General Article

Advances in Methods and Practices in Psychological Science
April-June 2025, Vol. 8, No. 2, pp. 1–22

DOI: 10.1177/25152459251328273
www.psychologicalscience.org/AMPPS

The Design of Current Replication Studies: A Systematic Literature Review on the Variation of Study Characteristics

Jerome Hoffmann[1], Mathias Twardawski[2], Johanna M. Höhs[3], Anne Gast[3], Steffi Pohl[4], and Marie-Ann Sengewald[1]

[1]Leibniz Institute for Educational Trajectories, Bamberg, Germany; [2]Department of Psychology, Ludwig-Maximilians-Universität München, München, Germany; [3]Department of Psychology, University of Cologne, Köln, Nordrhein-Westfalen, Germany; and [4]Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany

Abstract

In what aspects do replication studies differ from their primary studies? This question is central for providing insights into the reasons for the nonreplicability of psychological effects. So far, research on potential explanations for the nonreplicability of effects has mainly focused on publication bias and methodological challenges related to measurement error or statistical inference. The recently developed causal-replication framework directs attention toward controlling for differences in study characteristics, including variations in treatment conditions, outcome measures, recruitment, causal estimates, time, location, population, and setting. To contribute to this aim, we conducted a systematic literature review to investigate the design practices of current replication studies. We preregistered the assessment of study characteristics in a detailed review protocol and investigated the available information and intended or unintended variations across primary and replication studies. To do this, we compiled a database of studies that aimed to replicate a causal effect of a clearly stated primary study and that were published in impactful social- and cognitive-psychological journals between January 2017 and August 2022. Our review results highlight that compared with the primary study, authors of replication studies predominantly focus on controlling specific study characteristics in (i.e., methods, procedures, analysis) while often neglecting other study characteristics, such as population or setting. Furthermore, the results indicate that in most replication studies, multiple study characteristics are varied in the study comparison or are insufficiently reported. Accordingly, we discuss prevalent variations, reporting standards, and strategies for planning future replication studies.

Keywords

replication, effect heterogeneity, external validity, review protocol, metascience, open data, preregistration

Received 5/24/24; Revision accepted 2/27/25

In the last decade, the scientific community has recognized the importance of replication research for a better understanding of the reasons for effect heterogeneity (e.g., Nosek et al., 2022). For this goal, it needs to be explicitly stated in which way a replication is similar to a primary study and in which way it differs. Only when all sources of effect heterogeneity are considered can clear conclusions regarding replicability be drawn and a description of the conditions under which study results replicate be possible. Suggestions for a systematic design of replication studies traditionally focus on repeating the procedures of a primary study as closely as possible (Schmidt, 2009). However, prominent large-scale replication projects, such as the Open Science Collaboration (OSC; 2015) and the Many Labs Replication Project

Corresponding Author:
Jerome Hoffmann, Leibniz Institute for Educational Trajectories, Bamberg, Germany
Email: jerome.hoffmann@lifbi.de

(Klein et al., 2014), have revealed significant heterogeneity in various psychological effects despite carefully controlling the material and implementation procedures between a primary study and its replication.

These results sparked further discussions regarding the factors contributing to low replicability in psychological research. Next to methodological challenges concerning measurement error and statistical inference (e.g., Fiedler & Prager, 2018; Lewandowsky & Oberauer, 2020; Loken, & Gelman, 2017), unintended differences between the characteristics of primary and replication studies provide possible explanations (e.g., Nosek & Errington, 2017; Van Bavel et al., 2016). For instance, even with the greatest effort, it will not be possible to exactly replicate an existing study because, for example, the time, the population considered, the sampling strategy, or the laboratory in which the study is conducted can differ. All these differences can cause effect heterogeneity because it is well known in causal-inference theory (i.e., factors that threaten the external validity of effect estimates; Cronbach & Shapiro, 1982; Shadish et al., 2002).

Recently, the causal-replication framework (CRF; Steiner et al., 2019; Wong & Steiner, 2018) provided definitions and assumptions that formally specify all potentially relevant study characteristics for replicability across studies. We apply the CRF to examine which study characteristics are reported and which study characteristics are typically varied or held constant in studies that focus on the replication of causal effects. For this, we compiled a database of recently published replication studies and their primary studies in two psychological fields (i.e., cognitive and social psychology) that play pivotal roles in discussions surrounding the replication crisis (e.g., OSC, 2015). Using all available information sources (i.e., the replication and primary studies and supplemental materials), we assessed the availability of information on the different study characteristics and their intended or unintended variation according to the preregistered criteria.

Our results describe the current reporting standards for study characteristics and the prevalent differences between primary and replication studies. In addition, we provide comparisons between the two psychological disciplines and regarding other structural indicators of the replication studies (including author overlap between replication and primary study, preregistration of the replication, type of replication, and time gap to the primary study). Finally, we discuss different practices and systematic design approaches of current and future replication studies.

The CRF

The CRF (Steiner et al., 2019; Wong & Steiner, 2018) provides a general theoretical basis, which we use to specify relevant study characteristics. Based on causal theory (i.e., the potential-outcomes framework; Rubin, 1974, 2005), the CRF encompasses five assumptions, two replication assumptions and three individual study assumptions, for the replication of a well-defined causal effect in the sense of a treatment-control contrast across two or more studies (for the descriptions of all assumptions and examples for causes of effect heterogeneity, see Table 1). Effect heterogeneity can occur when one or more of the CRF assumptions are not met.

The individual study assumptions define requirements for causal inference in a single study, including the correct identification, estimation, and reporting of the effect of interest. These assumptions can be addressed by using designs and analyses for causal-effect estimation in (quasi)experiments (see e.g., Shadish et al., 2002). Explanations for effect heterogeneity relating to these assumptions refer to researchers' degrees of freedom in the treatment assignment and analysis methods in each study. For instance, variation of treatment conditions within or between participants, the selection of control variables and the specification of modeling assumptions, or methodological issues, such as measurement error, can affect the individual study results and their replicability (see e.g., Stanley & Spence, 2014). Open-science principles and preregistration aim to improve the transparency and reproducibility of the experimental design, data, and analysis in a study (National Academies of Sciences, Engineering, and Medicine, 2018).

Although causal inference in each study, as indicated by the individual study assumptions, is a prerequisite to compare effects between studies, additional variations can occur between studies such that the effects differ. The CRF thus defines additional replication assumptions that describe the necessary conditions for reproducing the same effect across studies. The first replication assumption addresses the stability of both the treatment and outcome across studies. This corresponds to a traditional understanding of replications (e.g., Schmidt, 2009) by closely following the procedures of an original study and by using identical materials. Large-scale replication projects (e.g., Klein et al., 2014; OSC, 2015) often address this assumption by collaborating with the original authors to ensure a high level of fidelity to the material and procedure of the original studies. The second assumption pertains to the equivalence of the causal estimand (i.e., the effect of interest for the specific treatment-control contrast given the population and setting under investigation). In particular, intervention research highlights that the characteristics of the recruited participants and the setting and timing for conducting a study can result in different effect estimands (e.g., description of the external validity of effects in Shadish et al., 2002). Increasing attention has also been given to these study characteristics in current standards for conducting

Table 1. Assumptions of the Causal-Replication Framework (Steiner et al., 2019)

Design assumptions for replicating a causal effect		Examples of variations that can contribute to effect heterogeneity
Replication assumptions across studies	Treatment and outcome stability	
	No variation in treatment and control conditions	→ Changes to the study material (e.g., instructions, stimuli, vignettes), control group (e.g., active or waitlist control), . . .
	No variation in outcome measures	→ Different instruments, assessment modes, or changes of response scale, . . .
	No mode-of-study-selection effects	→ Motivational differences in case of selection with or without incentives, . . .
	No peer, spillover, or carryover effects	→ Learning effects across studies by identical researchers, participants, . . .
	Equivalence in the causal estimand	
	Same causal quantity of interest	→ Different effects (e.g., average treatment effect, intent-to-treat effect), . . .
	Identical effect-generating processes	→ Social change, impact of historical events (e.g., war, pandemics), . . .
	Identical distribution of population characteristics	→ Different samples (e.g., college students, clinical sample, or provider sample) and different inclusion and exclusion criteria, . . .
	Identical distribution of setting variables	→ Different electronic equipment, different physical settings (e.g., field, online, or laboratory), different social setting, . . .
Individual study assumptions	Unbiased identification of effects	→ Causal inference bias (e.g., selection bias, attrition, noncompliance), . . .
	Unbiased estimation of effects	→ Analysis bias (e.g., incorrect model assumptions, measurement error), . . .
	Correct reporting	→ Mistakes in publications, unreproducible documentation, . . .

replication research (see e.g., Brandt et al., 2014; LeBel et al., 2018). For instance, Brandt et al. (2014) proposed specifying differences in the setting, remuneration, and participant populations next to the procedures and material of a study in their replication recipe.

Reporting and Variation of Study Characteristics

The CRF offers guidance on factors that potentially cause effect heterogeneity. To better understand the conditions under which effects replicate, it is important to report on all these aspects such that it can be described in which ways a replication study resembles the primary study and in which ways it differs. Variations of study characteristics are not inherently a problem but can be intended, depending on the aim of a replication study. Direct replications require equivalent study characteristics in comparison with the respective primary study for recovering the same effect with new data to accumulate evidence for the existence of an effect (e.g., LeBel et al., 2017; Schmidt, 2009). In contrast, conceptual replications vary study characteristics to inform about the generalizability and boundary conditions of an effect (Borsboom et al., 2021; LeBel et al., 2017). A few examples addressing the generalizability are whether a theory holds for all people or just for people in Western countries or whether a certain intervention works only in in-person settings or also in an online setting.

Contemporary debates about the replicability of psychological findings highlight that the generalization of study results is often limited because of restricted samples and the specific conditions under which effects were found (Bauer, 2023; Yarkoni, 2022). Although this supports the impact of the replication assumptions for describing sources of effect heterogeneity, the CRF furthermore highlights the benefit of a systematic variation of such study characteristics for explaining effect heterogeneity in the comparison of two studies (Steiner et al., 2019; Wong et al., 2022). Specifically, when relaxing only a single replication assumption while ensuring that all other assumptions are met, strong causal inference for the impact of the variation of study characteristics on effect heterogeneity is possible. For example,

if two studies on reducing prejudice were conducted at the same time, with the same procedures, using the same outcome measures, recruiting participants with the same characteristics, focusing on the same causal quantity, but varying the instructions of the intervention group, then one could attribute differences in effect estimates between the two studies to the altered instructions. In practice, varying one study characteristic (i.e., relaxing one assumption) is easy, but it is challenging to control the impact of other study characteristics, that is, to satisfy all other assumptions by ensuring that no unintended variations are prevalent. For example, when evaluating the impact of the setting (laboratory vs. online setting) for a study on facial recognition, usually not only the setting is varied but also the participants that one may recruit or the electronic equipment because of differences in characteristics of the monitors used. Thus, if differences in effect estimates between a primary study and a replication study are found, one does not know whether these are due to the setting, the participants, the electronic equipment, or a combination of these. Each involved variation can amplify or reduce effect differences between studies. In addition, interaction effects among the variations are possible, too. Thus, in a comparison of two studies (i.e., one replication and primary study), a joint variation of multiple study characteristics provides only limited insights into the reasons for effect heterogeneity. Most recently, Wong et al. (2022) used the CRF for designing conceptual replication studies that control for unintended variations between studies per design for evaluating effect heterogeneity of a teacher training in educational research. We now apply the CRF to other disciplines to study the current replication practices.

Next to the discipline and the type of replications (i.e., direct or conceptual), other structural indicators may be relevant for the reporting and variation of study characteristics in replication research. For instance, preregistration was introduced as a central tool for enhancing the reporting and design of empirical studies, including replications (e.g., Hardwicke & Wagenmakers, 2023; Nosek et al., 2018). Furthermore, overlap of the authors of the primary and replication studies is associated with more successful replications (e.g., Lemons et al., 2016; Makel et al., 2012; Makel & Plucker, 2014), maybe because study characteristics can be more readily kept constant in replication studies. However, alternative explanations are also possible, such as the time differences between the replication and primary study that may also affect the available information on a primary study. A more comprehensive examination of study characteristics is warranted as a first step to understand the prevalent variations in replication research and possible associations with disciplinary or other structural indicators in replication research.

Research Questions

We aimed to investigate practices for designing replication studies in a systematic literature review. Several literature reviews on replication studies have been conducted that focused on publication and success rates of replication studies but did not assess differences in study characteristics between primary and replication studies (e.g., Cook et al., 2016; Lemons et al., 2016; Makel et al., 2012; Perry et al., 2022). Our review extends the current knowledge on replication research by providing first insights into (a) how well different study characteristics are reported and (b) which study characteristics are typically varied or held constant.

We focus on replication studies that aim to replicate a causal effect reported in a specific primary study and that have been published in recent years—after the first large-scale replication studies were conducted (Klein et al., 2014; OSC, 2015), after suggestions for systematic replications were proposed (Brandt et al., 2014; Schmidt, 2009), and after unintended differences in direct replications were highlighted (Nosek & Errington, 2017; Van Bavel et al., 2016). We include two psychological disciplines (i.e., cognitive and social psychology) that received much attention following "failed" replication research (e.g., Brandt et al., 2014; OSC, 2015; Schmidt, 2009). We deemed it possible that the reporting and variation of study characteristics may differ between the disciplines. In addition, we provide comparisons for author overlap between studies, preregistration of the replication, type of the replication, and time gap to the primary study.

Disclosures

Preregistration, data, and materials

The methods of this literature review were preregistered on the OSF (https://osf.io/yxgc8). We followed this preregistration and report deviations when necessary. Furthermore, the data, analysis script to reproduce the results, and additional material are available at https://osf.io/yv5ns/.

Method

Transparency and reporting

The present research conforms to the PRISMA guidelines for reporting systematic reviews (Page et al., 2021). In line with the descriptive focus of the review, we report on information sources, eligibility criteria, search strategy, selection process, data items, and data-collection process. The data management, selection process, and data-collection process were managed using CADIMA (Version 2.2.3; Kohl et al., 2018), a free web tool for

Table 2. List of Journals Included in the Literature Review

Cognitive-psychology journals		Social-psychology journals	
Journal of Experimental Psychology: General	(5.498)	*Journal of Personality and Social Psychology*	(8.460)
Journal of Memory and Language	(4.521)	*British Journal of Social Psychology*	(6.920)
Journal of Cognition	(3.85)	*Social Psychological and Personality Science*	(5.316)
Thinking & Reasoning	(3.537)	*Personality and Social Psychology Bulletin*	(4.560)
Cognitive Psychology	(3.468)	*European Journal of Social Psychology*	(3.930)
Journal of Experimental Psychology: Learning, Memory, and Cognition	(3.140)	*Sex Roles: A Journal of Research*	(3.812)
Journal of Experimental Psychology: Human Perception and Performance	(3.077)	*Social and Personality Psychology Compass*	(3.798)
Memory & Cognition	(2.482)	*Journal of Experimental Social Psychology*	(3.532)
Psychological Research	(2.424)	*Social Psychology*	(3.444)
Attention, Perception, & Psychophysics	(2.157)	*Group Processes & Intergroup Relations*	(2.708)
The Quarterly Journal of Experimental Psychology	(2.138)	*Journal of Social and Personal Relationships*	(2.681)
Visual Cognition	(1.875)	*Social Psychology of Education: An International Journal*	(2.614)
		Journal of Cross-Cultural Psychology	(2.577)
		International Review of Social Psychology	(2.5)
		Social Psychology Quarterly	(2.163)
		Comprehensive Results in Social Psychology	(3.9[a])

Note: Journal impact factors are presented in parentheses.
[a]CiteScore (Scopus).

conducting literature reviews. We performed data preparation, analyses, and visualization in R (Version 4.3.0; R Core Team, 2023) using the packages *ggplot2* (Version 3.4.2; Wickham, 2016) and *readxl* (Version 1.4.2; Wickham & Bryan, 2023).

Database

Information sources. As sources for the replication studies, we preregistered a list of impactful, peer-reviewed journals for cognitive- and social-psychological research from the Scimago Journal & Country Ranking. Ultimately, we selected 12 journals to represent cognitive-psychological research and 16 journals to represent social-psychological research. For a list of the selected journals, see Table 2. We based the selection on an examination of the descriptions and author guidelines of potential journals to assess whether they were representative of the respective fields and ensured that they did not explicitly exclude replication studies from publication. A more detailed description of the journal-selection process is presented in the preregistered review protocol.

Eligibility criteria. We preregistered eight eligibility criteria to specify the range of studies we aimed to include in the literature review: (a) The article was published in one of the selected journals; (b) the article was written in English; (c) the article was published between January 2017 and August 2022; (d) the article was an empirical article with primary data; (e) the article included a replication study as one of the main topics of the article; (f) the replication study was an experiment or quasiexperiment with a treatment-control comparison for applying the CRF; (g) the replication study did not employ functional imaging methods, such as PET or functional MRI, or other techniques measuring brain activity, such as EEG; and (h) the replication study focused on human research.

To summarize, we focused on current studies in one of the two psychological fields that aimed to conduct a replication of a specific primary study with new data. Criteria f through h above enhance the comparability of the studies according to characteristics that can affect the replicability of the research results. Moreover, we considered (quasi)experimental research with causal effects and excluded other research traditions that are prevalent in neurocognitive studies and animal research.

Search strategy. We employed a similar search strategy as previous literature reviews on replication studies (e.g., Lemons et al., 2016; Makel et al., 2012). Specifically, we used APA PsycInfo as the primary database to identify potentially relevant records from the journals listed in

Table 2 between January 2017 and August 2022. One journal (i.e., *Comprehensive Results in Social Psychology*) was not covered by PsycInfo, so we used the advanced search function on the journal's website. Our literature search focused on articles that included the term "replicat*" in their title, abstract, or keywords.

Selection process. In the first step, we investigated the title and abstract of each identified record. We applied the criteria publication language (Criteria b), that the article included an empirical study with original data (Criteria d), and that the article included a replication study (Criteria e). To this end, we distributed the records among four independent raters (J. Hoffmann, M.-A. Sengewald and two trained student assistants). Records that met all specified criteria were deemed eligible for the next screening step. In the second step, we screened the full-text articles for all eligibility criteria. Two independent raters (J. Hoffmann and a trained student assistant) were assigned to evaluate the articles, and 20% of the reports were independently screened by both raters to determine metrics of interrater agreement. If all the eligibility criteria were satisfied, the respective articles were considered eligible for data extraction. Because of limitations on the availability of research personnel, we specified a limit of 60 articles (i.e., 30 articles for cognitive psychology and 30 articles for social psychology) to be investigated in the comprehensive data collection. We identified these articles by random sampling from all eligible articles. The selected articles underwent an expert rating to assess their fit to the research fields as a form of quality control.

Data collection and analysis

As the first step in the data-collection process, we collected structural indicators from the selected articles. The structural indicators included the publication year of the primary article and replication article, overlap in authorship (i.e., at least one author worked on both articles), research field as indicated by the journal (i.e., cognitive psychology or social psychology), the type of replication as reported by the authors (i.e., direct or conceptual), and whether the replication study had been preregistered. We used these data to describe the replication studies in our database and to conduct additional analyses on the availability of information and equivalence of study characteristics.

We had preregistered the concrete information that we wanted to collect regarding the availability of information and the equivalence of study characteristics. For this, we identified eight study characteristics and more detailed subcharacteristics that potentially affect replicability in psychological research according to the replication assumptions of the CRF (Steiner et al., 2019). For a comprehensive list of the coded information, see Table 3. To specify the stability of the treatment and outcome across studies, we assessed information on the treatment conditions, outcome measures, and recruitment strategy. As relevant information for the equivalence of the causal estimand across studies, we considered the specification of the effect of interest (i.e., causal quantity) and characteristics for the timing, location, population, and setting under which the studies were conducted.

We collected all data manually by first inspecting the method section of the respective replication study. If necessary and applicable, we used information from the complete article and from preregistrations and supplementary material. As an additional source of information, we took the primary study's article whenever necessary to directly compare methods and procedures between the replication and its primary study. In case multiple replication studies were reported in a single journal article, we included only one replication per primary study to control for similarities in the design of replication series of the same authors on the same primary studies. For this, we included only the replication study that was reported first in the article for a specific primary study. Because of the extensive amount of information, all records were single-coded by J. Hoffmann except for a small subset of articles (i.e., two articles for each psychological field) that were coded by M.-A. Sengewald to verify the data-collection process.

Availability of information. We coded the availability of information for each subcharacteristic into three categories of where we had to look to obtain the necessary information: (a) replication article alone (including preregistration and supplemental material), (b) replication article and primary article, and (c) no sufficient information. We rated replication article alone if the authors of the replication study described how a subcharacteristic was implemented in the primary study and the replication study or if they claimed equivalence or variation of subcharacteristics in the article or supplementary material (including preregistrations). If the authors indicated equivalence in the replication study or supplementary material, we additionally checked the primary article for contrary evidence. If the equivalence of a subcharacteristic could not be determined from the replication article alone, we used information from the primary article as an additional source of information. In case the primary article was necessary to determine equivalence or variation between studies, we rated the availability of information as replication article and primary article. The rating of no sufficient information was given only if we were not able to determine the equivalence of a subcharacteristic based on all available information sources.

Table 3. List of Information Coded for Analyses

Structural indicator	Relevant information			Coded information
Year of publication	Years in which replication study and primary study were published			Publication years
Overlap in authorship	At least one author worked on both articles			Two levels: yes/no
Research field	As indicated by journal (see Table 2)			Two levels: cognitive/social psychology
Type of replication	As indicated by an explicit author statement			Two levels: direct/conceptual replication
Preregistration	The replication study has been preregistered			Two levels: yes/no
Study characteristic	**Relevant information**	**Criteria for equivalence**	**Criteria for variation**	**Coded information**
Treatment conditions				
Content	Underlying effect or psychological processes investigated in the study	The same psychological processes are under investigation (e.g., social dilemma, stimulus contingency, training tasks)	Different psychological processes are under investigation	Availability of information with three levels: replication article alone/replication and primary article/no sufficient information Equivalence with four levels: equivalent/varied (intentional)/varied (unintentional)/no sufficient information
Material	Treatment materials, such as stimuli, vignettes, or intervention	The same treatment materials are being used	Treatment material is varied between studies (including simple translations, recreations, adaptations)	
Delivery	Implementation of the study, including instructions, instructors, procedure, and protocol	Same implementation of the treatment	Implementation differs between studies	
Outcome				
Construct	Main outcome under investigation, for example, prejudice, vigilance, or reaction times	The same construct is measured as indicated by author statement or instrument used	Different constructs are used (even if they are related, e.g., aggression vs. frustration)	Availability of information with three levels: replication article alone/replication and primary article/no sufficient information Equivalence with four levels: equivalent/varied (intentional)/varied (unintentional)/no sufficient information
Instrument	How the outcome was measured, that is, which questionnaire or method was used	The same instrument or method is being used in both studies	Different instruments or methods are being used (including simple translations, short forms, different measurement scales, or other obvious variations)	

(continued)

Table 3. *(continued)*

Study characteristic	Relevant information	Criteria for equivalence	Criteria for variation	Coded information
Recruitment strategy				
Incentives	Compensation of participants	Both studies use the same incentives (e.g., course credit, monetary)	Incentives differ between both studies	Availability of information with three levels: replication article alone/ replication and primary article/no sufficient information Equivalence with four levels: equivalent/varied (intentional)/varied (unintentional)/no sufficient information
Advertisement	Study advertisement to participants	The study is promoted or described equally to participants in both studies	Description or promotion of the studies differs	
Causal quantity	Methods of data analysis and statistical effect of interest, that is, main effect or interaction effect	The same statistical effect is considered in both studies as indicated by author statement or analysis methods used	Different statistical effects are considered in both studies	Availability of information with three levels: replication article alone/ replication and primary article/no sufficient information Equivalence with four levels: equivalent/varied (intentional)/varied (unintentional)/no sufficient information
Timing	Time frame in which the data were collected; if no information is available, the publication year is taken as proxy	Data were collected in the same year	Data were collected in different years	Availability of information with three levels: replication article alone/ replication and primary article/no sufficient information Equivalence with four levels: equivalent/varied (intentional)/varied (unintentional)/no sufficient information
Location	Where the data were collected, that is, which country, city, or university	Data were collected in the same country	Data were collected in different countries	Availability of information with three levels: replication article alone/ replication and primary article/no sufficient information Equivalence with four levels: equivalent/varied (intentional)/varied (unintentional)/no sufficient information
Population				
Criteria	Target population or exclusion criteria for participants	The same eligibility criteria apply in both studies	Eligibility criteria differ between studies	Availability of information with three levels: replication article alone/ replication and primary article/no sufficient information Equivalence with four levels: equivalent/varied (intentional)/varied (unintentional)/no sufficient information
Characteristics	Descriptive statistics of demographic variables (i.e., gender and age)	The samples do not differ in terms of population characteristics	Comparisons of relevant population characteristics reveal differences	

(continued)

Table 3. *(continued)*

Study characteristic	Relevant information	Criteria for equivalence	Criteria for variation	Coded information
Setting				
Social	Whether participants were subjected to the same social setting (e.g., individual or group setting, under supervision, or virtual interaction)	The same social setting is used in both studies	Different social settings are used in both studies	Availability of information with three levels: replication article alone/ replication and primary article/no sufficient information Equivalence with four levels: equivalent/varied (intentional)/varied (unintentional)/no sufficient information
Physical	Whether the study was implemented online or in a laboratory, also including test setup, software, and topic-specific conditions	The same physical setting is used in both studies	Different physical settings are used in both studies	
Device for data collection	Which method was used to collect data (e.g., interview, pen and paper, or computer)	The same device is used for data collection	Different device for data collection is used	

For the formal details of the analysis regarding the availability of information ratings for the different study characteristics, see Appendix A. We first calculated the proportion of studies that provided sufficient information on each subcharacteristic. Then, we calculated an average score (i.e., we averaged the proportions of the respective subcharacteristics that refer to a common study characteristic). Furthermore, we determined an availability index for each study characteristic from the average scores that can take on values between 0 and 1. An availability index of 1 for a specific study characteristic indicates that all respective subcharacteristics were sufficiently reported based on all available information sources. A value of 0 indicates that no sufficient information was available for all subcharacteristics in all replication studies.

Next to investigating the available information across all replication studies in our sample, we also classified the studies based on the structural indicators in different subgroups (i.e., cognitive vs. social psychology, author overlap vs. no author overlap, preregistration vs. no preregistration, conceptual vs. direct replication, and small time gap vs. large time gap between primary study and replication study). To compare the results between the subgroups, we calculated log odds ratios (*OR*s) for each study characteristic. We considered log *OR* greater than 0.36 in absolute terms as substantial differences (Sánchez-Meca et al., 2003). We report no log *OR* when the availability index of a study characteristic was 1 or 0 in at least one subgroup. In this case, the availability ratings had no variance, and log *OR* could not be calculated.

Equivalence of study characteristics. We coded the equivalence between the replication study and primary study for each subcharacteristic and used the preregistered list of criteria to assess four categories: equivalent, varied (intended), varied (unintended), or no sufficient information. An intended variation had to be explicitly reported in the article; otherwise, we rated the subcharacteristic as unintentionally varied. Table 3 provides a summary of the criteria for coding the equivalence and highlights slight modifications from the preregistration. Modifications covered a slightly altered wording of some criteria to clarify the meaning and changes of the criteria for two characteristics (i.e., timing of data collection and population characteristics) to be less strict.

The analysis of the equivalence ratings was similar to the availability ratings; for the formal details, see Appendix A. We first determined how many study characteristics were equivalent or varied or lacked sufficient information per study, for which we used the equivalence ratings of the subcharacteristics. Only study characteristics for which all subcharacteristics were rated equivalent were considered equivalent. If at least one subcharacteristic was rated varied, we considered the respective study characteristic to be varied. Otherwise, we rated the study characteristic as no sufficient information. Second, we calculated the proportions of the equivalent ratings for all subcharacteristics given that

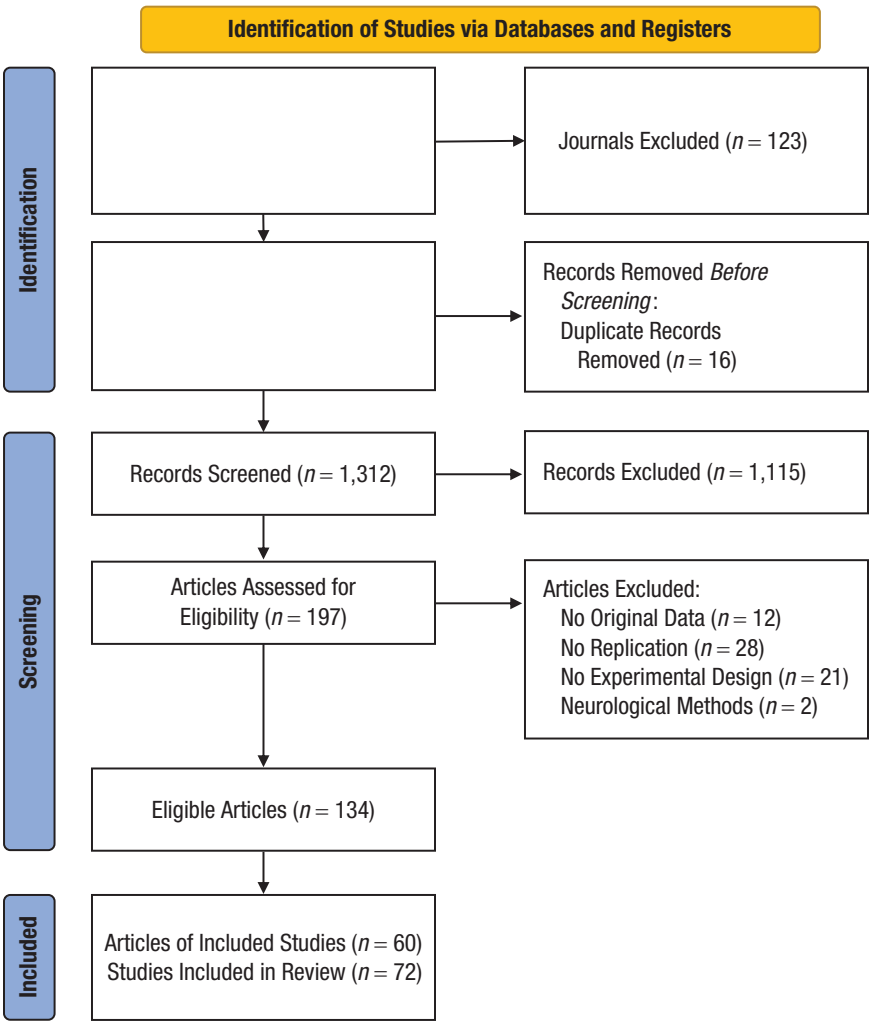

Fig. 1. Flow diagram of literature search and selection process. Template retrieved from Page et al. (2021).

sufficient information was available. Then, we obtained an average score for each study characteristic. Third, we calculated the proportion of intended variations among the varied subcharacteristics and study characteristics.

We also calculated an equivalence index for every study characteristic. The equivalence index describes the probability of a study characteristic to be equivalent to the respective primary studies given that sufficient information is available. It can take on values between 0 and 1: An equivalence index of 1 indicates that a study characteristic was rated as equivalent in all subcharacteristics and all studies that provided sufficient information; a value of 0 indicates that it was rated as varied in all subcharacteristics and all studies that provided sufficient information. Next to investigating the equivalence across all replication studies in our sample, we also compared the results in the subgroups based on the structural indicators. Similar to the analysis of the availability ratings, we calculated and interpreted the log *OR* for each study characteristic.

Results

Literature search

The literature search was conducted on August 11, 2022. For a flow chart of the selection process in accordance with the PRISMA criteria (Page et al., 2021), see Figure 1. A total of 1,328 records were identified in the literature search. The extracted records were screened for duplicates, and identified duplicates were manually excluded, resulting in 1,312 records that were assessed for eligibility. In the first step of the selection process, we excluded 1,115 records that did not fit our criteria, resulting in

Table 4. Descriptive Statistics on Availability of Information and Equivalence Ratings

Study characteristic	Available information	Proportion of equivalence[a]	Proportion of intended variation[b]
Treatment conditions	97.2%	74.8%	22.6%
Content	100.0%	100.0%	—
Material	97.2%	65.7%	29.2%
Delivery	94.4%	57.4%	17.2%
Outcome	97.9%	92.2%	0.0%
Construct	100.0%	100.0%	—
Instrument	95.8%	84.1%	0.0%
Recruitment	27.8%	42.5%	0.0%
Incentives	44.4%	40.6%	0.0%
Advertisement	11.1%	50.0%	0.0%
Causal quantity	98.6%	87.3%	0.0%
Timing	100.0%	0.0%	1.4%
Location	59.7%	48.8%	40.9%
Population	52.8%	35.5%	4.1%
Criteria	73.6%	47.2%	3.6%
Characteristics	31.9%	8.7%	4.8%
Setting	54.2%	70.9%	2.9%
Social	16.7%	83.3%	0.0%
Physical	68.1%	57.1%	4.8%
Device for data collection	77.8%	80.4%	0.0%

Note: We summarized the ratings across all studies in our sample.
[a]Proportion of equivalent ratings given that sufficient information was available.
[b]Proportion of intended variation given that variation occurred.

197 records left for the full-text screening of corresponding articles. Of the 197 articles retrieved for a full-text screening, 40 articles were screened independently by two raters. The overall agreement to include an article was 77.5%, based on disagreements on the criteria original data (97.5% agreement), replication study (82.5% agreement),[1] and study design (90% agreement). All other criteria had perfect agreement. One hundred thirty-four articles passed the entire screening process. From these, we randomly selected 60 articles (i.e., 30 articles for cognitive psychology and 30 articles for social psychology).

Descriptive statistics

The sample of 60 articles included 77 replication studies in total. Five replication studies were excluded because it was impossible to identify a unique primary study because the primary article included multiple studies. Out of the 72 replication studies, 32 (44.4%) were published in cognitive-psychology journals, and 40 (55.6%) were published in social-psychology journals. In total, 10 (13.9%) replication studies had an overlap in authorship, and 35 (48.6%) were preregistered. Forty (55.6%) replication studies self-classified as a direct replication, 10 (13.9%) self-classified as a conceptual replication, and 22 (30.6%) did not explicitly state the replication type. The replication studies were conducted between 1 year and 46 years apart from the respective primary study; median time difference was 7.5 years.

Availability of information

For a summary of the proportions of studies providing sufficient information, see Table 4. For detailed information on the proportions of availability ratings for every study and subcharacteristic, see Appendix B (Table B1). Study characteristics with the most information available were timing (100%),[2] causal quantity (98.6%), outcome (97.9%), and treatment conditions (97.2%). The other study characteristics were less often sufficiently reported: Location was sufficiently reported in 59.7% of studies, population was sufficiently reported in 52.8% of studies, setting was sufficiently reported in 54.2% of studies, and recruitment was sufficiently reported in 27.8% of studies. These results are in line with the traditional practice of conducting replication studies, putting a focus on the repetition of material and analysis methods and the reporting thereof. The available information can also vary between subcharacteristics of the same study characteristics. For example, in the case of population, the target population was reported in 73.6% of studies, and

Table 5. Odds Ratios of Availability Ratings by Study Characteristic and Subgroup Comparison

Study characteristic	Cognitive psychology (n = 32) vs. social psychology (n = 40)	Overlap (n = 10) vs. no overlap (n = 62)	Preregistration (n = 35) vs. no preregistration (n = 37)	Conceptual replication (n = 10) vs. direct replication (n = 40)	Small gap (36) vs. large gap (36)
Treatment conditions	0.79 (–0.23)		0.94 (–0.06)	0.74 (–0.30)	0.19[a] (–1.65)
Outcome				0.24[a] (–1.43)	2.03[a] (0.71)
Recruitment	0.58[a] (–0.54)	1.92[a] (0.65)	1.89[a] (0.64)	0.18[a] (–1.74)	0.57[a] (–0.56)
Causal quantity					
Timing					
Location	0.38[a] (–0.97)	1.01 (0.01)	1.29 (0.25)	0.33[a] (–1.10)	0.56[a] (–0.58)
Population	0.58[a] (–0.54)	1.80[a] (0.59)	0.90 (–0.11)	1.23 (0.20)	1.12 (0.11)
Setting	0.54[a] (–0.61)	0.96 (–0.04)	1.99[a] (0.69)	0.60[a] (–0.50)	1.04 (0.04)

Note: Logarithmic *OR*s are reported in parentheses. Cells highlighted in light gray indicate comparisons with a limited database in one subgroup. Cells highlighted in dark gray indicate that availability ratings had no variance in at least one of the subgroups. Overlap = studies with overlap in authorship; no overlap = studies without overlap in authorship; small gap = small time difference between primary study and replication study (< 7.5 years); large gap = large time difference between primary and replication study (> 7.5 years).
[a]Substantial difference with $|\log(OR)| > 0.36$.

specific sample characteristics (i.e., age and gender) were reported in 31.9% of studies. Regarding the setting, sufficient information on the physical setting (68.1%) and the device for data collection (77.8%) was usually reported, whereas sufficient information on the social setting was reported in 16.7% of studies.

We investigated differences in the availability index between different subgroups (i.e., cognitive psychology vs. social psychology, author overlap vs. no author overlap, preregistration vs. no preregistration, conceptual replication vs. direct replication, and a small time gap vs. large time gap between primary study and replication study) and found several substantial differences in terms of *OR* ($|\log(OR)| > 0.36$; Sánchez-Meca et al., 2003). For detailed results, see Table 5. Note that replication studies in cognitive psychology had lower availability scores in recruitment ($OR = 0.58$), location ($OR = 0.38$), population ($OR = 0.58$), and setting ($OR = 0.54$) compared with replication studies in social psychology. Preregistered studies had higher availability scores in recruitment ($OR = 1.89$) and setting ($OR = 1.99$) compared with studies that had not been preregistered. Furthermore, replication studies that had been conducted closer in time to the primary study had lower availability scores in treatment conditions ($OR = 0.19$), recruitment ($OR = 0.57$), and location ($OR = 0.56$) and higher availability scores in outcome ($OR = 2.03$) compared with replication studies that were conducted further apart from the primary study. The comparison of the overlap in authorship and the replication types was limited in our database because the subgroups of studies with an overlap in authorship and conceptual replications include just 10 studies each. Thus, we do not further refer to these comparisons.

Equivalence of study characteristics

When calculating the number of equivalent study characteristics per study, we found that 0 to 5 characteristics were equivalent ($M = 2.46, SD = 1.01$), 1 to 6 characteristics varied ($M = 3.33, SD = 1.27$), and 0 to 5 characteristics lacked sufficient information ($M = 2.21, SD = 1.17$). Furthermore, in every study in our sample, multiple study characteristics either varied or lacked sufficient information.

For summarized information on the equivalence and intended variation of study characteristics, see Table 4. For the detailed proportions of equivalence ratings for every study and subcharacteristic, see Appendix B (Table B2). The study characteristics with the highest proportions of equivalence, given sufficient information, were outcome (92.2%), causal quantity (87.3%), and treatment conditions (74.8%), closely followed by setting (70.9%). The lowest proportions of equivalence ratings were observed in location (48.8%), recruitment (42.5%), population (35.5%), and timing (0%). We also assessed the methods that were used to keep study characteristics equivalent, such as translation, back-translation methods for equivalent material; the implementation of a comparable incentive scheme; or the recreation of the original setting conditions. However, only a minority of studies reported on such methods (for more information on these results, see the OSF material). Some of the variations in replication studies were intended because the authors investigated the generalizability of the effect under investigation. Among varied study characteristics, we recorded intended variations for location (40.9%), treatment conditions (22.6%), population (4.1%), setting

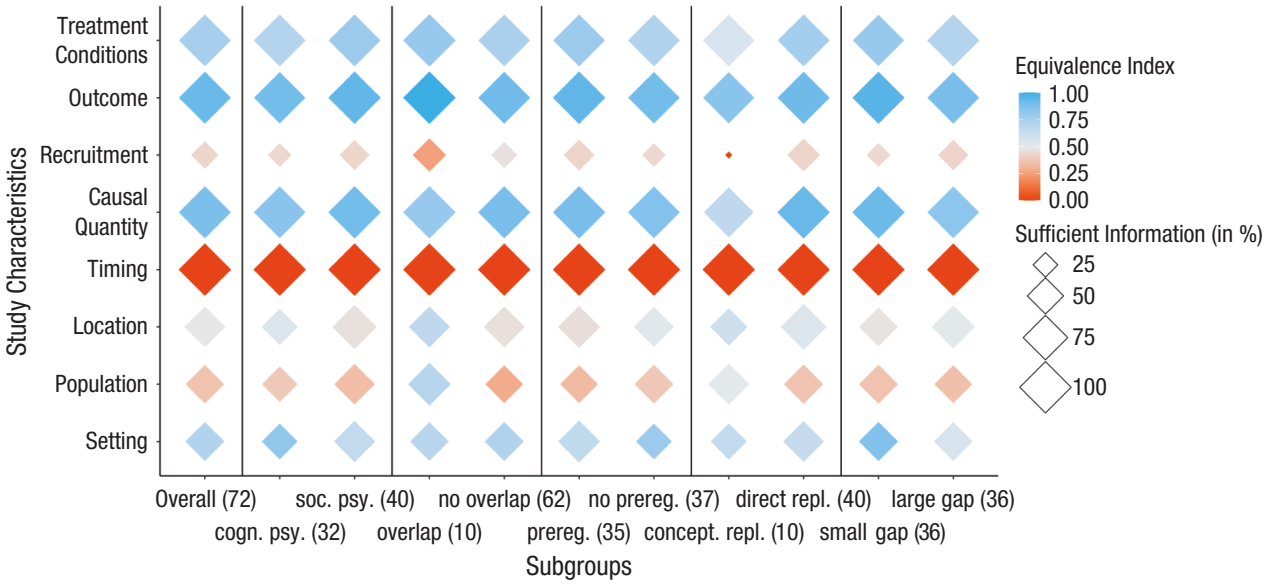

Fig. 2. Equivalence index dependent on study characteristics and subgroups. Number of studies in each group is in parentheses. Cogn. psy. = cognitive psychology; soc. psy. = social psychology; overlap = studies with overlap in authorship; no overlap = studies without overlap in authorship; prereg. = preregistered replication studies; no prereg = not preregistered replication studies; concept. repl. = conceptual replications; direct repl. = direct replications; small gap = small time difference between primary and replication study (< 7.5 years); large gap = large time difference between primary and replication study (> 7.5 years).

(2.9%), and timing (1.4%). However, we recognized most of the observed variations to be unintended. Figure 2 summarizes the equivalence indices for all replication studies and all subgroup comparisons in relation to the available information.

We observed several substantial group differences in terms of *OR*s for predicting the equivalence of study characteristics (for detailed results, see Table 6). In the comparison of the research fields, setting ($OR = 2.49$) was more often rated equivalent in replication studies published in cognitive-psychology journals, and treatment conditions ($OR = 0.63$), outcome ($OR = 0.61$), and causal quantity ($OR = 0.58$) were less often rated equivalent in replication studies published in cognitive-psychology journals compared with studies published in social-psychology journals. Preregistered studies were more often equivalent in the treatment conditions ($OR = 1.46$) and outcome ($OR = 1.80$); however, they were less often equivalent in setting ($OR = 0.51$) than not-preregistered studies. Treatment ($OR = 1.67$), outcome ($OR = 2.92$), causal quantity ($OR = 2.28$), and setting ($OR = 5.18$) were more often equivalent in studies that had a smaller time gap to the primary study compared with studies with a larger time gap.

Discussion

The ongoing discourse surrounding the replicability of psychological effects can be enriched by investigating differences in study characteristics between primary studies and replication studies (including the treatment conditions, outcome measures, recruitment, causal estimates, time, population, location, and setting) that may affect replicability. With our literature review, we pursued this goal by assessing the current practices for designing replication studies and describing the variation of study characteristics and how researchers report on them.

We identified eight study characteristics and various subcharacteristics that could potentially cause heterogeneity of an effect of interest by applying the CRF (Steiner et al., 2019; Wong & Steiner, 2018). Thereby, we translated the theoretical assumptions under which replication success can be expected into concrete study characteristics that can be assessed. Our review was preregistered along with our search, screening, and coding criteria in a review protocol on OSF. We applied the review protocol to replication research in two different psychological fields. For this, we identified 12 impactful journals from the field of cognitive psychology and 16 impactful journals from social psychology. Among the 1,312 research articles published from January 2017 to August 2022 in these journals, a total of 134 research articles (10.21%) contained a replication study according to our criteria (i.e., requiring an explicitly stated replication purpose for a previously published and clearly referenced study). We assessed the availability of information on the different study characteristics and their equivalence to the primary study in a random sample of 60 articles (i.e., 30 for each respective psychological field).

Table 6. Odds Ratios of Equivalence Ratings by Study Characteristic and Subgroup Comparison

Study characteristic	Cognitive psychology (n = 32) vs. social psychology (n = 40)	Overlap (n = 10) vs. no overlap (n = 62)	Preregistration (n = 35) vs. no preregistration (n = 37)	Conceptual replication (n = 10) vs. direct replication (n = 40)	Small gap (n = 36) vs. large gap (n = 36)
Treatment conditions	0.63[a] (–0.46)	1.41 (0.35)	1.46[a] (0.38)	0.39[a] (–0.95)	1.67[a] (0.51)
Outcome	0.61[a] (–0.49)		1.80[a] (0.59)	0.52[a] (–0.66)	2.92[a] (1.07)
Recruitment	1.02 (0.02)	0.38[a] (–0.97)	0.92 (–0.09)		1.09 (0.09)
Causal quantity	0.58[a] (–0.55)	0.52[a] (–0.66)	1.25 (0.22)	0.16[a] (–1.82)	2.28[a] (0.82)
Timing					
Location	1.32 (0.28)	2.35[a] (0.86)	0.76 (–0.28)	1.31 (0.27)	0.90 (–0.11)
Population	1.18 (0.17)	5.63[a] (1.73)	0.83 (–0.18)	1.75[a] (0.56)	1.03 (0.03)
Setting	2.49[a] (0.91)	0.89 (–0.12)	0.51[a] (–0.67)	1.04 (0.04)	5.18[a] (1.64)

Note. Logarithmic *OR*s are reported in parentheses. Cells highlighted in light gray indicate comparisons with a limited database in one subgroup. Cells highlighted in dark gray indicate that equivalence ratings had no variance in at least one of the subgroups. Overlap = studies with overlap in authorship; no overlap = studies without overlap in authorship; small gap = small time difference between primary study and replication study (< 7.5 years); large gap = large time difference between primary study and replication study (> 7.5 years).
[a]Substantial difference with $|\log(OR)| > 0.36$.

Insights about current replication designs

Our results show that replication studies typically follow a traditional understanding of exactly repeating the procedures of a primary study (Schmidt, 2009). Given that sufficient information was available, outcome, analysis methods, and treatment conditions were the characteristics most frequently kept equivalent across studies, closely followed by the setting. All these characteristics were rated equivalent in more than 70% of replication studies compared with fewer than 50% for the other study characteristics (i.e., location, recruitment, population, and timing). However, the subcharacteristics varied considerably in terms of equivalence to the primary study. For example, although the same underlying processes as in the primary studies were investigated in all replication studies, in more than a third of studies, the authors used different study materials and changed procedural details. Furthermore, although the target population was equivalent in almost half of the replication studies, as defined by inclusion criteria, fewer than 10% of studies were equivalent in terms of age and gender of the sample.

One important caveat to these results is that they are based on studies that provided sufficient information on the equivalence of study characteristics.[3] In terms of reporting transparency, we observed large differences between study characteristics. Although the equivalence of timing of data collection, analysis methods, outcome, and treatment conditions were sufficiently reported in virtually all studies, the equivalence of location of data collection, setting, and population were sufficiently reported in fewer than half of the studies, and the equivalence of recruitment strategy was sufficiently reported in just about a quarter of the studies. As with the equivalence of study characteristics, the availability of information varied among subcharacteristics. For example, we found information on the equivalence of the target population in about 75% of studies, whereas for age and gender, we found information in only about one-third of studies. Another example is the setting because the equivalence of physical setting and the device used for data collection were reported in the majority of studies, whereas for the social setting, this was the case in just one out of six studies.

The results further reveal differences between the replication studies in social psychology and cognitive psychology that match disciplinary differences. Thus, more information was available on all assessed study characteristics in social psychology, the field in which guidelines for replication research have been prominently proposed (e.g., Brandt et al., 2014). Accordingly, the treatment conditions, outcome, and analysis method were more often equivalent in social-psychological studies, too. In contrast, the equivalence of setting characteristics (i.e., the device for data collection and physical and social settings) was more the focus of replication studies in cognitive psychology. This may be plausible because cognitive psychology draws more on basic cognitive processes and might therefore often require a distraction-free environment, which might contribute to a stronger motivation to control the setting compared with social psychology. Examinations of other structural indicators revealed that author overlap and preregistration foster the reporting of study characteristics; however, this does not translate to the equivalence of study characteristics.

Note that all replication studies varied or did not sufficiently report multiple study characteristics simultaneously. When study characteristics are not reported, this can imply that these are theoretically not relevant for a specific effect of interest. For instance, there might be a consensus that fundamental cognitive processes, as typically studied in cognitive psychology, are independent of certain study characteristics, such as population characteristics, thus, making it unnecessary to consider. However, it can also imply that there are blind spots in the design of replication studies that need to be further considered to understand aspects causing effect heterogeneity. Especially when different causes of effect heterogeneity vary simultaneously, it becomes impossible to draw strong inference about which study characteristic affects replication success or failure when comparing two studies (e.g., Coyne et al., 2016; Hedges & Schauer, 2019; Steiner et al., 2019; Wong & Steiner, 2018). A systematic variation of study characteristics that describes specific intended differences between studies and rules out alternative explanations for effect heterogeneity would provide strong evidence for understanding the reasons for replication success or failure.

Implications for planning replication studies

A more detailed assessment and reporting of study characteristics that potentially cause effect heterogeneity across studies would be very helpful for describing the conditions under which certain effects replicate. In this review, we used global study characteristics and provided a general overview of the different assumptions of the CRF. These assumptions can enrich the scientific discourse by offering a structured approach to potential sources of effect heterogeneity and provide a foundation to develop reporting standards that can guide future research (see also the PICO framework in medicine; Icahn School of Medicine at Mount Sinai, n.d.). However, such standards require concrete considerations of all relevant study characteristics for a specific effect of interest. When all potential sources of variability are clear, this would offer more comprehensive insights into the replicability of study results and facilitate systematic comparisons across studies. In addition, such information can be used in meta-analysis for investigating predictors for the heterogeneity of a specific effect across multiple studies (see e.g., Hedges & Olkin, 1985; Hedges & Schauer, 2019). Furthermore, the documentation of variations between studies directly adds to the open-science principles (National Academies of Sciences, Engineering, and Medicine, 2018).

In addition, if all replication studies follow a systematic design, specific variations between studies can be ruled out as possible explanations for effect heterogeneity, thus offering insights into the impact of specific variations on effect heterogeneity. Along these lines, Holzmeister et al. (2024) used various large-scale replication attempts to investigate population, design, and analytic heterogeneity as causes of effect heterogeneity (see also Olsson-Collentine et al., 2020). They considered preregistered many-lab studies (e.g., Klein et al., 2014, 2018) for population heterogeneity because constant procedures in all studies controlled for this source of effect heterogeneity. Likewise, many-condition studies or metastudy approaches (e.g., Baribault et al., 2018; DeKay et al., 2022; Huber et al., 2023) randomly assign participants to different variations of an experiment such that design heterogeneity can be investigated. These replication designs substantially improve the interpretability of the impact of variations between studies on effect heterogeneity by intentionally and systematically introducing variations between studies. Multiple variations between studies are possible and interaction effects between specific variations can be investigated because the replication designs ensure a systematic variation of study characteristics in the series of replication studies.

On a similar note, other methodological developments, such as the CRF (e.g., Steiner et al., 2019), provide the formal background for differentiating between different causes of effect heterogeneity and designs for a systematic variation of study differences (Wong et al., 2022). For concrete applications, subject-matter theory is required for detailed insights into the study characteristics that may cause heterogeneity of the effect of interest, such as person characteristics beyond demographics and more details on the specific setting and location. When planning post hoc replication studies to investigate the replicability of existing results, it is challenging to control all study characteristics. For example, in our database, the time of data collection varied in every replication study, and differences in setting or person characteristics are typically difficult to avoid as well. These variations in study characteristics can then be investigated only as a compound because it is impossible to disentangle the impact of certain study characteristics on effect heterogeneity in case of multiple variations. As an alternative, the CRF suggests prospective replication designs in which primary studies and replication studies are planned together to control unintended variations between studies (e.g., Wong et al., 2022; Wong & Steiner, 2018).

Limitations

Our results provide detailed insights into the reporting and design of replication research and relate the current practices to systematic replication attempts that can enhance the understanding of reasons for effect heterogeneity. However, for concrete applications of our results, several limitations need to be considered.

First, the considered study characteristics and subcharacteristics are relatively broad. We specified and operationalized them in a way that we could apply them to a variety of replication studies investigating different effects and phenomena in social psychology and cognitive psychology. This might neglect relevant study characteristics for a specific effect of interest. For example, we operationalized the location of data collection as the respective country in which the data were assessed. For some effects, such as an intergroup effect that involves specific minorities, this operationalization may be too broad because effects could also depend on local differences in a country (e.g., an urban or rural environment). Thus, when planning a replication study, one may need to adjust our criteria to relevant effect moderators for the specific effect of interest. Our global criteria can be considered as an initial set with easily accessible characteristics with which we could already discover substantial differences in the reporting and design of replication studies.

Second, we focused on specific replication studies based on our preregistered inclusion criteria. Thus, we described replication research that explicitly states the respective primary study and that investigates a causal effect in an (quasi)experiment. Following these criteria, we found a prevalence of replication studies in the considered journals of 10.2%. Because experimental studies are rather common in cognitive psychology and social psychology, only 10.7% of replication studies were excluded based on a nonexperimental study design. The inclusion criteria allowed us to consider all available information for a study comparison (i.e., the replication study and primary study plus all supplemental materials) and facilitated the investigation of the equivalence of study characteristics according to the CRF. Thus, our results represent rather controlled research conditions and do not represent replication practices for other study types, such as observational research.

Third, we limited ourselves to the coding of 60 research articles (i.e., 30 from social-psychology journals and 30 from cognitive-psychology journals). This limit was specified in our preregistered review protocol and was based on our resources for the extensive manual-coding process. In total, we went through 72 replication studies and 72 primary studies and coded information on the reporting and the equivalence of eight study characteristics, each with one to three subcharacteristics. We are confident that the overall pattern of reporting and design practices can be illustrated with these replication studies given that they were randomly selected from 134 identified articles that included replication studies in the considered journals. For subsequent investigations of the replication studies, we provide the documented database on the OSF.

Conclusion

In this literature review, we used a comprehensive list of study characteristics to evaluate to what extent these were varied, kept constant, or not reported on in replication studies. Our results suggest that in current replication studies, researchers mainly focus on specific study characteristics, such as the treatment, outcome, and analysis methods. These characteristics are generally well documented and kept constant. Other study characteristics, such as population, setting, location, and timing of data collection, are less well documented and more often varied in replication studies. For more insights into the reasons of effect heterogeneity, we suggest careful reporting and systematic variation or control of relevant study differences. A transparent documentation of the different sources of effect heterogeneity and systematic designs that control for some variations in replication studies can substantially enhance the evidence on the replicability of effects in different fields of psychology.

Appendix A

Formal description of the analysis of availability of information ratings

To calculate an availability index for each study characteristic, we first recoded the availability ratings V as follows:

$$V = \begin{cases} 0, \textit{no sufficient information} \\ 1, \textit{otherwise} \end{cases}.$$

That means that every subcharacteristic rated replication article alone or replication article and primary article obtained the value of 1 and subcharacteristics for which no sufficient information was available obtained the value of 0.

For each study characteristic i and each subgroup j, we then calculated an availability index AI from the availability variable V as follows:

$$AI_{ij} = \frac{1}{k_i \times p_j} \sum_{w=1}^{k_i} \sum_{z=1}^{p_j} V_{wz}, \tag{1}$$

where k_i denotes the number of subcharacteristics that refer to a study characteristic i and p_j denotes the number of studies in subgroup j.

To compare availability indices between a subgroup j and another subgroup h, we calculated odds ratios (OR) for each study characteristic for different subgroup comparisons, whereby the ORs were calculated as follows:

$$OR_i = \frac{\left(\frac{AI_{ij}}{1 - AI_{ij}}\right)}{\left(\frac{AI_{ih}}{1 - AI_{ih}}\right)}, \textit{with } j \neq h, \tag{2}$$

where OR_i refers to the *OR* of study characteristic i and AI_{ij} and AI_{ib} refer to the availability index of study characteristic i in subgroups j or h, respectively.

Formal description of the analysis of equivalence ratings

We determined how many study characteristics were equivalent, varied, or lacked sufficient information per study, for which we used the equivalence ratings of the subcharacteristics. Only study characteristics for which all subcharacteristics were rated equivalent were considered equivalent. If at least one subcharacteristic was rated varied, we considered the respective study characteristic to be varied. Otherwise, we rated the study characteristic as no sufficient information:

$$EC_{in} = \begin{cases} equivalent, & if\ \forall G_{kin} = equivalent \\ varied, & if\ \exists G_{kin} = varied \\ no\ sufficient\ information, & otherwise \end{cases}$$

where G_{kin} denotes the equivalence indicator of subcharacteristic k of study characteristic i in study n and EC_{in} denotes the resulting equivalence indicator of study characteristic i in study n. We then calculated the number of equivalent, varied, and insufficiently reported study characteristics per study and the respective means and standard deviations.

To calculate an equivalence index for each study characteristic, we recoded the equivalence ratings of subcharacteristics. This was done as follows:

$$U = \begin{cases} 1, equivalent \\ 0, varied \\ NA, no\ sufficient\ information \end{cases}.$$

That means that every subcharacteristic rated equivalent obtained the value 1, every subcharacteristic rated varied irrespective of the intention obtained the value 0, and every subcharacteristic rated no sufficient information obtained a missing value. We computed the equivalence index *EI* for each study characteristic i and each group j similarly to the availability index:

$$EI_{ij} = \frac{1}{k_i \times p_j - m_{ij}} \sum_{w=1}^{k_i} \sum_{z=1}^{p_j} U_{wz}, \quad (4)$$

where k_i denotes the number of subcharacteristics that refers to a study characteristic i, p_j refers to the number of studies in group j, and m_{ij} refers to the number of missing values because of no-sufficient-information ratings.

Appendix B

Table B1 and Table B2 include descriptive information on the proportions of availability ratings and equivalence ratings, respectively. The information was coded in every replication study for each subcharacteristic. The values pertaining to the study characteristics resemble the mean value of the respective subcharacteristics.

Table B1. Proportions of Availability Ratings by Study Characteristics and Subcharacteristics

Study characteristic	Replication study	Replication study and primary study	No sufficient information
Treatment conditions	85.6%	11.6%	2.8%
Content	97.2%	2.8%	0.0%
Material	83.3%	13.9%	2.8%
Delivery	76.4%	18.1%	5.6%
Outcome	72.9%	25.0%	2.1%
Construct	76.4%	23.6%	0.0%
Instrument	69.4%	26.4%	4.2%
Recruitment	10.4%	17.4%	72.2%
Incentives	15.3%	29.2%	55.6%
Advertisement	5.6%	5.6%	88.9%
Causal quantity	40.3%	58.3%	1.4%
Timing	25.0%	75.0%	0.0%
Location	43.1%	16.7%	40.3%
Population	21.5%	31.3%	47.2%
Criteria	37.5%	36.1%	26.4%
Characteristics	5.6%	26.4%	68.1%
Setting	31.5%	22.7%	45.8%
Social setting	11.1%	5.6%	83.3%
Physical setting	47.2%	20.8%	31.9%
Device for data collection	36.1%	41.7%	22.2%

Table B2. Proportions of Equivalence Ratings by Study and Subcharacteristics

Study characteristic	Equivalent	Varied		No sufficient information
		Intended	Unintended	
Treatment conditions	72.7%	5.6%	19.0%	2.8%
Content	100.0%	0.0%	0.0%	0.0%
Material	63.9%	9.7%	23.6%	2.8%
Delivery	54.2%	6.9%	33.3%	5.6%
Outcome	90.3%	0.0%	7.6%	2.1%
Construct	100.0%	0.0%	0.0%	0.0%
Instrument	80.6%	0.0%	15.3%	4.2%
Recruitment	11.8%	0.0%	16.0%	72.2%
Incentives	18.1%	0.0%	26.4%	55.6%
Advertisement	5.6%	0.0%	5.6%	88.9%
Causal quantity	86.1%	0.0%	12.5%	1.4%
Timing	0.0%	1.4%	98.6%	0.0%
Location	29.2%	12.5%	18.1%	40.3%
Population	18.8%	1.4%	32.6%	47.2%
Criteria	34.7%	1.4%	37.5%	26.4%
Characteristics	2.8%	1.4%	27.8%	68.1%
Setting	38.4%	0.5%	15.3%	45.8%
Social setting	13.9%	0.0%	2.8%	83.3%
Physical setting	38.9%	1.4%	27.8%	31.9%
Device for data collection	62.5%	0.0%	15.3%	22.2%

Transparency

Action Editor: Pamela Davis-Kean
Editor: David A. Sbarra

Author Contributions

Jerome Hoffmann: Conceptualization; Data curation; Formal analysis; Investigation; Project administration; Visualization; Writing – original draft; Writing – review & editing.

Mathias Twardawski: Conceptualization; Funding acquisition; Investigation; Writing – review & editing.

Johanna M. Höhs: Conceptualization; Investigation; Writing – review & editing.

Anne Gast: Conceptualization; Funding acquisition; Writing – review & editing.

Steffi Pohl: Conceptualization; Funding acquisition; Writing – review & editing.

Marie-Ann Sengewald: Conceptualization; Data curation; Funding acquisition; Investigation; Project administration; Writing – original draft; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (German Research Foundation); 464369680.

Open Practices

This article has received the badges for Open Data and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Jerome Hoffmann https://orcid.org/0009-0000-1570-819X
Marie-Ann Sengewald https://orcid.org/0000-0003-4155-394X

Notes

1. Disagreements mainly referred to an ambiguity for identifying the primary study. Some of the replications aimed at replicating a general paradigm instead of the results of a specific primary study or referred to different primary studies. As a result, three of the seven articles were excluded, and four articles were considered as replications because the respective primary study could be identified.
2. We obtained information on timing in 100% of studies because we took the publication year as a proxy for the year of data collection if no date was mentioned in the studies.
3. Note that a low reporting transparency reflects our ability to determine equivalence or variation of study characteristics with the given information in the primary study and replication study. Consequently, if, for example, a replication study reported descriptive statistics on age and gender but the primary study is lacking this information, we rated these characteristics as no sufficient information.

References

References marked with an asterisk indicate studies included in the literature review for data extraction.

*Agauas, S. J., Jacoby, M., & Thomas, L. E. (2020). Near-hand effects are robust: Three OSF pre-registered replications of visual biases in perihand space. *Visual Cognition*, *28*(3), 192–204. https://doi.org/10.1080/13506285.2020.1751763

*Anvari, F., Olsen, J., Hung, W. Y., & Feldman, G. (2021). Misprediction of affective outcomes due to different evaluation modes: Replication and extension of two distinction bias experiments by Hsee and Zhang (2004). *Journal of Experimental Social Psychology*, *92*, Article 104052. https://doi.org/10.1016/j.jesp.2020.104052

*Ball, F., Groth, R.-M., Agostino, C. S., Porcu, E., & Noesselt, T. (2020). Explicitly versus implicitly driven temporal expectations: No evidence for altered perceptual processing due to top-down modulations. *Attention, Perception, & Psychophysics*, *82*(4), 1793–1807. https://doi.org/10.3758/s13414-019-01879-1

*Balzarini, R. N., Dobson, K., Chin, K., & Campbell, L. (2017). Does exposure to erotica reduce attraction and love for romantic partners in men? Independent replications of Kenrick, Gutierres, and Goldberg (1989) study 2. *Journal of Experimental Social Psychology*, *70*, 191–197. https://doi.org/10.1016/j.jesp.2016.11.003

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*, *115*(11), 2607–2612. https://doi.org/10.1073/pnas.1708285114

Bauer, P. J. (2023). Generalizations: The grail and the gremlins. *Journal of Applied Research in Memory and Cognition*, *12*(2), 159–175. https://doi.org/10.1037/mac0000106

Borsboom, D., Van Der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., Van 'T, & Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

*Burnham, B. R. (2020). Are liberals really dirty? Two failures to replicate Helzer and Pizarro's (2011) study 1, with meta-analysis. *Journal of Personality and Social Psychology*, *119*(6), e38–e42. https://doi.org/10.1037/pspa0000238

*Calderon, S., Mac Giolla, E., Ask, K., & Granhag, P. A. (2020). Subjective likelihood and the construal level of future events: A replication study of Wakslak, Trope, Liberman, and Alony (2006). *Journal of Personality and Social Psychology*, *119*(5), e27–e37. https://doi.org/10.1037/pspa0000214

*Chabris, C. F., Heck, P. R., Mandart, J., Benjamin, D. J., & Simons, D. J. (2019). No evidence that experiencing physical warmth promotes interpersonal warmth. *Social Psychology*, *50*(2), 127–132. https://doi.org/10.1027/1864-9335/a000361

*Chen, J., Kwan, L. C., Ma, L. Y., Choi, H. Y., Lo, Y. C., Au, S. Y., Tsang, C. H., Cheng, B. L., & Feldman, G. (2021). Retrospective and prospective hindsight bias: Replications and extensions of Fischhoff (1975) and Slovic and Fischhoff (1977). *Journal of Experimental Social Psychology*, *96*, Article 104154. https://doi.org/10.1016/j.jesp.2021.104154

Cook, B. G., Collins, L. W., Cook, S. C., & Cook, L. (2016). A replication by any other name: A systematic review of replicative intervention studies. *Remedial and Special Education*, *37*(4), 223–234. https://doi.org/10.1177/0741932516637198

Coyne, M. D., Cook, B. G., & Therrien, W. J. (2016). Recommendations for replication research in special education: A framework of systematic, conceptual replications. *Remedial and Special Education*, *37*(4), 244–253. https://doi.org/10.1177/0741932516648463

*Crawford, J. T., Fournier, A., & Ruscio, J. (2019). Does subjective SES moderate the effect of money priming on socioeconomic system support? A replication of Schuler and Wänke (2016). *Social Psychological and Personality Science*, *10*(1), 103–109. https://doi.org/10.1177/1948550617740941

Cronbach, L. J., & Shapiro, K. (1982). *Designing evaluations of educational and social programs*. Jossey-Bass.

*Dang, J., Liu, X., Xiao, S., Mao, L., Chan, K. T., Li, C., Lin, M., Liu, Z., Luo, Y., Sun, Y., Wu, Y.-H., & Schiöth, H. B. (2021). The beauty of the zero: Replications and extensions of the hidden-zero effect in delay discounting tasks. *Social Psychological and Personality Science*, *12*(4), 544–549. https://doi.org/10.1177/1948550620929454

DeKay, M. L., Rubinchik, N., Li, Z., & De Boeck, P. (2022). Accelerating psychological science with metastudies: A demonstration using the risky-choice framing effect. *Perspectives on Psychological Science*, *17*(6), 1704–1736. https://doi.org/10.1177/17456916221079611

*Ekelund, M., & Ask, K. (2021). Stigmatization of voluntarily childfree women and men in the UK. *Social Psychology*, *52*(5), 275–286. https://doi.org/10.1027/1864-9335/a000455

*Essien, I., Stelter, M., Kalbe, F., Koehler, A., Mangels, J., & Meliß, S. (2017). The shooter bias: Replicating the classic effect and introducing a novel paradigm. *Journal of Experimental Social Psychology*, *70*, 41–47. https://doi.org/10.1016/j.jesp.2016.12.009

*Ferguson, C. J., Gryshyna, A., Kim, J. S., Knowles, E., Nadeem, Z., Cardozo, I., Esser, C., Trebbi, V., & Willis, E. (2022). Video games, frustration, violence, and virtual reality: Two studies. *British Journal of Social Psychology*, *61*(1), 83–99. https://doi.org/10.1111/bjso.12471

Fiedler, K., & Prager, J. (2018). The regression trap and other pitfalls of replication science—Illustrated by the report of the Open Science Collaboration. *Basic and Applied Social Psychology*, *40*(3), 115–124. https://doi.org/10.1080/01973533.2017.1421953

*Forscher, P. S., Mitamura, C., Dix, E. L., Cox, W. T. L., & Devine, P. G. (2017). Breaking the prejudice habit: Mechanisms, timecourse, and longevity. *Journal of Experimental Social Psychology*, *72*, 133–146. https://doi.org/10.1016/j.jesp.2017.04.009

*Frătescu, M., Van Moorselaar, D., & Mathôt, S. (2019). Can you have multiple attentional templates? Large-scale replications of Van Moorselaar, Theeuwes, and Olivers (2014) and Hollingworth and Beck (2016). *Attention, Perception, & Psychophysics*, *81*(8), 2700–2709. https://doi.org/10.3758/s13414-019-01791-8

*Genschow, O., Westfal, M., Crusius, J., Bartosch, L., Feikes, K. I., Pallasch, N., & Wozniak, M. (2021). Does social psychology persist over half a century? A direct replication of Cialdini et al.'s (1975) classic door-in-the-face technique. *Journal of Personality and Social Psychology*, *120*(2), e1–e7. https://doi.org/10.1037/pspa0000261

*Gunther, K. L., & McKinney, M. R. (2020). Poor peripheral binding depends in part on stimulus color. *Attention, Perception, & Psychophysics*, *82*(7), 3606–3617. https://doi.org/10.3758/s13414-020-02086-z

*Gyurkovics, M., Kovacs, M., Jaquiery, M., Palfi, B., Dechterenko, F., & Aczel, B. (2020). Registered Replication Report of Weissman, D. H., Jiang, J., & Egner, T. (2014). Determinants of congruency sequence effects without learning and memory confounds. *Attention, Perception, & Psychophysics*, *82*(8), 3777–3787. https://doi.org/10.3758/s13414-020-02021-2

*Haaf, J. M., Rhodes, S., Naveh-Benjamin, M., Sun, T., Snyder, H. K., & Rouder, J. N. (2021). Revisiting the remember–know task: Replications of Gardiner and Java (1990). *Memory & Cognition*, *49*(1), 46–66. https://doi.org/10.3758/s13421-020-01073-x

Hardwicke, T. E., & Wagenmakers, E.-J. (2023). Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nature Human Behaviour*, *7*(1), 15–26. https://doi.org/10.1038/s41562-022-01497-2

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hedges, L. V., & Schauer, J. M. (2019). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods*, *24*(5), 557–570. https://doi.org/10.1037/met0000189

*Hoeben Mannaert, L. N., Dijkstra, K., & Zwaan, R. A. (2017). Is color an integral part of a rich mental simulation? *Memory & Cognition*, *45*(6), 974–982. https://doi.org/10.3758/s13421-017-0708-1

Holzmeister, F., Johannesson, M., Böhm, R., Dreber, A., Huber, J., & Kirchler, M. (2024). Heterogeneity in effect size estimates. *Proceedings of the National Academy of Sciences*, *121*(32), Article e2403490121. https://doi.org/10.1073/pnas.2403490121

*Hoogeveen, S., Wagenmakers, E.-J., Kay, A. C., & Van Elk, M. (2018). Compensatory control and religious beliefs: A Registered Replication Report across two countries. *Comprehensive Results in Social Psychology*, *3*(3), 240–265. https://doi.org/10.1080/23743603.2019.1684821

Huber, C., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Weitzel, U., Abellán, M., Adayeva, X., Ay, F. C., Barron, K., Berry, Z., Bönte, W., Brütt, K., Bulutay, M., Campos-Mercade, P., Cardella, E., Claassen, M. A., Cornelissen, G., Dawson, I. G. J., . . . Holzmeister, F. (2023). Competition and moral behavior: A meta-analysis of forty-five crowd-sourced experimental designs. *Proceedings of the National Academy of Sciences*, *120*(23), Article e2215572120. https://doi.org/10.1073/pnas.2215572120

Icahn School of Medicine at Mount Sinai. (n.d.). *Evidence-based medicine: The PICO framework*. https://libguides.mssm.edu/ebm/ebp_pico

*Kidd, D., & Castano, E. (2019). Reading literary fiction and theory of mind: Three preregistered replications and extensions of Kidd and Castano (2013). *Social Psychological and Personality Science*, *10*(4), 522–531. https://doi.org/10.1177/1948550618775410

*Kim, J., Schlegel, R. J., Seto, E., & Hicks, J. A. (2019). Thinking about a new decade in life increases personal self-reflection: A replication and reinterpretation of alter and Hershfield's (2014) findings. *Journal of Personality and Social Psychology*, *117*(2), e27–e34. https://doi.org/10.1037/pspp0000199

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., 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: A "many labs" replication project. *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, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., 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

Kohl, C., McIntosh, E. J., Unger, S., Haddaway, N. R., Kecke, S., Schiemann, J., & Wilhelm, R. (2018). Online tools supporting the conduct and reporting of systematic reviews and systematic maps: A case study on CADIMA and review of existing tools. *Environmental Evidence*, *7*, Article 8. https://doi.org/10.1186/s13750-018-0115-5

*Kumar, A. A., Balota, D. A., & Steyvers, M. (2020). Distant connectivity and multiple-step priming in large-scale semantic networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(12), 2261–2276. https://doi.org/10.1037/xlm0000793

*Landwehr, K. (2017). Titchener's ⊥ with its lines tilted—A partial replication and extension of Cormack and Cormack (1974). *Attention, Perception, & Psychophysics*, *79*(1), 223–229. https://doi.org/10.3758/s13414-016-1205-5

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology*, *113*(2), 254–261. https://doi.org/10.1037/pspi0000106

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

*Lehmann, G. K., & Calin-Jageman, R. J. (2017). Is red really romantic?: Two pre-registered replications of the red-romance hypothesis. *Social Psychology*, *48*(3), 174–183. https://doi.org/10.1027/1864-9335/a000296

Lemons, C. J., King, S. A., Davidson, K. A., Berryessa, T. L., Gajjar, S. A., & Sacks, L. H. (2016). An inadvertent concurrent

replication: Same roadmap, different journey. *Remedial and Special Education*, *37*(4), 213–222. https://doi.org/10.1177/0741932516631116

*Levering, K. R., Conaway, N., & Kurtz, K. J. (2020). Revisiting the linear separability constraint: New implications for theories of human category learning. *Memory & Cognition*, *48*(3), 335–347. https://doi.org/10.3758/s13421-019-00972-y

Lewandowsky, S., & Oberauer, K. (2020). Low replicability can support robust and efficient science. *Nature Communications*, *11*, Article 358. https://doi.org/10.1038/s41467-019-14203-0

*Li, X., Xiong, Z., Theeuwes, J., & Wang, B. (2020). Visual memory benefits from prolonged encoding time regardless of stimulus type. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(10), 1998–2005. https://doi.org/10.1037/xlm0000847

*Lin, Y., Sasin, E., & Fougnie, D. (2021). Selection in working memory is resource-demanding: Concurrent task effects on the retro-cue effect. *Attention, Perception, & Psychophysics*, *83*(4), 1600–1612. https://doi.org/10.3758/s13414-020-02239-0

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*(6), 304–316. https://doi.org/10.3102/0013189X14545513

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

*Maquestiaux, F., Arexis, M., Chauvel, G., Ladoy, J., Boyer, P., & Mazerolle, M. (2021). Ebbinghaus visual illusion: No robust influence on novice golf-putting performance. *Psychological Research*, *85*(3), 1156–1166. https://doi.org/10.1007/s00426-020-01298-0

*Masson, M. E. J., Rabe, M. M., & Kliegl, R. (2017). Modulation of additive and interactive effects by trial history revisited. *Memory & Cognition*, *45*(3), 480–492. https://doi.org/10.3758/s13421-016-0666-z

*Mayiwar, L., & Lai, L. (2019). Replication of Study 1 in "Differentiating social and personal power" by Lammers, Stoker, and Stapel (2009). *Social Psychology*, *50*(4), 261–269. https://doi.org/10.1027/1864-9335/a000388

National Academies of Sciences, Engineering, and Medicine. (2018). *Open science by design: Realizing a vision for 21st century research*. National Academies Press. https://doi.org/10.17226/25116

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., & Errington, T. M. (2017). Making sense of replications. *eLife*, *6*, Article e23383. https://doi.org/10.7554/eLife.23383

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., 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*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Olsson-Collentine, A., Wicherts, J. M., & Van Assen, M. A. L. M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, *146*(10), 922–940. https://doi.org/10.1037/bul0000294

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

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*, Article 89. https://doi.org/10.1186/s13643-021-01626-4

*Papenmeier, F., & Timm, J. D. (2021). Do group ensemble statistics bias visual working memory for individual items? A registered replication of Brady and Alvarez (2011). *Attention, Perception, & Psychophysics*, *83*(3), 1329–1336. https://doi.org/10.3758/s13414-020-02209-6

*Pekár, J., & Kinder, A. (2020). The interplay between non-symbolic number and its continuous visual properties revisited: Effects of mixing trials of different types. *Quarterly Journal of Experimental Psychology*, *73*(5), 698–710. https://doi.org/10.1177/1747021819891068

Perry, T., Morris, R., & Lea, R. (2022). A decade of replication study in education? A mapping review (2011–2020). *Educational Research and Evaluation*, *27*(1–2), 12–34. https://doi.org/10.1080/13803611.2021.2022315

*Philipp-Muller, A., & MacDonald, G. (2017). Avoidant individuals may have muted responses to social warmth after all: An attempted replication of MacDonald and Borsook (2010). *Journal of Experimental Social Psychology*, *70*, 272–280. https://doi.org/10.1016/j.jesp.2016.11.010

*Pitteri, M., Marchetti, M., Grassi, M., & Priftis, K. (2021). Pitch height and brightness both contribute to elicit the SMARC effect: A replication study with expert musicians. *Psychological Research*, *85*(6), 2213–2222. https://doi.org/10.1007/s00426-020-01395-0

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

*Ren, D., Wesselmann, E. D., & Van Beest, I. (2021). Seeking solitude after being ostracized: A replication and beyond. *Personality and Social Psychology Bulletin*, *47*(3), 426–440. https://doi.org/10.1177/0146167220928238

*Röseler, L., Schütz, A., Blank, P. A., Dück, M., Fels, S., Kupfer, J., Scheelje, L., & Seida, C. (2021). Evidence against subliminal anchoring: Two close, highly powered, preregistered, and failed replication attempts. *Journal of Experimental Social Psychology*, *92*, Article 104066. https://doi.org/10.1016/j.jesp.2020.104066

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331. https://doi.org/10.1198/016214504000001880

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, *8*(4), 448–467. https://doi.org/10.1037/1082-989X.8.4.448

*Schacherer, J., & Hazeltine, E. (2019). How conceptual overlap and modality pairings affect task-switching and mixing costs. *Psychological Research*, *83*(5), 1020–1032. https://doi.org/10.1007/s00426-017-0932-0

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

*Schneegans, S., Harrison, W. J., & Bays, P. M. (2021). Location-independent feature binding in visual working memory for sequentially presented objects. *Attention, Perception, & Psychophysics*, *83*(6), 2377–2393. https://doi.org/10.3758/s13414-021-02245-w

*Schöpper, L.-M., Singh, T., & Frings, C. (2020). The official soundtrack to "Five shades of grey": Generalization in multimodal distractor-based retrieval. *Attention, Perception, & Psychophysics*, *82*(7), 3479–3489. https://doi.org/10.3758/s13414-020-02057-4

*Schulze, C., James, G., Koehler, D. J., & Newell, B. R. (2019). Probability matching does not decrease under cognitive load: A preregistered failure to replicate. *Memory & Cognition*, *47*(3), 511–518. https://doi.org/10.3758/s13421-018-0888-3

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

*Siem, B., Neymeyer, L., & Rohmann, A. (2021). Entertainment education as a means to reduce anti-Muslim prejudice – For whom does it work best? An extended replication of Murrar and Brauer (2018). *Social Psychology*, *52*(1), 51–60. https://doi.org/10.1027/1864-9335/a000432

*Stamkou, E., Homan, A. C., Van Kleef, G. A., & Gelfand, M. J. (2022). The spatial representation of leadership depends on ecological threat: A replication and extension of Menon et al. (2010). *Journal of Personality and Social Psychology*, *123*(3), e1–e22. https://doi.org/10.1037/pspa0000304

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, *9*(3), 305–318. https://doi.org/10.1177/1745691614528518

*Stanley, M. L., Yang, B. W., & De Brigard, F. (2018). No evidence for unethical amnesia for imagined actions: A failed replication and extension. *Memory & Cognition*, *46*(5), 787–795. https://doi.org/10.3758/s13421-018-0803-y

*Stefani, M., Sauter, M., & Mack, W. (2020). Delayed disengagement from irrelevant fixation items: Is it generally functional? *Attention, Perception, & Psychophysics*, *82*(2), 637–654. https://doi.org/10.3758/s13414-019-01926-x

Steiner, P. M., Wong, V. C., & Anglin, K. (2019). A causal replication framework for designing and assessing replication efforts. *Zeitschrift Für Psychologie/Journal of Psychology*, *227*(4), 280–292. https://doi.org/10.1027/2151-2604/a000385

*Tian, Q., Liu, X., Zhou, J., & Sun, T. (2021). Do animals' minds matter less, when meat gets personal? Replications of Piazza and Loughnan (2016) in China. *Social Psychological and Personality Science*, *12*(3), 417–425. https://doi.org/10.1177/1948550620920982

*Tomko, L., & Proctor, R. W. (2017). Crossmodal spatial congruence effects: Visual dominance in conditions of increased and reduced selection difficulty. *Psychological Research*, *81*(5), 1035–1050. https://doi.org/10.1007/s00426-016-0801-2

*Travis, S. L., Dux, P. E., & Mattingley, J. B. (2017). Re-examining the influence of attention and consciousness on visual afterimage duration. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(12), 1944–1949. https://doi.org/10.1037/xhp0000458

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*, *113*(23), 6454–6459. https://doi.org/10.1073/pnas.1521897113

*Watkins, H. M., & Brandt, M. (2019). The moral landscape of war: A registered report testing how the war context shapes morality's constraints on default representations of possibility. *Journal of Experimental Social Psychology*, *85*, Article 103843. https://doi.org/10.1016/j.jesp.2019.103843

Wickham, H. (2016). *ggplot2*. Springer. https://doi.org/10.1007/978-3-319-24277-4

Wickham, H., & Bryan, J. (2023). *readxl: Read excel files* (R package Version 1.4.2). https://CRAN.R-project.org/package=readxl

Wong, V. C., Anglin, K., & Steiner, P. M. (2022). Design-based approaches to causal replication studies. *Prevention Science*, *23*, 723–738. https://doi.org/10.1007/s11121-021-01234-7

Wong, V. C., & Steiner, P. M. (2018). *Replication designs for causal inference* (EdPolicyWorks Working Paper Series, Issue 62). EdPolicyWorks. https://education.virginia.edu/documents/epw62-replication-designs2018-04pdf

*Xie, X., Liu, L., & Jaeger, T. F. (2021). Cross-talker generalization in the perception of nonnative speech: A large-scale replication. *Journal of Experimental Psychology: General*, *150*(11), e22–e56. https://doi.org/10.1037/xge0001039

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, Article e1. https://doi.org/10.1017/S0140525X20001685

*Zhang, Y., Chen, X., Liu, Z., Zhang, Y., Jiang, T., You, X., & Luo, Y. (2021). Nice guys finish last? The effect of lay theories on prosocial actors' motivation and future benefits. *Social Psychological and Personality Science*, *12*(1), 91–98. https://doi.org/10.1177/1948550620902282

*Ziano, I., Xiao, Q., Yeung, S. K., Wong, C. Y., Cheung, M. Y., Lo, C. Y. J., Yan, H. C., Narendra, G. I., Kwan, L. W., Chow, C. S., Man, C. Y., & Feldman, G. (2021). Numbing or sensitization? Replications and extensions of Fetherstonhaugh et al. (1997)'s "Insensitivity to the value of human life." *Journal of Experimental Social Psychology*, *97*, Article 104222. https://doi.org/10.1016/j.jesp.2021.104222

*Ziano, I., Yao, J. D., Gao, Y., & Feldman, G. (2020). Impact of ownership on liking and value: Replications and extensions of three ownership effect experiments. *Journal of Experimental Social Psychology*, *89*, Article 103972. https://doi.org/10.1016/j.jesp.2020.103972

*Zickfeld, J. H., Van De Ven, N., Schubert, T. W., & Vingerhoets, A. (2018). Are tearful individuals perceived as less competent? Probably not. *Comprehensive Results in Social Psychology*, *3*(2), 119–139. https://doi.org/10.1080/23743603.2018.1514254