AMERICAN PSYCHOLOGICAL ASSOCIATION

Clinical Psychology: Science and Practice

ISSN: 0969-5893

2021, Vol. 28, No. 2, 210–221
https://doi.org/10.1037/cps0000013

Replication Target Selection in Clinical Psychology: A Bayesian and Qualitative Reevaluation

Merle-Marie Pittelkow, Rink Hoekstra, Julie Karsten, and Don van Ravenzwaaij
Faculty of Behavioral and Social Sciences, University of Groningen

Low replication rates and ill-informed selection of replication targets can harm clinical practice. We take a pragmatic approach to suggest which studies to replicate in clinical psychology. We propose a 2-step process entailing a quantitative assessment of evidence strength using Bayes factors, and a qualitative assessment concerning theory and methodology. We provide proof of concept on a series of published clinical studies. We included 75 studies with 94 individual effects. Step 1 yielded 42 effects (45%) with Bayes Factors suggesting ambiguous evidence or absence of an effect. These 42 effects were qualitatively assessed by 2 raters. We illustrate their decision process and discuss advantages and disadvantages of the proposed steps.

Public Health Significance Statement
This study provides a pioneering constructive approach to dealing with the replication crisis in clinical psychology. We propose and discuss selection criteria for studies in need of replication.

Keywords: replication crisis, replication target selection, clinical psychology, Bayesian statistics, transparency

For the past decade, there has been growing concern that a substantial proportion of the published literature reports spurious findings: a phenomenon coined the replication crisis.

The replication crisis has been the topic of debate in both peer-reviewed (e.g., Ioannidis, 2005; Nosek et al., 2015) and non-peer-reviewed outlets (e.g., Mullarkey, 2019). Yet, clinical psychology has only recently become subject to the replication crisis conversation (Hengartner, 2018; Tackett, Brandes, King, et al., 2019; Tackett, Brandes, & Reardon, 2019; Tackett et al., 2017). The long absence of clinical psychology in the replication crisis debate is surprising, given the implications of low replicability in clinical research. Basing treatment recommendations and clinical practice on clinical trials that might fail to replicate may lead to sub-optimal clinical outcomes, misinformed decisions regarding reimbursement, policy making, and further funding (Leichsenring et al., 2017).

Until recently, evidence regarding the presence of a replication crisis primarily stemmed from discussion and inspection of either broad-themed journals or social and personality psychology outlets (Tackett et al., 2017). During the past couple of years, efforts increased to expand the discussion on the replication crisis to issues specific to clinical psychology (see for example Tackett & Miller, 2019). Some of this budding literature focuses on increasing reproducability of *future* studies by advocating the use of open science practices (e.g., Nutu et al., 2019), and methodological rigor (e.g., Reardon et al., 2019), and some studies question the reliability of *previously published* studies (e.g., Cuijpers, 2016; Reardon et al., 2019; Sakaluk et al., 2019; Tackett & Miller, 2019). Clearly, there is a question of which published findings we can trust in clinical psychology, and which studies require additional corroboration through replication studies (Hengartner, 2018).

This paper presents a practical approach to selecting replication targets in clinical psychology by combining a quantitative Bayesian re-analysis with qualitative and theoretical considerations. We illustrate the application of our suggested approach by using a sample of studies from the *Journal of Consulting and Clinical Psychology*. This journal was chosen as it is prominent in clinical psychology and publishes studies concerned with the diagnosis and treatment of mental disorders. We will demonstrate how our suggestions can be used to create a shortlist of studies for which replication might be most useful.

Replication Crisis in Clinical Psychology

One indicator of the replication crisis in clinical psychology is the systematic overestimation of the average efficacy of psychotherapy in the published literature due to publication bias (Driessen et al., 2015). For instance, 24% of trials evaluating the efficacy

Merle-Marie Pittelkow https://orcid.org/0000-0002-7487-7898
Rink Hoekstra https://orcid.org/0000-0002-1588-7527
Julie Karsten https://orcid.org/0000-0001-5727-0245
Don van Ravenzwaaij https://orcid.org/0000-0002-5030-4091
Correspondence concerning this article should be addressed to Merle-Marie Pittelkow, Faculty of Behavioral and Social Sciences, University of Groningen, 1/2 Grote Kruisstraat, 9712 TS Groningen, the Netherlands. Email: m.pittelkow@rug.nl

of a variety psychological treatments for major depressive disorder funded by the National Institutes of Health were not published, all of which yielded statistically nonsignificant results. This resulted in an overestimation of the efficacy of psychotherapy of approximately 25% (Driessen et al., 2015). Similar overestimation of efficacy due to unpublished trials has been reported for cognitive behavioral therapy for major depressive disorder in adults (Cuijpers et al., 2010).

Another reason why published effects in clinical psychology may not reflect true underlying effects is that treatments are oftentimes directly compared without an adequate placebo condition. In addition to the publication bias, these comparisons might be influenced by allegiance bias (bias in favor of the main author's psycho-therapeutic allegiance; Luborsky, 1999), sponsorship bias (bias in favor of results expected by a sponsor; Cristea et al., 2017), and various other sources of biases including unblinded outcome assessors (Khan et al., 2012) or small sample sizes (Cuijpers et al., 2010). Taken together, a variety of biases appear to distort findings in the clinical literature - with a detailed discussion given by (Hengartner, 2018). Overall, it appears that problems inherent to the replication crisis distort clinical psychological science too.

Replication

One way to improve trust in *previously published* findings is through conducting replications. There are two types of replications: direct and conceptual. A direct replication refers to a study utilizing the exact methodology as a previous study, whereas a conceptual replication allows for adjustments of methodological aspects with the aim to expand the knowledge regarding the reported effect. Academia is not only challenged by low replicability, but also by low rates of attempted replications, regardless of their outcome, and clinical psychology is no exception. Though clearly needed, replication studies are scarce in psychology. A previous investigation estimated an overall replication rate of only 1.07% (Makel et al., 2012). Potential reasons for these low rates of replications include limited funding and increased difficulty to publish. Combined, low replicability and a low rate of replications can have profound implications on clinical practice (Hengartner, 2018; Tackett et al., 2017).

The replication crisis undermines the quest for evidence-based treatments, as research findings might overestimate effects. Nonetheless, replication in clinical research involving patient samples appears even more challenging than laboratory or online studies in terms of costs, hidden variables, and ethical considerations (Gelman & Geurts, 2017). Furthermore, treatment approaches and treatment studies are plentiful. Currently, replication of every study is not feasible, meaning that only a subset of studies can be selected for replication, assuming limited means.

But which studies to select as replication targets? At present, the selection process appears biased (for a comprehensive overview see Laws, 2016). Laws suggests that researchers might choose to replicate a "cheap and easy" study, as they are quick to set up and execute. Others might opt to replicate a surprising, unexpected, or curious findings. This might go as far as deciding to replicate a study, because the hypothesis was thought to be improbable. While these reasons are understandable and might sometimes even be justifiable, they are prone to bias and are neither systematic nor transparent.

With increasing awareness regarding the need to replicate in psychology (e.g., Zwaan et al., 2018), more and more authors argue the need to clearly justify selection of replication targets (for an overview of recent developments please refer to Isager et al., 2020). Some authors suggest the use of cost-benefit analysis (Coles et al., 2018) or Bayesian decision-making (Hardwicke et al., 2018), while others suggest random selection to be most appropriate (Kuehberger & Schulte-Mecklenbeck, 2018).

Whatever the reasoning behind selecting a replication target might be, we believe that the selection should be systematic and transparent. For example, Field and colleagues (2019) developed a clearly formulated set of criteria to allow authors to illustrate and justify their decision. However, we are not aware of a set of criteria tailored to clinical psychological studies specifically. We believe that providing a framework (i.e., a set of criteria) to systematically and transparently assess the need for replication in clinical psychology would inform researchers as to the necessity and benefits of replicating a certain study, aid the process of establishing evidence-based treatments, and make replications easier and more frequent.

Statistical Considerations

But how does one decide which studies are most in need of replication? One practical way to do this is by evaluating the strength of the statistical evidence in published clinical studies. Published studies for which strength of evidence is low may be in need of replication. In a first attempt to analyze evidence regarding published studies, Sakaluk et al. (2019) conducted a meta-scientific review of selected empirically supported treatments (ESTs). To this end, they used rates of misreporting, estimates of statistical power, R-Index values, and Bayes Factors (*BF*). Their results indicate low evidential support for some, but not all, ESTs under scrutiny. More than half of the ESTs classified as strong by the American Psychological Association (52%) performed poorly across most or even all of the considered metrics of evidential value. This is worrying. Given the existence of gold-standard therapeutic interventions for which the available evidence is relatively low, one may wonder what the state of affairs is for the evidential load of more fundamental clinical psychological intervention research. Identifying studies with weak evidential support for treatment effects and recommending their replication at an early stage might prevent recommendation of weakly supported therapeutic interventions.

At present, *Null-Hypothesis Significance Testing* (NHST) is most commonly employed to formally test whether an effect of interest exists. In NHST, *p*-values quantify the probability of obtaining the observed data x or more extreme data X under the assumption that the null hypothesis (typically, H_0: no effect) is true:[1]

$$p = pr(X \geq |x| \ |H_0) \tag{1}$$

[1] This formula concerns one-sided tests. For two sided test: $p = 2 \times Pr(X \geq |x| \ |H_0)$ holds

Overreliance on the *p*-value is problematic. In the clinical literature, *p*-values are used for establishing evidential value. For example, the threshold of $p < 0.05$ (in two studies) served as one of the original Division 12 criteria for identifying empirically supported treatments (Chambless & Hollon, 1998). Similarly, the Food and Drug Administration, in charge of endorsing medical treatments in the U.S., defines substantial evidence for efficacy as given by "at least two adequate and well-controlled studies, each convincing on their own" (U.S. Food and Drug Administration, 1998, p. 3). A recent simulation study, however, demonstrated how the endorsement criterion of two *p*-values below .05 can be misleading, suggesting efficacy in the absence of an effect (van Ravenzwaaij & Ioannidis, 2017). This was especially true when the true effect size and sample size were small and the number of trials large (van Ravenzwaaij & Ioannidis, 2017), a scenario we oftentimes see in the clinical literature. This in turn might lead to incorrect endorsement decisions.

An additional shortcoming of *p*-values is that they are frequently misinterpreted by students and academic psychologists alike (Haller & Krauss, 2002; Oakes, 1986). Common misconceptions include the illusion that the probability of making a wrong decision is known when rejecting the null hypothesis and that the *p*-value indicates *reliability* of the obtained effects (Gigerenzer, 2004). These assumptions wrongfully overestimate the confidence in results supported by a significant *p*-value and undermine the actual need for replication.

For over half a century, NHST and the reliance on *p*-values has been criticized heavily (for an overview, see Kline, 2013; Morey et al., 2016; van Ravenzwaaij & Ioannidis, 2017; Wagenmakers, 2007). One major criticism is that *p*-values fail to quantify evidence in favor of the null hypothesis (Hoekstra et al., 2018). The *p*-value disregards the probability of an event under the alternative hypothesis (Rouder et al., 2009). To obtain evidence in favor or against the null hypothesis one needs to determine under which hypothesis an observed event would be most likely. In other words, one needs to compare the likelihood of the observed effect under the null hypothesis to the likelihood of the observed effect under the alternative hypothesis.

The Bayesian framework offers a practical alternative to NHST (Wagenmakers, 2007). In contrast to *p*-values, *BF*s allow researchers to quantify evidence in favor of either hypothesis (Gronau et al., 2019; Jeffreys, 1961; Rouder et al., 2009; Van Ravenzwaaij et al., 2019). In the Bayesian framework, the predictive evidence of two competing hypotheses is compared (for an elaborate discussion of Bayesian statistics we refer the interested reader to Etz et al., 2018) and the resulting ratio is referred to as the *BF*. In other words, the *BF* represents the ratio of the probability of the observed data *D* under the assumption that the alternative hypothesis is true compared to the probability of an observed data *D* under the assumption that the null hypothesis is true:

$$BF = \frac{\Pr(D|H_a)}{\Pr(D|H_0)} \tag{2}$$

A *BF* of 10 indicates that the data is 10 times more likely to have occurred under the alternative than under the null hypothesis. When the *BF* is 1, the data is equally likely under the two hypotheses, that is, the data does not favor one hypothesis over the other. A *BF* of $\frac{1}{10}$ indicates that the data is 10 times more likely to have occurred under the null hypothesis than under the alternative hypothesis.

While lower *p*-values might correspond to larger *BF*s, the relationship is not one-on-one. Consider for instance the following examples:

$t(10) = 4.59$, $p = .001$, $BF = 32.5$
$t(1000000) = 3.29$, $p = .001$, $BF = 0.25$

In both examples, the *p*-value is exactly .001, but evidential strength differs due to different sample sizes. In the first instance, the *BF* shows strong relative evidence for the alternative hypothesis, whereas in the second instance, the *BF* points toward the null hypothesis.

For purposes of the present study, the *BF* allows us to identify studies for which the statistical evidence in favor of an effect is ambiguous (our criterion is $BF < 3$) or for which the statistical evidence even favors the absence of an effect ($BF < \frac{1}{3}$; Jeffreys, 1961). As such, the *BF* can not only quantify evidence in favor or against the null hypothesis but also serve as a gradual decision criterion. With the typical practice of comparing *p*-values to a fixed threshold of 0.05, such a such a distinction between evidence for either hypothesis would not be possible.

Qualitative Considerations

The quality of published clinical studies should not be judged by the available statistical evidence alone. Some studies with strong statistical evidence might have used a participant sample that does not generalize well to the population or might have conducted a study in the lab that does not properly answer the underlying research question. Alternatively, some studies with comparatively weaker statistical evidence might be methodologically sound and theoretically important studies on a rare population for which it is simply difficult to obtain the requisite number of participants for adequate statistical power. In this article, we will follow the general approach put forward by Field et al. (2019) and argue that studies most in need of replication are those for which the statistical evidence is ambiguous, but for which the original methodology and theoretical relevance is sound. Original studies with strong evidence are less likely to need additional corroboration. Replication studies for which the original methodology may not have been optimal or for which the theoretical relevance was comparatively low may not have as much merit as replication studies for which the methodology and theoretical relevance was comparatively high. Such considerations are important in times where limited funding means replicating every single study may not be feasible.

The Present Study

We combine a Bayesian re-analysis with a qualitative reevaluation (cf. Field et al., 2019). We chose to focus on the Journal of Consulting and Clinical Psychology because of its prominence in the field of clinical psychology and its focus on treatment and prevention of psychological disorders. While there is consensus regarding the need to replicate, there is still criticism regarding the selection of studies to be replicated (Tackett et al., 2017). In order to combat this, we aim to ease the decision process regarding replication by proposing a two-step hierarchical approach. First, evidence load of a selection of articles is quantified using *BFs*.

Second, we propose qualitative criteria to assess need for replication and make recommendations based on these criteria. Following an implementation of the scoring system, we illustrate the use and feasibility of these criteria by illustrating how two independent raters would apply them.

Method

Literature Search and Sample

The online database PsycInfo was used to extract a total of 533 articles published in the *Journal of Consulting and Clinical Psychology* between 2012 and 2016 [2]. Subsequently, four independent investigators selected articles for inclusion. Studies were included if their main finding was supported by a *t*-statistic, as it is comparatively easy to calculate *BF*s when one has the test statistic and degrees of freedom available. This comprised one-sided and two-sided paired- and independent-samples *t*-tests, as well as *t*-tests for slopes in linear regression models. We also included *F*-test results with a degree of freedom of 1, as they are conceptually equivalent to *t*-tests. *F*-values were transformed into *t*-values using $t = \sqrt{F}$. We identified main findings as those findings that were presented as most important in the abstract, or those presented as primary finding in the Results section. If there was (a) more than one main effect supported by a *t*-test, or (b) more than one *t*-statistic supporting a main effect, all *t*-statistics were considered. For each analysis that was included, we extracted sample size, *t*-value, *p*-value, and degrees of freedom. In total, 78 articles comprising 97 *t*-statistics were included in the re-analysis. Upon closer inspection we noted that one study reported a chi-square test. Therefore, this study was excluded from the subsequent analysis. For two studies the *t*-statistic supported secondary and not the main finding. Consequently, these studies were excluded from subsequent analysis. Thus, the final sample included 75 studies comprising 94 effects. An overview of the studies is available in the online supplementary material on OSF.

Bayes Factor Reanalysis

Analysis was conducted in R, which can freely be downloaded from https://cran.r-project.org/, using the "BayesFactor" package version 0.9.12-4.2 (Morey et al., 2018).

We calculated degrees of freedom for 27 studies (36%) that did not report them (following Eisenhauer, 2008). We also calculated exact *p*-values using the reported *t*-statistics and degrees of freedom. For the majority of the studies the discrepancy between reported and calculated *p*-values was minor and did not influence the decision. For five studies however (6.67%), our calculations indicated *p*-values larger than .05 while the authors reported statistically significant results (i.e., S15, S24, S33, S47, and S67). Finally, we calculated *BF*s using the reported *t*-statistics and sample size. If more than one *t*-statistic was reported, multiple *BF*s were calculated corresponding to one *t*-statistic each. Calculation was customized depending on the type of test. If not otherwise specified, *t*-tests were considered to be two-sided. The *BF*s reported here either (1) compared the hypothesis that the standardized effect size (for one-sample *t*-tests) or mean difference (for two-sample *t*-tests) is 0 to the alternative that they are not 0; or (2) compared the hypothesis that a regression coefficient did not differ significantly from 0 to the alternative that they did. By default, a non-informative Jeffreys prior is used on the variance of the normal population and a Cauchy prior with scale parameter of $r = \frac{\sqrt{2}}{2}$ is used on the standardized effect size (Morey et al., 2018). For our purpose, we preferred this default prior over a more informed prior, as these default priors are widely used in psychology (e.g., Morey et al., 2018; The JASP Team, 2018). The resulting *BF*s served as an indication for the strength of evidence. Though we acknowledge that clear cutoffs are arbitrary, we follow Jeffreys classification of evidential strength lower than 3 "not worth more than a bare mention" (e.g., Jeffreys, 1961) and as such selected all studies with a *BF* below 3 for further analysis.

Qualitative Considerations

Studies selected based on their *BF* were further inspected by two of the authors independently to judge the need for replication. Articles were judged subjectively among their clinical and scientific relevance as well as their quality. This process was hierarchical in nature. First, need of replication based on the clinical and scientific relevance was judged. If a study was considered to be relevant, quality was assessed to determine which studies to select. An overview of the selection criteria was listed in Table 1. We would like to stress that the aim of this article was not to judge the quality of the raters' judgements or to reach consensus regarding replication targets but to illustrate how two people would independently use the proposed criteria and how this would allow for more transparency in the process of selecting replication targets.

Relevance

The *Journal of Consulting and Clinical Psychology* publishes articles with various foci, including the treatment and prevention of psychological disorders. Consequently, the journal's scope includes theory-based intervention studies, studies aimed at understanding the mechanism underlying development and maintenance of psychological disorders, and studies regarding the effectiveness of psychological interventions. Ultimately, all these studies should serve to benefit patients and practitioners and advance knowledge about psychological diseases.

To assess the clinical and scientific relevance we therefore considered several criteria. First, we distinguished studies with an intervention from those which did not implement an intervention. We assumed that intervention studies would have more wide-reaching real life consequences, as they can be translated into clinical advice, and policies are based on them. Second, we inspected whether the study considered a clinical sample. If so, we further considered the severity of the condition under investigation by judging the associated level of impairment and whether the condition under investigation was not yet well-researched. A study investigating a clinical sample could make stronger implications regarding the processes and outcomes in pathological samples. Thus, we considered studies with clinical samples to be more clinically relevant. Further, we argued that research regarding conditions which are not yet well understood is more scientifically relevant as compared to studies regarding well researched psychopathology.

[2] Data collection was performed in early 2017 and spanned the previous five years of publications.

Table 1
Overview of the Selection Criteria

Label	Criteria
Relevance	
Clinical relevance	Intervention study (yes/no)
	Clinical sample (yes/no)
	Severity of condition (low/medium/severe)
Scientific relevance	Evidence base (small/substantial/ large)
Quality	
Theory	Scientific Background sound (yes/no)
	Clear rational (yes/no)
Methodology	CONSORT criteria met (yes/no)
	Statistical method appropriate (yes/no)
Interpretation	Interpretation and conclusion follow logically (yes/no)
	Generalizability (limited/good)
	Robustness (limited/good)

Quality

The CONSORT (CONsolidated Standards of Reporting Trials; Schulz et al., 2010) statement was used to guide the assessment of the studies' quality. Originally, the CONSORT statement refers to a set of guidelines for reporting parallel group randomized trials. At present, a variety of extensions exist making the CONSORT guidelines applicable for a variety of research designs. Though the guidelines are not intended to assess studies' quality (Schulz et al., 2010), they have been previously proven useful in assessing quality of studies (Falci & Marques, 2015). The CONSORT checklist includes items concerned with the theoretical background of the study, methodology, statistical analysis, and interpretation. For an overview see online supplementary material on OSF. Here, the primary focus was on three areas: (1) theoretical background, (2) methodology, (3) interpretation.

Theory

We adopted item 2a from the CONSORT checklist focusing on scientific background and rationale. We labeled a study in need of replication if it had a strong theoretical framework but failed to find substantial evidence in favor of the alternative hypothesis. A strong theoretical base points toward high clinical and scientific relevance and replication could add substantially to the existing literature. Studies for which the underlying theory was "shaky" and rationales were unclear were not considered for replication. In sum, we considered studies where replication could lead to a refinement of the theory and strengthen trust in the claimed effect, or result in a refutation of the proposed effect and serve as evidence against the theory.

Methodology

We adopted a selection of CONSORT checklist items to guide judgment of methodology. These included: description of the trial and study design (item 3a), changes in method (item 3b), sample selection (item 4a), setting and location (item 4b), operationalization (item 6a/b), randomization (item 8a/b), blinding (item 11a/b), and methods for statistical analysis (item 12a/b, 15, 17a). Additionally, sample heterogeneity and missing data were considered. If studies were limited in their methodology and/or the operationalization was inappropriate, we classified the study as in potential need of replication.

Interpretation

Lastly, we considered the interpretation of results. Special attention was paid to limitations (item 20), consistency with results (item 22), generalizability (item 21), and robustness of the finding. Studies for which the conclusions reach beyond the scope of the empirical evidence should not serve as a base for future research but should require conceptual replication to justify the claims in the conclusion. Such conclusions could be overstatements of effects or interpretation beyond the scope of the obtained results. In both cases it remains to be established whether the true effect equals the claimed effect. Conceptual replication could help to clarify the nature of the "true" effect. Studies low in generalizability might benefit from varied replication to increase knowledge regarding the context specificity of effects. Robustness was considered as an additional measure of confidence in the claimed effect.

Results

All information necessary to reproduce our analysis can be found on OSF (osf.io/xd2fk/).

Bayes Factors

Among the 94 effects studied, *p*-values ranged from 1.88×10^{-42} to .086. In total, six studies reported statistically nonsignificant *p*-values. In one study the main findings was the absence of a difference supported by a statistically nonsignificant effect (i.e., S66). *BF*s ranged from 0.40 to 5.33×10^{40}. The relationship between *p*-values and *BF*s is illustrated in Figure 1. Overall, there were 40 effects (42.55%), for which the *BF*s indicated weak or no evidence for the claimed effect. Seven effects (7.45%) had a *BF* lower than 1 (range 0.40–0.98) indicating support for the null hypothesis and 33 effects (35.11%) had a *BF* between 1 and 3 (range 1.13–3.00) indicating inconclusive evidence for either hypothesis. The other 54 effects (57.45%) had *BF*s greater than 3, indicating at least substantial evidence for the alternative hypothesis.

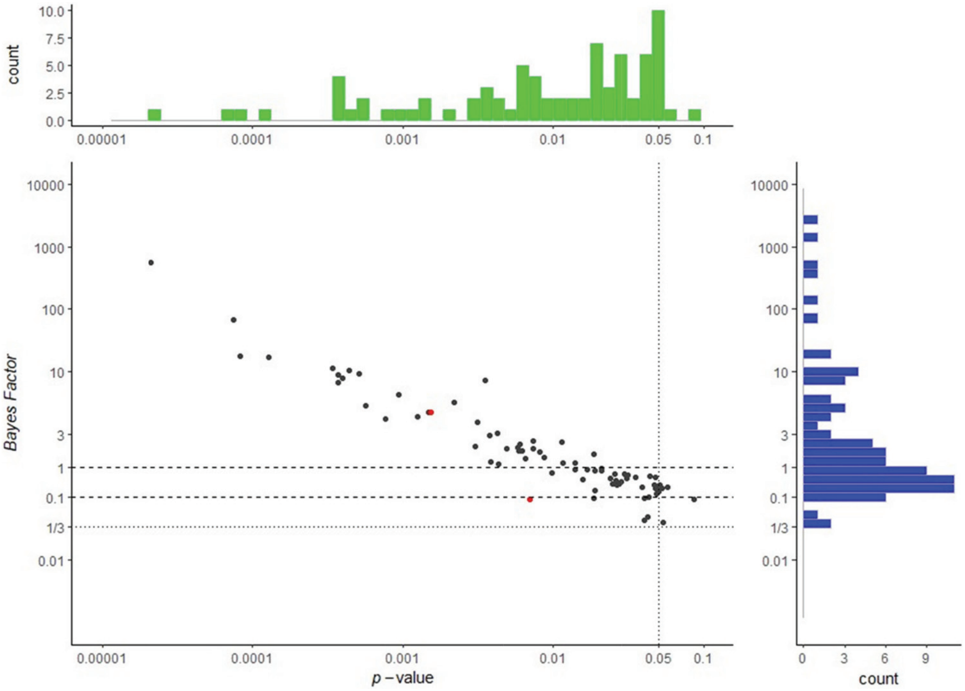
Qualitative Consideration

Based on their *BF* being lower than three, 36 studies corresponding to the 40 effects were included in the qualitative re-analysis. As one of the studies (i.e., S34) was a meta-analytic review and the the criteria were designed to evaluate primary studies, this particular study was excluded from the qualitative reevaluation leaving a total of 35 studies.

Figure 2 provides an overview of the decision process regarding the need for replication for both raters. Rater 1 [R1] suggested replication of 16 studies (21.33% of the original data set), while rater 2 [R2] suggested replication of 23 studies (30.67% of the original data set). Overall, the raters agreed in their decision as to whether to suggest replication or not in 25 cases (71.43%).

Regarding the suggestions, there are four types of (dis)agreement between the two raters: (a) agreement on the suggestion whether or not to replicate *and* the decisive criterion (k = 10), (b) agreement on the suggestion whether or not to replicate but *not* the decisive criterion (k = 15), (c) disagreement on the

Figure 1
Relationship Between p-Values and BFs



Note. BFs below 1/3 indicate evidence in favor of the null hypothesis. *BFs* between 1/3 and 3 indicate ambiguous evidence. *BFs* above 3 indicate evidence in favor of the alternative hypothesis. Two meta-analytic studies are highlighted in red. Histograms of *BFs* (blue) and *p*-values (green) are displayed next to the plot. See the online article for the color version of this figure.

suggestion whether or not to replicate *and* the decisive criterion (k = 9), and (d) disagreement on the suggestion whether or not to replicate but *not* the decisive criterion (k = 1). See also Table 2. In the following section, we discuss an example of each type of (dis)agreement. We do so with the aim to illustrate how the criteria can be applied. We would like to stress that these studies were chosen at random if possible (i.e., through a random number generator in R) and our aim is neither to criticize the authors nor their work, but to illustrate a decision process regarding the need to replicate (rather than, say, selecting the study that is easiest to execute instead). For more detailed information about the decision process regarding all 35 studies we refer the interested reader to the Excel table provided in the online supplementary material on OSF.

Agreement on the Suggestion Whether or Not to Replicate as Well as the Decisive Criterion (Example: S1)

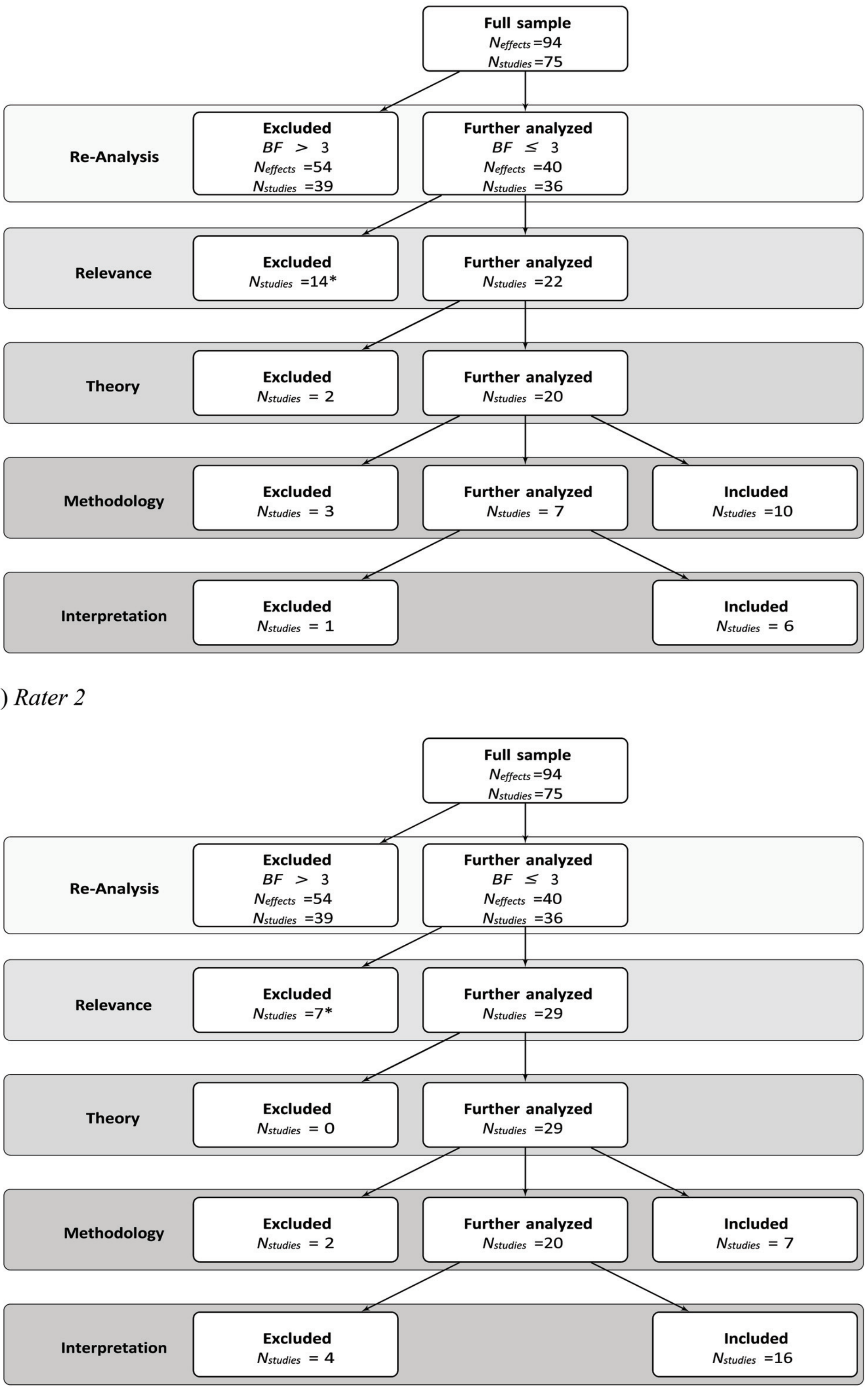
Both raters suggested replication of a two-arm cluster randomized controlled trial (RCT) comparing low-income patients with newly diagnosed Major Depressive Disorder (MDD) receiving either motivational interviewing (MI) in addition to standard treatment or standard treatment only. The authors hypothesized that patients receiving MI would on average show greater improvement of depressive symptoms and higher remission rates. Both raters judged the study to be clinically relevant as it included an intervention administered to a clinical sample with a medium disease burden. R1 argued that" *Though the field of depression is widely researched, the authors make a convincing point that outcomes of primary care show room for improvement*," which is mirrored in R2's conclusion that" *Implications for treatment* [are] *substantial*." Moreover, the two raters independently judged the study to have sound underlying theory, a clear rational, and no methodological flaws. However, both raters reached the conclusion that the interpretation was overstated given the findings of the present study. The results suggested only a modest benefit of MI over no MI at 36 weeks follow up with greater improvement in depressive symptoms and a higher remission rate for the MI group. The authors subsequently conclude that this interaction was indicative of MI to result in substantial, significant, and clinically meaningful changes, which both raters judged to be an over-interpretation. Consequently, both raters independently reached the decision to suggest a conceptual replication.

Agreement on the Suggestion Whether or Not to Replicate but Not the Decisive Criterion(Example: S72)

Both raters suggested replication of a RCT concerned with the effectiveness of sleep-directed hypnosis (hypCPT) as an addition to cognitive processing theory (CPT) in improving sleep impairments in patients with Post Traumatic Stress Disorder (PTSD). The raters agreed on clinical relevance based on the study including an intervention and a clinical sample with severe disease burden. Moreover, both raters judged the study to have a clear rational and theoretical underpinning: The authors acknowledged a lack of a

Figure 2

Flow Chart Providing an Overview of the Qualitative Revaluation Process



(a) *Rater 1*

(b) *Rater 2*

Note. * One study was excluded as it was not a primary study.

Table 2
Results of the Qualitative Reevaluation

	Rater 1		Rater 2		
Study number	Replication suggested?	Criterion	Replication suggested?	Criterion	Match?
1	Yes	Interpretation	Yes	Interpretation	Yes
3	Yes	Methodology	Yes	Methodology	Yes
7	No	Theory	Yes	Interpretation	No
9	No	Methodology	No	Relevance	Decision only
10	No	Theory	Yes	Interpretation	No
11	Yes	Methodology	Yes	Interpretation	Decision only
12	No	Relevance	No	Relevance	Yes
15	Yes	Methodology	Yes	Interpretation	Decision only
18	Yes	Interpretation	Yes	Methodology	Decision only
23	Yes	Methodology	Yes	Interpretation	Decision only
24	Yes	Methodology	No	Methodology	Reason only
25	Yes	Interpretation	Yes	Methodology	Decision only
26	No	Relevance	No	Relevance	Yes
28	No	Relevance	No	Relevance	Yes
32	No	Methodology	No	Methodology	Yes
33	Yes	Methodology	Yes	Interpretation	Decision only
37	Yes	Methodology	Yes	Methodology	Yes
38	No	Relevance	No	Interpretation	Decision only
41	No	Relevance	No	Interpretation	Decision only
42	No	Relevance	No	Interpretation	Decision only
43	Yes	Interpretation	Yes	Interpretation	Yes
45	No	Relevance	Yes	Interpretation	No
47	Yes	Methodology	Yes	Interpretation	Decision only
49	No	Methodology	Yes	Interpretation	No
51	No	Relevance	Yes	Interpretation	No
52	No	Relevance	No	Relevance	Yes
54	Yes	Interpretation	No	Relevance	No
59	No	Interpretation	No	Relevance	Decision only
62	No	Relevance	Yes	Interpretation	No
65	No	Relevance	Yes	Interpretation	No
66	No	Relevance	Yes	Methodology	No
67	Yes	Methodology	Yes	Interpretation	Decision only
69	No	Relevance	No	Interpretation	Decision only

specific evidence-based theory regarding sleep impairments in patients with PTSD, but they provide an elaborate review of the existing literature from which their rational follows logically. Nonetheless, R1 did question as to why the specific intervention was chosen. While R1 did not identify methodological flaws, R2 suggested that the high drop-out rate (41%) resulted in a small sample (i.e., $N_{hypCPT} = 25$ vs. $N_{CPT} = 29$ at posttreatment and $N_{hypCPT} = 35$ vs. $N_{CPT} = 30$ at follow-up), decreasing power and increasing the chance of a *Type 2 error* (Leppink et al., 2016). Combined with the clinical relevance of this particular study, R2 thus suggested replication with a larger sample to ensure stability of the effect. Meanwhile, R1 suggested replication based on the interpretation of the study. R1 was concerned with the authors suggesting hypnosis as a clinical tool for patients despite the limited evidence base at this point. Further investigation of the efficacy of hypCPT appears warranted before making treatment recommendations.

Disagreement on the Suggestion Whether or Not to Replicate but Not the Decisive Criterion (Example: S24)

This study utilized data from an intervention trial. In the original trial, students, who violated campus alcohol policies, were randomly assigned to one of four conditions (a) a counselor delivered brief motivational interview (BMI), (b) a computer delivered intervention called Alcohol Edu for Sanctions (EDU), (c) a computer delivered intervention called Alcohol 101 Plus, or (d) a delayed intervention group. This study focused on the first two interventions. The aim was to (a) determine whether social networks influenced the extent to which college students initiated or maintained reductions in drinking following the interventions, and (b) to explore which intervention was most effective for students with riskier social networks.

On this particular trial, the two raters disagreed regarding the suggestion to replicate, while justifying their decision with the same decisive criterion. R1 judged clinical relevance of the study to be unclear, as the study did not include a clinical sample and the condition under investigation was not rated as severe. Nonetheless, R1 judged it to be scientifically relevant based on the primary focus of this study being identification of moderators of long-term success of these interventions. R1 felt this had clear scientific relevance as it could inform future studies regarding the efficacy of these interventions for subgroups. On the other hand, R2 concluded that the study was clinically relevant as excessive drinking is highly prevalent in the target group. Moreover, R2 reasoned that drinking problems during college times could have long-lasting implications on their later life-situation. Overall, both raters judged the study to be relevant. Additionally, both raters thought that the theoretical background appeared to be sound and a rational was clearly stated: The authors explored various potential mechanisms through which one's social network might impact drinking behavior and how exactly the proposed intervention could intervene.

Regarding methodology, opinions varied substantially. While R2 did not mention methodological shortcomings and did not suggest replication, R1 criticized the methodology and suggested replication with potential changes in the study's methodology and analysis strategy to increase generalizability and robustness of the results. First, R1 suggested to reduce the differences in time-investment between the two interventions. The BMI was delivered in person and took approximately one hour to complete, while the EDU was delivered by computer and took approximately two hours to complete. While the original study included control conditions for the difference in modalities, R1 did not feel like the difference in time-investment was appropriately considered. R1 suggested that one could engage patients in the BMI condition for an additional hour, maybe with a filler task, to overcome this potential bias. Second, R1 suggested to adapt the outcome variable (i.e., alcohol consumption during the past month) to remove the potentially obscuring influence of recall bias (Althubaiti, 2016).

One potential adaptation would be to use ecological momentary assessment (EMA). The use of EMA would prevent recall bias by sampling participants' behavior in real time and in the subject's natural environment. Third, R1 suggested exploration of reasons for missing data and application of a different missing-data strategy. Listwise deletion assumes that the data are missing completely at random (MCAR; Baraldi & Enders, 2010). If the assumption of MCAR is violated, which it commonly is, casewise deletion leads to biased estimates.

Alternatively, multiple imputation or maximum likelihood estimation are recommended (Schafer & Graham, 2002). Based on these perceived shortcomings, R1 suggested replication based on methodology.

Disagreement on the Suggestion Whether or Not to Replicate as Well as the Decisive Criterion (Example: S7)

This example concerned a RCT comparing the effectiveness of cognitive behavioral social skills training (CBSST) in improving negative symptoms and defeatist attitudes with an active psychosocial control condition, namely goal-focused supportive contact (GFSC), in patients with schizophrenia and schizoaffective disorder. Here, the raters differed in both their suggestion to replicate as well as the decisive criterion. Both raters judged the study to be clinically relevant as it included an intervention in a clinical sample with severe disease burden. Moreover, both raters suggested that the evidence base was large. Based on an inspection of the theory, R1 concluded that no replication was needed as R1 perceived the study itself to be a conceptual replication of previous studies. R2 considered not only the theoretical background but also methodology and interpretation and concluded that" *results are promising given the older population. As the available literature for this population is limited and somewhat conflicted, replication would be valuable*." In other words, R2 was not convinced that the evidence was strong enough yet to make clinical suggestions based on the presented results.

Discussion

The present article proposed a method for selecting clinical psychological studies most in need of replication based on a two-step process. In the first step, evidence in published articles was subjected to a Bayesian reanalysis. This Bayesian reanalysis allowed a subdivision in (a) compelling evidence for the alternative hypothesis (i.e., there likely is an effect), (b) ambiguous evidence (i.e., there may or may not be an effect), and (c) evidence in favor of the null hypothesis (i.e., there likely is no effect). As a second step, studies with ambiguous evidence or evidence in favor of the null hypothesis were evaluated among a set of qualitative criteria to assess potential need for replication. The need for replication was judged by considering clinical and scientific relevance and the quality of the study. Quality of the study was further divided into concerns regarding the theory, methodology, and interpretation. We recommend applying each of these sequentially, so as to retain studies as possible candidates for replication in a step-by-step process based on theory, methodology and interpretation. We illustrated the application or our proposed method by having two raters independently make recommendations based on these criteria.

We demonstrated our method on 75 articles published in the *Journal of Consulting and Clinical Psychology* between 2012 and 2016. We applied a default *BF* analysis to these articles. Variability in *BF*s [$0.4 - 5.3 \times 10^{10}$] was comparable to that observed for the more general journal *Psychological Science* [$0.1 - 1.9 \times 10^{10}$] (Field et al., 2019) but smaller than variability observed for studies cited to support evidence-based treatments [$1.0 - 1.4 \times 10^{32}$] (Sakaluk et al., 2019). Results indicated that for almost half of the re-analysed studies, the *BFs* suggested ambiguous evidence or even evidence in favor of the null hypothesis as opposed to the alternative hypothesis. This proportion was slightly larger than what Field and colleagues observed for general psychological studies (Field et al., 2019). Field and colleagues reported almost half of their *BF*s to lie between 1 and 5. Please note that the re-analysis here served as a proof of concept and as such contained a convenient sample of clinical psychological studies across many different topics. We believe the method presented here is well positioned to assess need for replication for studies in a specific subfield, such as studies investigating a certain disorder or a particular treatment model.

The combination of the Bayesian reanalysis and our qualitative assessment led R1 to select 16 and R2 to select 23 studies considered to be most suitable for replication. We hasten to say that our assessment of a study benefiting from replication should not be taken to imply that we believe the study itself is of poor quality. Rather, we believe the underlying theory sound, but the present statistical strength of evidence in need of further corroboration.

Moreover, labeling a study as not in need of replication should not be understood as an advise against replication. Many of the studies we scrutinized presented compelling statistical evidence for the focal effect. In those cases further studies building on those results, often labeled conceptual replications or follow-up studies (e.g., Schmidt, 2009), are perhaps more valuable than more direct replications. We stress that we believe there is always merit in replicating studies, our tool is simply one that helps prioritize replication targets in conditions of sparse resources. One could make a strong case for conducting conceptual replications of studies with weak methodology but compelling statistical evidence. The qualitative criteria presented here can also be applied in such cases to make the decision process more systematic and transparent.

Likewise, we would like to point out that the present approach cannot speak to the 'truth' value of a specific effect. While the present approach includes operationalization and assessment of the concept under investigation as criteria, assessment of the validity of a given measurement was not always possible. One obstacle was a lack of transparency about measurement decisions, also criticized by Flake and Fried (2020). While the structural validity of measurements employed in clinical psychology is a pressing issue, it is not one that the present method can assess. The raters used all available information to judge whether the measure employed was appropriate and suggested alternatives whenever needed (see online supplementary material on OSF for more details).

The present tool did not mandate which type of replication to suggest. Nonetheless, the raters sometimes specified which type of replication they perceived to be most appropriate. Generally, the raters suggested direct replication. Whenever the methodology was considered weak or the sample and thus generalizability of the results limited, the rater suggested conceptual replication specifically. Overall, the tool allowed for great flexibility and varying levels of detail regarding the specific suggestions.

During the selection process we noticed that the CONSORT statement was well suited to guide the assessment of the qualities of studies as many authors appeared to have employed these guidelines themselves. The CONSORT guidelines are specifically designed to target clinical intervention and prevention studies, thus applying directly to our target sample.

Overall, we are satisfied with the present criteria. Nonetheless, we noticed a few downsides: the grading process was rather time-consuming, and we were unable to make a ranking within the studies we suggest for replication. Moreover, while the raters assessed theory by the best of their abilities and knowledge, the raters have a general background in clinical psychology and lack specific expertise in the scrutinized studies. This was to some extent a necessary evil: the selection of study topics was very broad. The raters took a pragmatic approach and focused on the evidence presented by the authors to support the particular aim of the study and whether the rational was clear and logical. This is in line with our operationalization of the criterion (i.e., *Scientific Background sound (yes/no)*; *Clear rational (yes/no)*). For future use of this method for assessment of what studies to replicate in a specific clinical area, we recommend soliciting raters with in-depth expertise into the relevant area. Overall, we perceive the theoretical and methodological considerations regarding the need for replication to be crucial. The time it takes to screen studies along our proposed criteria is more than easily earned back by replicating only those studies that most require corroborating evidence, thus spending the time it takes to conduct a study most effectively.

Some might criticize that the process of developing and applying these criteria was subjective and question the benefit of using this particular approach. We would like to stress that our main intention is to encourage discussion regarding how to choose studies for replication. To our knowledge, no clear criteria exist and the ones we propose here should serve as a starting point in the development of universal decision criteria for replication.

Revisions of and additions to the present method are imaginable. One potential consideration, not yet (explicitly) included in the present suggestions, is the incorporation of misreporting in the original article. A case can be made for not suggesting replication of studies with ambiguous evidence coupled with misreporting, as we might suspect questionable research practices to account for the ambiguity of evidence. However, misreporting can have many potential causes and not all suggest questionable intent (e.g., typos). Therefore, we refrained from including this criterion in the present approach.

Another consideration is to evaluate effect size more explicitly when selecting a replication target. While a *BF* can only state that an effect exists, it does not speak to the practical relevance of this effect. The present approach however mirrors our belief that only once the presence of an effect is established does it make sense to investigate the size of that effect. We first establish presence or absence of an effect using a Bayesian re-analysis before assessing practical relevance, which *can* include considerations of effect size. Overall, we are aware that the present suggestions are the starting point of a long and interesting discussion regarding replication target selection involving the scientific community.

The subjectivity of the process of replication target selection inherently means that different people end up with a different selection. In the present study we demonstrated a relatively high level of agreement (i.e., 71.43%) given the subjective nature of this decision process. Moreover, a clear advantage of applying this structured approach is that we can neatly follow the decision process of individuals and compare them, be it in cases of agreement or disagreement. The criteria offer transparency and openness regarding the selection of replication targets. We believe our proposal for selecting studies most in need of replication presents a clear advance over either selecting studies at random or selecting those studies for which replication is quick and easy.

Where to go from here? To validate criteria and potentially expand their application, the Delphi technique could be employed to reach consensus regarding which criteria to consider when judging need for replication. The Delphi technique aims to combine expert opinions in a systematic manner over a number of times to produce trustworthy information or data (i.e., criteria) that can be used for an intended purpose (Fink et al., 1984; Mbakwe et al., 2016). Through a series of questionnaires, experts (including clinical researchers and practitioners) could be asked to judge the usefulness of the proposed criteria. Consensus could be used to provide useful guidelines for researchers to apply when judging the need to replicate an individual study.

To our knowledge, we are the first ones to propose a set of criteria to prioritize which clinical psychology studies to replicate in case of sparse resources. Nonetheless, the results presented in this study are not without caveats. First, we limited our re-analysis to studies published in the *Journal of Consulting and Clinical Psychology* that supported their main effect with a *t*-statistic. Focusing on a single journal and statistical test only limits generalizability of our quantitative results (i.e., distribution of Bayes Factors and number of studies suggested in need of replication). Studies published in other journals might have different strengths and weaknesses. Moreover, considering the crisis of trust in study results an increase in more sophisticated statistical methods has been observed over the past years. Thus, we encourage investigations at different journals and at studies that employ more complex experimental designs to see if the proposed approach lends itself to clinical psychological studies in general.

Second, development of the qualitative selection criteria was based on a combination of the CONSORT statement and our own assessment. Use of the CONSORT statement limits applicability of these criteria to clinical intervention and prevention studies. There are many merits to different kinds of study designs, but here we chose to limit our investigation to clinical intervention and prevention studies. Further, the overarching themes of the qualitative criteria, namely relevance, theory, methodology, and interpretation, apply universally to scientific publications and we are positive that these could be applied to other areas of psychology as well. For example, Field and colleagues (2019) applied a similar methodology to articles published in *Psychological Science*. Employing a similar combination of quantitative (i.e., $\frac{1}{3} \leq BF < 3$) and qualitative selection criteria (i.e., theoretical importance, relevance, and insufficient investigation), they were able to select 3 studies out of 57 potential candidates to replicate. Overall, it appears that a combination of both statistical, theoretical, and methodological criteria is useful to a variety of fields in psychology. Here, we demonstrated usefulness of a variation of these criteria specifically targeting clinical, psychological intervention studies. We would like to encourage the use and refinement of these criteria to judge the need of replication in clinical and other areas of psychology, and science.

Third, during the quantitative and qualitative re-analysis we noticed that the present approach was not suited for meta-analytic effects. Future studies employing a similar methodology might want to limit themselves to primary studies only.

Lastly, the outcome of the present reanalysis is dependent on the model and chosen threshold. We have chosen a default specification of the prior probabilities used in our Bayesian reanalysis. Our data is available on OSF and a sensitivity analysis exploring alternative priors is provided in the online supplementary material on OSF. We encourage additional analysis, for example using more informed priors. Moreover, we chose a *BF* threshold of 3 (and 1/3) based on Jeffreys heuristics, but alternative choices are also defensible. For example, both the Journal of Experimental Social Psychology and Cortex consider 6 as threshold of compelling Bayesian evidence.

Concluding Statement

The present article provides a concrete approach with worked example on how to systematically and transparently justify selection of one or several replication targets. We hope that our work inspires discussion about selection criteria for replication in clinical psychology under conditions of limited resources.

References

Althubaiti, A. (2016). Information bias in health research: Definition, pitfalls, and adjustment methods *Journal of Multidisciplinary Healthcare*, *9*, 211–217. https://doi.org/10.2147/JMDH.S104807

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*(1), 5–37. https://doi.org/10.1016/j.jsp.2009.10.001

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*(1), 7–18. https://doi.org/10.1037/0022-006X.66.1.7

Coles, N. A., Tiokhin, L., Scheel, A. M., Isager, P. M., & Lakens, D. (2018). The costs and benefits of replication studies. *Behavioral and Brain Sciences*, *41*, Article e124. https://doi.org/10.1017/S0140525X18000596

Cristea, I. A., Gentili, C., Pietrini, P., & Cuijpers, P. (2017). Sponsorship bias in the comparative efficacy of psychotherapy and pharmacotherapy for adult depression: Meta-analysis *The British Journal of Psychiatry*, *210*(1), 16–23. https://doi.org/10.1192/bjp.bp.115.179275

Cuijpers, P. (2016). *Are all psychotherapies equally effective in the treatment of adult depression? The lack of statistical power of comparative outcome studies*. BMJ Publishing Group. https://doi.org/10.1136/eb-2016-102341

Cuijpers, P., Smit, F., Bohlmeijer, E., Hollon, S. D., & Andersson, G. (2010). Efficacy of cognitive–behavioural therapy and other psychological treatments for adult depression: Meta-analytic study of publication bias *The British Journal of Psychiatry*, *196*(3), 173–178. https://doi.org/10.1192/bjp.bp.109.066001

Driessen, E., Hollon, S. D., Bockting, C. L. H., Cuijpers, P., & Turner, E. H. (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), Article e0137864. https://doi.org/10.1371/journal.pone.0137864

Eisenhauer, J. G. (2008). Degrees of freedom. *Teaching Statistics*, *30*(3), 75–78.

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, *25*(1), 219–234. https://doi.org/10.3758/s13423-017-1317-5

Falci, S. G. M., & Marques, L. S. (2015). Consort: When and how to use it. *Dental Press Journal of Orthodontics*, *20*(3), 13–15. https://doi.org/10.1590/2176-9451.20.3.013-015.ebo

Field, S. M., Hoekstra, R., Bringmann, L., & van Ravenzwaaij, D. (2019). When and why to replicate: As easy as 1, 2, 3? *Collabra*: *Psychology*, *5*(1), 46. https://doi.org/10.1525/collabra.218

Fink, A., Kosecoff, J., Chassin, M., & Brook, R. H. (1984). Consensus methods: Characteristics and guidelines for use. *American Journal of Public Health*, *74*(9), 979–983. https://doi.org/10.2105/ajph.74.9.979

Flake, J. K., & Fried, E. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–565.

Gelman, A., & Geurts, H. M. (2017). The statistical crisis in science: How is it relevant to clinical neuropsychology? *The Clinical Neuropsychologist*, *31*(6-7), 1000–1014. https://doi.org/10.1080/13854046.2016.1277557

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2019). Informed Bayesian t-tests. *The American Statistician*, *74*(2), 1–14. https://doi.org/10.1080/00031305.2018.1562983

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research Online*, *7*(1), 1–20.

Hardwicke, T. E., Tessler, M. H., Peloquin, B. N., & Frank, M. C. (2018). A Bayesian decision-making framework for replication. *Behavioral and Brain Sciences*, *41*, Article e132. https://doi.org/10.1017/S0140525X18000675

Hengartner, M. P. (2018). Raising awareness for the replication crisis in clinical psychology by focusing on inconsistencies in psychotherapy research: How much can we rely on published findings from efficacy trials? *Frontiers in Psychology*, *9*, 256. https://doi.org/10.3389/fpsyg.2018.00256

Hoekstra, R., Monden, R., van Ravenzwaaij, D., & Wagenmakers, E.-J. (2018). Bayesian reanalysis of null results reported in medicine: Strong yet variable evidence for the absence of treatment effects. *PLoS ONE*, *13*(4), e0195474. https://doi.org/10.1371/journal.pone.0195474

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Isager, P. M., Aert, R. C. M. v., Bahník, Š., Brandt, M., DeSoto, K. A., Giner-Sorolla, R., Krueger, J., Perugini, M., Ropovik, I., Veer, A. v., Vranka, M. A., & Lakens, D. (2020). *Deciding what to replicate: A formal definition of "replication value" and a decision model for replication study selection*. PsyArXiv. https://doi.org/10.31222/OSF.IO/2GURZ

The JASP Team. (2018). JASP (Version 0.8.6) [Computer software]. https://jasp-stats.org/

Jeffreys, H. (1961). *Theory of probability*. Oxford University Press.

Khan, A., Faucett, J., Lichtenberg, P., Kirsch, I., & Brown, W. A. (2012). A systematic review of comparative efficacy of treatments and controls for depression. *PLoS ONE*, *7*(7), e41778. https://doi.org/10.1371/journal.pone.0041778

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd.). American Psychological Association. https://doi.org/10.1037/14136-000

Kuehberger, A., & Schulte-Mecklenbeck, M. (2018). Selecting target papers for replication. *Behavioral and Brain Sciences*, *41*, Article e139. https://doi.org/10.1017/S0140525X18000742

Laws, K. R. (2016). Psychology, replication & beyond. *BMC Psychology*, *4*(1), 30. https://doi.org/10.1186/s40359-016-0135-2

Leichsenring, F., Abbass, A., Hilsenroth, M. J., Leweke, F., Luyten, P., Keefe, J. R., Midgley, N., Rabung, S., Salzer, S., & Steinert, C. (2017). Biases in research: risk factors for non-replicability in psychotherapy and pharmacotherapy research. *Psychological Medicine*, *47*(6), 1000–1011. https://doi.org/10.1017/S003329171600324X

Leppink, J., Winston, K., & O'Sullivan, P. (2016). Statistical significance does not imply a real effect. *Perspectives on Medical Education*, *5*(2), 122–124. https://doi.org/10.1007/s40037-016-0256-6

Luborsky, L. (1999). The researcher's own therapy allegiances: A" wild card" in comparisons of treatment efficacy. *Clinical Psychology: Science and Practice*, *6*(1), 95–106. https://doi.org/10.1093/clipsy/6.1.95

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Mbakwe, A. C., Saka, A. A., Choi, K., & Lee, Y. J. (2016). Alternative method of highway traffic safety analysis for developing countries using delphi technique and Bayesian network. *Accident Analysis and Prevention*, *93*, 135–146. https://doi.org/10.1016/j.aap.2016.04.020

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). Package "bayesfactor" [R package version 0.9.12-4.2]. https://cran.r-project.org/web/packages/BayesFactor/index.html

Mullarkey, M. (2019). Banishing the unicorns: Moving toward a more replicable clinical science. https://medium.com/@mullarkey.mike/banishing-the-unicorns-movingtoward-a-more-replicable-clinical-science-fdd6e64c1f1b

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys M., ... Yarkoni, T. (2015). Promoting an open research culture Author guidelines for journals could help to promote transparency, openness, and reproducibility *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nutu, D., Gentili, C., Naudet, F., & Cristea, I. A. (2019). Open science practices in clinical psychology journals: An audit study. *Journal of Abnormal Psychology*, *128*(6), 510–516. https://doi.org/10.1037/abn0000414

Oakes, M. W. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Wiley.

Reardon, K. W., Smack, A. J., Herzhoff, K., & Tackett, J. L. (2019). An N-pact factor for clinical psychological research. *Journal of Abnormal Psychology*, *128*(6), 493–499. https://doi.org/10.1037/abn0000435

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Sakaluk, J. K., Williams, A. J., Kilshaw, R. E., & Rhyner, K. T. (2019). Evaluating the evidential value of empirically supported psychological treatments (ESTs): A meta-scientific review. *Journal of Abnormal Psychology*, *128*(6), 500–509. https://doi.org/10.1037/abn0000421

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90–100. https://doi.org/10.1037/a0015108

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials *BMC Medicine*, *8*(1), 18. https://doi.org/10.1186/1741-7015-8-18

Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's replication crisis and clinical psychological science.. *Annual Review of Clinical Psychology*, *15*(1), 579–604. https://doi.org/10.1146/annurev-clinpsy-050718-095710

Tackett, J. L., Brandes, C. M., & Reardon, K. W. (2019). Leveraging the Open Science Framework in clinical psychological assessment research. *Psychological Assessment*, *31*(12), 1386–1394. https://doi.org/10.1037/pas0000583

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., Oltmanns, T. F., & Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science*, *12*(5), 742–756. https://doi.org/10.1177/1745691617690042

Tackett, J. L., & Miller, J. D. (2019). Introduction to the special section on increasing replicability, transparency, and openness in clinical psychology. *Journal of Abnormal Psychology*, *128*(6), 487–492. https://doi.org/10.1037/abn0000455

U.S. Food and Drug Administration. (1998). *Guidance for industry: E9 statistical principles for clinical trials*. https://www-fda-gov.proxy-ub.rug.nl/regulatory-information/search-fda-guidance-documents/e9-statistical-principles-clinical-trials

van Ravenzwaaij, D., & Ioannidis, J. P. A. (2017). A simulation study of the strength of evidence in the recommendation of medications based on two trials with statistically significant results. *PLoS ONE*, *12*(3), e0173184. https://doi.org/10.1371/journal.pone.0173184

Van Ravenzwaaij, D., Monden, R., Tendeiro, J. N., & Ioannidis, J. P. A. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC Medical Research Methodology*, *19*, Article 71. https://doi.org/10.1186/s12874-019-0699-7

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Improving social and behavioral science by making replication mainstream: A response to commentaries. *Behavioral and Brain Sciences*, *41*, Article e157. https://doi.org/10.1017/S0140525X18000961

Received June 8, 2020
Revision received October 29, 2020
Accepted December 10, 2020 ■