, ideally, including many effects from many studies, meta-analysis can (for an unbiased literature) provide a more accurate estimate of both effect size and effect size heterogeneity. Modern meta-analysis has its origins in the meta-analytic study of clinical phenomena: Glass (1976) in fact coined the term "meta-analysis" in the context of research into psychotherapy efficacy (Glass 2015, Smith & Glass 1977). Statistical integration of results from multiple studies is much older (Pearson 1904), but more recent work by Glass and colleagues, as well as others (Shadish & Lecy 2015), has helped spark modern meta-analysis by leveraging effect sizes to integrate findings across heterogenous research designs (Gurevitch et al. 2018). In sum, clinical psychology was central to establishing the lingua franca of replicability studies in terms of effect size, power, and related considerations.

Parallels between historical meta-analysis and the current replicability crisis are informative. Much like now, concerns, as well as methods for addressing them, were driven by skepticism about

specific scientific claims—in that case, claims of psychotherapy efficacy (Shadish & Lecy 2015). Various lines of evidence have confirmed long-standing concerns that the psychotherapy research literature itself has been affected by publication bias (e.g., Cuijpers et al. 2010, Driessen et al. 2015, Sohn 1996). This pattern of findings suggests that research on replicability is coming full circle. Whether or not prospective registered replication projects will follow remains to be seen. On the one hand, psychotherapy research has, perhaps, one of the most thorough meta-analytic evidence bases in all of science, and the existence of NIH and other agency records has allowed for retrospective registered replication analyses. On the other hand, registered replication projects have followed other meta-analytic demonstrations of publication bias in other fields, and there have been counterarguments regarding the magnitude of psychotherapy publication bias (Niemeyer et al. 2013), as well as observations that other types of bias (e.g., measurement error) might counter publication bias per se (Staines & Cleland 2007).

Clinical Assessment Gaps Between Research and the Field

Additional areas of concern for clinical psychology research are the assessment gaps between research and applied settings and the extensive heterogeneity in forms of assessment across studies. Problems with defining optimal practices for diagnoses notwithstanding (as discussed previously), ideal, structured assessment practices are rarely, if ever, applied in the same fashion in the laboratory and in the clinic due to practical limitations (Reardon et al. 2017). Further, varying assessment approaches (e.g., prospective versus retrospective) have been shown to drastically alter lifetime diagnostic prevalence estimates even between academic studies (Moffitt et al. 2010). These problems likely contribute to discrepancies between field and academic settings in the prevalence rates for many disorders (Garb 2005); these findings imply nongeneralizability and nonreplication of inclusion criteria and mental health outcomes across contexts.

Assessment gaps between clinical and academic studies are evidenced in divergent disorder prevalence estimates reported in these two settings. The breadth of those gaps varies widely by disorder. Some research suggests that primary care practitioners underdiagnose some disorders (e.g., early personality disorder) when comparing field prevalence rates and data from epidemiological surveys (Conway et al. 2017, Garb 2005). Research has also suggested that prevalence rates are inflated for other disorders (e.g., pediatric bipolar disorder) in the clinic relative to the lab due to an overreliance on potentially biased clinical judgment instead of structured assessment, to possible malingering, and to the incentives to assign diagnoses created by pharmaceutical companies (Frances & Widiger 2012, Garb 2005, Jenkins et al. 2011). On the other hand, field trials and epidemiological research have also been criticized for not capturing mid-range diagnoses of the kind that generally present in a primary care setting, so the potential underestimation of population prevalence rates in research is also a possible contributor to these gaps (Jones 2012). Whereas the factors contributing to these discrepancies (e.g., malingering) vary by disorder, one primary commonality is the lack of structured assessments in primary care contexts. As discussed, this can lead to both under- and overdiagnosis, though the full implementation of structured assessment is impeded by sizeable practical constraints. In evaluating the replicability of various effects in clinical psychology, these methodological features warrant close consideration.

The degree of structure in the assessment of psychopathology is not the only measurement factor that may contribute to heterogeneity in disorder prevalence estimates; methodological differences between studies may also contribute to assessment discrepancies. One meta-analysis of prevalence rates for a variety of childhood disorders found that estimates varied substantially as a function of sampling frame (e.g., schools or households), sample representativeness, and type of diagnostic interview (Polanczyk et al. 2015). Other researchers have found vastly different lifetime prevalence rates as a function of retrospective versus prospective study designs, with prospective

rates approximately doubling retrospective rates (Moffitt et al. 2010). This substantial method variance has immense implications for the replicability of clinical science, as prospective studies are extremely time- and resource-intensive. Therefore, much descriptive psychopathology research is conducted retrospectively, possibly yielding estimates that may in turn face greater replication problems.

Generalizability of Randomized Controlled Trials to the Field/Community Settings

Although RCTs provide the gold standard of evidence for treatment efficacy, often, little attention is paid to how the design of RCTs affects the external validity or generalizability of their findings. As Shadish and colleagues (2002, p. 18) noted, "readers of experimental results are rarely concerned with what happened in that particular, past, local study. Rather, they usually aim to learn either about theoretical constructs of interest or about a larger policy." The essential irony of many RCTs designed to test efficacy is that maximizing internal validity (the degree to which a treatment is known to cause changes in the outcome) often comes at the cost of external validity (the degree to which effects may be generalized to broader populations and contexts) (Shadish et al. 2002). There has long been concern about the differences between populations eligible for RCTs and the general patient population (Rothwell 2005) in mental health intervention research (Stuart et al. 2015). One review of exclusion criteria for studies of substance use disorders (SUDs) estimated that between 63% and 68% of individuals with SUDs would be excluded from RCTs for reasons such as prior treatment, low motivation, other substance use, or medical problems (Moberg & Humphreys 2017). On the other hand, evidence for the treatment of many other disorders contradicts this claim. For example, a few studies have shown that most patients seen in outpatient community settings would be eligible for at least one published RCT (Stirman et al. 2003, 2005). These and other studies from inpatient settings have suggested that ineligibility of patients in community settings is often due to disorders that are in partial remission or are milder compared with those observed in RCTs (Stirman et al. 2003, 2005; von Wolff et al. 2014). Some research has suggested that the most common exclusion criteria seen in RCTs for depression are unlikely to influence the generalizability of their effects (van der Lem et al. 2012).

However, the generalizability of RCTs is influenced not only by client characteristics (i.e., how well they match the population for whom there is evidence of efficacy) but also by the degree to which the characteristics of a treatment delivered in an RCT match those delivered in applied settings. Here, the gap between research and practice seems to be much wider. One study of the efficacy of treatments for children in usual care settings reported that the typical effect size of treatment was nearly half or less than that reported in prior meta-analyses of RCTs (Weisz et al. 1995). This study also noted substantial differences between the context of RCT delivery and community settings. RCTs frequently rely on highly trained therapists with homogenous and intensive training in the treatment, high levels of supervision from treatment experts, and relatively low caseloads of patients with similar disorders; by contrast, psychotherapists in the community carry a high caseload of patients with a wide variety of diagnoses, and they get relatively little supervision (Weisz et al. 1995) that is provided sparingly (approximately an hour per week) and rarely focuses on evidence-based interventions (Dorsey et al. 2013, 2017).

EXTENDING CURRENT PROPOSALS FOR REFORM TO CLINICAL PSYCHOLOGY RESEARCH

Since the most recent instantiation of psychology's replication crisis in 2011, the rate at which tools and incentives for increasing replicability have been developed has been staggering.

However, the majority of these efforts have been initiated by psychologists within a limited range of subfields, and therefore concerns have been raised as to their applicability to areas outside of mainstream social, cognitive, and personality psychology (Finkel et al. 2015, Tackett et al. 2017). In addition to this tension, more fundamental issues raise additional concerns about generalizability. For example, crucial questions around defining and operationalizing replication may not translate directly to much clinical psychological research, and related statistical and analytic issues may necessitate a broader discussion on how to better capture a variety of psychological findings, including those from the clinical domain. We briefly discuss three domains of proposal for reform: applying proposed tools and solutions to clinical psychological research, defining and operationalizing replication in clinical psychological research, and extending relevant statistical debates and concerns to clinical psychology.

Proposed Solutions and Tools for Reform

Here we briefly describe five nonexclusive (and nonexhaustive) routes for increasing replicability: (*a*) open materials, (*b*) open data, (*c*) (pre)registration, (*d*) Registered Reports, and (*e*) multisite collaborations for both replications and original research. We then provide an overview of various incentives for engaging with these practices and how they relate to clinical science. This list is only a selection of current efforts toward reform, but it will perhaps serve as a starting point for those interested in enhancing the openness, transparency, and reproducibility of their research.

Open materials. Open materials are most typically materials that study authors make available by posting all digitally shareable protocols, measures, and other applicable study tools to a publicly available online repository. Open materials allow for a more thorough evaluation of a study and facilitate standardization and replicability in future efforts using the same materials. However, clinical psychology materials are, on average, not as readily shareable as those in some other subfields. First, monetization is a much more central issue to clinical psychological assessments than to similar measures in other subfields (Yates & Taub 2003). Second, public access to assessments for psychopathology is especially problematic from an ethical standpoint due to the increased likelihood of malingering and diagnoses by unqualified persons; indeed, test security is explicitly listed under the American Psychological Association's (APA) Ethical Principles Section 9.11 (Am. Psychol. Assoc. 2017). Third, measure integrity can be compromised if assessment materials are made available to those whom they are designed to assess. Neuropsychological tests, for example, are often tightly controlled to protect against coaching and practice effects (Calamia et al. 2012), and re-norming these materials with sufficient power is enormously resource intensive. However, many clinical psychology materials are arguably not different from those in other subfields in terms of the extent to which they may be shared. Experimental paradigms such as those designed to evoke fear or reward responses, scripts for parent-child interaction tasks, and instructions for daily diary completion are all examples of materials that do not carry the ethical and practical weight of measures such as those described above. As such, open materials in clinical psychology are a nuanced issue, and they warrant further attention as the field joins the reproducibility conversation.

Open data. Open data are comprehensive study data posted to an online repository that can be accessed by individuals outside of the study team. The term "open data" includes both fully publicly available data and data posted to trusted third-party repositories that restrict access to researchers who meet criteria demonstrating proper handling of sensitive data—what is called protected access (Blohowiak et al. 2018). Open data are intended to facilitate verification and increase the

reliability of a project's reported results, especially because it is often practically impossible to obtain study data otherwise, even when they are described as "available on request" (Wicherts et al. 2006). One major concern with sharing clinical data is identifiability, as clinical psychology data generally include sensitive health and legal information (e.g., endorsement of criminal behavior), low base-rate conditions (e.g., schizophrenia), and rare populations (e.g., individuals aged over 89 years). Aside from trusted data repositories, another proposed solution is the application of the so-called safe harbor method, which is a deidentification method originally developed for medical data that substantially limits privacy risks by removing 18 features such as precise ages over 89 years or exact zip codes in areas with populations of less than 20,000 (Walsh et al. 2018). In addition, clinical psychology can benefit by learning from efforts from other areas of social science research that are grappling with similar concerns (e.g., relationship research on sensitive topics) (Joel et al. 2018).

A second major concern about open data among clinical psychology researchers is the incentive for conducting time- and resource-intensive data collection when open data are normative. Although this problem (the so-called tragedy of the commons) is applicable to all of psychology, it is even more salient for data that require years, even decades, of effort and funding on the part of investigators to collect, given that incentives generally favor researchers who publish papers over those who collect data (or perhaps, they favor those who get grants over those who publish papers over those who collect data). Some proposed solutions to this problem include data embargoes, whereby researchers can publish data only after a prespecified period of time in which they plan to publish on the data set themselves, and venues allowing for the publication of open data sets so that researchers may benefit from citations when the data are used (Munafò et al. 2017).

Preregistration. In a strict sense, preregistration entails public posting of time-stamped records of study design, hypotheses, outcomes, and data analysis pipelines prior to the execution of a study, including prior to the start of data analysis. Preregistration is intended to clearly delineate which findings in a study were predicted a priori and thus clarify the diagnosticity of inferential statistical tests (Nosek et al. 2018). Preregistration of primary and secondary outcomes is now mandated for clinical trials by the United States government as well as many medical journals (see also International Committee of Medical Journal Editors in De Angelis et al. 2004). Despite this push within the broader medical field, clinical psychology has been slow to adopt preregistration. This may be due to a focus in psychology's replicability conversation on strict preregistration—that is, one in which all aspects of a study can be neatly delineated prior to study commencement. Strict preregistration is a standard that is not achievable for much research outside of lab-based studies with undergraduates, which may result in researchers opting out completely (Finkel et al. 2015, Tackett et al. 2018). The lack of uptake in clinical psychology specifically may also be influenced by the more common usage of archival data sets, the protracted nature of data collection and hypothesis development in larger and longer studies, and the unpredictability of sampling due to base rates, difficulties with recruitment, and additional challenges encountered in clinical research (see Tackett et al. 2017 for a review). There is a growing recognition of these problems within the replicability conversation, however, and some researchers have warned against throwing out the proverbial baby with the bathwater (Tackett et al. 2018). Indeed, we urge clinical psychological researchers to consider various options along the registration continuum, as many possibilities exist beyond strict preregistration. Recent recommendations can be summed up as follows: Be as transparent as is feasible, as some basis for evaluating credibility is better than none at all.

Registered Reports. Registered Reports (RRs) are a type of empirical article that undergoes peer review at the design phase of the project, with in-principle acceptance awarded before the results

are known by authors and reviewers to ensure the publication of the project regardless of whether the findings support the hypotheses or not. This initiative is intended to reduce publication bias and the selective reporting of significant results. RRs are essentially detailed preregistrations of studies that are evaluated by peer reviewers, updated based on reviewer feedback, and, barring deviations from the agreed-upon study design and analytic plan or author withdrawal, published upon completion. First implemented in 2013, RRs are currently offered as an article type in 108 journals across psychology, neuroscience, and medicine. At the time of writing, however, very few clinical psychology journals have yet confirmed RRs as an accepted article type. Barriers to RRs for clinical psychologists overlap with those for preregistration and those for open materials and data, as these are often requirements for RRs. The use of existing data sets has also been raised as a concern for the interface between RRs and clinical psychology, but over 50% of journals that offer RRs accept submissions based on existing data (see **https://cos.io/rr/**). RRs have been cited as ideal solutions to many problems with replicability, but this is an initiative that requires further reconciliation with the needs of clinical science in the future.

Multisite collaborations. Multisite collaborations are projects initiated and maintained by researchers at multiple institutions for the purposes of conducting both large-scale replications and original research. Multisite collaborations are intended to increase diversity and statistical power in research, especially among populations that are hard to recruit, by pooling resources between labs. The last several years have seen the rise of many impactful multisite collaborations, including the Human Connectome Project (Van Essen et al. 2013), the Psychological Science Accelerator (Moshontz et al. 2018), and StudySwap (Chartier & McCarthy 2018). Still other efforts exist for the purpose of conducting large-scale replications, including the Many Labs (Klein et al. 2014) and ManyBabies (Frank et al. 2017) projects as well as Registered Replication Reports, which is a type of article similar to RRs and focused on multisite replications (**https://www.psychologicalscience.org/publications/replication**). Multisite collaboration is especially important for resource intensive research, and it is therefore optimally applicable to clinical science. Although multisite collaborations have become commonplace for original research in clinical psychology (for example, in neuroimaging research), large-scale replication efforts for clinical research findings have not yet taken hold. Given the obstacles that clinical psychology faces in doing exact replications (see Tackett et al. 2017), collaborations present rich opportunities to integrate replication in an efficient manner. However, without structural incentives for engaging in these efforts, their uptake is likely to be slow.

All researchers are motivated by structural incentives and norms, and changes in these factors are likely to have an effect on the future engagement of clinical psychology researchers with reproducibility and open science (Lilienfeld 2017). In perhaps one of the most powerful shifts in structural incentives, the NIH announced that a discussion of "Rigor and Reproducibility" would be required in the Significance/Approach sections of NIH grant applications submitted after May 25, 2016, with plans to soon apply these standards to institutional training grants, institutional career development grants, and individual fellowships (NIH 2015). Although this change is promising, it is unclear how it will be operationalized and what direct effect it will have on scientific practices.

Beyond funding structures, journals—including several high-impact clinical science journals—have also begun adopting standards of openness, including the Transparency and Openness Promotion (TOP) Guidelines (see **https://cos.io/top/**). The TOP Guidelines are gradients of enforcement of eight open science standards/practices (e.g., preregistration, replication, code transparency), with four tiers of stringency. Some journals have also begun implementing a badge system, whereby individual articles are displayed with acknowledgments of engagement with open

science practices including open data, open materials, and preregistration. At the time of writing, at least two clinical psychology journals (*Clinical Psychological Science* and *Collabra: Clinical Psychology*) have adopted badges promoting open science practices. Since implementing badges in psychology journals, data sharing has risen from approximately 3% to 40% of articles in participating journals (with no change in nonparticipants), and open materials have likewise increased, though to a lesser extent (Kidwell et al. 2016). Moreover, open access to clinical psychological research has expanded in the past decade, and the NIH now mandates that all peer-reviewed manuscripts published on NIH-funded research be made publicly available (NIH 2008). Clinical psychology researchers have been relatively slow to adopt reproducibility and open science; however, structural incentives and norms enforced by funding agencies and clinical psychology journals may advance this engagement substantially. As such, it is imperative that clinical scientists join the conversation on developing tools, incentives, and norms around these issues to overcome obstacles to implementation.

Defining and Operationalizing Replication in Clinical Psychological Research

An ongoing challenge in the replication crisis is how best to define replicability. For example, what aspects of a design must be replicated in order to consider something a direct versus a conceptual replication (Zwaan et al. 2018)? Even researchers at the forefront of the reform movement have struggled to answer this question when working primarily from an experimental cognitive/social psychological framework, in which constraining direct replications is likely more feasible. Multiple barriers emerge that prevent direct replication in clinical psychology (Tackett et al. 2017), and a truly direct replication may never be possible (Tackett & McShane 2018). Some critical examples of barriers in clinical psychological research are the variable nature of clinical samples across sites and labs; the changing nature of diagnostic constructs, which results in evolving definitions and measurements; and the use of statistical methods (i.e., prioritization of effect size estimation over NHST) that require more nuanced approaches to replication than dichotomous success/failure decisions allow (Tackett & McShane 2018, Tackett et al. 2017). Highly relevant and generalizable conversations are occurring within epidemiology, with parallels in terms of primary concerns and challenges for solutions (Lash et al. 2018).

As the broader reform movement considers the best way to operationalize replication success and failure in certain types of psychological research, members of the clinical psychology community must directly address the question: By what criteria should we define a replication attempt? And how do we operationalize the outcomes of these attempts? Given the nature of much clinical psychological research, we may be well served by considering replication as a latent construct and working toward construct validation efforts—a space in which many clinical psychology researchers are quite comfortable (Cronbach & Meehl 1955). This will require an ongoing effort to define the construct and an iterative feedback process between measurement attempts and construct refinement. Such an approach would also likely be best served by more complex and nuanced measurement, including, for example, the use of intervals or reaction ranges in place of dichotomous (e.g., success/failure) or point estimates, as well as the use of hierarchical/meta-analytic models to account for effects across studies, sites, and samples (Tackett & McShane 2018).

One potential starting point is a comprehensive review of previous attempts in the clinical psychology literature, which certainly exist (e.g., see the thoughtful analysis by Sher et al. 2011 on how the "cat's cradle" pattern in latent trajectory analysis may be artificially recovered from longitudinal substance abuse data sets). We see this as an exciting and incredibly important area for further consideration as clinical psychology becomes more and more engaged with replication and other reform efforts.

Extending Statistical Concerns to Clinical Psychological Research

Concerns about the applicability of existing replication paradigms to clinical psychology extend to statistical methods. Differences in the types of questions and designs that constitute clinical psychological science lead to different replicability concerns, which in turn have implications for how replication is conceived of and operationalized in clinical psychology versus other areas.

For instance, in contemporary discussions of replicability, a great deal of focus has been placed on *p*-values, especially in terms of two general concerns: how they are interpreted vis-à-vis an observed sample or result, and their vulnerability to manipulation in the form of *p*-hacking. The former concern, about *p*-value interpretation, derives from the insufficiency of *p*-values as an index of the replicability or veridicality of a given result (Colquhoun 2017): *p*-values represent $P(X|h)$, the probability of an observed result X at least as extreme as the observed result, given a hypothesis of interest h, when what is needed is instead an estimate of the probability of a hypothesis h given an observed data set or result X, or $P(h|X)$. The latter concern about *p*-hacking arguably derives in part from a focus in the literature on NHST in its various forms.

These two general concerns about *p*-values apply to nearly, if not all, areas of science, including clinical psychology. However, certain characteristics of clinical psychological research relative to other areas of psychology lend additional emphasis to these concerns. For instance, in clinical psychology, the ubiquity of observational designs and the focus on applied utility lend additional weight to the arguments for moving away from NHST and toward effect size estimation; when faced with large, highly powered epidemiological samples, or comparisons of novel versus existing treatments, significance per se might be less important than effect size and cost-benefit considerations (McShane et al. 2018a). Similarly, the heterogeneity of clinical populations and extremely low base rates of many clinical phenomena amplify concerns about the use of *p*-values in making inferences about findings, as $P(h|X)$ depends on the probabilities of the hypotheses being examined in the population at hand.

At the danger of overgeneralizing, it might be argued that clinical research designs are often more expensive than the types of designs that have dominated replicability discussions in other areas of psychology due to participant recruitment costs, large sample sizes, and other considerations (Tackett et al. 2017). Therefore, some approaches to studying replicability, such as collaborative preregistered replications, might be more difficult or expensive in clinical psychology than in other areas of psychology. This may increase the weight placed on meta-observational approaches to replicability, especially meta-analysis, which has always fundamentally concerned itself with issues of replicability and generalizability of results. The fields of clinical psychology and psychiatry are likely to benefit substantially from improvements in meta-analysis. These fields might achieve a better understanding of how inferences derived from traditional meta-analytic approaches coincide with, or can be better calibrated with, those of collaborative preregistration approaches, which might not always be feasible.

The expense of clinical designs also has implications for understanding how publication bias might occur in clinical psychology versus other fields. Although participants might be relatively expensive either in number or in recruitment costs, the measures are sometimes still relatively cheap to administer, which might give rise to so-called measure hacking, whereby results are manipulated by selectively reporting results for specific measures or variants of measures. The possibility of this phenomenon gives weight to the importance of developing and using well-validated and justified measures (Marshall et al. 2000). It also underscores the importance of understanding the psychometric structures of clinical constructs, as these structures point to potential similarities and points of convergent replication in designs. Even if findings cannot be replicated across participants, they might be replicated across different measures of the same constructs, or across

different constructs that operate equivalently within a theory (McShane et al. 2018b). Traditional psychometric considerations—such as those about reliability, convergent and discriminant validity, and psychometric structure—help define the contours of replicability expectations in terms of what measures a finding might or might not be expected to replicate across (Markon 2015). The next stages require more specific efforts to outline quantitative criteria—more likely in the form of possible intervals rather than precise point estimates—for replicability and psychometric standards from a construct validity perspective, which will more accurately capture much research in clinical psychology (Tackett et al. 2018).

The types of topics pursued in clinical research shape questions about replicability in other ways. Structural modeling is an important focus of research in clinical science, for example, but it is sometimes unclear how to define replicability of structural findings. Schieber and colleagues (2017) have shown that a given definition of structural equivalence can result in causally distinct structures being identified as equivalent, and that the set of structures identified as equivalent can vary depending on the definition used. This ambiguity derives from how different patterns in causal paths can be weighted: Studies that employ different weighting methods can come to different conclusions about the extent to which one structure is replicating another. Moreover, a given well-fitting structural model only represents one of many possible well-fitting models, and it should not be taken as confirmation of any one model as empirical truth (Tomarken & Waller 2003).

However, identifying replicable causal structures can also be difficult regardless of how replicability is defined. One emerging problem is the identification of replicable or generalizable details in causal structures in which a number of variables might be correlated due to multiple mechanisms simultaneously, including direct effects, indirect effects, dynamic processes, and unmeasured third variables. In much the same way that multicollinearity in regression creates challenges in distinguishing unique from shared components of correlated predictive effects, in causal networks or systems, it becomes difficult to isolate direct causal relationships between variables controlling for other direct, indirect, and latent effects in the broader causal system. Although some evidence suggests that broad features of causal networks might generalize across samples (Fried et al. 2018), other research has suggested that specific causal pathways, isolated from the broader structural features, are often poorly replicable (Bulteel et al. 2018, Forbes et al. 2017). These difficulties in the identification of generalizable structural pathways create challenges in the interpretation of causal systems above and beyond broad superordinate features. Research is needed to identify mechanisms for increasing the statistical power with which unique causal pathways can be identified.

LEVERAGING IMPLEMENTATION SCIENCE TO IMPROVE CREDIBILITY IN CLINICAL PSYCHOLOGICAL RESEARCH AND BEYOND

In addition to considering areas where the clinical psychological evidence base may be weak as well as ways to bring the open science/replicability conversation to clinical psychology, it is time to think more broadly about the ways in which bringing clinical psychology to the conversation can help shape and improve the field. One highly salient example is the area of intervention/implementation science, which provides a broad framework aimed at better understanding how to get people to change—that is, the ways in which individuals (and groups/organizations) initially adopt new strategies, and how these adaptive changes can be maintained. In the following section, we discuss in greater depth the ways in which the open science/replicability conversation may benefit from leveraging the existing evidence base on implementation science.

Diffusion and Dissemination of Best Practices Are Likely to Be Ineffective

Several lines of evidence suggest that there is a substantial gap between the optimal and the actual practice of research methods in psychological science. First, although some tutorials on research methods are among the most highly cited manuscripts in the field (Sharpe 2013, Sternberg 1992), most articles introducing new statistical methods are not highly cited even years after their publication (Altman & Goodman 1994). The majority of articles in applied psychology cite one or zero quantitative articles (Mills et al. 2010). Despite the popularity of tutorials, applied researchers exhibit fundamental misunderstandings of methodological principles and statistical ideas, such as *p*-values and NHSTs (Nickerson 2000), Cronbach's alpha (Schmitt 1996, Sijtsma 2009), and confidence intervals (Belia et al. 2005, Cumming et al. 2004); similarly, they often misunderstand the assumptions and interpretation of regression (Williams et al. 2013) and analyses of covariance (Miller & Chapman 2001, Westfall & Yarkoni 2016). Indeed, statistical myths are common enough to have books dedicated to correcting them (e.g., Lance & Vandenberg 2008).

Instead of applying statistical thinking to data analyses (Chance 2002), researchers rely on heuristics and rules of thumb to guide many decisions about data analysis (what have been called mindless rituals) (see Gigerenzer 2004, Gigerenzer & Gaissmaier 2011), regardless of the evidence in support of such heuristics or the degree to which they have been shown to be misleading (Greenland et al. 2016). For example, researchers have frequently used rules of thumb to determine sample size (Cohen 1990), model fit (Marsh et al. 2004), or test reliability (Sijtsma 2009). Gigerenzer (2004) described the mindless ritual of NHST in which researchers begin with a null hypothesis (such as no correlation or no mean differences), fail to specify the predictions of an alternative hypothesis, and use $p < 0.05$ to reject the null hypothesis and accept whatever version of their hypothesis they favor. Recent survey data indicated that, when psychology researchers used rules of thumb to estimate the necessary sample size to examine a typically sized effect, they routinely overestimated power and underestimated the required sample size (Bakker et al. 2016). Rituals, rules, and magic numbers can be helpful checks against common errors in human judgment (Kahneman & Tversky 1974), but when they are overused in a research context that largely ignores developments in the quantitative literature, they can become quickly outdated.

There are frequent errors in the application and reporting of statistical models. Several studies have estimated that between 10% and 20% of the psychological literature reports incorrect significance tests (Bakker & Wicherts 2011, Berle & Starcevic 2007, García-Berthou & Alcaraz 2004, Nuijten et al. 2016), frequently in favor of the researcher's hypotheses (Bakker & Wicherts 2011). Another recent review estimated that nearly half of the examined articles in social psychology report means or standard deviations that are inconsistent with the sample size (Brown & Heathers 2017). The methodological literature is rife with reviews of common errors or misapplications of research methods. Researchers struggle to properly adjust for covariates (Lord 1967, Miller & Chapman 2001, Westfall & Yarkoni 2016), to use power analysis to inform study design (Cohen 1962, Sedlmeier & Gigerenzer 1989), to establish construct validity (Campbell & Fiske 1959, Fiske & Campbell 1992), to test for mediation (Shrout & Bolger 2002, Zhao et al. 2010), to test exploratory factor models (Fabrigar et al. 1999), and to judge the fit of structural equation models (Jackson et al. 2009, McNeish et al. 2018). In short, if a method of analysis is popular, it is easy to find a review of frequent reporting and analysis errors in its application.

The gap between what is optimal scientific practice and what is standard practice likely arises in part because there is no systematic model for the implementation of best practices in research methods. To date, efforts to improve scientific practice have focused on diffusion, dissemination, and to a lesser extent, implementation. Diffusion, or the passive delivery of information (such as publishing articles in methodology journals), is likely to be the least effective method, given how

few methodological papers are cited in substantive articles (Mills et al. 2010). Dissemination, or targeted knowledge delivery, may be achieved through methodological tutorials for applied audiences (i.e., Grimm 2007, King et al. 2018), which are highly cited (Sharpe 2013), or in workshops and preconferences. These relatively passive methods of increasing knowledge may inadvertently contribute to the problems with the application or interpretation of statistical models described above. Because innovation is highly incentivized, it may be that new statistical methods are rapidly adopted with little attention to the marginal improvement in inferences they might provide compared to existing methods or to the degree to which they increase the risk of error. For example, Bauer (2007) described the rapid adoption of growth mixture models despite the numerous problems with their application and interpretation (such as misidentification of classes and dependency of class solutions on study design) (see King et al. 2018, Sher et al. 2011).

Finally, some structural efforts at implementation (or purposeful and targeted efforts at changing practices) have been made, whereby funding agencies, practice organizations (such as the APA or the Association for Psychological Science), or journals have provided recommendations or requirements around the reporting of scientific studies. For example, the APA Publication Manual provides standards for reporting many basic statistics (Am. Psychol. Assoc. 2010), and editorial boards have attempted to mandate or incentivize statistical reporting practices (Finch et al. 2004, Ioannidis 2018, Kidwell et al. 2016). However, without systematic change at the organizational level, structural efforts will only last as long as the reformers are in charge (Finch et al. 2004).

Perhaps the solution would be to improve graduate and postgraduate training in research methods. However, evidence suggests that there is an increasing gap between the training that doctoral students receive and the research methods that are commonly applied in current research. The typical graduate statistical training sequence has been repeatedly noted to cover the basics of experimental and correlational analysis, requiring a mean of 1.2 semesters of statistics training, and to do a poor job of covering more advanced topics (such as latent variables, measurement, or quasi-experimental designs) (Aiken et al. 1990, 2008). Although there is a proliferation of research methods workshops, there is no evidence that they actually improve participants' methodological skills in a way that substantively enhances their research practices. Adult and professional education research suggests that the impact of workshops is limited (Lyon et al. 2011). Passive learning is well known to be ineffective at producing changes in practice or skill mastery (Beidas & Kendall 2010, Herschell et al. 2010). Evidence suggests that learning is influenced not only by the specific practices of teaching, but also by the multiple contexts in which learning occurs (Bransford et al. 2000), such that optimal learning in a new method will occur by building on existing knowledge and practicing the new skills in the context in which they will be applied (Lovett & Greenhouse 2000). Coupled with the evidence of a shortage of quantitative psychologists (Aiken et al. 2008, Golinski & Cribbie 2009), this increases the likelihood that applied researchers will be called upon to use or evaluate research methods for which they lack expertise.

Leveraging the Advent of Implementation Science in Health Care for Open Science Reform

The notion of a gap between typical practice and optimal practice has been widely studied in other fields, perhaps most notably in the area of health care (Chambers et al. 2013). Studies of medical interventions suggest that it can take decades for evidence-based practices (EBPs) to appear in typical practice (Contopoulos-Ioannidis et al. 2008, Morris et al. 2011). For example, one study estimated that around 10% of therapists used manuals for EBPs in community mental health settings (Becker et al. 2013). Other work has suggested that although many therapists apply elements of EBPs in their standard practice, these are used in doses too low to have

therapeutic effect (Garland et al. 2010). This suggests that when typical individuals obtain treatment, either they are unlikely to receive an evidence-based treatment (Zima et al. 2005) or what they do receive will be ineffective. This is at least in part because the evidence base is developed by therapists with low caseloads who are closely supervised and treat homogeneous classes of patients, whereas clinicians in the community have heterogeneous and high caseloads and receive low levels of supervision (Weisz et al. 1995). It is difficult for therapists to choose the appropriate EBP for each case from among the hundreds that are potentially available, and it is likely impossible for any single therapist to master a variety of EBPs for all clients (Chorpita et al. 2007).

Implementation science is a relatively new field in health care that offers frameworks, methods, and outcomes that can be usefully applied to enhance the quality of research methods. The implementation science perspective is grounded in the evidence that the diffusion and dissemination of research evidence and best practices do little to change behaviors or systems that maintain the status quo, at least in part because they only attempt to increase knowledge about better practices, which only exert one influence on behavior change (Joyce & Showers 2002). Indeed, the implementation science perspective recognizes that implementation is multiply determined by characteristics of the intervention, of the individual, and of the interpersonal, organizational, and macro-level context (Damschroder et al. 2009). Successful implementation is characterized not only by changes in the final outcome (such as more reliable and valid research findings or improved patient health outcomes), but also by specific outcomes of the implementation strategies, such as increases in fidelity, widespread adoption, and increases in perceived acceptability, feasibility, and appropriateness (Lewis et al. 2015, Proctor et al. 2011). Successful implementation is achieved by using strategies that target a variety of implementation outcomes at multiple levels (Powell et al. 2015).

The principles of implementation science may be readily applied to the improvement of research methods in psychology. First, the multilevel determinants perspective recognizes that there are multiple systems and stakeholders at many levels that influence the implementation of effective research practices. Implementation will be more effective if applied researchers understand and believe that the new method is valuable, that they can apply and interpret the method themselves, that the method meets their needs and resources, and that their peers (other researchers, editors, reviewers, and funding agencies) expect that method to be applied. Implementation will be more widespread in contexts (e.g., research groups and programs, journals, professional organizations) that expect, support, and reward the implementation of a method and that have policies and incentives that promote better practices.

It will be critical to consider the degree to which the characteristics of a new method, such as its adaptability, complexity, design, relevance for given problems, and cost, influence implementation. The widespread adoption of macros, spreadsheets, and online tools for common statistical procedures (Dawson 2014, Preacher et al. 2006) highlights the importance of reducing complexity and cost as a way of increasing implementation. Kirk (1996) noted that some statistical indicators (such as R^2 as a measure of effect size in regression) were more commonly included in statistical output than others. Thus, implementation will be improved if software packages can present output in a way that conforms to best reporting practices for psychologists while discouraging poor practices.

It will be important to measure the outcomes of implementation efforts to understand their successes and failures (Lewis et al. 2015, Proctor et al. 2011). Effectiveness (i.e., whether or not the method solves a specific problem well) has long been the central focus of methodological research. However, effective interventions may not be implemented if they are not viewed by stakeholders as acceptable, appropriate, and feasible (Crable et al. 2018). Implementation success should also

take into account the degree to which the methods are adopted, penetrate research in a given field, and are used sustainably with fidelity. Theoretically, implementation outcomes are the precursor to the effectiveness of interventions in practice (Proctor et al. 2011). If implementation efforts are not successful, a quantitative intervention may be effective in that it produces appropriate and accurate answers, but it may not be viewed as acceptable, feasible, or sustainable, which might lead to low adoption; alternatively, the intervention may be widespread but conducted with poor fidelity, which would impair its effectiveness. In other words, successful implementation is defined by the widespread, appropriate, and accurate use of a method.

Finally, there is a wide variety of strategies that can be borrowed from successful implementation efforts in health care (Powell et al. 2015). Formal implementation blueprints can be used to guide implementation efforts and to identify change targets, timelines, and measures of progress or success. Organizational efforts can work to incentivize best practices—an example would be the badge system that is used to recognize research that conforms to open science practices (Kidwell et al. 2016)—or to develop best practice guidelines that can be used to guide the work of applied researchers (Proctor et al. 2009). Some work has suggested, for example, that providing guidelines for reporting the results of RCTs was related to improvements in statistical reporting (Turner et al. 2012). An emerging area of interest is how the design and usability of interventions and implementation strategies may impede or facilitate widespread implementation (Lyon & Koerner 2016). Finally, it will be critical to design and test effective methods of training, supervision, and consultation that are based on current best practices (Dorsey et al. 2013, Lyon et al. 2011).

TOWARD A MORE TRANSPARENT AND REPRODUCIBLE CLINICAL PSYCHOLOGICAL SCIENCE

In the short period of time—several years—since we published an article questioning the absence of clinical psychology from the replicability conversation (Tackett et al. 2017), substantial strides have been made toward reform efforts by the clinical psychology community. Clinical psychology has now seen the emergence of badges promoting open science and transparency in two field-specific journals; a special issue on the topic is forthcoming in the *Journal of Abnormal Psychology*; discussion panels and symposia have emerged at field-specific and field-wide conferences (e.g., the Association for Psychological Science convention); and an initial effort at establishing a database of interested clinical psychologists was met with resounding success. It is clear that a sector of the clinical psychology community is ready and willing to tackle these problems head-on, but it is also clear that clinical psychology as a subfield is lagging behind other areas of psychological science in these efforts.

This situation places us in a critical position at the current time, and we need more engagement from the clinical psychology community across many levels—from examining and charting the potential weak points in our literature base, to defining and operationalizing replication on both theoretical and statistical levels, to applying (and modifying) existing tools to promote a more open, transparent, and reproducible science. There are also great opportunities to consider the many ways in which we, as a field, have already been tackling these issues, even if not specifically under an open science framework, and to leverage our existing knowledge and expertise to accelerate and facilitate the open science reform movement writ large.

As clinical psychology continues to engage with this conversation, we will undoubtedly encounter new challenges. There are concerns discussed somewhat informally that those advancing the open science and reform movement may not be open-minded, judicious, or considerate. This is a larger challenge than we can address here—certainly, a movement that is largely predicated on criticizing an established structure will come across as antagonistic at times. Nonetheless,

this does not mean that high levels of antagonism are necessary or even particularly helpful in advancing these efforts. Perhaps this is a domain where the increased involvement of individuals with clinical training might be of some utility to the broader movement.

There are other tensions currently, for example in approaches to the peer-review and editorial process. If editors or reviewers are overly critical of theories or results that contradict their own while being overly accommodating to those that they favor, this might point to an existing bias in the peer-review system. A transparent author may inadvertently provide extra ammunition for reviewers to identify potential flaws in the methods or results of the study (regardless of their verisimilitude). One potential solution to this issue is RRs because after the initial peer review of the study introduction and methods, reviewers and editors may not rescind their decision to publish the article on the basis of the results obtained, effectively eliminating that source of bias. Other forms of preregistration may also protect against reviewer prejudice. When methods and analytical strategies are prespecified and time-stamped in a registration document, potential post hoc reviewer criticisms are considered exploratory and bear the burden of (registered) empirical investigation to present a strong case against final results, rather than simply requesting authors use the reviewer's preferred form of analysis or collect more data. Beyond registration, other solutions may be employed, such as signed reviews, open (but anonymous) reviews, the use of preprints for feedback from a wider scientific audience, and monitoring for this sort of reviewer behavior by editors. None of these approaches is a panacea; each has benefits and drawbacks, and some of these are discussed in detail elsewhere (Ross-Hellauer et al. 2017). These and other issues are actively being discussed in the open science community. Clinical psychologists have an opportunity to weigh in and to be part of working through some of these potentially thorny obstacles.

In sum, we hope that we have provided some acceleration to the ongoing conversation about open science, transparency, and replicability within the field of clinical psychological science. It is our assessment that clinical psychology has many strengths to leverage, including aspects of our previous methods and practices that have served us well, as well as expertise and knowledge that will benefit the open science and reform movement more broadly. However, there remains much work to be done, and with a small minority of clinical psychologists actively engaging in these efforts, it is clear that this movement has not yet taken hold in our subfield. We call out to our clinical colleagues to join us in these challenging efforts and to facilitate a movement that will result in better science and a more trustworthy evidence base.

SUMMARY POINTS

1. Although clinical psychological scientists have been slow to engage in the replicability conversation, their involvement has increased in recent years.
2. Identifying weak points in the clinical psychology evidence base is a critical need; problematic areas likely include research bases that are reliant on small sample sizes (e.g., in clinical neuroscience and intervention research), gaps between diagnostic assessment in the field and in research settings, and generalizability of randomized controlled trials to field and community settings.
3. A number of problematic methods and practices likely impair the replicability of clinical psychological findings, including low sample size and power, diagnostic unreliability,

publication bias, the use of questionable research practices, and variable adherence to registration guidelines (e.g., for intervention trials).

4. Areas to consider in a replication context include the use of meta-analysis (which has its historical roots in the field of clinical psychology), the best way to define and operationalize replication in clinical psychology, and the extent to which current statistical concerns (e.g., the problematic use of null hypothesis significance testing) generalize to clinical psychological research.
5. Many current proposals for improvement and reform can be adapted for clinical psychological research, including the use of open materials, open data, (pre)registration, Registered Reports, and multisite collaborations for both replications and original studies.
6. The clinical psychological expertise and knowledge base can also be leveraged to facilitate advances in the open science movement; one example is to consider existing work from intervention/implementation science as a tool to enhance the adoption and maintenance of open science methods.

FUTURE ISSUES

1. Explicit efforts are needed to identify weak spots in the clinical science research base. This mapping may occur by examining research topics with evidence for limited replicability or generalizability or by identifying the fields most associated with known problems that limit replicability (e.g., small sample sizes, reduced power, problematic measurement, publication bias, etc.).
2. A better understanding is needed of known lab-to-field gaps, such as the diagnostic discrepancy between research and field settings and the problematic generalizability of randomized controlled trials to the field and to community settings.
3. As clinical psychologists increasingly engage with current open science and reform tools, these solutions will most likely need to be adapted to better suit the types of research designs, samples, and methodological/analytic approaches that clinical psychologists typically use. This tailoring will most expediently follow from early attempts at adapting existing tools and solutions in current research efforts.
4. Clinical psychologists need to directly engage with persistent questions around how to define a replication attempt and a replication outcome (e.g., success/failure), tackling this question at both the theoretical and the methodological level.
5. Clinical psychologists should leverage their existing knowledge base to enhance open science and reform efforts, for example by applying principles from implementation science to these domains.
6. The replication movement is in need of greater involvement and engagement by clinical psychological researchers. These reforms will take effort, and they will benefit most from input coming from researchers across different domains of clinical psychological science.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Abram SV, DeYoung CG. 2017. Using personality neuroscience to study personality disorder. *Personal. Disord. Theory Res. Treat.* 8:2–13

Aiken LS, West SG, Millsap RE. 2008. Doctoral training in statistics, measurement, and methodology in psychology: replication and extension of Aiken, West, Sechrest, and Reno's 1990 survey of PhD programs in North America. *Am. Psychol.* 63(1):32–50

Aiken LS, West SG, Sechrest L, Reno RR, Roediger HL III, et al. 1990. Graduate training in statistics, methodology, and measurement in psychology: a survey of PhD programs in North America. *Am. Psychol.* 45(6):721–34

Altman DG, Goodman SN. 1994. Transfer of technology from statistical journals to the biomedical literature. *JAMA* 272(2):129–32

Am. Psychol. Assoc. 2010. *Publication Manual of the American Psychological Association*. Washington, DC: Am. Psychol. Assoc. 6th ed.

Am. Psychol. Assoc. 2017. Ethical principles of psychologists and code of conduct. *American Psychological Association*. **https://www.apa.org/ethics/code/**

APA Publ. Commun. Board Work. Group J. Article Rep. Stand. 2008. Reporting standards for research in psychology: Why do we need them? What might they be? *Am. Psychol.* 63(9):839–51

Bakker M, Hartgerink CHJ, Wicherts JM, van der Maas HLJ. 2016. Researchers' intuitions about power in psychological research. *Psychol. Sci.* 27:1069–77

Bakker M, van Dijk A, Wicherts JM. 2012. The rules of the game called psychological science. *Perspect. Psychol. Sci.* 7(6):543–54

Bakker M, Wicherts JM. 2011. The (mis)reporting of statistical results in psychology journals. *Behav. Res. Methods* 43(3):666–78

Bauer DJ. 2007. Observations on the use of growth mixture models in psychological research. *Multivar. Behav. Res.* 42(4):757–86

Becker EM, Smith AM, Jensen-Doss A. 2013. Who's using treatment manuals? A national survey of practicing therapists. *Behav. Res. Ther.* 51(10):706–10

Beidas R, Kendall P. 2010. Training therapists in evidence-based practice: a critical review of studies from a systems contextual perspective. *Clin. Psychol. Sci. Pract.* 17:1–30

Belia S, Fidler F, Williams J, Cumming G. 2005. Researchers misunderstand confidence intervals and standard error bars. *Psychol. Methods* 10(4):389–96

Berle D, Starcevic V. 2007. Inconsistencies between reported test statistics and *p*-values in two psychiatry journals. *Int. J. Methods Psychiatr. Res.* 16(4):202–7

Blohowiak BB, Cohoon J, de-Wit L, Eich E, Farach FJ, et al. 2018. *Badges to acknowledge open practices*. Open Sci. Framew. Proj., Cent. Open Sci., Charlottesville, VA. **https://osf.io/tvyxz/**

Bransford JD, Brown AL, Cocking RR. 2000. *How People Learn: Brain, Mind, Experience, and School*. Washington, DC: Natl. Acad. Press. Expand. ed.

Brown NJL, Heathers JAJ. 2017. The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Soc. Psychol. Personal. Sci.* 8(4):363–69

Bulteel K, Mestdagh M, Tuerlinckx F, Ceulemans E. 2018. VAR(1) based models do not outpredict AR(1) models in current psychological applications. *Psychol. Methods* 23(4):740–56

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, et al. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14(5):365–76

Calamia M, Markon K, Tranel D. 2012. Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *Clin. Neuropsychol.* 26(4):543–70

Campbell DT, Fiske DW. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56(2):81–105
Chambers DA, Glasgow RE, Stange KC. 2013. The dynamic sustainability framework: addressing the paradox of sustainment amid ongoing change. *Implement. Sci.* 8(1):117
Chance BL. 2002. Components of statistical thinking and implications for instruction and assessment. *J. Stat. Educ.* 10:3
Chartier CR, McCarthy RJ. 2018. *StudySwap: a platform for inter-lab replication, collaboration, and research resource exchange*. Open Sci. Framew. Proj., Cent. Open Sci., Charlottesville, VA. **https://osf.io/rd37b/**
Chorpita BF, Becker KD, Daleiden EL. 2007. Understanding the common elements of evidence-based practice: misconceptions and clinical examples. *J. Am. Acad. Child Adolesc. Psychiatry* 46(5):647–52
Cohen J. 1962. The statistical power of abnormal-social psychological research: a review. *J. Abnorm. Soc. Psychol.* 65(3):145–53
Cohen J. 1990. Things I have learned (so far). *Am. Psychol.* 45(12):1304–12
Colquhoun D. 2017. The reproducibility of research and the misinterpretation of *p*-values. *R. Soc. Open Sci.* 4(12):171085
Contopoulos-Ioannidis DG, Alexiou G, Gouvias TC, Ioannidis JP. 2008. Life cycle of translational research for medical interventions. *Science* 321:1298–99
Conway CC, Tackett JL, Skodol AE. 2017. Are personality disorders assessed in young people? *Am. J. Psychiatry* 174(10):1000–1
Crable EL, Biancarelli D, Walkey AJ, Allen CG, Proctor EK, Drainoni ML. 2018. Standardizing an approach to the evaluation of implementation science proposals. *Implement. Sci.* 13(1):71
Cremers HR, Wager TD, Yarkoni T. 2017. The relation between statistical power and inference in fMRI. *PLOS ONE* 12:e0184923
Cronbach LJ, Meehl PE. 1955. Construct validity in psychological tests. *Psychol. Bull.* 52:281–302
Cuijpers P. 2016. Are all psychotherapies equally effective in the treatment of adult depression? The lack of statistical power of comparative outcome studies. *Evid.-Based Ment. Health* 19(2):39–42
Cuijpers P, Sijbrandij M, Koole S, Huibers M, Berking M, Andersson G. 2014. Psychological treatment of generalized anxiety disorder: a meta-analysis. *Clin. Psychol. Rev.* 34(2):130–40
Cuijpers P, Smit F, Bohlmeijer E, Hollon SD, Andersson G. 2010. Efficacy of cognitive-behavioural therapy and other psychological treatments for adult depression: meta-analytic study of publication bias. *Br. J. Psychiatry* 196(3):173–78
Culverhouse RC, Saccone NL, Horton AC, Ma Y, Anstey KJ, et al. 2018. Collaborative meta-analysis finds no evidence of a strong interaction between stress and 5-HTTLPR genotype contributing to the development of depression. *Mol. Psychiatry* 23(1):133–42
Cumming G, Williams J, Fidler F. 2004. Replication and researchers' understanding of confidence intervals and standard error bars. *Underst. Stat.* 3(4):299–311
Cybulski L, Mayo-Wilson E, Grant S. 2016. Improving transparency and reproducibility through registration: the status of intervention trials published in clinical psychology journals. *J. Consult. Clin. Psychol.* 84(9):753–67
Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. 2009. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implement. Sci.* 4(1):50
Dawson JF. 2014. Moderation in management research: what, why, when, and how. *J. Bus. Psychol.* 29(1):1–19
De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, et al. 2004. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *N. Eng. J. Med.* 351(12):1250–51
Dorsey S, Pullmann MD, Deblinger E, Berliner L, Kerns SE, et al. 2013. Improving practice in community-based settings: a randomized trial of supervision—study protocol. *Implement. Sci.* 8:89
Dorsey S, Pullmann MD, Kerns SEU, Jungbluth N, Meza R, et al. 2017. The juggling act of supervision in community mental health: implications for supporting evidence-based treatment. *Adm. Policy Ment. Health* 44(6):838–52

Driessen E, Hollon SD, Bockting CL, Cuijpers P, Turner EH. 2015. Does publication bias inflate the apparent efficacy of psychological treatment for major depressive disorder? A systematic review and meta-analysis of US National Institutes of Health–funded trials. *PLOS ONE* 10(9):e0137864

Duncan LE, Keller MC. 2011. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *Am. J. Psychiatry* 168(10):1041–49

Dwan K, Gamble C, Williamson PR, Kirkham JJ, Rep. Bias Group. 2013. Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review. *PLOS ONE* 8(7):e66844

Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Methods* 79(3):272–99

Fanelli D. 2010. Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLOS ONE* 5(4):e10271

Finch S, Cumming G, Williams J, Palmer L, Griffith E, et al. 2004. Reform of statistical inference in psychology: the case of *Memory & Cognition*. *Behav. Res. Methods Instrum. Comput.* 36(2):312–24

Finkel EJ, Eastwick PW, Reis HT. 2015. Best research practices in psychology: illustrating epistemological and pragmatic considerations with the case of relationship science. *J. Personal. Soc. Psychol.* 108(2):275–97

Fiske DW, Campbell DT. 1992. Citations do not solve problems. *Psychol. Bull.* 112(3):393–95

Forbes MK, Wright AG, Markon KE, Krueger RF. 2017. Evidence that psychopathology symptom networks have limited replicability. *J. Abnorm. Psychol.* 126(7):969–88

Fraley RC, Vazire S. 2014. The N-pact factor: evaluating the quality of empirical journals with respect to sample size and statistical power. *PLOS ONE* 9(10):e109019

Frances AJ, Widiger T. 2012. Psychiatric diagnosis: lessons from the DSM-IV past and cautions for the DSM-5 future. *Annu. Rev. Clin. Psychol.* 8(1):109–30

Frank MC, Bergelson E, Bergmann C, Cristia A, Floccia C, et al. 2017. A collaborative approach to infant research: promoting reproducibility, best practices, and theory-building. *Infancy* 22(4):421–35

Freedman R, Lewis DA, Michels R, Pine DS, Schultz SK, et al. 2013. The initial field trials of DSM-5: new blooms and old thorns. *Am. J. Psychiatry* 170:1–5

Fried EI, Eidhof MB, Palic S, Costantini G, Huisman-van Dijk HM, et al. 2018. Replicability and generalizability of posttraumatic stress disorder (PTSD) networks: a cross-cultural multisite study of PTSD symptoms in four trauma patient samples. *Clin. Psychol. Sci.* 6(3):335–51

Garb HN. 2005. Clinical judgment and decision making. *Annu. Rev. Clin. Psychol.* 1:67–89

García-Berthou E, Alcaraz C. 2004. Incongruence between test statistics and *P* values in medical papers. *BMC Med. Res. Methodol.* 4:13

Garland AF, Brookman-Frazee L, Hurlburt MS, Accurso EC, Zoffness RJ, et al. 2010. Mental health care for children with disruptive behavior problems: a view inside therapists' offices. *Psychiatr. Serv.* 61(8):788–95

Gigerenzer G. 2004. Mindless statistics. *J. Socio-Econ.* 33(5):587–606

Gigerenzer G, Gaissmaier W. 2011. Heuristic decision making. *Annu. Rev. Psychol.* 62:451–82

Glass GV. 1976. Primary, secondary, and meta-analysis of research. *Educ. Res.* 5(10):3–8

Glass GV. 2015. Meta-analysis at middle age: a personal history. *Res. Synth. Methods* 6(3):221–31

Golinski C, Cribbie RA. 2009. The expanding role of quantitative methodologists in advancing psychology. *Can. Psychol.* 50(2):83–90

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, et al. 2016. Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *Eur. J. Epidemiol.* 31(4):337–50

Grimm KJ. 2007. Multivariate longitudinal methods for studying developmental relationships between depression and academic achievement. *Int. J. Behav. Dev.* 31(4):328–39

Gurevitch J, Koricheva J, Nakagawa S, Stewart G. 2018. Meta-analysis and the science of research synthesis. *Nature* 555(7695):175–82

Herschell AD, Kolko DJ, Baumann BL, Davis AC. 2010. The role of therapist training in the implementation of psychosocial treatments: a review and critique with recommendations. *Clin. Psychol. Rev.* 30(4):448–66

Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. 2009. Publication bias in clinical trials due to significance of trial results. *Cochrane Database Syst. Rev.* 2009:MR000006

Ioannidis JPA. 2018. The proposal to lower *P* value thresholds to .005. *JAMA* 319(14):1429–30

Jackson DL, Gillaspy JA, Purc-Stephenson R. 2009. Reporting practices in confirmatory factor analysis: an overview and some recommendations. *Psychol. Methods* 14(1):6–23
Jenkins MM, Youngstrom EA, Washburn JJ, Youngstrom JK. 2011. Evidence-based strategies improve assessment of pediatric bipolar disorder by community practitioners. *Prof. Psychol. Res. Pract.* 42(2):121–29
Joel S, Eastwick PW, Finkel EJ. 2018. Open sharing of data on close relationships and other sensitive social psychological topic: challenges, tools, and future directions. *Adv. Methods Pract. Psychol. Sci.* 1:86–94
John LK, Loewenstein G, Prelec D. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* 23(5):524–32
Jones KD. 2012. A critique of the DSM-5 field trials. *J. Nerv. Ment. Dis.* 200(6):517–19
Joyce B, Showers B. 2002. *Student Achievement Through Staff Development*. Alexandria, VA: Assoc. Superv. Curric. Dev.
Kahneman D, Tversky A. 1974. Judgment under uncertainty: heuristics and biases. *Science* 185(4157):1124–31
Kidwell MC, Lazarević LB, Baranski E, Hardwicke TE, Piechowski S, et al. 2016. Badges to acknowledge open practices: a simple, low-cost, effective method for increasing transparency. *PLOS Biol.* 14(5):e1002456
Kim H, Eaton NR. 2015. The hierarchical structure of common mental disorders: connecting multiple levels of comorbidity, bifactor models, and predictive validity. *J. Abnorm. Psychol.* 124(4):1064–78
King KM, Littlefield AK, McCabe CJ, Mills KL, Flournoy J, Chassin L. 2018. Longitudinal modeling in developmental neuroimaging research: common challenges, and solutions from developmental psychology. *Dev. Cog. Neurosci.* 33:54–72
Kirk RE. 1996. Practical significance: a concept whose time has come. *Educ. Psychol. Meas.* 56(5):746–59
Klein RA, Ratliff KA, Vianello M, Adams RB, Bahník Š, et al. 2014. Investigating variation in replicability: a "Many Labs" replication project. *Soc. Psychol.* 45(3):142–52
Lance CE, Vandenberg RJ. 2008. *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences*. New York: Taylor & Francis
Lash TL, Collin LJ, Van Dyke ME. 2018. The replication crisis in epidemiology: snowball, snow job, or winter solstice? *Curr. Epidemiol. Rep.* 5(2):175–83
Leichsenring F, Abbass A, Hilsenroth MJ, Leweke F, Luyten P, et al. 2017. Biases in research: risk factors for non-replicability in psychotherapy and pharmacotherapy research. *Psychol. Med.* 47(6):1000–11
Lewis CC, Fischer S, Weiner BJ, Stanick C, Kim M, Martinez RG. 2015. Outcomes for implementation science: an enhanced systematic review of instruments using evidence-based rating criteria. *Implement. Sci.* 10(1):155
Lilienfeld SO. 2017. Psychology's replication crisis and the grant culture: righting the ship. *Perspect. Psychol. Sci.* 12:660–64
Lord FM. 1967. A paradox in the interpretation of group comparisons. *Psychol. Bull.* 68(3):304–5
Lovett MC, Greenhouse JB. 2000. Applying cognitive theory to statistics instruction. *Am. Stat.* 54(3):196–206
Lyon AR, Koerner K. 2016. User-centered design for psychosocial intervention development and implementation. *Clin. Psychol. Sci. Pract.* 23(2):180–200
Lyon AR, Stirman SW, Kerns SEU, Bruns EJ. 2011. Developing the mental health workforce: review and application of training approaches from multiple disciplines. *Adm. Policy Ment. Health* 38(4):238–53
MacCallum RC, Zhang S, Preacher KJ, Rucker DD. 2002. On the practice of dichotomization of quantitative variables. *Psychol. Methods* 7:19–40
Markon KE. 2015. Ontology, measurement, and other fundamental problems of scientific inference. *Psychol. Inq.* 26(3):259–62
Marsh HW, Hau KT, Wen Z. 2004. In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's 1999 findings. *Struct. Equ. Model.* 11(3):320–41
Marshall M, Lockwood A, Bradley C, Adams C, Joy C, Fenton M. 2000. Unpublished rating scales: a major source of bias in randomised controlled trials of treatments for schizophrenia. *Br. J. Psychiatry* 176(3):249–52
McNeish D, An J, Hancock GR. 2018. The thorny relation between measurement quality and fit index cutoffs in latent variable models. *J. Personal. Assess.* 100(1):43–52

McShane BB, Gal D, Gelman A, Robert C, Tackett JL. 2018a. Abandon statistical significance. *Am. Stat.* In press

McShane BB, Tackett JL, Bockenholt U, Gelman A. 2018b. Large scale replication projects in contemporary psychological research. *Am. Stat.* In press

Meehl PE. 1990. Why summaries of research on psychological theories are often uninterpretable. *Psychol. Rep.* 66(1):195–244

Miller GA, Chapman JP. 2001. Misunderstanding analysis of covariance. *J. Abnorm. Psychol.* 110(1):40–48

Mills L, Abdulla E, Cribbie RA. 2010. Quantitative methodology research: Is it on psychologists' reading lists? *Tutor. Quant. Methods Psychol.* 6(2):52–60

Miloyan B, Fried EI. 2017. A reassessment of the relationship between depression and all-cause mortality in 3,604,005 participants from 293 studies. *World Psychiatry* 16:219–20

Moberg CA, Humphreys K. 2017. Exclusion criteria in treatment research on alcohol, tobacco and illicit drug use disorders: a review and critical analysis. *Drug Alcohol Rev.* 36(3):378–88

Moffitt TE, Caspi A, Taylor A, Kokaua J, Milne BJ, et al. 2010. How common are common mental disorders? Evidence that lifetime prevalence rates are doubled by prospective versus retrospective ascertainment. *Psychol. Med.* 40(6):899–909

Morris ZS, Wooding S, Grant J. 2011. The answer is 17 years, what is the question: understanding time lags in translational research. *J. R. Soc. Med.* 104(12):510–20

Moshontz H, Campbell L, Ebersole CR, IJzerman H, Urry HL, et al. 2018. The Psychological Science Accelerator: advancing psychology through a distributed collaborative network. *Adv. Methods Pract. Psychol. Sci.* 1:501–15

Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, et al. 2017. A manifesto for reproducible science. *Nat. Hum. Behav.* 1(1):21

Munder T, Brütsch O, Leonhart R, Gerger H, Barth J. 2013. Researcher allegiance in psychotherapy outcome research: an overview of reviews. *Clin. Psychol. Rev.* 33(4):501–11

Nickerson RS. 2000. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol. Methods* 5(2):241–301

Niemeyer H, Musch J, Pietrowsky R. 2013. Publication bias in meta-analyses of the efficacy of psychotherapeutic interventions for depression. *J. Consult. Clin. Psychol.* 81(1):58–74

NIH (Natl. Inst. Health). 2008. Revised policy on enhancing public access to archived publications resulting from NIH-funded research. *Natl. Inst. Health*. **https://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-033.html**

NIH (Natl. Inst. Health). 2015. Implementing rigor and transparency in NIH & AHRQ research grant applications. *Natl. Inst. Health*. **https://grants.nih.gov/grants/guide/notice-files/NOT-OD-16-011.html**

Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. 2018. The preregistration revolution. *PNAS* 115(11):2600–6

Nuijten MB, Hartgerink CHJ, van Assen MALM, Epskamp S, Wicherts JM. 2016. The prevalence of statistical reporting errors in psychology (1985–2013). *Behav. Res. Methods* 48(4):1205–26

Open Sci. Collab. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716

Pearson K. 1904. Report on certain enteric fever inoculation statistics. *BMJ* 3:1243–46

Polanczyk GV, Salum GA, Sugaya LS, Caye A, Rohde LA. 2015. Annual research review: a meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. *J. Child Psychol. Psychiatry* 56(3):345–65

Powell BJ, Waltz TJ, Chinman MJ, Damschroder LJ, Smith JL, et al. 2015. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. *Implement. Sci.* 10(1):21

Preacher KJ, Curran PJ, Bauer DJ. 2006. Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *J. Educ. Behav. Stat.* 31(4):437–48

Proctor EK, Landsverk J, Aarons G, Chambers D, Glisson C, Mittman B. 2009. Implementation research in mental health services: an emerging science with conceptual, methodological, and training challenges. *Adm. Policy Ment. Health* 36(1):24–34

Proctor EK, Silmere H, Raghavan R, Hovmand P, Aarons G, et al. 2011. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Adm. Policy Ment. Health* 38(2):65–76

Rapport MD, Orban SA, Kofler MJ, Friedman LM. 2013. Do programs designed to train working memory, other executive functions, and attention benefit children with ADHD? A meta-analytic review of cognitive, academic, and behavioral outcomes. *Clin. Psychol. Rev.* 33(8):1237–52

Reardon KW, Mercadante EJ, Tackett JL. 2017. The assessment of personality disorder: methodological, developmental, and contextual considerations. *Curr. Opin. Psychol.* 21:39–43

Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, et al. 2013. DSM-5 field trials in the United States and Canada, part II: test-retest reliability of selected categorical diagnoses. *Am. J. Psychiatry* 170(1):59–70

Ross-Hellauer T, Deppe A, Schmidt B. 2017. Survey on open peer review: attitudes and experience amongst editors, authors, and reviewers. *PLOS ONE* 12(12):e0189311

Rothwell PM. 2005. External validity of randomised controlled trials: "To whom do the results of this trial apply?" *Lancet* 365(9453):82–93

Schieber TA, Carpi L, Díaz-Guilera A, Pardalos PM, Masoller C, Ravetti MG. 2017. Quantification of network structural dissimilarities. *Nat. Commun.* 8:13928

Schmitt N. 1996. Uses and abuses of coefficient alpha. *Psychol. Assess.* 8(4):350–53

Sedlmeier P, Gigerenzer G. 1989. Do studies of statistical power have an effect on the power of studies? *Psychol. Bull.* 105(2):309–16

Shadish WR, Cook T, Campbell D. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin

Shadish WR, Lecy JD. 2015. The meta-analytic big bang. *Res. Synth. Methods* 6(3):246–64

Sharpe D. 2013. Why the resistance to statistical innovations? Bridging the communication gap. *Psychol. Methods* 18(4):572–82

Sher KJ, Jackson KM, Steinley D. 2011. Alcohol use trajectories and the ubiquitous cat's cradle: cause for concern? *J. Abnorm. Psychol.* 120:322–35

Shrout PE, Bolger N. 2002. Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychol. Methods* 7(4):422–45

Sijtsma K. 2009. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 74(1):107–20

Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22(11):1359–66

Smith ML, Glass GV. 1977. Meta-analysis of psychotherapy outcome studies. *Am. Psychol.* 32(9):752–60

Sohn D. 1996. Publication bias and the evaluation of psychotherapy efficacy in reviews of the research literature. *Clin. Psychol. Rev.* 16(2):147–56

Soto CJ. 2018. *How replicable are links between personality traits and consequential life outcomes? The Life Outcomes of Personality Replication Project*. Presented at the European Conference on Personality, Zadar, Croat., July 17–21. **https://osf.io/9dta7/**

Staines GL, Cleland CM. 2007. Bias in meta-analytic estimates of the absolute efficacy of psychotherapy. *Rev. Gen. Psychol.* 11(4):329–47

Sternberg RJ. 1992. *Psychological Bulletin's* top 10 "hit parade." *Psychol. Bull.* 112(3):387–88

Stirman SW, DeRubeis RJ, Crits-Christoph P, Brody PE. 2003. Are samples in randomized controlled trials of psychotherapy representative of community outpatients? A new methodology and initial findings. *J. Consult. Clin. Psychol.* 71(6):963–72

Stirman SW, DeRubeis RJ, Crits-Christoph P, Rothman A. 2005. Can the randomized controlled trial literature generalize to nonrandomized patients? *J. Consult. Clin. Psychol.* 73(1):127–35

Stuart EA, Bradshaw CP, Leaf PJ. 2015. Assessing the generalizability of randomized trial results to target populations. *Prev. Sci.* 16(3):475–85

Tackett JL, Brandes CM, Reardon KW. 2018. Leveraging the Open Science Framework in clinical psychological assessment research. *Psychol. Assess.* In press

Tackett JL, Lilienfeld SO, Patrick CJ, Johnson SL, Krueger RF, et al. 2017. It's time to broaden the replicability conversation: thoughts for and from clinical psychological science. *Perspect. Psychol. Sci.* 12(5):742–56

Tackett JL, McShane BB. 2018. Conceptualizing and evaluating replication across domains of behavioral research. *Behav. Brain Sci.* 41:e152

Tomarken AJ, Waller NG. 2003. Potential problems with "well fitting" models. *J. Abnorm. Psychol.* 112:578–98

Turner L, Shamseer L, Altman DG, Schulz KF, Moher D. 2012. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Syst. Rev.* 1:60

van der Lem R, de Wever WWH, van der Wee NJA, van Veen T, Cuijpers P, Zitman FG. 2012. The generalizability of psychotherapy efficacy trials in major depressive disorder: an analysis of the influence of patient selection in efficacy trials on symptom outcome in daily practice. *BMC Psychiatry* 12:192

Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K. 2013. The WU-Minn Human Connectome Project: an overview. *NeuroImage* 80:62–79

von Wolff A, Jansen M, Hölzel LP, Westphal A, Härter M, Kriston L. 2014. Generalizability of findings from efficacy trials for chronic depression: an analysis of eligibility criteria. *Psychiatr. Serv.* 65(7):897–904

Walsh CG, Xia W, Li M, Denny JC, Harris PA, Malin BA. 2018. Enabling open-science initiatives in clinical psychology and psychiatry without sacrificing patients' privacy: current practices and future challenges. *Adv. Methods Pract. Psychol. Sci.* 1(1):104–14

Wampold BE, Mondin GW, Moody M, Stich F, Benson K, Ahn H. 1997. A meta-analysis of outcome studies comparing bona fide psychotherapies: empirically, "all must have prizes." *Psychol. Bull.* 122(3):203–15

Weisz JR, Donenberg GR, Han SS, Weiss B. 1995. Bridging the gap between laboratory and clinic in child and adolescent psychotherapy. *J. Consult. Clin. Psychol.* 63(5):688–701

Westfall J, Yarkoni T. 2016. Statistically controlling for confounding constructs is harder than you think. *PLOS ONE* 11(3):e0152719

Wicherts JM, Borsboom D, Kats J, Molenaar D. 2006. The poor availability of psychological research data for reanalysis. *Am. Psychol.* 61(7):726–28

Williams M, Grajales CAG, Kurkiewicz D. 2013. Assumptions of multiple regression: correcting two misconceptions. *Pract. Assess. Res. Eval.* 18(11):1–14

Yates BT, Taub J. 2003. Assessing the costs, benefits, cost-effectiveness, and cost-benefit of psychological assessment: We should, we can, and here's how. *Psychol. Assess.* 15(4):478–95

Zhao X, Lynch JG Jr., Chen Q. 2010. Reconsidering Baron and Kenny: myths and truths about mediation analysis. *J. Consum. Res.* 37(2):197–206

Zima BT, Hurlburt MS, Knapp P, Ladd H, Tang L, et al. 2005. Quality of publicly-funded outpatient specialty mental health care for common childhood psychiatric disorders in California. *J. Am. Acad. Child Adolesc. Psychiatry* 44(2):130–44

Zwaan RA, Etz A, Lucas RE, Donnellan MB. 2018. Making replication mainstream. *Behav. Brain Sci.* 41:e120