Clinical Psychology in Europe

PsychOpen GOLD

Scientific Update and Overview

Check for updates

Research Into Evidence-Based Psychological Interventions Needs a Stronger Focus on Replicability

Helen Niemeyer[1], Christine Knaevelsrud[1], Robbie C. M. van Aert[2], Thomas Ehring[3]

[1] *Department of Clinical Psychological Intervention, Freie Universität Berlin, Berlin, Germany.* **[2]** *Department of Methodology and Statistics, Tilburg University, Tilburg, the Netherlands.* **[3]** *Department of Psychology, LMU Munich, Munich, Germany.*

Clinical Psychology in Europe, 2023, Vol. 5(3), Article e9997, https://doi.org/10.32872/cpe.9997

Received: 2022-07-29 • **Accepted:** 2023-07-17 • **Published (VoR):** 2023-09-29

Handling Editor: Cornelia Weise, Philipps-University of Marburg, Marburg, Germany

Corresponding Author: Helen Niemeyer, Division of Clinical Psychological Intervention, Department of Education and Psychology, Freie Universität Berlin, Schlossstr. 27, 12163 Berlin, Germany. Phone: 0049-30-838-54798. E-mail: helen.niemeyer@fu-berlin.de

Abstract

Background: It is a precondition for evidence-based practice that research is replicable in a wide variety of clinical settings. Current standards for identifying evidence-based psychological interventions and making recommendations for clinical practice in clinical guidelines include criteria that are relevant for replicability, but a better understanding as well refined definitions of replicability are needed enabling empirical research on this topic. Recent advances on this issue were made in the wider field of psychology and in other disciplines, which offers the opportunity to define and potentially increase replicability also in research on psychological interventions.

Method: This article proposes a research strategy for assessing, understanding, and improving replicability in research on psychological interventions.

Results/Conclusion: First, we establish a replication taxonomy ranging from direct to conceptual replication adapted to the field of research on clinical interventions, propose study characteristics that increase the trustworthiness of results, and define statistical criteria for successful replication with respect to the quantitative outcomes of the original and replication studies. Second, we propose how to establish such standards for future research, i.e., in order to design future replication studies for psychological interventions as well as to apply them when investigating which factors are causing the (non-)replicability of findings in the current literature.

Keywords

replicability, evidence-based interventions, criteria development

Highlights

- Refined replicability criteria used to identify empirically supported treatments are proposed.
- Concrete steps for refining replication in research on psychological interventions are proposed.
- A taxonomy of direct to conceptual replication adapted to research on interventions is provided.

Recent years have seen an increased focus on conceptual approaches to the replicability of research findings, and a growing number of empirical investigations on this issue, in the areas of psychology (Klein et al., 2014; Klein et al., 2018; Open Science Collaboration [OSC], 2015), economics (e.g., Camerer et al., 2016), epidemiology (e.g., Kaltiala-Heino, Työläjärvi, & Lindberg, 2019; Zisook et al., 2007) and medicine (Errington, Denis, Perfito, Iorns, & Nosek, 2021). Replicability refers to "the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected" (Bollen, Cacioppo, Kaplan, Kronsnick, & Olds, 2015; p. 3). Research related to psychological interventions has not paid the same level of attention to recent conceptual developments of replicability (Tackett et al., 2017) as seen in other fields. Yet the strong emphasis on providing evidence-based treatments in clinical psychology and psychiatry (e.g., Tolin et al., 2015) demands that clinical practice should be directly informed and guided by the best available empirical evidence on the efficacy of interventions, as typically collected in randomized controlled trials (RCTs). A precondition for evidence-based practice is that the research is replicable in a wide variety of clinical settings in order to demonstrate high external validity.

Low replicability in a research field may be partly due to so-called "hidden moderators" (Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016), which prevent the effect from being observed in a replication due to an (unobserved) moderator. Examples include characteristics of the clinical population to which the intervention is offered, treatment-related moderators, or differences in contextual variables. In other words, a study might be successfully replicated in a research outpatient clinic but not in a regular community clinic. Identifying hidden moderators is crucial in order to critically evaluate the generalizability of treatment effects to different clinical settings. "Direct" and "conceptual" are labels for replication studies depending on the similarity to the original study (LeBel et al., 2018; Zwaan, Etz, Lucas, & Donnellan, 2018). Direct replication studies allow to investigate the replicability of a study result, whereas conceptual replications serve to determine the generalizability. The relevance of replication categories has been shown in other fields, such as economics (Fiala, Neubauer, & Peters, 2022; Peters, Langbein,

& Roberts, 2018), where different replication rates were found depending on the definition of the replication studies. In order to define the similarity between original and replication study consensus on the most important characteristics is necessary. The "constraints on generality" criteria (COG; Simons, Shoda, & Lindsay, 2017) help to explicitly determine the targeted population and the study procedures in order to define a direct replication as well as to identify hidden moderators in conceptual studies. A COG statement overcomes the ambiguity of classifying replications as direct or conceptual post hoc because it specifies the target populations for the original claim (Simons et al., 2017; Simons, Shoda, & Lindsay, 2018).

In addition, non-replicability of effects may also be caused by questionable research practices (QRPs; John, Loewenstein, & Prelec, 2012). QRPs comprise a range of activities that are not a research field´s best practices, such as flexibly analyzing data until the results are significant (called *p*-hacking; Whitt et al., 2022) or hypothesizing after the results are known (called HARKing; John et al., 2012). They cause an overrepresentation of statistically significant results in the literature. Performing multiple analyses in combination with selectively reporting statistically significant results increases the number of false-positive findings in the published literature (Forstmeier, Wagenmakers, & Parker, 2017; Simmons, Nelson, & Simonsohn, 2011) and biases effect size estimation. Other factors that may cause non-replicability are reporting errors or sampling error. Importantly, in a given case of non-replicability, more than one factor can be expected to be relevant (Nosek et al., 2022).

Closely related to replicability is reproducibility. Reproducibility is obtained when the reanalysis of the original data using the same procedures arrives at the same result (Maassen et al., 2020). This is also referred to as computational or analytic reproducibility (LeBel et al., 2018). Reproducibility in psychology was investigated by Artner and colleagues (2021) who found that 70% of the reported statistical results were reproducible. When comparing reproducibility rates across disciplines, it is important to note that the definitions of replicability and reproducibility differ across disciplines (Artner et al., 2021). To date, reproducibility attempts are highly uncommon in research on psychological interventions (see also, Sandve, Nekrutenko, Taylor, & Hovig, 2013).

Do Current Research Standards Pay Enough Attention to Replicability?

Current standards for investigating psychological interventions, identifying evidence-based interventions, and making recommendations for clinical practice in clinical guidelines include criteria that are relevant for the issue of replicability. For example, the criteria for empirically supported treatments (ESTs; David, Lynn, & Montgomery, 2018) were laid down by the American Psychological Association's (APA) Division 12 in the early 1990s (Chambless & Hollon, 1998; for a recent revision, see Tolin et al., 2015).

According to these criteria, treatment effects must have been demonstrated in several independent studies, and a systematic evaluation of the methodological quality of studies as well as risk of bias needs to have been conducted, e.g., using the Cochrane risk-of-bias tool (ROB; Sterne et al., 2019) or the Grading of Recommendations, Assessment, Development and Evaluations (GRADE; Guyatt et al., 2008), consisting of six domains (e.g., risk of bias, [im-]precision of effect estimates). The need to critically assess study quality and the risk of bias has also led to the development of specific reporting standards for clinical trials, such as the Consolidated Standards of Reporting Trials (CONSORT; Schulz et al., 2010), and for reporting systematic reviews and meta-analyses, such as the Preferred Reporting Items for Systematic Reviews and Meta-Analyses" statement (PRISMA; Moher et al., 2015) or the "Meta-Analysis Reporting Standards" (MARS; American Psychological Association, 2020).

However, despite these important advances, the criteria used to identify ESTs and/or recommend clinical interventions for clinical guidelines currently have not yet been updated in line with the recent advances on replicability in the wider field of psychology and in other disciplines (Errington et al., 2021). Although the reporting standards and rating schemes address some of the variables that are relevant to assess (the lack of) replicability in studies on psychological interventions (i.e., pre-specification of the hypotheses and statistical methods, examining publication bias and heterogeneity), they neither include all of the relevant aspects nor do they make an explicit distinction between different types of replication (e.g., direct versus conceptual replications) or specify statistical criteria for a successful replication. A refinement of the criteria for replication in research on psychological interventions and specific suggestions for their application are therefore required. Moreover, an assessment of QRPs, reporting error and demands for pre-registration are currently not included in the quality assessment of clinical studies.

Currently there are only few investigations of the replicability of studies on psychological interventions. One exception is Sakaluk et al. (2019) who systematically examined the evidential value of treatments that have been classified as ESTs by standard criteria. They also applied Schimmack's replicability index (R-index, Schimmack, 2016), which focuses on statistical significance, and statistical power, as well as Bayesian meta-analysis. Results showed that statistical power and replicability estimates were low. Moreover, differences in the level of empirical support according to EST criteria did not parallel differences in indices of statistical power or replicability. Based on their analysis, the authors argued that higher methodological standards are necessary in research on psychological interventions, including sufficient statistical power and standards for reporting descriptive and inferential statistics.

In line with Sakaluk and colleagues (2019) as well as with the recommendations developed in other areas of psychology and beyond (Ioannidis, 2008; Valentine, 2009), we suggest that there is a need to enhance the replicability of research into psychological

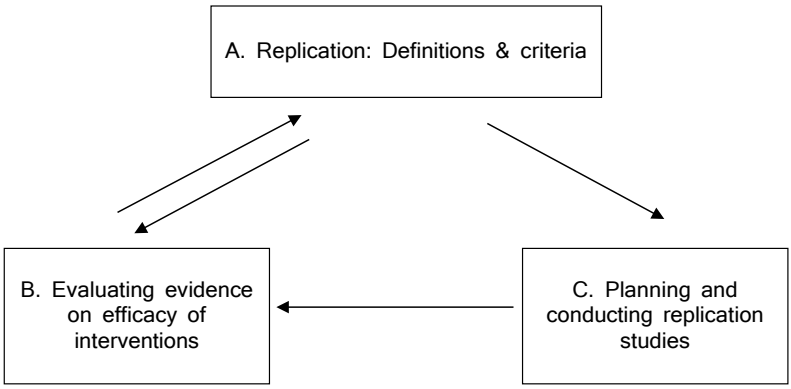
interventions and therefore propose to refine the definition of and criteria for replicability in this field. To this aim, some of the developments and resources from other areas will be adopted and, if necessary, adapted to the specificities of research on psychological interventions, as well as the given criteria and definitions refined.

Proposing a Research Strategy for Assessing, Understanding, and Improving Replicability in Research Evaluating Psychological Interventions

To improve the current situation, we propose progress in three interrelated areas (A – C; see Figure 1). The concrete steps that need to be taken are described in the following sections.

Figure 1

A Strategy for Assessing, Understanding, and Improving Replicability in Research on Psychological Interventions



A. Replication: Definitions and Criteria

First, the definition of replication currently used in research on psychological interventions is refined, based on a taxonomy of different study design types of replication, study characteristics that increase the trustworthiness of results, and statistical criteria for (un-)successful replication. At a minimum, we suggest three aspects to be crucial:

1. Taxonomy of Replication

Refining replication in research on psychological interventions is a complex endeavor. The definition of replication as aiming to duplicate the results of an original study by applying the same procedures to a new sample (Bollen et al., 2015) provides no specific

criteria as to what constitutes "the same procedure" with respect to the characteristics of an original study. Similarly, the EST criteria that treatment effects need to be demonstrated in several independent studies do not specify any details of the study designs of the required independent studies (Tolin et al., 2015).

Attempts to refine the concept of replication have been made in other areas of psychology and social sciences. We adopt the approach of LeBel et al. (2018) who provide a replication taxonomy ranging from direct to conceptual replication, depending on the degree of similarity between an original and a replication study according to several design facets, such as the operationalization of the independent and dependent variables, or investigator independence. To investigate the replicability of a treatment effect, direct replications are necessary. Conceptual replications cannot falsify the hypothesis of replicability, but can, on the other hand, help to evaluate the boundary conditions of treatment effects, the generalizability of intervention effects to different contexts, and/or the mechanisms of change underlying treatment effects. They can help to answer the question of whether (and which) hidden moderators are a cause of low replicability in "combination" with direct replications.

In order to define the characteristics that need to be identical for a study to qualify as direct replication, the constraints on generality criteria (COG; Simons et al., 2017) are applied. The COG criteria provide a general scheme for which characteristics of study participants (the targeted population), study material and procedures, and the temporal specificity of an effect are necessary to be kept the same for a replication study to be an exact replication. Principles for choosing variables for the COG should be known empirical or theoretical boundary conditions, conditions that are tied to the substance of the study, and factors that experts consider to be important.

The taxonomy suggested by LeBel and colleagues (2018) combined with the COG results in a continuum from direct to conceptual replication that can be pre-specified. The dimensions underlying the classification of replication types should include procedural details (e.g., diagnostic instruments, blinding of assessors, unconcealed allocation/risk of bias), statistical methods, contextual variables (e.g., cultural context), therapist-related factors (manual adherence), and researcher-related factors (e.g., allegiance, conflicts of interest), all of which are also potential moderator variables.

Consider, for instance, a case in which a newly developed intervention for depression is first tested against a waitlist condition (WL) and is found to be superior. A subsequent study replicates the initial study, but compares the same intervention to treatment as usual (TAU). A direct replication of the newly developed intervention for depression would need to consist of a second comparison to WL, whereas the use of a different control condition (or treatment delivery in a natural setting, or applying the intervention over the internet etc.) constitutes a conceptual replication that already tells us something about the generalizability of the intervention effects and the relative efficacy of the new treatment. As another example, we might consider a case in which a new 12-session

treatment for panic disorder is favorably tested against WL. A subsequent study also compares this new treatment to WL but uses a protocol that involves only 10 sessions, is conducted in a different country, and examines a slightly older patient population; and this second study does not find the treatment to be efficacious. Is this a failed replication study? Due to the lack of clear criteria, we are not currently able to provide a definitive answer to this question. With so many changes at once, we will never know why it did not replicate. Therefore, we need the changes to be decided on and documented more specifically; ideally, replication studies should change on one dimension at a time, so that differences in effects can be clearly attributed.

Incentives for authors for the use of a COG statement integrated into the taxonomy by LeBel and colleagues (2018) could be a protection from overly broad claims, a higher likelihood of successful replications, and inspiring follow-up studies that built upon the findings. Editors and reviewers could request a COG statement. Incentives for editors could be to have an equivalent measure to evaluate all papers, and for reviewers to have a measure for quality control, whereas for readers it helps to learn about the generality of the claims of a study (Simons et al., 2017).

2. Study Characteristics That Increase the Trustworthiness of Results

Although some important methodological factors are included in current standards of study quality assessment, there is evidence that many intervention studies fall short of characteristics that increase the trustworthiness of results. Moreover, QRPs and publication bias distort the literature and limit the replicability of studies. In addition to the existing guidelines we propose to include the following issues:

- An assessment of reporting errors should be conducted. For consistency checks of *p* values, "statcheck" can be applied (Epskamp & Nuijten, 2016).
- Pre-registration should be mandatory. The study design and analysis plan need to be pre-specified and saved in a public registry or published prior to data collection. Pre-registration is a measure to enhance transparency, document timestamped decisions, helping to differentiate between confirmatory and exploratory analyses, and for reducing *p*-hacking and HARKing. Alternatively, registered reports (RRs) are a sensible publishing format that reduces QRPs and publication bias because in RRs the peer review is conducted prior to the data collection. This emphasizes the research question and the quality of methodology instead of the significance of the results (Chambers & Tzavella, 2022). Checklists for pre-registration and recommendations for RRs have been developed in the wider field of psychology to enhance the quality of reports and pre-registrations (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016). Developments in adjacent fields are ahead, such as in biomedical research where journals banded together to make registration mandatory (Siebert et al., 2020; ClinicalTrials.gov). Registered reports and replication reports are a promising format also for clinical psychological journals.

- A systematic assessment of whether the information provided in a pre-registration is sufficient should always be conducted and should be considered in the EST criteria or guidelines.
- It should be assessed whether the final study report matches the pre-registered plan. We do acknowledge, though, that this places an extra burden on reviewers, who need to spend more time reviewing a manuscript. To reduce this burden journals can invite specialized reviewers to specifically review open science aspects of the manuscript, such as whether the pre-registration matches the final study report or checking any shared materials.
- Open data and open materials should become standard to enhance transparency. Replication studies benefit to a large extent from open data and materials. However, it should be noted that open data and materials is not a prerequisite for replicating studies (Buzbas, Devezer, & Baumgaertner, 2023). If highly sensitive data present challenges to open data principles, restricted access to data, e.g. according to the different access categories of the German Psychological Association (DGPs[1]), is also a viable alternative. This is in line with the standards of the American Psychological Association (2020), which invites researchers to share their data. It should be motivated if data cannot be shared due to ethical or legal constraints, e.g. due to participant confidentiality or missing consent. Open material and sensitive material with restricted access can both be stored in repositories, such as the Open Science Framework (OSF; osf.io).

3. Criteria for Successful Replication

As described in the taxonomy of replication, exact versus conceptual replication studies provide different information in case of replication success or failure. For example, when a conceptual replication study shows a failure of replication, this might be the result of hidden moderators. However, criteria are necessary for determining when (both direct and conceptual) replication studies are a success or failure. This conceptual issue has also not been explicitly addressed in mental health research to date, i.e. what defines a successful replication with respect to the statistical outcome of both the original and the replication study. That is, in addition to the definition of the study design as direct or conceptual replication, we propose criteria for the comparison of the quantitative results of an original and a replication study and the assessment of the replication of the study results as successful or failure, which are currently missing in research on psychological interventions.

Recent large-scale replication studies have proposed and comparatively evaluated different criteria, such as statistical significance, i.e., a study is deemed to be replicated if both the original study and the replication are statistically (non-)significant, or the

1) https://zwpd.transmit.de/images/zwpd/dienstleistungen/ethikkommission/vorlage_opendata_v1.docx

direction of both effect estimates is the same (OSC, 2015). However, an application of criteria for (un)successful replication in research on psychological interventions is lacking (see also Nosek et al., 2022).

Given that multiple statistical options to determine replication success exist (OSC, 2015; Zwaan, Etz, Lucas, & Donnellan, 2018) and that there is no consensus for one particular method, we provide a short overview of the most relevant ones: Both original and replication studies are statistically (non-)significant, the direction of both effect estimates is the same, the original effect falls within the confidence interval of the replication, original and replication result are combined and significance is assessed (OSC, 2015), statistical consistency between the original study and replications is evaluated in multisite replication projects (Mathur & VanderWeele, 2020), the small telescopes approach (Simonsohn, 2015), sceptical *p*-value (Held, 2020), and replication Bayes factor (Ly, Etz, Marsman, & Wagenmakers, 2019). These criteria represent the currently most prominent options for evaluating replicability. Recently, a comparison of seven approaches (significance, small telescopes, classical and Bayesian meta-analysis, Bayes factor and replication Bayes factor, as well as skeptical *p*-value (Held, 2020) has been conducted (Muradchanian, Hoekstra, Kiers, & van Ravenzwaaij, 2021). According to the authors, Bayesian metrics as well as meta-analytic methods were found to perform slightly better than the other approaches in terms of true and false positives rates. That is, a positive replication result is observed when the underlying true effect is non-zero or when the true effect is practically zero under different levels of publication bias in a simulation study. When evaluating replicability in research on psychological interventions, we suggest applying multiple methods, all of which should be preregistered before conducting the study. Researchers should come to conclusions based on the results of all the methods, as they perform quite similarly. Moreover, applying more methods also provides more information.

All criteria presented in the three categories taxonomy of replication, study characteristics that increase the trustworthiness of results, and criteria for successful replication are provided in an info box (see Table 1). We exemplarily propose up to three specific criteria for each COG subdomain. This list is not exhaustive, because study designs and research foci differ considerably. We recommend that researchers adapt the COG specifically to the study designs that are utilized in their research domains.

Table 1

Info Box for Replication Studies in Clinical Psychology

Overall domains / Subdomains
1. Taxonomy of replication: Constraints on generality (COG)
Participants[a] • Diagnoses • Symptom severity • Comorbidity
Materials / stimuli[a] • Manual used • Adherence to manual • Therapist training / supervision
Procedure[a] • Primary and secondary outcomes • Type of assessment (e.g., clinician-based vs. self-rated) • Type of allocation
Historical / temporal specificity[b] • Changes in diagnostic criteria (e.g. in DSM) • Common use of cellphones or internet access for app- and browser-based interventions / blended approaches
2. Study characteristics that increase the trustworthiness of results
Scales[c] for quality assessment used (according to study type)
Are reporting errors absent in the study?
Preregistration • Is a study pre-registered or is it a registered report? • Are there sufficient details in the pre-registration/registered report? • Do the analyses in the pre-registration match those in the final study report?
3. Criteria for successful replication: Methods to consider
Are the data and study materials openly available?
Are both original and replication study statistically significant?
Are the effect sizes of both the original and replication study in the same direction?
Does the effect size of the original study lie in the CI of the replication?
Is the meta-analytic effect size of combining the original and replication study statistically significant?

Overall domains / Subdomains
Is the effect size of the original study consistent with the replications in a multisite replication project (Mathur & VanderWeele, 2020)?
Small telescopes approach (Simonsohn, 2015): Is the replication effect size not significantly smaller than an effect size that would have 33% statistical power based on the sample size of the original study?
Replication Bayes factor (Verhagen & Wagenmakers, 2014; Wagenmakers, Verhagen, & Ly, 2016: Is there more evidence that the effect size of the replication is a null effect compared to the effect observed in the original study?

Note. DSM = Diagnostic and Statistical Manual; CI = Confidence interval.
[a]The proposed specific criteria are exemplary and not exhaustive. [b]This category takes into account that norms and standards change over time, and studies should be evaluated according to the respective historical period. [c]The quality assessment should be conducted according to the specific scale that is used.

B. Evaluating Evidence on Efficacy of Interventions

Beyond establishing standards for future research, it is also important to understand which factors are causing the (non-)replicability of findings in the current literature by systematically investigating moderators of treatment effects. Specifically, the relative contributions of the different variables outlined in Section A to replication success (outcome) are of interest, e.g. study quality, the type of replication design, and contextual variables. Pre-registration and a taxonomy of replication should also be systematically integrated into the classification of ESTs, clinical guidelines, and meta-analyses to enhance the transparency and methodological comparability. In addition, differences between preregistered/replicated studies and other studies should be studied.

Moderator analyses can best be addressed with meta-analytic methods. For example, the efficacy of some interventions may be highly dependent on context variables, e.g., successful replication may only be demonstrated in very direct replication designs and may have low generalizability to different contexts. Other interventions may be more context-independent, with effects being replicated even in less strict settings regarding patient or therapist characteristics or modes of treatment delivery. That is, the criteria for replication outlined above should be related to the evaluation of studies as ESTs and considered when summarizing studies in meta-analyses. Importantly, findings from this line of research can then be useful to further refine the replication concept and criteria (A). For example, if a particular therapist characteristic is not relevant for determining the replicability, it no longer needs to be taken into account when evaluating whether a study is a direct or conceptual replication.

Moderators can also include variables that are typically used to address meta-scientific questions, for example whether a study was pre-registered or provides open data. Thus, investigating pre-registration as moderator in meta-analyses against the background of replicability can shed light on whether pre-registered studies differ from

non-pre-registered studies not only in terms of treatment efficacy and study quality, but also in the replicability of their results.

C. Planning and Conducting New Replication Studies

The new definitions and criteria (A) should be used to design future replication studies for psychological interventions in order to test the consistency of treatment effects by means of direct replication studies, as well as the generalizability of findings to varying contexts on the basis of an explicit taxonomy of replication. To guide future replication research, the taxonomy of different types of replication, including the relevant dimensions of similarity vs. dissimilarity of research design features and a COG statement, tailored to research on psychological interventions, should be applied. Researchers should start by directly replicating an original treatment effect in order to investigate whether the effect exists. Then, to examine the generalizability and detect hidden moderators, they should move on to conceptual replication studies, in which they modify important aspects of the study design (e.g., treatment manual used, characteristics of treatment delivery, definition of outcome, comparison condition, and contextual factors). Depending on how many and which variables in the COG are kept equal, the similarity of replication studies along the continuum from direct to conceptual replications should be varied. Thereby it can be determined in a direct replication whether an effect exists, and its boundary conditions and mechanisms can be identified in conceptual replications. Thus, the distinction between direct and conceptual replication studies will be helpful for assessing the heterogeneity of findings for a particular intervention. That is, conceptual replications will test whether the proposed constraints on generality are accurate, leading to a more refined understanding of the robustness of effects. A systematic program of research should evaluate how the size of an effect varies as a function of those constraints (Simons et al., 2018).

An important first step is to conduct an exact replication study to confirm the result of the original study. Second, in order to identify the most important hidden moderators assessed conceptual replications and also meta-analyses should be conducted, once a sufficient number of replication studies has been conducted where as rule-of-thumb can be used that 5 to 10 studies are needed per included moderator in a meta-analysis (van Houwelingen, Arends, & Stijnen, 2002). An agreed set of quality standards and criteria based on the COG concept that must be included in clinical trial reports should be established and constantly refined. The criteria and quality standards will inform future replication studies, and should also be taken into account by experts evaluating the current state of evidence of an intervention, e.g. when developing clinical guidelines or establishing EST.

In the long term, the adoption of COG statements will lead to a more cumulative understanding of the scope of the effects of psychological interventions.

Conclusion

The current gold standard in evidence-based psychological treatments can be criticized for not paying sufficient attention to replicability. The current discussion surrounding replicability and reproducibility (Ioannidis, 2012; Munafò et al., 2017) offers the opportunity to define and potentially increase replicability also in mental health research. The development of an explicit concept and taxonomy of replication will enable the classification of studies investigating clinical interventions with respect to their similarity with original studies and will aid in planning and conducting replication studies in the future. The criteria themselves need to be continuously updated based on advances in replicability research in other areas and informed by emerging evidence regarding (moderators of) replicability in mental health research.

However, also a number of limitations have to be noted. Even if an effect is true, it is possible to fail to replicate due to seemingly innocuous differences in the implementation of the study (i.e. due to "hidden moderators"). Small variations in studies are unavoidable and exact replication is strictly impossible. Baribault and colleagues (2018) suggest to randomize variables that may be moderators of an effect in replication studies in order to test the robustness and generalizability of an effect. They propose a random selection of potential moderators, that is characteristics of the design that are not supposed to make a difference. If characteristics do not affect the results, this means that the results are more generalizable and to alter minor things should not matter. This is suggested for experimental research, e.g. different implementations of the same stimulus could be used to study whether the results are robust. However, as a large number of studies is necessary for this approach, it is not applicable to RCTs on psychological interventions. Compared to research in social psychology, studies in research on psychological interventions are much more costly and time-consuming, which makes it more difficult to study replicability. The question of how much money and effort researchers should spend on studying replicability given that conducting such studies is expensive in clinical psychology is related to the decision when to move on to other research topics, because studying replicability means at the same time that less scientific progress with respect to new findings will be made. This demonstrates that not all recommendations from social psychology are applicable in clinical psychology.

Based on this we would like to invite the readers to engage in discussions about the concrete criteria and next steps that we proposed. Designing replication studies should be based on empirical evidence and on theoretical predictions (Simons et al., 2018) and considered to be a collective research enterprise.

Funding: This research received funding from the Berlin University Alliance, an excellence initiative of the German Research Foundation (312_OpenCall_3).

Acknowledgments: We would like to thank Sarah Mannion for language editing.

Competing Interests: The authors have declared that no competing interests exist.

References

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods, 26*(5), 527–546. https://doi.org/10.1037/met0000365

American Psychological Association. (2020). *Publication manual of the American Psychological Association: The official guide to APA style* (7th ed.). American Psychological Association.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White, C. N., De Boeck, P., & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences of the United States of America, 115*(11), 2607–2612. https://doi.org/10.1073/pnas.1708285114

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Kronsnick, J. A., & Olds, J. L. (2015). *Social, behavioral, and economic sciences perspectives on robust and reliable science.* National Science Foundation. https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

Buzbas, E. O., Devezer, B., & Baumgaertner, B. (2023). The logical structure of experiments lays the foundation for a theory of reproducibility. *Royal Society Open Science, 10*(3), Article 221042. https://doi.org/10.1098/rsos.221042

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour, 6*, 29–42. https://doi.org/10.1038/s41562-021-01193-7

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

David, D., Lynn, S. J., & Montgomery, G. H. (2018). *Evidence-based psychotherapy: The state of the science and practice.* Wiley.

Epskamp, S., & Nuijten, M. B. (2016). *Statcheck: Extract statistics from articles and recompute* p *values* (R package Version 1.2.2) [Computer software]. https://CRAN.R-project.org/package=statcheck

Errington, T. M., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Challenges for assessing replicability in preclinical cancer biology. *eLife, 10*, Article e67995. https://doi.org/10.7554/eLife.67995

Fiala, N., Neubauer, F., & Peters, J. (2022). *Do economists replicate?* (Ruhr Economic Papers, No. 939). http://hdl.handle.net/10419/250076

Forstmeier, W., Wagenmakers, E.-J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings – A practical guide. *Biological Reviews of the Cambridge Philosophical Society, 92*(4), 1941–1968. https://doi.org/10.1111/brv.12315

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal, 336*(7650), 924–926. https://doi.org/10.1136/bmj.39489.470347.AD

Held, L. (2020). A new standard for the analysis and design of replication studies. *Journal of the Royal Statistical Society Series A: Statistics in Society, 183*(2), 431–448. https://doi.org/10.1111/rssa.12493

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7*(6), 645–654. https://doi.org/10.1177/1745691612464056

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kaltiala-Heino, R., Työläjärvi, M., & Lindberg, N. (2019). Gender dysphoria in adolescent population: A 5-year replication study. *Clinical Child Psychology and Psychiatry, 24*(2), 379–387. https://doi.org/10.1177/1359104519838593

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., . . . Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology, 45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, S., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. B., Busching, R., . . . Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443–490. https://doi.org/10.1177/2515245918810225

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science, 1*(3), 389–402. https://doi.org/10.1177/2515245918787489

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2019). Replication Bayes factors from evidence updating. *Behavior Research Methods, 51*(6), 2498–2508. https://doi.org/10.3758/s13428-018-1092-x

Maassen, E., van Assen, M. A. L. M., Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PLoS ONE, 15*(5), Article e0233107. https://doi.org/10.1371/journal.pone.0233107

Mathur, M. B., & VanderWeele, T. J. (2020). New statistical metrics for multisite replication projects. *Journal of the Royal Statistical Society Series A: Statistics in Society, 183*(3), 1145–1166. https://doi.org/10.1111/rssa.12572

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., & the PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews, 4*(1), Article 1. https://doi.org/10.1186/2046-4053-4-1

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*(1), Article 0021. https://doi.org/10.1038/s41562-016-0021

Muradchanian, J., Hoekstra, R., Kiers, H., & van Ravenzwaaij, D. (2021). How best to quantify replication success? A simulation study on the comparison of replication success metrics. *Royal Society Open Science, 8*(5), Article 201697. https://doi.org/10.1098/rsos.201697

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology, 73*(1), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods, 48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Peters, J., Langbein, J., & Roberts, G. (2018). Generalization in the tropics – Development policy, randomized controlled trials, and external validity. *The World Bank Research Observer, 33*(1), 34–64. https://doi.org/10.1093/wbro/lkx005

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

Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS Computational Biology, 9*(10), Article e1003285. https://doi.org/10.1371/journal.pcbi.1003285

Schimmack, U. (2016). *The replicability-index: Quantifying statistical research integrity.* https://wordpress.com/post/replication-index.wordpress.com/920

Schulz, K. F., Altman, D. G., Moher, D., & the CONSORT Group. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *Trials, 11*(1), Article 32. https://doi.org/10.1186/1745-6215-11-32

Siebert, M., Gaba, J. F., Caquelin, L., Gouraud, H., Dupuy, A., Moher, D., & Naudet, F. (2020). Data-sharing recommendations in biomedical journals and randomised controlled trials: An audit of journals following the ICMJE recommendations. *BMJ Open, 10*(5), Article e038887. https://doi.org/10.1136/bmjopen-2020-038887

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2018). Constraints on generality statements are needed to define direct replication. *Behavioral and Brain Sciences, 41*, Article e148. https://doi.org/10.1017/S0140525X18000845

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., . . . Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ, 366*, Article l4898. https://doi.org/10.1136/bmj.l4898

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

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*(4), 317–338. https://doi.org/10.1111/cpsp.12122

Valentine, J. C. (2009). Judging the quality of primary research. In H. Cooper, L. V. Hedges, & J. D. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (Vol. 2, pp. 129-146). Russel Sage Foundation.

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 of the United States of America, 113*(23), 6454–6459. https://doi.org/10.1073/pnas.1521897113

van Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine, 21*(4), 589–624. https://doi.org/10.1002/sim.1040

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General, 143*(4), 1457–1475. https://doi.org/10.1037/a0036731

Wagenmakers, E.-J., Verhagen, J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods, 48*(2), 413–426. https://doi.org/10.3758/s13428-015-0593-0

Whitt, C. M., Miranda, J. F., & Tullett, A. M. (2022). History of replication failures in psychology. In W. O'Donohue, A. Masudo, & S. Lilienfeld (Eds.), *Avoiding questionable research practices in applied psychology* (pp. 73-97). Springer.

Zisook, S., Rush, A. J., Lesser, I., Wisniewski, S. R., Trivedi, M., Husain, M. M., Balasubramani, G. K., Alpert, J. E., & Fava, M. (2007). Preadult onset vs. adult onset of major depressive disorder: A replication study. *Acta Psychiatrica Scandinavica, 115*(3), 196–205. https://doi.org/10.1111/j.1600-0447.2006.00868.x

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*, Article e120. https://doi.org/10.1017/S0140525X17001972