It's Time to Broaden the Replicability Conversation: Thoughts for and from Clinical Psychological Science

Jennifer L. Tackett[1], Scott O. Lilienfeld[2], Christopher J. Patrick[3], Sheri L. Johnson[4], Robert F. Krueger[5], Joshua D. Miller[6], Thomas F. Oltmanns[7], & Patrick E. Shrout[8]

[1]Northwestern University, Department of Psychology

[2]Emory University, Department of Psychology

[3]Florida State University, Department of Psychology

[4]University of California – Berkeley, Department of Psychology

[5]University of Minnesota, Department of Psychology

[6]University of Georgia, Department of Psychology

[7]Washington University, Department of Psychology

[8]New York University, Department of Psychology

Manuscript in press, *Perspectives on Psychological Science*

Author Note

Jennifer L. Tackett, Department of Psychology, Northwestern University; Scott O. Lilienfeld, Department of Psychology, Emory University; Sheri L. Johnson, Department of Psychology, University of California – Berkeley; Robert F. Krueger, Department of Psychology, University of Minnesota; Joshua D. Miller, Department of Psychology, University of Georgia; Thomas F. Oltmanns, Department of Psychology, Washington University; Christopher J. Patrick, Department of Psychology, Florida State University; Patrick E. Shrout, Department of Psychology, New York University.

Correspondence concerning this article should be addressed to Jennifer L. Tackett, Department of Psychology, Northwestern University, Evanston, IL, 60208. E-mail: jennifer.tackett@northwestern.edu

Abstract

Psychology is in the early stages of examining a crisis of replicability stemming from several high-profile failures to replicate studies in experimental psychology. This important conversation has largely been focused on social psychology, with some active participation from cognitive psychology. Nevertheless, several other major domains of psychological science – including clinical science – have remained insulated from this discussion. The goals of this article are to (a) examine why clinical psychology and allied fields, such as counseling and school psychology, have not been central participants in the replicability conversation, (b) review concerns and recommendations that are less (or more) applicable to or appropriate for research in clinical psychology and allied fields, and (c) generate take-home messages for scholars and consumers of the literature in clinical psychology and allied fields, as well as reviewers, editors, and colleagues from other areas of psychological science.

It's Time to Broaden the Replicability Conversation: Thoughts for and from Clinical Psychological Science

It is both an exciting and a challenging time to be a psychological scientist. The field is undergoing a process of deep self-examination, with a long-overdue, critical look at our scientific practices and the credibility of our knowledge base (Lilienfeld & Waldman, in press; Pashler & Wagenmakers, 2012). This process has been fueled by a series of high-profile replication failures in psychological science as well as in other fields, such as medicine and molecular genetics (Hagger & Chatzisarantis, 2016; Prasad, Citu, & Ioannidis, 2012), several high-profile cases of fraud in psychology, and growing concerns regarding the pervasiveness of questionable research practices (QRPs). These developments have prompted an important and evolving conversation concerning the rationale for our commonly accepted scientific practices and the robustness of findings in psychological science.

Although the replicability conversation affects all of psychological science, we contend that most ongoing efforts have been overly insular. Much of the discourse has been limited to laboratory-based experiments with convenience samples typical of those in social and cognitive psychology, to the exclusion of the many other domains of psychological research (e.g., Finkel, Eastwick, & Reis, 2015), including those in the broad domain of individual differences psychology. Our goal in this article is straightforward: To broaden the replicability conversation to encompass another key domain of psychological science, namely, clinical science, and to engage clinical scientists in the conversation. Clinical science's large-scale exclusion from the replicability conversation is perhaps surprising given that clinical psychologists frequently collaborate with scholars in domains of psychiatry, neuroimaging, and molecular genetics in which replication difficulties have been observed and widely

discussed (e.g., Button et al., 2013; Ioannidis & Trikalinos, 2005). Our discussion and recommendations in the current manuscript are most pertinent to such disciplines as clinical psychology, counseling psychology, school psychology, psychiatry, social work, psychiatric nursing, epidemiology, and other areas intersecting with applied mental health sciences.

At the same time, a key goal of our article is to highlight—for the entire field of psychological science—how insularity in the ongoing conversation is impeding efforts at field-wide reform. Proposed solutions must be sufficiently comprehensive and nuanced to apply to the enormously diverse spectrum of research that psychological scientists conduct. Hence, we draw readers' attention to underappreciated ways in which the ongoing replicability conversation is missing the mark, and underscore the relevance of these omissions for our collective ability to move forward toward a more reproducible science.

A Brief Summary of the Replicability Crisis in Psychology

The replicability crisis has been discussed at length in both peer-reviewed (e.g., Ioannidis, 2005; Spellman, 2015) and non-peer-reviewed (e.g., Hilgard, 2016) outlets in recent years; hence, our review of this topic is necessarily brief and limited to key concepts that provide a context for readers. This crisis has been brought to the fore by two prominent events: (1) widely publicized replication failures, particularly in social and cognitive psychology (e.g., Donnellan, Lucas, & Cesario, 2015; Klein et al., 2014; Open Science Collaboration, 2015) and (2) new attention to questionable research practices (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). A third issue to which we accord less attention is outright fraud (i.e., falsification and fabrication of data), which has always been deplored and is not considered further in this article.

Failures to replicate occur when a statistical test of an association fails to emerge as significant in analyses of new data collected to parallel data that previously yielded a significant statistical effect. As a notable example of this, the Open Science Collaboration (Nosek et al., 2015) arranged for 100 associations sampled from three top tier journals—comprised of social and cognitive psychological studies—to be tested in replication studies, and found that only 36% of the associations were statistically significant in the new studies. The literature is already replete with commentaries suggesting reasons for this low rate of replication, including statistical factors (Maxwell, Lau & Howard, 2015), contextual considerations (Van Bavel et al, 2016), and variation in study designs and measures (Stroebe & Strack, 2014). The Open Science authors showed that replication success was positively related to the size of the original effect and negatively related to the degree to which the original result was deemed surprising. Moreover, the effect sizes found in the Open Science Collaboration report were about half the magnitude of those reported in the original articles, corroborating a broader phenomenon known as the decline effect (Schooler, 2011).

An influential article by Simons, Nelson and Simonsohn (2011) initiated a somewhat separate conversation, delineating questionable research practices (QRPs) that can dramatically increase the likelihood of a false positive result. Some examples include (a) the use of multiple alternate variables, (b) adaptive stopping rules when collecting data, (c) multiple alternate statistical analyses and (d) exclusion or refinement of study conditions or sample strata. All of these practices are justifiable in certain contexts, but become problematic when they are performed in an exploratory fashion but later reported as confirmatory (predicted). Related terms, reflecting the many decisions researchers make when working through data analyses (Ioannidis, 2005), include data-contingent analysis, “the garden of forking paths” (Gelman & Loken, 2013), and “researcher

flexibility in data analysis". The statistical tests stemming from QRPs and "p-hacking" (that is, efforts intended to inflate opportunities to observe a *p*-value below the designated threshold for significance) tend to be biased, and the Type I error rate often increases dramatically.

This recent discussion also emphasizes a lack of transparency in our scientific culture. For example, hypothesizing after the results are known ("HARKing", Kerr, 1998) refers to presenting hypotheses as "a priori" when they were actually developed alongside or following data analysis. HARKing presumably happens commonly, in part, as a response to our field's preference for findings framed as being generated in the context of justification rather than the context of discovery (see Reichenbach, 1938). Many QRPs and the incentive structures that inadvertently promote them, such as intense publication and grant pressures, come into play in individual studies, such that problems created by the use of small sample sizes may be compounded by, and also increase the likelihood of, p-hacking (Bakker, van Dijk, & Wicherts, 2012). The collective result of these practices, in a cultural milieu that has almost surely reinforced them, is an unacceptably large proportion of psychological findings that fail to replicate (Nosek et al., 2015).

One salutary consequence of these sobering realizations within the field has been a renewed interest in improving the conduct of psychological research. Indeed, many efforts have been dedicated to better understanding the root causes of these problems while moving toward improvement and field-wide change (e.g., Finkel, Eastwick, & Reis, 2015; Spellman, 2015; Vazire, 2014). Researchers have offered a number of overarching recommendations in the spirit of improving the replicability of psychological findings (e.g., the Transparency and Openness Promotion, or TOP, guidelines; Nosek et al., 2015; Cumming, 2014). We will discuss many of these recommendations in

subsequent sections, as we highlight the extent to which these recommendations are important for increasing reproducibility of clinical psychological science, but also pose problems for clinical psychological science and are in need of revision in the psychological science community.

Before doing so, however, we examine potential reasons why clinical psychological science has not been an active player in the broader conversation on replicability in psychological science—reasons which, themselves, may inform the replicability conversation if better understood. We then review five recommendations for improving replicability of research results from the viewpoint of clinical science. For each of these recommendations, we discuss barriers to implementation from a clinical science perspective and broader lessons for the reproducibility conversation in psychological science. Then, we offer suggestions and potential alternatives for clinical science research.

Why has Clinical Psychology Been Largely Missing from the Conversation?

It's Not Me, It's You

Much of the conversation concerning psychology's replication crisis has taken place in broad-themed journals such as *Perspectives on Psychological Science*, *Psychological Science*, and *American Psychologist*, although it has also been an active focus of discussion in social and personality outlets (e.g., *Journal of Personality and Social Psychology, Social Psychology, Social Psychological and Personality Science*, and *European Journal of Personality*). Perhaps surprisingly, however, meta-scientific articles on replicability have yet to appear in major clinical psychology outlets. We argue that remaining outside the conversation is a missed opportunity both for clinical scientists and for the broader field of psychological science.

For one thing, the practical implications of false positives in clinical science are substantial. Though in this article we focus primarily on implications of the replicability debate for clinical research, this debate raises serious questions concerning how the profession should move forward with empirically supported assessment and intervention techniques, given that many published research findings bearing on these methods may be false or at least overestimated in magnitude. Indeed, the current criteria for empirically supported therapies (ESTs) –psychological interventions that are deemed to work for specific disorders – require positive results from only two well-conducted independently replicated randomized control trials, regardless of the number of negative results. In light of the broader replicability debates in psychology, the current criteria for ESTs would appear to be woefully lenient, because the presence of only two positive outcome studies, especially in the context of negative outcome studies, could well be attributable to chance (see Tolin, McKay, Forman, Klonsky, & Thombs, 2015, for a more rigorous approach to operationalizing ESTs that takes into account the full body of research evidence). In light of this concern and many others that we will discuss, one could argue that replicability is, or should be, of particular concern for clinical science. With this goal in mind, we next consider potential reasons why clinical science has remained largely insulated from these recent debates.

Large-scale failed replication efforts are not unique to psychological science, a point made especially salient in recent replication failures of biomedical research by large pharmaceutical (e.g., Bayer) and biotechnology (e.g., Amgen) firms (Baker, 2016; Owens, 2011). Nevertheless, the conversations regarding how to change scientific practices and conventions have not been distributed evenly across scientific disciplines and sub-disciplines. In psychology, there has been considerably more attention to these issues in social psychology (e.g., Finkel et al., 2015) and in general forums such as APS

annual conventions. Clinical, counseling and health psychologists have been far less involved in these conversations and less involved in proposing field-wide remedies.

Methodological Approaches in Clinical Psychological Science

Perhaps some subfields such as clinical science have been less involved in the replicability conversation because their methods and analytic practices are less problematic than in other domains. Indeed, there are initial indications that replicability problems may be especially marked in certain domains, such as social psychology (Bakker et al., 2012; Fraley & Vazire, 2014; John et al., 2012; Open Science Collaboration 2015; Pritschet, Powell, & Horne, 2016). Specifically, researchers have identified differences among subfields—often pointing to particular problems in social psychology—in the average sample sizes employed in published research (Fraley & Vazire, 2014), in the reported use of certain QRPs (John et al., 2012), in replication success rate (Open Science Collaboration, 2015), and in the use of flexible data interpretation such as characterizing results as "marginally significant" or "trend level" (Pritschet et al., 2016). We are not entirely persuaded by this argument, however. For one thing, this conclusion could in part reflect "detection bias": certain domains may show less replicable findings than others merely because they have subjected their findings to higher scrutiny. Moreover, many articles in clinical psychology journals examine what are traditionally regarded as relatively rare conditions, such as dissociative identity disorder or trichotillomania, and thus are marked by small sample sizes and accompanying low statistical power (Davison & Lazarus, 2007). For these reasons, it seems unlikely that clinical psychology would be immune from serious replicability problems of its own. Furthermore, rather than focusing on which subfields have "won" or "lost" these battles, it will ultimately be more productive to delineate the major sources of replication difficulties across subfields.

Clinical psychological science is a diverse area that encompasses both experimental and correlational approaches, with the latter being more common than the former. Experiments are often carried out to test manualized interventions designed to help clients with a given disorder. Ideally, these interventions exert strong effects; small effects leave clients with substantial residual pathology and affiliated distress. This state of affairs is in contrast to most social or cognitive manipulations, in which small and perhaps even tiny effects may be of considerable theoretical importance. This distinction is relevant to the current conversation given that replication is more likely when effects are large in magnitude.

Furthermore, the issues of replication in correlational designs differ from those in experimental designs. Clinical scientists typically focus on individual difference variables, particularly those related to psychopathology, such as psychiatric symptoms or signs, psychiatric diagnoses, trauma, early childhood adversity, genetic vulnerability, or personality traits. These differences cannot be manipulated, pragmatically or ethically. As a consequence, substantial reliance on correlational approaches in clinical scientific research is inevitable. Some have argued that correlational research often prioritizes measurement and estimation of effect sizes over the presence or absence of statistically significant effects (e.g., Funder, 2016). Such methodological differences, including relative emphasis on correlational versus experimental approaches, may partially explain the presence of field-based differences in QRPs that adversely impact the replicability of findings.

Methodological differences in clinical psychological science, compared with other subfields of psychology, may make the determination of "what counts" as a successful or failed replication even more challenging. Disciplines prioritizing a dichotomous approach to significance testing may regard the question of a "successful"

versus “unsuccessful” replication as clearer; in contrast, when estimation of effect sizes is the primary goal, new criteria must be adopted to determine when a replicated effect replicates. In addition, “focal hypotheses”, which are often the emphasis of pre-registration efforts and tests of replicability/reproducibility, are less typical in correlational and descriptive research. It is often unclear how to apply an entire set of specific hypotheses to a large correlation matrix, or even whether such an approach is advisable (e.g., Rothman, 1990; Perneger, 1998).

Also, differences in measurement across disciplines may influence replicability rates. Clinical psychological science, consistent with correlational psychological research more broadly, typically relies on aggregate measures, such as questionnaires or interviews, and may be more likely to use multiple measurements and approaches such as latent variable modeling, all of which may have a bottom-line impact on enhancing replicability (e.g., Funder, 2016).

Clinical Scientists are Accustomed to Imperfection and Hard-to-Collect Data

Most clinical psychological scientists are accustomed to noisy and imperfect data, especially because they typically study individuals with psychopathology, many of whom can be quite variable in their laboratory task performance from session to session, and in some cases even within sessions (e.g., see Williams, Strauss, Hultsch, Hunter, & Tannock, 2007, for evidence of this problem among participants with attention-deficit hyperactivity disorder). Differences in the identification and recruitment of participants are also relevant. Sampling procedures in clinical research are probably more variable than in most other areas of psychological science. Clinical scientists are often faced with the task of identifying clinical participants who come from idiosyncratic settings, such as a specific clinic to which the researcher has access, and typically cannot rely heavily on broad convenience samples, such as undergraduates or community samples recruited via

Amazon's Mechanical Turk. As such, clinical scientists often assume that undetected moderators, such as severity and length of illness, medication status, and treatment responsiveness, are operating in virtually any sample they collect.

Along with this layer of complexity comes the presumption that these major differences in sampling will lead effect sizes to fluctuate from study to study. This perspective may increase overall comfort with publishing small and nonsignificant effects. All of this may have contributed to an expectation that any one study, from any one laboratory, is highly noisy and affected by a variety of undetected moderators. Hence, there may be less of an expectation of a "definitive" finding. This appreciation may in turn affect the research culture in such a way as to de-incentivize overstatements of any single study, and it may further decrease the value placed on any single finding. At the same time, this "messiness" in data renders it challenging to operationalize what qualifies as a replication success or failure, as this concept surely falls along a continuum from direct to conceptual replications, or using other terminology, literal to constructive replications (e.g., Lykken, 1968).

Related to the nature of the data clinical scientists typically work with, some of the QRPs and concerns about researcher flexibility in data collection and analysis discussed widely in the literature (e.g., Simmons et al., 2011) appear to be less relevant to clinical psychological science. For example, aiming to recruit a certain number of participants per "cell" (i.e., a specific category or condition that is balanced across other categories or conditions in a study) is relevant primarily for experimental work, particularly in research designs in which random assignment is possible. In contrast, this approach is less applicable to a substantial majority of clinical psychological studies, especially in the psychopathology domain. Clinical research that relies on cells is often much more constrained by recruitment and sampling than is research using more

accessible or larger samples. Similarly, the concept of "data termination", that is, flexibility in deciding when to cease running new participants in a study, is often constrained by availability of participants in clinical science research, more so than ideal statistical practices or principles.

Another potential result of these sampling constraints and differences is a research culture that less often incorporates multiple studies into a single manuscript, resulting in less of a "culture of replication" within or across clinical psychology laboratories. For one thing, collecting hard-to-acquire data, such as measures from low base-rate psychopathological samples, poses a practical impediment to generating separate complementary datasets in a reasonable amount of time. For another, a focus on effect size estimation and measurement leads researchers to prioritize collecting data from the largest possible sample size. Thus, in clinical science research, it may be more typical to pool all available data into a single analysis, a practice demonstrated to boost the likelihood of replicability (Bakker et al., 2012; Schimmack, 2012). One potential benefit of these cultural differences is that it may decrease the likelihood of publication bias stemming from the "file drawer effect" (Rosenthal, 1979).

A further feature of clinical science is that it has historically held close ties to community psychology (e.g., Sarason, 1981), with its emphasis on variable social contexts. When clinical scientists examine large epidemiological surveys that contain observations from different ethnic groups, regions, social classes, and ages, they are typically open to variation across associations that are interpreted as attributable to meaningful moderators rather as than failures to replicate. The consideration of random effects in population surveys and in meta-analyses of clinical trials is not unusual in clinical science.

In sum, important differences in clinical psychological science practices (e.g., more reliance on correlational designs and effects sizes, greater comfort with messy or noisy data) may account for its absence in the overall replicability conversation, while pointing to potential differences in reproducibility between areas of research in psychological science. Many of these conjectures could be studied empirically, and may spur important meta-science topics moving forward (e.g., on delineation of an exploratory-confirmatory spectrum of research in practice).

Where Do We Go From Here?

Challenges and Recommendations for Clinical Science

We next discuss five major domains of recommendations for enhancing replicability in clinical psychological science; we refer to these broader categories as "goals." We consider potential challenges that clinical scientists face in implementing each of these goals, and provide suggestions for addressing these challenges (see Table 1), including some that have been presented elsewhere (e.g., Asendropf et al., 2013). Rather than reiterate what others have written about these topics, we focus on the extent of "misfit" these recommendations present for clinical science research, alongside constructive recommendations for reducing this misfit.

Goal #1: Reduce the Prevalence of QRPs

One pressing need is to reduce the prevalence of QRPs in clinical psychological science, as well as psychological science more broadly. Clinical scientists would benefit from the active and productive conversation that is well underway regarding the nature, extent, and impact of QRPs on the replicability of psychological research findings.

To facilitate this endeavor, we need to ascertain the prevalence of QRPs in clinical psychological science. For example, it might be common to select one symptom scale

over another as an outcome if the former indicates a larger effect. In contrast, it might be less common to do interim analyses on the desired effect as the data are being collected.

In conjunction with this endeavor, efforts need to be made to alter the current incentive structure, which inadvertently promotes and rewards QRPs (Patrick & Hacjak, 2016). For example, the grant-driven culture in clinical psychology, especially for researchers operating in research-intensive universities and medical schools, may be something of a recipe for QRPs. If investigators know that their continued grant funding is potentially contingent on obtaining positive results, the temptation to engage in p-hacking and HARKing may prove difficult to resist (Lilienfeld, in press).

Barriers for clinical science. An impediment to reducing QRPs is the awareness that many, if not most, of us have at times engaged in research practices that have made our work less reproducible. In accepting that these practices are widespread and have been integrated into our broader research culture, we can adopt a forward-thinking perspective that can reduce their prevalence. The multi-faceted nature of clinical psychological science suggests that a nuanced and discipline-specific approach is needed to appreciate which QRPs affect clinical science most markedly, and the best ways to curtail them. Certain areas of clinical psychological science may encounter distinctive barriers to achieving this goal. For example, many clinical psychological scientists collect extremely large datasets intended to serve as archival data for many years, even decades. The conceptualization of QRPs in this context is more challenging than in a standard experimental design, with a single sample collected to test a primary research question.

Potential steps for clinical science. A first general recommendation for all clinical psychological scientists is to keep up-to-date regarding research on p-hacking, HARKing, and other QRPs, so as to better understand how their decision-making and

investigative approaches may affect the replicability of their findings. Similarly, it would behoove clinical scientists to use this knowledge to inform behavior as reviewers, editors, mentors, and co-authors, seeking to improve the quality and replicability of research in clinical psychological science. Many new online tools facilitate learning about various QRPs, and the extent to which they affect our own research programs. For example, it is easy to examine one's own evidence base with tools like www.shinyapps.org/apps/p-checker/ (Schönbrodt, 2015) – although such exercises also point out the limitations of current QRP definitions when applied to other areas of research, such as clinical science (e.g., some tests rely exclusively on focal hypothesis tests, which as noted earlier are less common in correlational research). In addition, it is important to be appropriately cautious about new tools that have not undergone extensive validation.

A second recommendation is to report all relevant dependent variables in our studies, either in the text of the article itself or, for the growing number of journals that permit it, in supplemental materials. In many studies, particularly large-scale projects, it is not feasible to report *every* dependent variable, but is certainly reasonable to report all those expected to relate to major dependent variables of interest. This reality raises questions concerning the classification of a failure to report all dependent measures in a study as a formal QRP (e.g., see John et al., 2012). Still, in keeping with the suggestion to report all dependent indicators relevant to our target construct, if more than one measure of depressive symptomatology was used in a study, all should be reported along with clear justification for which measures were used in the analyses and why. In addition, in such instances, a citation to the larger study should always be provided. This practice is much more feasible now with the advent of websites such as the Open Science Framework, which provide a means for archiving such information when a specific

publication does not exist. Similarly, results that rely on the use of covariates should document whether the results change when the covariates are excluded.

A third and particularly important recommendation is for clinical scientists to pre-register all study hypotheses to mitigate outcome reporting bias and minimize the odds of HARKing (see Lindsay, Simons, & Lilienfeld, in press, for guidelines) and to develop more of a culture of replication in their laboratories by making replicability issues a central feature of graduate mentoring and laboratory discussions. Although preregistration of hypotheses and analyses is not a panacea, it helps to minimize the likelihood of QRPs, as well as HARKing, by ensuring that investigators are reasonably explicit about which analyses they initially intended to perform.

A fourth recommendation is to engage in thoughtful discussion with fellow researchers and students about how *p*-hacking and QRPs can emerge in clinical science research. Inflation of researcher degrees of freedom in analyses and reporting can assume many forms. For example, certain QRPs may frequently arise with respect to unique sampling constraints or analytic decisions, such as researcher degrees of freedom in creating thresholds for factor loadings or correlation interpretations, decisions made in constraining structural equation models, or the development of study-specific coding schemes. They may also come into play when making difficult post-hoc decisions regarding which, if any, disordered participants to exclude from analyses on the basis of apparent inattention or misunderstanding of experimental instructions.

Considering these questions from different research traditions presents a number of challenges. While some commentaries have proposed clear demarcation between exploratory and confirmatory analyses as a solution, it can be difficult to differentiate where exploration ends and confirmation begins in projects where data collection is ongoing over the course of years or decades (Finkel et al, 2015; Gelman & Loken, 2013).

At the same time, however, the progressive-sequential approach used in smaller-scale research (i.e., with new studies building on findings from p studies) has been linked to a higher rather than lower prevalence of QRPs (Bakker et al., 2012). Thus, if recommendations are to move the field to more intensive data collection of larger samples, enhanced measurement, and more generalizable samples, some consideration of the exploratory/confirmatory continuum is needed.

Goal #2: Pre-registration and Open Data

Another increasingly common recommendation for improving the replicability of psychological science findings, which we endorse, is to (a) pre-register hypotheses, design, and analyses and (b) make collected data publicly available, or open access (e.g., Gelman & Loke, 2013; Jonas & Cesario, 2015; Lindsay et al., in press; Spellman, 2015). Some have suggested that pre-registration may be a comprehensive "meta-recommendation", mitigating a number of replicability problems simultaneously (e.g., Bishop, 2013; Chambers, 2014). The field has been answering this call, with new resources being made available to facilitate researchers. One salient example is the Open Science Framework, which provides researchers with a platform to preregister hypotheses, data collection, and data analysis, along with infrastructure to make de-identified data publicly available following the study (https://osf.io). Pre-registration is growing increasingly accessible and is far more flexible than many realize.

Barriers for clinical science. Still, several challenges to implementing preregistration, some of which are especially relevant to clinical science, are worth noting (see also Gelman & Loken, 2013). First, many clinical science studies rely on extensive, resource-intensive datasets. For example, descriptive psychopathology research relying on community-drawn participants often requires hundreds, if not thousands, of participants to yield adequate variance in the constructs of interest. Such participants can

be expensive to recruit and frequently undergo hours of structured clinical interviews, questionnaires, laboratory measures, and other tasks. The extensive intellectual efforts required to plan a study of this magnitude, acquire requisite funding, and organize and supervise the collection of such data may increase investigators' reluctance to making their data publicly available. If such open access were mandated, it could dissuade researchers from investing the necessary resources to embark on challenging data collection endeavors, which would impede scientific progress. Thus, as a field, clinical psychological science will need to afford thoughtful discussion to the proper balance between openness/transparency and investment of researcher resources.

A second challenge is that mid-course adjustments to sample procedures are not always foreseeable. For example, certain clinics may become more difficult to sample from than expected; recruitment of participants who meet diagnostic thresholds may be more challenging than expected; or new sampling and recruitment approaches may need to be introduced as a study proceeds to achieve an adequate overall sample size. Such often unanticipated factors, which are related to the nature of populations with which clinical scientists often work, may pose obstacles to pre-registering a specific sampling approach or alter subsequent analyses. Nevertheless, such issues are addressable through use of modified pre-registration schemes that permit flexibility in relation to unforeseen events and barriers.

A third challenge is that hypothesis development often occurs alongside data collection, analysis, and dissemination (Gelman & Loken, 2013). This complexity creates a barrier to full preregistration of study hypotheses and analyses, as researchers often do not know all hypotheses that will be tested *a priori*. In time-intensive studies where knowledge of a phenomenon evolves alongside ongoing collection of data, it may be unrealistic and scientifically inadvisable for researchers to test only those hypotheses

formulated prior to data collection. This iterative process between theory and data is a fundamental component of a growing science (Cronbach & Meehl, 1955; Tellegen & Waller, 2008) that needs to be accommodated through a more flexible pre-registration approach that permits hypotheses and analytic plans to be modified over time – with appropriate documentation, such as attachment of time-stamps to each change in hypotheses and analytic plan (see Lindsay et al., in press).

A fourth and final barrier to making data openly accessible or publicly available relates to the sensitive nature of data typically collected in clinical science studies. Personal health information, including details regarding stressful and even traumatic life events as well as lifetime diagnoses for a variety of mental health conditions, is sensitive and must be carefully protected. The sensitive nature of such data is exacerbated by the study of relatively low base-rate conditions, such as schizophrenia; higher base-rate clinical conditions, such as major depressive disorder, will still often present in only a minority of participants in any given study. As such, publicly available data of this sort are more likely to be difficult to fully de-identify, and the consequences of identifying participants potentially more consequential. For example, a dataset might contain only one 24-year old African American with schizophrenia, which could automatically identify this individual in a modestly sized outpatient clinic or psychiatric hospital.

Potential steps for clinical science. One recommendation for addressing issues of the foregoing types is for clinical science researchers to become more creative in adapting the basic principles underlying pre-registration and open access data. For example, researchers may wish to pre-register plans for the creation of large datasets prior to data collection, understanding that pre-registration of all hypotheses for such studies is often unrealistic. That is, certain aspects of pre-registration may still be applicable, although which aspects of each study can easily be pre-registered may differ

for clinical research. The key is that pre-registration can happen even *after* the data are already collected. However, this option places an onus on researchers to refrain from evaluating hypotheses prior to pre-registration, a temptation that may become stronger after data have been collected and are readily available on computers. Avoiding such temptation is necessary if pre-registration is a goal. Similarly, researchers must describe hypotheses as entirely confirmatory only if they have not yet conducted preliminary analyses testing these ideas.

As a step in this direction, existing pre-registration systems (e.g., https://aspredicted.org or https://osf.io/k5wns/) can be adapted for use in experimental and correlational psychopathology research. Modifications to these protocols should be documented, and become part of the larger conversation on pre-registration of psychological research (e.g. through published accounts of the necessary modifications for a specific study).

A third recommendation is that clinical scientists should do a better job of documenting possible moderators in published work. The often idiosyncratic characteristics of participant samples used in clinical science studies make it crucial to comprehensively report all aspects of sampling procedures, along with describing potential moderators that set boundary conditions on the results. Transparency concerning the range of measures used to test a hypothesis, and specification of which measures did or did not conform to hypotheses, are critical in this regard.

Finally, given existing limits to open access, clinical scientists should consider other ways to make their data available such as providing access to editors and reviewers when submitting work for publication. Explicit statements can be included in publications to indicate when this has occurred, along with a clear rationale for limiting the open access of such data. While clinical scientists should generally be willing to share their

data with other scientists who seek to verify or replicate their results, certain precautions such as confidentiality contracts may be warranted in these situations. The OSF website can accommodate data sharing under generic circumstances, and advances to OSF capabilities are rapidly occurring. With increased demand from certain stakeholders, such as clinical psychological scientists, it is likely that technology will adapt to our specific needs, including the ability to protect privatized data from more general public users.

Goal #3: Independent Replication

As noted earlier, a significant contributor to the replicability crisis has been a string of widely publicized failures to replicate findings from prominent experimental psychology studies (e.g., Donnellan, Lucas, & Cesario, 2015; Klein et al., 2014; Open Science Collaboration, 2015). These failed replications and the adverse publicity associated with them have aroused active debate in the field, with the conversation at times becoming contentious and occasionally heated (e.g., see Letzter, 2016). Although few would deny that independent replications are a worthy endeavor, the field continues to debate thorny questions such as the key characteristics of a "suitable" replication attempt, the appropriate researchers to carry out such attempts, the amount of time one should reasonably spend on replication efforts versus the production of "new" science, and best practices for publicizing results from replication studies.

Barriers for clinical science. The barriers to conducting independent replications of clinical science research overlap largely with those to increasing clinical sample sizes in general. Such issues as recruitment constraints when studying low base rate conditions, sampling variations in clinical populations, the evolution of psychiatric diagnoses and the occasional introductions of novel ones (e.g., hoarding disorder or excoriation disorder in DSM-5; American Psychiatric Association, 2013), and inadequate funding will also affect the feasibility of conducting independent replications. The resource-intensive

nature of collecting clinical samples or community-based samples large enough to ensure adequate variance in clinical constructs may further increase the reluctance to invest such extensive resources in replications, particularly in replicating the work of others. A further barrier to conducting independent replications of clinical science studies is the sensitive nature of data-sharing, as previously discussed. Independent replications are preferably performed with access to raw data from the original study, and the sensitive nature of clinical science data (e.g., health information, diagnostic data) may hinder replication efforts.

Potential steps for clinical science. We recommend that clinical science researchers consider creative new approaches to the task of replication that address the aforementioned obstacles, rather than following older replication standards adopted by researchers in other areas. For example, more effort should be invested in replicating findings across existing datasets when possible, rather than treating each new investigation involving a separate sample as a distinct contribution to the literature. Similarly, clinical scientists should make more concerted efforts to break down barriers toward data sharing and be more communal and open in sharing their data. Considering ways to revise existing incentive structures will be critical for promoting these behaviors. For example, datasets could be cited in order to credit investigators for their efforts in creating them. As a vehicle for this, the OSF website allows for cataloguing of specific datasets or projects, which can serve as a referent for citation purposes. Undoubtedly, large-scale institutional changes will be needed for academic departments to recognize and reward such efforts.

A third recommendation is to build a replication component into studies. For examples, researchers can seek to include previous measures and protocols in their future data collections, including measures both from their lab and others'. The desire to do

something "novel" must be balanced against a need for continuity and generalizability (e.g., Patrick & Hajcak, 2016). Every new study affords an opportunity to expand the sample size of previous projects or generate a replication of prior findings, or both.

Goal #4: Self-examination and replicability operationalization

An important aspect of improving the replicability of clinical science efforts involves carefully examining its practices, literature, and findings to identify areas in need of reform. For example, researchers should be more wary of extremely large effect sizes, particularly in combination with higher p-values, small samples, and/or lack of generalizability across research teams (e.g., Bakker et al., 2012). Through this process, methods for establishing replicability of clinical science can evolve as well.

Barriers for clinical science. One potential barrier to self-examination in clinical psychological science is that much research of this type is correlational or descriptive in nature, and thus not so clearly subject to concerns that have been raised about experimental research. Some differences between correlational and experimental research approaches that should be considered regarding issues of replicability include feasibility of data-pooling over multiple study approaches, and utilization of effect size estimates as opposed to dichotomous hypothesis-testing. Another barrier to self-examination in clinical science is that methodological constraints inherent to some of this work, such as small sample sizes in the case of rare clinical conditions, may restrict the ability to evaluate the robustness of findings. Sample-specific variance will exacerbate problems with generalizability in smaller samples, limiting overarching claims about the robustness of findings for certain clinical phenomena or with certain populations.

Potential steps for clinical science. We recommend that clinical scientists compare methodological features of experimental and correlational approaches in order to identify specific procedural differences in these approaches (e.g., an emphasis on

hypothesis-testing) as well as broader systemic differences (e.g., an emphasis on measurement and the use of latent variables) that exert effects on reproducibility. In addition, we recommend that clinical scientists attend to parameters such as average sample size in evaluating published bodies of work, consider evidence for publication bias, and be appropriately skeptical of large effects that have not been consistently replicated. This recommendation is intertwined with our suggestion to move toward more large-scale collaboration and replication efforts, particularly in content areas such as functional brain imaging research or descriptive studies of low base rate conditions, in which collecting larger samples are not feasible. Although external and practical constraints may limit sample sizes, this limitation must be factored into the confidence clinical scientists place in published findings and conclusions. In addition to publishing critical analyses of extant research findings that consider such issues, we call on clinical psychological scientists to report failed replications of findings or paradigms, and formulate conceptual articles focusing on steps that can be taken to improve replicability.

We further recommend that clinical scientists assemble expert groups and consortia who can work to establish disciplinary standards for the discipline, as well as subfields within it. One example is the Hierarchical Taxonomy of Psychopathology (HiTOP) consortium, consisting of researchers from clinical psychology, psychiatry, epidemiology, genetics, and beyond, joining together to address overarching concerns regarding the nosology of mental disorders. (Kotov et al., 2016). In addition, we call for clinical scientists to keep themselves updated on ongoing conversations regarding replicability occurring in other disciplines such as social and cognitive psychology, statistics, economics, public health, and medicine. Many of these discussions are taking place through non-traditional mechanisms such as online ‘blogs’, the Facebook PsychMAP group, and other social media outlets, highlighting the need for clinical

scientists to participate more actively in informal discussions occurring taking place outside of formal peer-reviewed contexts.

Goal #5: Increase Power

Another recommendation for enhancing replicability is to improve statistical power by boosting sample size, increasing measurement precision, decreasing error in variable quantification (e.g., through latent variable modeling), and increasing the use of within-subject research designs. These strategies are by no means new – having emerged in part from mid-20th century clinical science research (e.g., Cohen, 1962), including now-classic work on the efficacy of psychotherapy (Smith & Glass, 1977) – but they remain greatly underutilized (Marzalek, Barber, Kohlhart, & Cooper, 2011; Sedlmeier & Gigerenzer, 1989). Nevertheless, as attention to these issues has risen to the fore, journals are adapting their policies and reviewing standards to heighten expectations of sufficient power and adequate sample sizes.

Barriers for clinical science. Distinct challenges exist to optimizing statistical power in clinical science research. First, as previously noted, recruiting clinical participants can be considerably more difficult, time-consuming, and resource-intensive than using healthy or convenience samples. In addition, screening participants for diagnostic inclusion criteria is often a complex process that requires extensive and ongoing resources. For example, questions of the following types might be posed about persons responding to a recruitment ad for adults suffering from clinical depression: How many clinically depressed people in the local area never saw the ad, perhaps because they were too depressed to watch television or read the newspaper? Among depressed persons who saw the ad, how many were unable or unwilling to place a call? How do the responders differ from those who did not respond? Beyond this, some individuals responding to such an ad will be excluded from testing based on eligibility pre-screening

responses, and others who participate in testing will be excluded from analyses for diagnostic (e.g., not clinically depressed) or other reasons (e.g., lab equipment problems). In addition, relative to non-depressed persons, clinically depressed individuals will be less likely to return for follow-up appointments (Eaton, Anthony, Tepper, & Dryman, 1992), further limiting the effective sample size for multi-session studies. Selection factors of these types will severely constrain the final sample size and limit its representativeness; additionally, the amount of time and resources needed to arrive at the final sample is likely to be substantial. The implications for replicability should be obvious; studies that follow different recruitment strategies could easily produce discrepant results.

In addition, some investigators may elect to recruit samples of individuals with largely "pure" cases of a condition, namely, those who do not meet diagnostic criteria for another major mental disorder. Such exclusion criteria will tend to further diminish sample sizes. In fairness, there may occasionally be good reasons to recruit pure samples, as this strategy enhances internal validity. Nevertheless, this strategy is likely to come at a substantial cost in many cases. It limits not only sample size, but also external validity given that comorbidity tends to be far more often the rule than the exception in psychopathology research (First, 2005). As a consequence, researchers who routinely adopt this strategy may want to reconsider it given its likely effects on both statistical power and the generalizability of results.

Clinical science researchers also face a host of challenges when it comes to sampling and recruitment. For example, a researcher studying schizophrenia may be able to interview only individuals diagnosed with this disorder within a given geographic region. Most clinical science researchers are dependent on the cooperation and accessibility of community stakeholders in facilitating and promoting access to

participants, and/or rely on recruitment from a given treatment center. Those stakeholders and treatment centers may have differential access to more or less severe subpopulations. These issues of geographical and recruitment narrowness limit the ability to access the full range of individuals with a given condition; this form of sampling error again often limits generalizability. Thus, sampling factors for clinical science are likely to present impediments to the replicability of findings.

The foregoing barriers are magnified by a further barrier: a paucity of funding. Given the additional resources needed to recruit clinical or clinically relevant samples, inadequate funding poses a more serious obstacle to research in this area compared to other areas in which convenience samples can be relied upon for data collection. A fourth barrier to increasing statistical power and robustness of results in clinical science research comes from the continually evolving nature of psychopathology constructs such as those introduced by the DSM, ICD, and the NIMH Research Domain Criteria (an effort to shift focus from diagnostic categories to neurally-informed dimensions undergirding psychopathology symptom clusters; see Inset al., 2010). Mental disorders, along with their putative indicators, are necessarily constructs; they are hypothetical attributes that cannot be perfectly ("operationally") defined by any set of observable referents (Cronbach & Meehl, 1955). Accordingly, diagnostic systems for conditions of this type are continually evolving as scientific knowledge progresses and as diagnostic conventions change. An enormous amount of resources can be devoted to a longitudinal study of psychopathology in which the criteria for the key diagnoses change mid-way through, rendering the previously collected data potentially less generalizable to data collected under the new system. Similarly, if widely accepted diagnostic constructs are of questionable reliability or validity, replication efforts will be misdirected.

A final and fifth barrier is the heavy reliance on a descriptive or correlational approach in clinical science design and methods. If this approach is emphasized over theory development, it may promote over-interpretation of study-specific findings, and limit scientific progress.

Potential steps for clinical science. As is becoming standard across the field, researchers should perform power analyses in advance of studies to inform their target sample size (although an overemphasis on power analysis may promote an ill-advised overreliance on null hypothesis significance testing; Cumming, 2014). Although it can be challenging to anticipate likely effect sizes, researchers should be as explicit as possible about their predicted effect sizes, recognizing that most effects in individual differences research in psychology are likely to be small to medium in magnitude (e.g., Gignac & Szodorai, 2016).

A second recommendation is to use multiple measurements of key constructs whenever possible, consistent with the time-honored principles of critical multiplism and multiple operationalizations of constructs (Shadish, 1986). Clinical scientists should strive to offer broader coverage of target constructs to offset the barriers posed by shifting diagnostic criteria and evolving operationalizations in the field, and seek to enhance power by aggregating across measures to form latent variables.

A third recommendation for enhancing statistical power is to move toward measurement harmonization across laboratories whenever possible. It is becoming increasingly common for studies to involve multiple laboratories and multiple principal investigators working together, and 'Big Data' analytic approaches (e.g., Srivastava, 2015). Clinical scientists can more effectively confront the challenges distinctive to their research field (e.g., low base-rate conditions, difficult-to-recruit samples) by working together to share measures and create overlap among datasets of different groups working

on similar topics. One example of this approach is the well-known and still ongoing National Prodrome Longitudinal Study (NAPLS), which has engaged scholars at eight large university-connected sites to collect shared data on youth at elevated risk for schizophrenia and other psychotic conditions (Addington et al., 2007). Two other points are related to this third recommendation. One is the creation and use of open source measures, such as the NIH Toolbox (www.healthmeasures.net) and the International Personality Item Pool (http://ipip.ori.org/), to break down barriers of copyright and researcher cost. The other is a focus on cross-site collaborations and pooling of data to overcome the sample-specific characteristics and small sample sizes that limit work in clinical science. In combination with measurement harmonization, this practice can move the field toward larger, shared assessment protocols, with room to incorporate key measures from different participating laboratories and opportunities to test for moderating effects of sample-specific characteristics on phenomena of interest.

A fourth recommendation is to encourage greater emphasis on dimensional, rather than categorical or case-control, designs. The increased statistical power afforded by dimensional quantification approaches has been noted by others (Markon, Chmielewski, & Miller, 2011), and the newly launched RDoC initiative (Insel et al., 2011) provides a renewed focus on dimensions of psychopathology rather than on categories created by the imposition of largely arbitrary diagnostic thresholds. Nonetheless, perhaps because of the hegemony of the DSM and other categorical models over research and clinical practice, clinical science and related disciplines (e.g., psychiatry) have been slow to adopt dimensional operationalizations and measures. Resistance remains, but it seems clear at this point that dimensional quantification and analysis approaches will move us toward a more replicable science.

What can the Replicability Conversation Learn from Clinical Science?

The broader field of psychological science can assimilate several take-home messages from a clinical science perspective on the matter of replicability; we summarize seven points here. In doing so, we highlight the extent to which progress on issues of reproducibility will be most impactful and transformative if psychological scientists listen and talk to those working in allied domains. Otherwise, the reproducibility conversation will be unnecessarily prolonged and less productive than it otherwise might be. Clinical psychological science is but one example of the types of misfit between realities and proposed remedies that is likely to occur across the broader field as the replicability conversation progresses.

1. **A one-size-fits-all approach is unlikely to work.** The replicability movement in psychological science is multifaceted. Different areas within psychological science bring their own methods and traditions, as well as their own problems and solutions. We must broaden the conversation to include all areas of psychological science, and generate solutions that will be feasible for the field as a whole.

2. **Flexible approaches to pre-registration are needed for clinical science research.** We have offered a number of suggestions for ways that clinical scientists can think about incorporating pre-registration into their research. Pre-registration is likely to assume different forms for different types of research, including clinical science research.

3. **Open data approaches must accommodate sensitive data.** Clinical science research frequently grapples with sensitive information, often accompanied by the use of small and narrowly defined samples, such as those comprising participants with a specific diagnosis drawn from a specific clinic. This constraint, which has been insufficiently appreciated in ongoing discussions regarding open science and replicability, necessitates more sensitive approaches to dealing with data, and creative ways of considering the goals of open data in the clinical science context. In addition,

clinical science research sometimes relies on extensive archival datasets that are typically designed to yield analyses over the course of years or decades. Thus, posting data openly on the web is not as straightforward for clinical science research as it is for many domains of psychological research.

4. Require better accounting for differences in sample size and uniformity in participants for clinical science research. All researchers should be required to report details about how they recruited participants and participation rates from people they approached. For example, when using a clinical sample, researchers should indicate how many people were being treated for the same problem at that clinic as well as how many refused or were unable to participate. Similar considerations apply to community samples. Investigators should also bear in mind the possibility that some features of the clinical problems they are studying, such as self-knowledge, social anxiety, and motivation, may influence whether people are willing to participate in research. Full presentation of these details will make it easier to compare results across studies. In other words, the type of information needed to fully ascertain the nature of the sample is often different in clinical science research than in other domains of psychological science. Reviewers and editors handling manuscripts on clinical topics should look for and expect this information to be reported.

5. Consider what form clinical science-relevant "badges" will take. Given these constraints, clinical science research may not often be the ideal fit for narrowly defined best practice "badges," which are designated markers assigned to publications that demonstrate adherence to a variety of best practices, especially open data, open materials/measures, and preregistration of hypotheses and analytic plans. The concept of badges could be expanded to better appreciate the areas in which clinical science research can demonstrate strengths; for example, *Clinical Psychological Science*

has recently instated a badge system that should inform this question. In addition, we could explicitly recognize attempts to increase power through the use of cross-site collaborations, large samples sizes, or integrated comprehensive measurement. Creative and flexible ideas should allow the field to better reward psychological scientists across the discipline for efforts to increase the replicability of our results.

6. **De-emphasize innovation as stand-alone criterion.** Although we do not discuss it extensively here, the broader incentive structure plays a major role in many of these replicability issues (Lilienfeld, in press; Nosek et al., 2015). Clinical psychological science exists in this world as well, and faces a similar need to balance innovation and marketability of our findings with the production of rigorous, incremental, and replicable psychological science. The heavy reliance on large federal grants to conduct much clinical psychological science may further exacerbate some of these problems, and reflects a powerful external motivation for many QRPs. A related point is the need to conceive of exploratory and confirmatory research on a continuum. We have highlighted several examples of how this dichotomy may be oversimplified, and reifying it serves to further encourage certain QRPs, such as by pressuring researchers to present work as confirmatory even if it was not. Both exploratory and confirmatory research should be valued by the field (as well as "pseudo-exploratory" research, such as hypotheses developed with some, but not full, information).

7. **Be more patient and think "meta-analytically."** Despite legitimate calls for larger sample sizes, there will inevitably be a need for small N studies of dissociative identity disorder, trichotillomania, paraphilias, and other low base rate clinical conditions. At the same time, in the case of such studies, we need to (a) be considerably more circumspect in our claims and not make confident proclamations until much more data have been collected, (b) be honest when a finding was not hypothesized and (c) place less

emphasis on the results of individual studies, and instead to think meta-analytically, that is, to regard each study as merely one data point in a large population of studies, many of which have yet to be conducted. Similarly, reviewers, editors, and other stakeholders should reward these honesty and circumspection in authors' discussion of their limitations of their findings.

Concluding Thoughts

In this manuscript, we have presented multiple reasons why clinical psychological science and allied fields, such as counseling psychology, school psychology, psychiatry, epidemiology, and social work, have much to learn from the replicability conversation, and why the broader field of psychological science in turn has much to learn from clinical psychological science as the replicability conversation moves forward. We have shown that a number of key recommendations from other domains of psychological science confront challenges from a clinical science perspective. These challenges must be considered and accommodated when formulating overarching guidelines for the field. By highlighting these challenges and offering potential remedies, we hope to contribute a clinical science perspective to the broader conversation. This examination also serves as a useful exercise regarding the potential benefits of adopting a broader, field-based perspective as we move forward to increase the replicability of psychological science.

References

Addington, J., Cadenhead, K. S., Cannon, T. D., Cornblatt, B., McGlashan, T. H., Perkins, D. O., ... & Heinssen, R. (2007). North American Prodrome Longitudinal Study: a collaborative multisite approach to prodromal schizophrenia research. *Schizophrenia bulletin*, *33*, 665-672.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders, fifth edition* (DSM-5). Washington, D.C.: Author.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., … Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108-119. DOI: 10.1002/per.1919

Baker, M. (2016). Biotech giant publishes failures to confirm high-profile science. *Nature, 530*, 141. DOI: 10.1038/nature.2016.19269

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7,* 543-554.

Bishop, D. (2013). Why we need pre-registration. [Web log post]. Retrieved from http://deevybee.blogspot.com/2013/07/why-we-need-pre-registration.html

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376.

Chambers, C. (2014). Psychology's "registration revolution". [Web log post]. Retrieved from https://www.theguardian.com/science/head-quarters/2014/may/20/psychology-registration-revolution

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *Journal of Abnormal and Social Psychology*, *65*, 145-152.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Cumming, G. (2014). The new statistics why and how. *Psychological Science*, *25*, 7-29.

Davison, G.C. & Lazarus, A. (2007). Clinical case studies are important in the science and practice of psychotherapy. In S.O. Lilienfeld and W.T. O'Donohue. (Eds.), *The great ideas of clinical science: 17 principles that every mental health professional should understand* (pp. 149–62). New York: Routledge.

Donnellan, M. B., Lucas, R. E., Cesario, J. (2015). On the association between loneliness and bathing habits: Nine replications of Bargh and Shalev (2012) Study 1. *Emotion, 15,* 109-119. DOI: 10.1037/a0036079

Eaton, W. W., Anthony, J. C., Tepper, S., & Dryman, A. (1992). Psychopathology and attrition in the epidemiologic catchment area surveys. *American Journal of Epidemiology*, *135*, 1051-1059.

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology, 108,* 275-297. DOI: 10.1037/pspi0000007

First, M. B. (2005). Mutually exclusive versus co-occurring diagnostic categories: the challenge of diagnostic comorbidity. *Psychopathology*, *38*, 206-210.

Fraley, R. C. & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS One.* DOI: 10.1371/journal.pone.0109019

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology

recommendations for research and educational practice. *Personality and Social Psychology Review*, *8*, 3-12.

Funder, D. (2016). Why doesn't personality psychology have a replication crisis? [Web log post]. Retrieved from https://funderstorms.wordpress.com/2016/05/12/why-doesnt-personality-psychology-have-a-replication-crisis/

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time.* Unpublished manuscript

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102,* 74-78. DOI: 10.1016/j.paid.2016.06.069

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A., & Zwienenberg, M. (2015). A multi-lab pre-registered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11, 546-573.

Henriques, G. (2003). The tree of knowledge system and the theoretical unification of psychology. *Review of General Psychology*, *7*, 150-182.

Hilgard, J. (2016, March 3). A reading list for the replicability crisis [Web log post]. Retrieved from http://crystalprisonzone.blogspot.com/2016/03/a-reading-list-for-replicability-crisis.html.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., ... & Wang, P. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, *167*, 748-751.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS*

Med, 2, 0696-0701. DOI: 10.1371/journal.pmed.0020124

Ioannidis, J. P., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, *58*, 543-549.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23,* 524-532. DOI: 10.1177/0956797611430953

Jonas, K. J., & Cesario, J. (2015). How can preregistration contribute to research in our field? *Comprehensive Results in Social Psychology*. DOI: 10.1080/23743603.2015.1070611

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2,* 196-217.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahnik, S., … Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology, 45,* 142-152. DOI: 10.1027/1864-9335/a000178

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R. … Zimmerman, M. (2016). The hierarchical taxonomy of psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Manuscript under review.*

Letzter, R. (2016, September 26). We talked to a psychologist in the center of a brutal firestorm in the field of psychology. *Business Insider*. http://www.businessinsider.com/susan-fiske-methodological-terrorism-qa-2016-9

Lilienfeld, S.O. (in press). Psychology's replicability crisis and the grant culture: Righting the ship. *Perspectives in Psychological Science*.

Lilienfeld, S.O., & Waldman, I.D. (in press). *Psychological science under scrutiny*: *Recent challenges and proposed solutions*. New York: Wiley.

Lindsay, S., & Simons, D.J., & Lilienfeld, S.O. (in press). Preregistration 101. *APS Observer*.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151-159.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19-40.

Markon, K. E., Chmielewski, M., & Miller, C. J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: a quantitative review. *Psychological Bulletin*, *137*, 856-879.

Marszalek, J. M., Barber, C., Kohlhart, J., & Cooper, B. H. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*, 331-348.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, 70(6), 487.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., … Yarkoni, T. (2015). Promoting an open research culture. *Science, 348,* 1422-1425. DOI: 10.1126/science.aab2374

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349,* 943.

Owens, B. (2011, September 5). Reliability of 'new drug target' claims called into question [Web log post]. Retrieved from http://blogs.nature.com/news/2011/09/reliability_of_new_drug_target.html.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528-530.

Patrick, C. J., & Hajcak, G. (2016). RDoC: Translating promise into progress. *Psychophysiology, 53,* 415-424. DOI: 10.1111/psyp.12612

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *British Medical Journal, 316,* 1236-1238.

Platt, J. R. (1964). Strong inference. *Science*, *146*, 347-353.

Prasad, V., Cifu, A., & Ioannidis, J. P. (2012). Reversals of established medical practices: Evidence to abandon ship. *Journal of the American Medical Association*, *307*, 37-38.

Pritschet, L., Powell, D., & Horne, Z. (2016). Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. Online advance publication in *Psychological Science.* DOI: 10.1177/0956797616645672

Reichenbach, H. (1938). *Experience and prediction*. Chicago: University of Chicago Press.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638-641.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology, 1,* 43-46.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17,* 551-566.

Schönbrodt, F. D. (2015). *P-checker: One-for-all p-value analyzer*. Retrieved from http://shinyapps.org/apps/p-checker/.

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, *470*, 437.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309-316.

Shadish, W. R. (1986). Planned critical multiplism: Some elaborations.

Behavioral Assessment. 8, 75-103.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Perspectives on Psychological Science, 22,* 1359-1366.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*, 76-80. DOI: 10.1177/1745691613514755

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752-760. http://dx.doi.org/10.1037/0003-066X.32.9.752

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science, 10,* 886-899.

Srivastava, S. (2015). Replicability in personality psychology, and the symbiosis between cumulative science and reproducible science. [Web log post]. Retrieved from https://hardsci.wordpress.com/2015/06/14/replicability-in-personality-psychology-and-the-symbiosis-between-cumulative-science-and-reproducible-science/

Sternberg, R. J. (2005). *Unity in psychology: Possibility or pipedream?* Washington DC: American Psychological Association.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. Perspectives on Psychological Science, 9(1), 59-71.

Tellegen, A., & Waller, N. G. (2008). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. *The SAGE handbook of personality theory and assessment*,*2*, 261-292.

Tolin, D. F., McKay, D., Forman, E. M., Klonsky, E. D., & Thombs, B. D. (2015). Empirically supported treatment: Recommendations for a new model. *Clinical Psychology: Science and Practice*, *22*, 317-338.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. Proceedings of the National Academy of Sciences, 201521897.

Vazire, S. (2014, December 1). Why I am optimistic [Web log post]. Retrieved from http://sometimesimwrong.typepad.com/wrong/2014/12/why-i-am-optimistic.html.

Williams, B. R., Strauss, E. H., Hultsch, D. F., Hunter, M. A., & Tannock, R. (2007). Reaction time performance in adolescents with attention deficit/hyperactivity disorder: Evidence of inconsistency in the fast and slow portions of the RT distribution*. *Journal of Clinical and Experimental Neuropsychology*, *29*, 277-289.