TITLE: Publication Bias in Academic Decision Making in Clinical Psychology

Abstract

Publication bias, favoring statistically significant or hypothesis-consistent findings, impedes the process of scientific knowledge production. However, questions remain about the mechanisms underlying publication bias, and work has focused on between-subjects designs rather than within-subjects experiments. This study employed a within-subject design based on Dual Process Theories to investigate selective publication and reception choices. Four online experiments presented 16 abstract vignettes to clinical psychology researchers with statistical significance and hypothesis-consistency as the treatment variables (n = 75 per experiment). Participants initially judged publication, reading, or citation likelihood, providing an intuitive evaluation. After this they rated the accompanying *Feeling of Rightness* (FOR), followed by a reconsidered judgment. Using multilevel models, we examined if intuitive judgments as well as changes between evaluations varied as functions of statistical significance and hypothesis-consistency. Statistically nonsignificant and hypothesis-inconsistent results were not <not / negatively> associated with publication, reading, and citation likelihoods. Changes in publication, reading, and citation likelihoods were <not> directly associated with statistically nonsignificant and hypothesis-inconsistent results. Furthermore, these associations were <not> mediated by FOR. Our results suggest that biased publication and reception choices are <not> prevalent during intuitive judgements. However, bias <does not decrease/ decreases> during reconsidered evaluations of initially biased publication and reception choices.

Research practice in psychology is primarily empirical, often quantitative, and hypothesis-driven, involving the observation of multiple cases and the use of statistical significance to infer population characteristics from these observations: It is crucial for enabling a culture of cumulative science that results are published as fully as possible and are not withheld or discarded without being published (Scheel, 2021). Otherwise, biased findings, replication problems, and a waste of resources may result (Wieschowski et al., 2019; OSC, 2015). The non-publication of study results due to characteristics other than study quality is referred to as publication bias (e.g., Lortie et al., 2007; Dirnagl & Lauritzen, 2010; Smulders, 2013).

Over 60 years ago, Sterling (1959) already documented a publication bias of statistically significant findings in several psychological journals including journals of clinical psychology. Greenwald (1975) documented the decision process leading to a publication bias, claiming that there are four choice points in the process, when a bias could enter: "(a) the researcher's formulation of a hypothesis, (b) his collection of data, (c) his evaluation of obtained results, and (d) an editor's judgment of a manuscript reporting the research results" (p. 6).

Three more recent cohort studies also reported substantial rates of non-publication. Fleming (2019) found that the results of 140 out of 450 clinical trials were not made available in a government registry. Similarly, of 1,509 clinical trials in Germany, 26% had not been published six years after completion (Wieschowski et al., 2018). Franco and colleagues (2014) found a strong negative association between null results and publication, insofar as 64.6% of null results were not written up and only 20.8% were published, whereas only 4.4% of statistically significant and hypothesis-consistent results were not written up and 61.1% were published. Reasons given included the belief that null results would not be accepted for publication, lack of interest in non-significant results, or shifting focus to other projects.

Furthermore, regarding the acceptance of manuscripts for publication, little to no bias towards statistically significant results has been found on the part of editors in retrospective studies of real-life settings of manuscript submission processes to journals (Lee et al. 2006; Okike et al. 2008; Olson et al. 2002; Timmer et al. 2002). Both Olson and colleagues (2002) and Lee and colleagues (2006) found that having statistically significant results did not improve the chance of a study being published after adjusting for other study characteristics. Okike and colleagues (2008) found that the only factor significantly associated with acceptance for publication was level of evidence. Timmer and colleagues (2002) found no statistically significant association between direction of study results and subsequent publication, although studies with non-significant findings were less likely to be published in high impact journals.

Experimental Studies on Publication Bias

While many methodological and review studies have focused on identifying publication bias in published studies (Borenstein, 2019; Carter et al. 2019; Fanelli, 2011; Niemeyer et al. 2020; Scheel et al., 2021), experimental studies on publication decisions are crucial for determining the variables that impact the preference for positive results (Elson et al., 2020). However, only few studies have experimentally investigated the influence of statistically non-significant and hypothesis-inconsistent results on the evaluations of researchers on the quality and publishability

of studies (Atkinson et al., 1982; Augusteijn et al., 2023; Chopra et al., 2022; Elson et al., 2020; Epstein, 1990; Mahoney, 1977).

While Mahoney (1977) as well as Atkinson et al. (1982) found significant preferences for positive compared to negative results in experimental manuscript studies, with reviewers and consulting editors more likely to recommend acceptance for positive results, other researchers found no statistically significant preference for positive results in relation to the rate of study acceptance (Epstein, 1990), the overall recommendation of a study in the review process (Elson et al., 2020), or several study quality ratings (Augusteijn et al., 2023). However, all these between-subject studies only investigated the rating of one experimentally varied stimulus (manuscript or abstract), which was pointed out as a clear limitation by Elson et al. (2020). The generalizability of the results to different stimuli may be constrained if only one experimentally varied stimulus is presented. The identified difference or absence of a difference between positive and negative results might result from an interaction between unique abstract characteristics and statistical significance or hypothesis-consistency. Chopra et al. (2022) stand out as an exception in this context, as their study utilizing a within-subjects design with researchers in the field of economics revealed a significant tendency to reject null results.

Non-Reception

The publication of certain study results might also be considered unfavorable because researchers may expect negative consequences for subsequent citation and for their own reputation. This brings the relationship between publication bias, research quality, and the reception of published studies into focus. Results that are statistically significant or hypothesis-consistent might serve a signaling function in the scientific system and generate attention, such that the perception of results is tied to these characteristics. The non-publication of null, inconclusive, or hypothesis-inconsistent results might therefore be attributable to both poorer chances of publication and lower expected reputational gains.

Questions of this kind regarding citations are discussed in meta-science under the term *citation bias* (Jannot et al. 2013). Reviews in this area have yielded mixed findings: While Callaham et al. (2002) found no effect of positive results on citation frequency, other studies found higher citation rates for articles with statistically significant or hypothesis-consistent results (Jannot et al. 2013; Duyx et al. 2017).

To date, little attention has been paid to reception-related factors other than citations like reading decision, downloads, or altmetrics[1]. Tenopir and King (2002) found that electronic publishing and new search and communication tools have expanded scientists' reading range across multiple journals. We assume that *reception* and/or *citation bias* occur in the context of a large supply of scientific literature, possibly due to time pressure and selective search strategies as well as reception decisions by researchers in the field. To the best of our knowledge, no experimental study has investigated the influence of statistical significance or hypothesis-consistency to study citation and reading behavior in scientists.

Decision-Making Processes: Two-Response Paradigm

Experimental studies on publication bias have focused on assessing final quality judgements of manuscripts and abstracts (Atkinson et al., 1982; Augusteijn et al., 2023; Elson et al., 2020;

[1] Altmetris are various metrics about the impact of papers, such as online views, downloads, recommendations and others (Haustein, 2016; Priem, 2010).

Epstein, 1990; Mahoney, 1977), but the role of decision-making dynamics in judgements on submitting, citing or reading a paper has not been investigated by scholars in this field. Dual Process Theories (DPT) provide a theoretical framework to understand the relationship between fast, automatic and more deliberate, analytic judgement processes (Evans, 2006; Evans & Frankish, 2009; Kahneman, 2003). To investigate when decision-making relies on fast and automatic (*Type 1*) or slow and deliberate (*Type 2*) processing, the interaction of these two processing systems has been investigated by two-response procedure experiments (Thompson et al., 2011; Thompson & Johnson, 2014). In two-response procedure experiments, participants are typically asked to provide three evaluations regarding a presented stimulus: they start with a quick and intuitive *stage 1 response* that is assumed to indicate Type 1 processing (Evans, 1996; Thompson et al., 2011). This is followed by a question about the *Feeling of Rightness* (FOR) related to the stage 1 response, which is defined as the "degree to which the first solution that comes to mind feels right" (Ackerman & Thompson, 2017, p. 608).[2] Lastly, participants provide their *stage 2 response*, during which they are allowed as much time as needed to reconsider their initial answer and provide a final answer, which is assumed to be indicative of Type 2 processing (Evans, 1996; Thompson & Johnson, 2014). Using this two-response paradigm, we propose to describe researchers' decision-making concerning statistical significance and hypothesis-consistency, based on four assumptions.

First, we assume that negative evaluations of abstracts due to statistically non-significant as well as hypothesis-inconsistent results (negative results) are prevalent in stage 1 responses when fast and automatic Type 1 processing takes place, since this type of processing relies on heuristics, mental shortcuts, and pattern recognition (Thompson et al., 2011, Thompson & Johnson, 2014). From our point of view, negative results can be viewed as salient patterns of research abstracts and papers, which give researchers the impression of reduced publication and reception success.

Second, we assume that negative results are directly associated with more positive evaluations in stage 2 compared to stage 1. This results from more intensive and deliberate cognitive processing in stage 2, which allows adjustment of initially biased decisions (Thompson et al., 2011, Thompson & Johnson, 2014). During deliberate thinking researchers might be more aware of heuristics such as the negative results bias in scientific decision-making and thus might re-evaluate their initial negative responses.

Third, we propose that negative results are linked to lower FOR judgments after the intuitive processing. This assumption is derived from experimental observations: (a) FOR after Type 1 processing indicates the need for additional analysis (Type 2 processes) (Thompson et al., 2011; Thompson & Johnson, 2014), and (b) lower FOR judgments are more likely to be triggered by conflict stimuli, which induce cognitive conflict in decision-making, compared to non-conflict stimuli (Thompson & Johnson, 2014, p. 226). In our perspective, negative results can be regarded as conflict stimuli, potentially evoking conflicting cognitions within researchers. Two relevant conflicting cognitions for negative results might compete: an immediate, intuitive cognition (Type 1) that posits "it is challenging to publish null results" versus a more reflective, analytical cognition (Type 2) suggesting that, based on ongoing discussions in the scientific community, "null results

[2] Some authors propose an alternative three-stage nomenclature, where conflict detection responses such as FOR judgments are labeled as Stage 2 and Type 2 processing responses are designated as Stage 3 (Pennycook et al., 2015). However, for the sake of clarity, we chose to use a nomenclature wherein the stage numbering (Evans, 1996) aligns with the numbering of processing types (Thompson et al., 2011).

are equally worthy of publication". We assume that this conflict is also relevant for reading and citing.

Fourth, we propose that lower FOR are associated with more positive evaluations in stage 2 compared to stage 1. This hypothesis is grounded in empirical observations within DPT-experiments, that lower FOR is associated with higher response changes between stage 1 and stage 2 (Wang & Thompson, 2019, p. 37).

Furthermore, FOR and response changes might additionally be influenced by other factors such as (1) the extent of knowledge about the detrimental consequences of publication bias for research quality; (2) the subjective pressure to publish, (3) or academic professional status. Researchers that are aware of the relevance of publication bias in their own research routine are expected to experience a lower FOR when rejecting non-significant results. Conversely, researchers who feel (more) pressured in the scientific systems to publish might avoid significant (and hypothesis-conforming) results to ensure that their results are published, resulting in higher FOR ratings and lower response changes. Furthermore, the general level of FOR and response changes might differ based on academic professional status.

Present Study

The present study aims to identify decision-making processes behind selective non-publication and non-reception. Four online experiments were conducted in which fictitious abstracts were presented in a within-subjects design with statistical significance and hypothesis-consistency as experimental treatment variables. Our nested within-subjects linear mixed model design enabled us to isolate the effect of the treatment variables statistical significance and hypothesis-consistency by estimating the between-abstract heterogeneity. Furthermore, we investigated decision-making processes using the two-response procedure based on DPT (Kahneman, 2003; Thompson, 2009). Specifically, we examined intuitive (Type 1 processing) and considered evaluations (Type 2 processing) of research abstracts, along with the accompanying FOR of intuitive evaluations, as functions of the experimental variation of statistical significance and hypothesis-consistency. In summary, our study design addresses three gaps in the current literature: (A) experimental within-subject design studies on publication bias in psychology, (B) experimental research on non-reception, (C) and the role of decision-making processes in non-reception and non-publication.

We assume that the decision-making processes that contribute to publication bias are likely relevant for all empirically oriented disciplines, but we focus on clinical psychology for two reasons: First, to have a rather homogeneous sample with respect to familiarity with the stimulus material (abstracts). Second, clinical psychology exerts a strong impact on the health care system and society, including in terms of costs (Doran & Kinchin, 2017; Gabbard et al., 1997; Knapp & Wong, 2020; McDaid et al., 2019). Clinical decision making and clinical practice should be evidence-based (Shapiro, 2002) and inform therapeutic work, and only efficacious treatments should be used and invested in (Niemeyer et al., 2013). For this purpose, it is also necessary to publish statistically non-significant findings as well as findings that do not confirm hypotheses. The replication crisis identified over a decade ago has led to a high sensitivity to this issue in social and personality psychology, but further reform is necessary in the field of clinical psychology (Tackett et al., 2019).

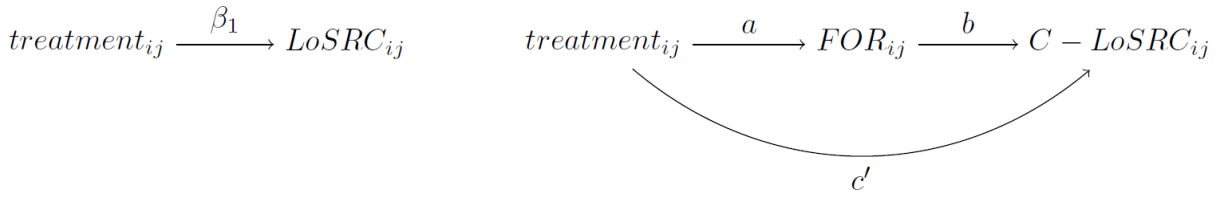
Hypotheses

For each combination of the response — evaluations regarding (1) publishability, (2) citation and (3) reading research papers based on research abstracts —and the treatment variable, either (1) statistical significance or (2) hypothesis-consistency, we test the three hypotheses HX.1, HX.2.1, and HX.2.2, resulting in a total of 3×2×3=18 hypotheses. HX.1 is examined using a multilevel model (MLM) and investigates intuitive responses regarding the *likelihood of submitting to publish, citing, or reading* (LoSRC) an abstract. However, HX.2.1 and HX.2.2 are analyzed using a multilevel mediation model (MLMM) and examine *changes in the likelihood of submitting to publish, citing, or reading* (C-LoSRC) an abstract, which are defined as the difference between considered responses in stage 2 and the intuitive responses in stage 1. The hypothesized relationships between the variables are shown in Figure 1.

- **HX.1**. *Effect (β_1) from Treatment on Intuitive Responses:* Researchers are more likely to give a negative evaluation of an abstract (report a lower LoSRC) if the abstract is not statistically significant or inconsistent with the hypothesis (meaning the experimental treatment has the value 1).
- **HX.2.1.** *Average Direct Effect (c') from Treatment on Response Change from stage 1 to stage 2.* Researchers are more likely to give a more positive considered evaluation than the initial intuitive evaluation in submitting to publish, citing, or reading (report a higher C-LoSRC) if the abstract is not statistically significant or inconsistent with the hypothesis (meaning the experimental treatment has the value 1).
- **HX.2.2**: *Average Causal Mediation Effect (a & b) from Treatment on Response Change from stage 1 to stage 2 via FOR:* Researchers are more likely to give a more positive considered evaluation than the initial intuitive evaluation in submitting to publish, citing, or reading (report a higher C-LoSRC) if the FOR of the initial evaluation was low, which itself will be more likely if the abstract is not statistically significant or inconsistent with the hypothesis (meaning the experimental treatment has the value 1)

Figure 1.

Hypothesized Relationships between Variables for HX.1 (left) as well as HX.2 (right)



Note. $LoSRC_{ij}$ = Intuitive response regarding the likelihood of submitting to publish, citing, or reading abstract *i* for individual *j* in stage 1; $C\text{-}LoSRC_{ij}$ = Response change regarding the likelihood of submitting to publish, citing, or reading abstract *i* for individual *j* between stage 2 and stage 1*;* FOR_{ij} = Feeling of Rightness regarding individual *j*'s LoSRC response with respect to abstract *i.*

Open Practices

The study was pre-registered with the Open Science Framework (OSF).[3] The data is available to the research community online in the spirit of open access. It is not possible to trace the data back to individual participants as we only recorded age in ranges and qualification levels in broad terms.

Preregistration

https://osf.io/ [will be added after revision]

Data, Materials, and Online Resources:

Open Data: https://osf.io/ [will be added after revision]

Open Materials: https://osf.io/ [will be added after revision]

All data and materials have been made publicly available at OSF and can be accessed at https://osf.io/ [will be added after revision] and https://osf.io/ [will be added after revision]. The protocol and analysis plans were preregistered at OSF and can be accessed at https://osf.io/ [will be added after revision]. Changes to the preregistered analyses are described in the text. This article has received badges for Open Data, Open Materials, and Preregistration.

More information about the Open Practices badges can be found at https://www.psychologicalscience.org/publications/badges. A preprint of the article was posted prior to publication: https://psyarxiv.com/ [will be added after revision].

Reporting

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Ethical Approval

The study protocol, materials and methods have been approved by the ethics committee of the Freie Universität Berlin.

Methods

Procedure

The experiments were programmed in jsPsych (De Leeuw, 2015) and ran on the platform '*Pavlovia.org'*. The experimental procedure for all four experiments is depicted in Figure 2. During the experiments, the participants were initially given an introduction to read, which was already announced in the e-mail invitation. In the introduction, they were told that we are interested in exploring how research abstracts are read and assessed in the context of routine procedures for literature review and preparing own publications in clinical psychological research practice (see Appendix G). Moreover, the researchers were told that we are interested in their spontaneous gut judgments and that it is therefore important to answer as quickly and honestly as possible. We also emphasized in the instruction that all abstracts presented are based on carefully conducted studies with sufficient statistical power.

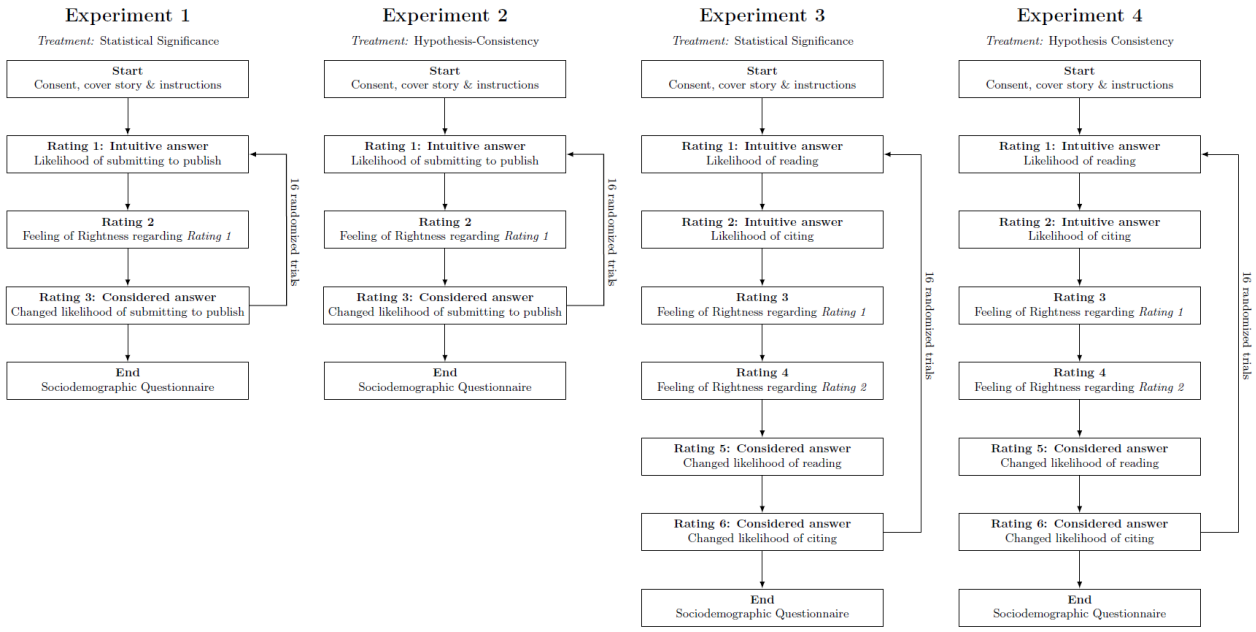
[3] All links will be adjusted after feedback by reviewers and editors

The participants were then presented with the fictitious abstracts. Out of a total of 32 abstracts per experiment (16 pairs of abstracts), each participant was presented with 16 different abstracts, with only one example presented per pair. Within pairs, either only statistical significance (experiments 1 and 3) or only hypothesis-consistency (experiments 2 and 4) was systematically varied. Otherwise, both abstracts in each pair were identical. The selection of abstracts to be presented was randomized and balanced. In experiments 1 and 2, participants' decisions in terms of publishability were assessed. Participants in experiments 3 and 4 decided how likely it is that they would first read and second cite the entire article based on the respective abstracts.

We adapted the two-response procedure as follows: Participants were instructed to provide an initial, intuitive response to the problem. This first response was to be given quickly, intuitively, and with a minimum of thought. This was done to maximize the probability that participants provided the first answer that came to mind. This was followed by an assessment of FOR. In the second stage the participants were allowed as much time as needed to reconsider their initial answer and provide a final answer. After rating 16 abstracts, they were asked to provide further sociodemographic and professional information. The reaction time in stage 1 and the change in the decision in stage 2 were recorded. Sensitivity analyses with respect to short reaction times in stage 2 were conducted. The design of the four experiments was identical except for the instructions, stimulus material, and questions. Participation in the study took approximately 20 minutes per experiment. Participants were randomly invited to participate in only one experiment.

Figure 2.

Experimental Procedure



Sample Recruitment Process

We conducted a power analysis (Westfall, Kenny & Judd 2014; refer to the Appendix G for the code and details) with an expected effect size of d = 0.55 for the main question (Chopra et al., 2022), using a sample of N = 75 participants and N = 16 stimuli per experiment, resulting in an

estimated power of 0.836. Online studies on publication bias with researchers typically experienced very low response rates, ranging from 3.5% (Chopra et al., 2022) to 5.9% (Augusteijn et al., 2023). In order to address the challenge of low response rates; we employed a double tracked sample recruitment process.

First, we utilized a pre-existing list of German Clinical Psychologists created for an ongoing meta-review on publication bias. The foundation of this researcher list was a compilation of state university departments or laboratories (N = 91) focused on research in clinical psychology, psychotherapy, or related fields in Germany. Laboratories in hospitals were excluded from this list. The inclusion spanned from pre-doctoral levels up to professors. We omitted universities in Berlin and Brandenburg from this approach, as we had already invited researchers from these regions to participate in an additional ongoing interview study. We gathered email addresses and the size of the population pool from the respective department websites, which resulted in a total of n = 1006 researchers in Germany.

Second, we aimed to extend our sample with researchers from countries with a high impact on psychological research in terms of publication and citation. According to O'Gorman and colleagues (2012) as well as Martinez-Garcia and colleagues (2012) researchers from the United States (US), United Kingdom (UK), Canada, Germany, Australia and Netherlands contribute the most publications and citations in the field of psychology. Therefore, we utilized the NCBI eUtils API to perform extensive correspondence mail scraping based on PubMed articles for US, UK, Canada, Australia, and Netherlands. Researchers often provide their correspondence email addresses in the affiliation section, which is frequently part of publicly available metadata. Thus, for each country, we conducted a PubMed search using the query: ((Mental Health Services[MeSH Terms]) OR (Psychotherapy[MeSH Terms]) OR (Psychology, Clinical[MeSH Terms])) AND (*Country*[Affiliation]) AND ("2015"[Date – Publication]: "2023"[Date – Publication]). This query yielded after removing duplicates n = 7797 email addresses for the United States, n = 2899 for the United Kingdom, n = 1490 for Canada, n = 1221 for Australia and n = 1055 for the Netherlands. Resulting in a total of n = 14462 mail addresses. A similar mail scraping approach using Web of Science was also recently used by another study on publication bias in psychology researchers (Augusteijn, 2023).

First, we sent an email to all the gathered researchers from Approach 1 (N = 1006), and a random sample of 5000 emails to researchers from Approach 2. After 2 weeks, we sent a reminder to the researchers to participate in our study. Subsequently, we drew random samples of researchers from each country until we reached our sample size of 75 participants per experiment.

Pilot Study

In a pilot phase, we tested whether the experiments run smoothly and evaluated the perceived realism of the abstracts. For this purpose, from our own clinical psychology department, we invited ten participants for each experiment to take part in the pilot phase. Five subjects participated fully in experiment 1, and two subjects participated in experiment 2, 3 and 4, respectively. Based on the results of the pilot study, ten abstracts had to be changed slightly in wording (see below).

Inclusion

The participants for the main study were currently employed at a university in Germany, the UK, Netherlands, the US, Canada, or Australia at a chair of clinical psychology, psychotherapy or in a comparable research group (also possible: scholarship for doctoral studies, active in research).

As the abstracts were presented in English, participants had to be able to understand academic English and feel comfortable with reading English abstracts.

Exclusion

We ran a plausibility check of the response times of the participants and excluded all responses connected to initial abstract reading time below 2 seconds.

Reimbursement

Participants did not receive reimbursement for taking part in the study.

Debriefing

Information about the aim of the study was provided after the entire data collection was completed to prevent socially desirable responding.

Data Security

All personal data is handled and stored in line with the requirements of the GDPR. The storage and backup of the research data was ensured during the project by the project leaders in cooperation with the responsible data processing officer of the computer media service of the university. For this purpose, the infrastructure of Freie Universität was used.

Materials

For each of the experiments, 16 case pairs of short abstracts have been generated. Each pair consists of two abstracts that are identical with the exception of the statistical significance or hypothesis-consistency of the results. The number of presented abstracts for the underlying counterbalanced design was determined using an a priori power analysis (Westfall et al., 2014). Stimuli were quasi-randomly presented to the participants (8 statistically non-significant and 8 statistically significant abstracts, or 8 hypothesis-consistent and 8 hypothesis-inconsistent abstracts).

To generate the abstracts, we first developed a table to create abstract titles containing different relevant aspects of clinical psychology research (population of interest, diagnosis, clinical intervention, intervention context or group comparison, outcome measure, mediating or moderating factors). This resulted in the creation of 60 fictitious research paper titles (e.g., “Quality of life effects of behavioral activation in teletherapy for patients with social anxiety”) in a first step.

Next, with the help of wordblot.ai, an AI-powered writing assistant, we used the 32 titles to generate 32 prototypical abstracts, which we fundamentally adapted based on the criteria of realism, readability, and structure (Background, Method, Results, Conclusion). Subsequently, a native English speaker proofread the abstracts according to readability and English language. Within our pilot study, we also asked colleagues to proofread and rate the abstracts according to realism (0% very unrealistic, 100% very realistic), resulting with a range of 10 – 99 % and an average of 67.8 %. With this in mind and on the basis of specific comments our colleagues made, we changed some wording in ten abstracts.

In the Background section of the abstracts, we used phrases like “It was hypothesized”/”We hypothesized that” for experiments 2 and 4, or phrases like “We investigated whether” for experiments 1 or 3. While the Method section remained the same for all variations, the Results section differed for experiments 2 and 4, with phrases like “Our results confirm” or “Our

hypotheses […] could not be confirmed" and for experiments 1 or 3 with phrases like "significantly predicted" or "did not significantly predict"/ "were non-significant". Finally, the Conclusion section was adjusted according to the results (e.g., "Our results might suggest that comparable to anxiety disorders, treatment success in OCD might depend on an individual's level of conscientiousness" or "Our results might suggest that there is no relevant relation between treatment success and an individual's level of conscientiousness in patients with OCD", see example in Appendix A). In order to resemble real abstracts as closely as possible, and to make it more difficult for participants to merely skim the abstracts, we did not include any subheadings for each section.

Abstract Length in Manuscript: The abstracts varying in statistical significance and the abstracts varying in hypothesis-consistency have a mean length of 155.2 (SD = 29.4) and 178.8 (SD=26.6) words, respectively.

The stimulus set of 64 abstracts was published after completion of the study for future replication studies [will be added after review]. In Appendix B, we provide a table of all abstracts including titles and designs.

Measures

For each abstract, we measured the amount of time taken to read the abstract, the response time (RT) to generate the first response (answer fluency), the likelihood of submitting to publish/reading/citing (0-100%), the Feeling of Rightness (FOR) rating (1-7, "When I gave the answer, I felt" - "Very uncertain " - " Very certain "), the second, considered response regarding the likelihood of submitting to publish/reading/citing (0-100%), and the difference between the two likelihood responses (see Fig. 2 for experimental procedures). Response times for all experiments were converted to $\log^{10}$ prior to analysis (RTs in the tables are reported in the original units). However, reaction time only represents a control variable for intuitive decision making in our design.

After rating 16 abstracts, the participants provided information regarding their sociodemographic and professional characteristics. The questions included their age, their gender, the country to which their respective research organization is affiliated (Germany, US, UK, Netherlands, Australia, Canada, or other), their position at the research organization, their psychotherapeutic practice, their reviewer history (number of peer reviews), their authorship experience (number of papers as first and last author), their familiarity with open science and publication bias (each rated from 1 = not at all familiar to 5 = very familiar), how pressured they feel to publish their research (from 1 = not at all pressured to 5 = very pressured), their work in other areas of psychology (several options provided, such as social psychology, biological psychology, etc.), and the number of their own published papers that contained statistically non-significant results and hypothesis-inconsistent results. Lastly, the researchers were presented with 6 items from the BFI-2-S conscientiousness subscale (Soto & John, 2017).

Implementation Period

The study was conducted in the second half of 2023. Data collection began in December 2023 and should be completed by January 2024. Data analysis was conducted in January 2024.

Statistical Analysis

HX.1 hypotheses are investigated using MLM, while HX.2.1 and HX.2.2 hypotheses are examined using MLMM. In all models positive results, statistical significant or hypothesis-

consistent results, are coded as 0 (*control*) and negative results, statistical non-significant or hypothesis-inconsistent results, are coded as 1 (*treatment).*

HX.1: Multilevel Models

In our MLM, we assumed that data points are nested both within participants and abstracts. We therefore proposed that the reported likelihood of submitting to publish (LoS), reading (LoR), or citing (LoC) an abstract is influenced both by participant and abstract characteristics. However, the experimental data structure is not perfectly hierarchical, as within each experiment we presented the same 16 abstracts to the participants, with the exception of the manipulation of statistical significance or hypothesis-consistency in the *Results* and *Discussion* section of the abstracts. Thus, we obtain a cross-classified data structure with a counterbalanced design, which can be adequately modeled with cross-classified multilevel modeling (Hox, 2002). To explore the association between statistical significance and LoS (H1.1), LoR (H3A.1) and LoC (H3B.1) as well as the association between hypothesis-consistency and LoS (H2.1), LoR (H4A.1) and LoC (H4B.1) we specified the following X.1 general model using the *lme4* (Bates et al., 2015) and *lmerTest* (Kuznetsova et al., 2017) package for linear mixed models with Restricted Maximum Likelihood (REML) estimation:

LoSRC ~ treatment + (treatment | participant) + (1 | abstract)

Our models conceptually subdivide the variation of LoSRC of participant *j* reading abstract *i* into random and fixed effects. The term *(treatment | participant)* indicates (a) a random slope for participants on the treatment variable and (b) a random intercept on the participant level, while (c) the term *(1 | abstract)* represents a random intercept on the abstract level. The term *treatment* captures the fixed regression effect for the manipulation of statistical significance or hypothesis-consistency.

HX.2.1 and HX.2.2: Multilevel Mediation Models

For our MLMM, we dropped the random intercept for abstracts, since causal multilevel mediation using the *mediation* package (Tingley et al., 2014) is only supported for two levels and cross-classified approaches are technically processed as three-level frameworks. Thus, we supposed that the *changes in the likelihood of submitting to publish from stage 1 to stage 2* (C-LoS = LoS_{t1} - LoS_{t0}), as well as the *changes in the likelihood of reading from stage 1 to stage 2* (C-LoR= LoR_{t1} - LoR_{t0}), and the *changes in the likelihood of citing from stage 1 to stage 2* (C-LoC = LoR_{t1} - LoR_{t0}) are influenced (a) directly by the treatment as suggested by HX.2.1 and (b) indirectly by the treatment via FOR as suggested by HX.2.2. Additionally, we assume that (c) the intercept of C-LoSRC differs between participants and (d) that the effects proposed in (a) and (b) differs between subjects in form of random slopes. Specifically, we explore the association between statistical significance and C-LoS (H1.2), C-LoR (H3A.2) and C-LoC (H3B.2) as well as the association between hypothesis-consistency and LoS (H2.2), C-LoR (H4A.2) and C-LoC (H4B.2). We specify the following X.2 basic models for all MLMM:

path_a = FOR ~ treatment + (treatment | subject)

path_b_c = C-LoSRC ~ FOR + treatment + (treatment | subject)

mediate(path_a, path_b_c, treat = "treatment", mediator = "FOR)

The mediation package reports estimates, 95% confidence intervals and p-values for essential mediation indices: Average Causal Mediation Effect (ACME), Average Direct Effect (ADE), Total Effect, and Proportion Mediated.

Further Model Specifications

The alpha level was set at 0.05. Normality of residuals as a function of model and level and homogeneity of variance in all models were examined for all MLM (Finch et al., 2019). Standardized beta coefficients were used as effect size measures for regression coefficients (Lorah, 2018) and interpreted as small (*St. Beta* ≤ 0.3), medium (0.3 ≤ *St. Beta* ≤ 0.5), and large (*St. Beta* ≤ 0.5). Since MLM as well as MLMM sometimes result in uninterpretable results due to lack of model convergence (Demian, 2019), we first dropped random slopes on the subject level and if models still failed to converge, we decided to report a 2x16 repeated measures ANOVA for MLM or a mediation model without level specification for MLMM models. For the R Code, assumption testing, and the further model results, see Appendix C, D and E.

Exploratory Analyses

We tested the influence of six categories of variables in exploratory analyses. An overview of all exploratory analyses is provided in Table 1.

Table 1

Overview on Exploratory Analyses

	Additive Term	Interaction Term	HX.1: MLM[a]	HX.2: MLMM[b]	Centering
Knowledge about PB[c]	X	X	X	X	GM[d]
Pressure to Publish	X	X	X	X	GM
Professional Status	X		X	X	NC[e]
Position	X		X	X	GM
Conscientiousness	X			X	GM
Country	X		X	X	NC

Note. [a] MLM = Multilevel Model; [b] MLMM= Multilevel Mediation Model; [c] PB = Publication Bias; [d] GM = Grand Mean Centering; [e] NC = No Centering.

Firstly, we examined the influence of *familiarity with publication bias on* LoSRC, FOR and C-LoSRC. Therefore, we included responses on an item on familiarity (*familiarity_PB*) with publication bias as a quasi-metric additive and interaction term *(familiarity_PB * treatment)* with treatment into the MLM and MLMM.

Secondly, we analyzed the influence of *pressure to publish* on LoSRC, FOR and C-LoSRC. Also, the pressure to publish variable *(pub_pressure*) was introduced as a quasi-metric additive and interaction term with treatment *(pub_pressure * treatment)* to MLM and MLMM.

Thirdly, we investigated the influence of *professional status group* on LoSRC in MLM. We separated the categorical variable *professional status group* (levels: pre-doctoral level, post-doctoral level, professor level) into two dichotomous predictors: *postdoc* and *professor.* Hence,

pre-doctoral level served as the reference category. We included additive terms for both status group predictors and the treatment into MLM and MLMM.

Fourthly, we investigated for both MLM and MLMM whether fatigue influences the decisions of researchers, since both motivation and concentration are likely to vary between the beginning, the middle and the end of evaluating 16 research abstracts. Therefore, we incorporated a *position* variable (numeric; values from 1-16; indexes when which abstract was presented) as an additive term into MLM and MLMM.

Fifthly, we investigated whether researcher's tendency to rate an abstract again after giving a first response is driven by conscientiousness rather than the treatment or FOR. Individuals scoring high on conscientiousness might exhibit a heightened likelihood of altering their initial responses. This propensity could potentially arise from conscientious individuals' inclination to feel a stronger sense of obligation in modifying their responses when prompted by instructions to reconsider the problem and provide a secondary response. Therefore, we introduced the BFI-2-S subscale conscientiousness as a covariate (*conscientiousness*) to the MLMM.

Sixth, we analyzed differences between researchers from different countries. We introduced additive regression terms for researchers from the Netherlands (*ned*), United Kingdom (*uk*), Canada (*can*), United States (*us*), Austrialia (*aus*) and other countries (*other*) into MLM and MLMM. Researchers from Germany represent the reference category. Controlling for country had not only the intention to control for national differences but also to take potential confounders of the two different recruitment processes into account.

Incorporating these exploratory analyses into the models resulted in the following *lme4* and *mediation* specifications:

Models X.1.C: Exploratory Specification for MLM stage 1 models

LoSRC ~ treatment * familiarity_PB + treatment * pub_pressure + postdoc + professor + position + nd + uk + us + can + aus + other + (treatment | subject) + (1 |abstract)

Models X.2.C: Exploratory Specification for MLMM stage 2 models

path_a = FOR ~ treatment * familiarity_PB + treatment * pub_pressure + (treatment | subject)

path_b_c = CLoSRC ~ treatment * familiarity_PB + treatment * pub_pressure + FOR + conscientiousness + position + nd + uk + us + can + aus + other + (treatment | subject)

mediate(path_a, path_b_c, treat = "treatment", mediator = "FOR)

Results

Descriptive Statistics

We invited XXXX participants to the study. In Pavlovia, only data of fully completed responses were saved as data sets. Because of technical [or unforeseen] difficulties, we had to exclude XX participants during the collection. We closed the experiment, as soon as we reached 75 analyzable participants per experiment. In total, 1,200 trials (75 participants x 16 abstracts) were completed per experiment in this study.

Over all experiments, XX (X %) participants identified as female, XX (X %) as male, and XX (X %) as diverse. XX (X %) were aged between 18-25, XX (X %) between 26-35, XX (X %) between 36-45, XX (X %) between 46-55, and XX (X %) over 55. XX (X %) were on predoctoral level, XX (X %) on postdoctoral level, and XX (X %) were professors. Table 2 shows sociodemographic data separately for the four experiments.

Median reading time for all abstracts was *Md* = XX.X seconds (*SD* = XX.X). There were <no significant differences | significant differences> between reading times in experiments in which statistical significance and hypothesis-consistency were varied, t(XXXX) = X.XX, p = .XXX. Due to too short reading times (<2 seconds) XX trials were excluded from further analysis (Experiment 1: XX; Experiment 2: XX; Experiment 3: XX; Experiment 4: XX).

Results for Stage 1

Over all abstracts, the mean LoS was XX % (*SD* = XX, *Range* = XX-XX%), the LoR was XX % (*SD* = XX, *Range* = XX-XX%), and LoR was XX % (*SD* = XX, *Range* = XX-XX%). Table 3 shows the LoSRC separately for significant and non-significant, and hypothesis-consistent and non-consistent results.

Table 2

Sociodemographic Data for all four Experiments

	Experiment 1 *n (%)*	Experiment 2 *n (%)*	Experiment 3 *n (%)*	Experiment 4 *n (%)*
Gender				
Female	47 (62.7)	47 (62.7)	47 (62.7)	47 (62.7)
Diverse	7 (9.3)	7 (9.3)	7 (9.3)	7 (9.3)
Male	21 (28.0)	21 (28.0)	21 (28.0)	21 (28.0)
Age				
18 - 25	11 (14.7)	11 (14.7)	11 (14.7)	11 (14.7)
26 - 35	31 (41.3)	31 (41.3)	31 (41.3)	31 (41.3)
36 - 45	21 (28)	21 (28)	21 (28)	21 (28)
46 - 55	7 (9.3)	7 (9.3)	7 (9.3)	7 (9.3)
> 55	5 (6.7)	5 (6.7)	5 (6.7)	5 (6.7)
Status Group				
Pre-doc	48 (64.0)	48 (64.0)	48 (64.0)	48 (64.0)
Post-doc	20 (26.7)	20 (26.7)	20 (26.7)	20 (26.7)
Professors	7 (9.3)	7 (9.3)	7 (9.3)	7 (9.3)
Total	75 (100)	75 (100)	75 (100)	75 (100)

Independent t-tests revealed <no significant differences | significant differences> differences between significant and non-significant abstracts for the LoS in Experiment 1, t(XXXX) = X.XX, p = .XXX, as well as for the LoR in Experiment 3, t(XXXX) = X.XX, p = .XXX, and the LoC in Experiment 3, t(XXXX) = X.XX, p = .XXX. Furthermore, significant differences were identified between consistent and non-consistent abstracts for the LoS in Experiment 2, t(XXXX) = X.XX, p = .XXX, as well as for the LoR in Experiment 4, t(XXXX) = X.XX, p = .XXX, and the LoC in Experiment 4, t(XXXX) = X.XX, p = .XXX, t(2248) = -13.82, p >.001. Boxplots showing the

differences in subjects' responses for significant and non-significant or consistent and non-consistent abstracts are shown in Figure 3.

Model Results

The results for the stage 1 MLM models (Model X.1) for Experiment 1 (Model 1.1), 2 (Model 2.1), 3 (Model 3A.1 and Model 3B.1)) and 4 (Model 4A.1 and Model 4B.1)) are presented in Table 4. Further results for stage 1 MLM models can be found in Appendix D. Graphical examination of residual plots (Appendix C) revealed that the assumption of normality of residuals for the random slope models is met for all models. Plots of standardized regression coefficients are seen in Figure 4 to visualize effect sizes.

Table 3

Descriptive Statistics for the Likelihood of Submitting for Publication, Reading, and Citing Significant and Non-Significant, and Hypothesis-Consistent and Non-Consistent Results

	Statistical significance		Hypothesis-consistency	
	Significant	Non-significant	Consistent	Non-consistent
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*
Likelihood (%) of				
Submitting for Publication	75.3 (17.3)	64.2 (19.7)	75.3 (17.3)	64.2 (19.7)
Reading	75.3 (17.3)	64.2 (19.7)	75.3 (17.3)	64.2 (19.7)
Citing	75.3 (17.3)	64.2 (19.7)	75.3 (17.3)	64.2 (19.7)

H1.1: Statistical Significance and Publishability In our main model for hypothesis H1.1, Model 1.1, the treatment variable statistical significance was < significantly | not significantly> associated with LoS (*Beta* = XX.XX, *p*=.XXX). On average the effect for statistical significance on LoS was <strong/medium/weak> (*St. Beta* = XX.XX), still for 95% of subjects a slope for statistical significance between XX.XX and XX.XX was identified ($\sigma^2_{\text{Statistical significance|Subject}}$ = XX.X). Furthermore, analysis of multilevel R^2 in Model 1.1 showed that fixed effects (*marginal* R^2 = *.XX*) explained XX% of the variance in LoS, while random effects and fixed effects together (*conditional* R^2 = .XX*)* accounted for XX% of the variance in LoS.

H2.1: Hypothesis-Consistency and Publishability In Model 2.1 (main model: random slope model) the factor hypothesis-consistency < significantly | not significantly> predicted LoS (Beta = XX.XX, p=.XXX). The effect for hypothesis-consistency on LoS was <strong/medium/weak> (*St. Beta* = XX.XX), however, 95% of subjects demonstrated a slope for hypothesis-consistency between XX.XX and XX.XX ($\sigma^2_{\text{Hypothesis-Consistency|Subject}}$ = XX.X). In Model 2.1 *marginal* R^2 was .XX, whereas *conditional* R^2 was .XX.

Figure 3

Simulated Data: Boxplots for Differences in the Likelihood of Submitting for Publication, Reading and Citing Significant and Non-Significant or Consistent and Non-Consistent Abstracts

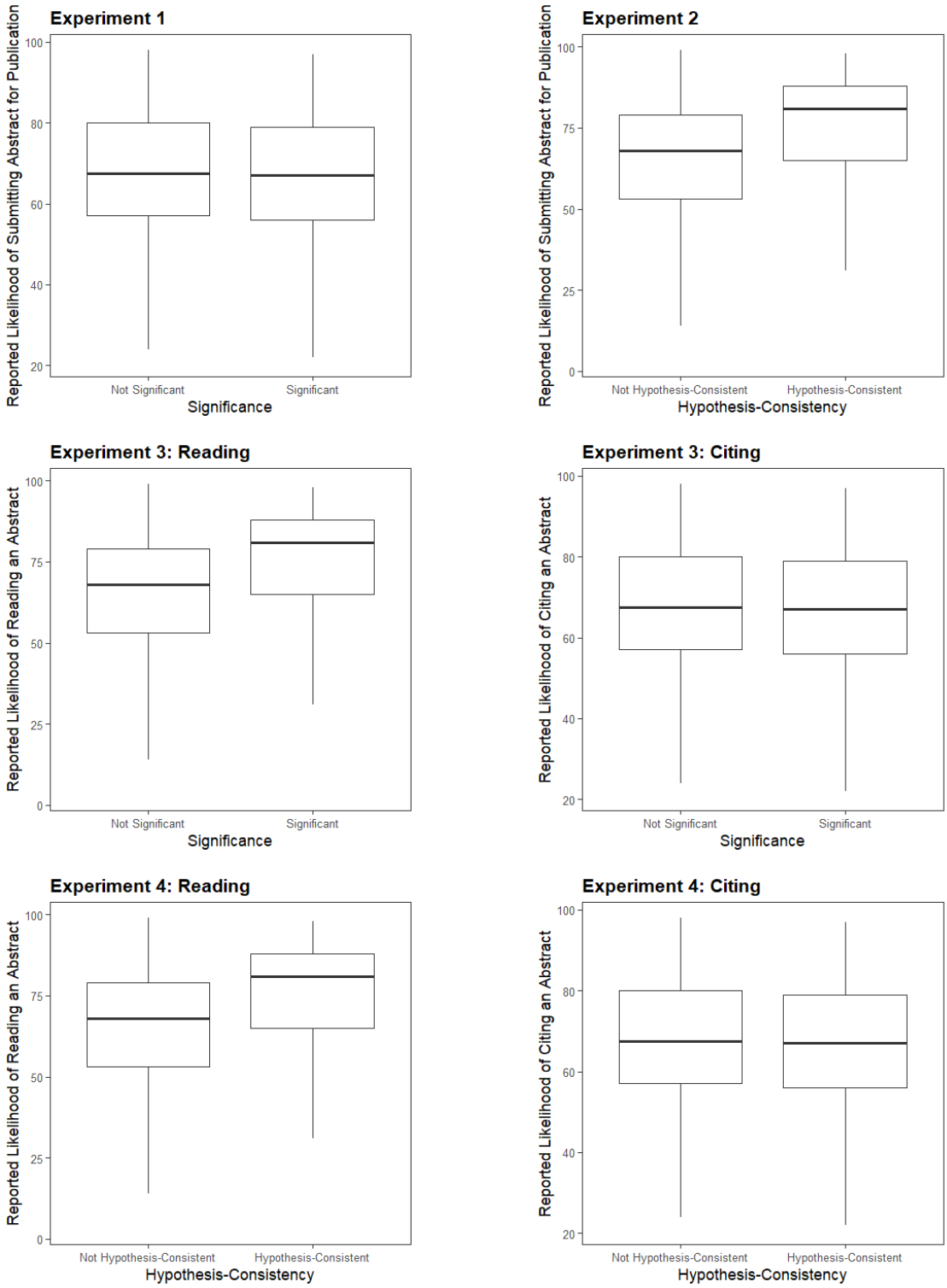


Table 4

Simulated Data: Model Parameters for All Stage 1 Random Slope Models

	Experiment 1 *Publishing* Model 1.1		Experiment 3 *Reading* Model 3A.1		Experiment 3 *Citing* Model 3B.1	
Fixed Effects	*Beta*	*SE*	*Beta*	*SE*	*Beta*	*SE*
Intercept	64.30***	1.76	64.30***	1.76	64.30***	1.76
Non-Significant Abstract	-11.13***	0.65	-11.13***	0.65	-11.13***	0.65
Random Effects	*SD*		*SD*		*SD*	
Subject: $\sigma^2_{(Intercept\|Subject)}$	12.49		12.49		12.49	
Item: $\sigma^2_{(Intercept\|Subject)}$	5.42		5.42		5.42	
Resid.: $\sigma^2_{(Resid.)}$	13.26		13.26		13.26	
Sig.|Subject: $\sigma^2_{(Sig.\|Subject)}$ [a]	1.44		1.44		1.44	
Model Quality	*Est.*		*Est.*		*Est.*	
R^2 Marginal	0.08		0.08		0.08	
R^2 Conditional	0.53		0.53		0.53	
	Experiment 2 *Publishing* Model 2.1		Experiment 4 *Reading* Model 4A.1		Experiment 4 *Citing* Model 4B.1	
Fixed Effects	*Beta*	*SE*	*Beta*	*SE*	*Beta*	*SE*
Intercept	64.30***	1.76	64.30***	1.76	64.30***	1.76
Inconsistent Abstract	-11.13***	0.65	-11.13***	0.65	-11.13***	0.65
Random Effects	*SD*		*SD*		*SD*	
Subject: $\sigma^2_{(Intercept\|Subject)}$	12.49		12.49		12.49	
Item: $\sigma^2_{(Intercept\|Item)}$	5.42		5.42		5.42	
Residuals: $\sigma^2_{(Residuals)}$	13.26		13.26		13.26	
HC|Subject: $\sigma^2_{(HC\|Subject)}$ [b]	1.44		1.44		1.44	
Model Quality	*Est.*		*Est.*		*Est.*	
R^2 Marginal	.08		.08		.08	
R^2 Conditional	.53		.53		.53	

Note. * $p \leq 0.05$ ** $p \leq 0.01$ and *** $p \leq 0.001$; [a] Sig. = Statistical significance; [b] HC = Hypothesis-Consistency.

H3A.1: Statistical Significance and Reception In the main model for Hypothesis H3.1, Model 3A.1, the variation of statistical significance of abstracts demonstrated <a significant | non-significant> relationship with LoR (Beta = XX.XX, p=.XXX). The effect for statistical significance on LoR was <strong/medium/weak> (*St. Beta* = XX.XX), yet 95% of subjects showed a slope for statistical significance between XX.XX and XX.XX ($\sigma^2_{Statistical\ significance|Subject}$ = XX.X). In Model 3A.1 *marginal* R^2 was .XX, while *conditional* R^2 was .XX.

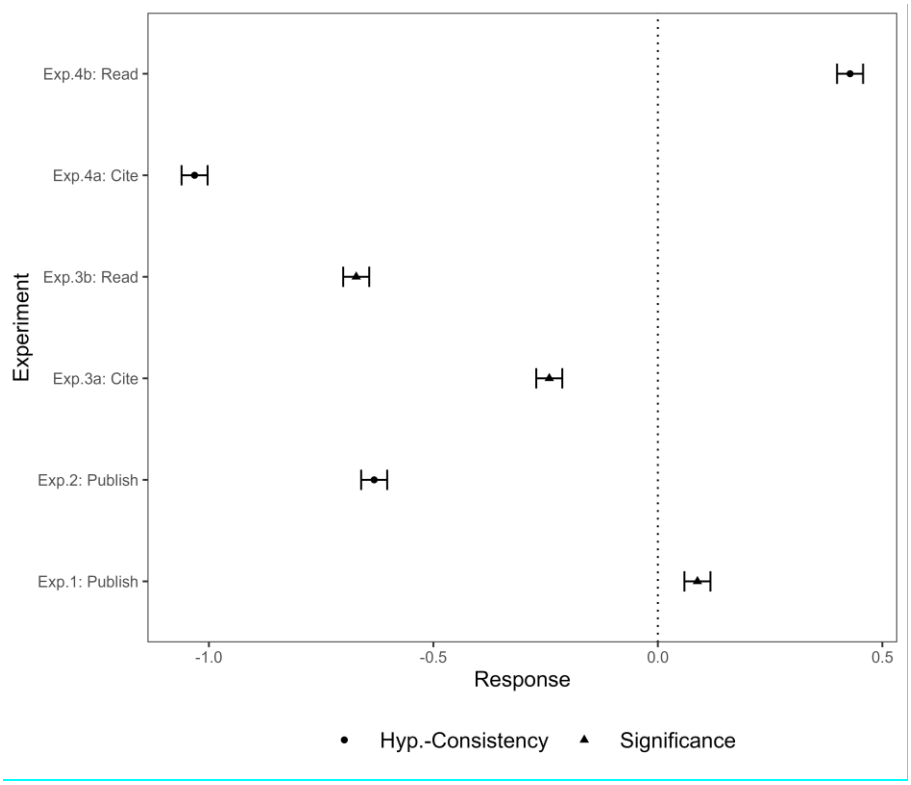
H3B.1: Statistical significance and Citation In the main model for Hypothesis H3B.1, Model 3B.1, the factor statistical significance was < significantly | not significantly> related to LoC (Beta = XX.XX, p=.XXX). On average the effect for statistical significance on LoC was <strong/medium/weak> (*St. Beta* = XX.XX): 95% of subjects exhibited a slope for the factor statistical significance between XX.XX and XX.XX ($\sigma^2_{Significance|Subject}$ = XX.X). In Model 3B.1 *marginal* R^2 was .XX, whereas *conditional* R^2 was .XX.

H4A.1: Hypothesis-Consistency and Reception In the main model for Hypothesis H4A.1, Model 4A.1, the hypothesis consistency of abstracts was < significantly | not significantly> associated with LoR (Beta = XX.XX, p=.XXX). The effect for hypothesis-consistency on LoR was <strong/medium/weak> (*St. Beta* = XX.XX): 95% of subjects demonstrated a slope for statistical significance between XX.XX and XX.XX ($\sigma^2_{\text{Hypothesis-Consistency|Subject}}$ = XX.X). In Model 4A.1 *marginal* R^2 was .XX, whereas *conditional* R^2 was .XX.

H4B.1: Hypothesis-Consistency and In the main model for Hypothesis H4B.1, Model 4B.1, hypothesis-consistency < significantly | not significantly> predicted LoC (Beta = XX.XX, p=.XXX). The effect for hypothesis-consistency on LoC was <strong/medium/weak> (*St. Beta* = XX.XX): 95% of subjects showed a slope for hypothesis-consistency between XX.XX and XX.XX ($\sigma^2_{\text{Hypothesis-Consistency|Subject}}$ = XX.X). In Model 4B.1 *marginal* R^2 was .XX, whilst *conditional* R^2 was .XX.

Figure 4

Plots of Standardized Regression Coefficients for the Treatment Variables Hypothesis-Consistency and Statistical significance for all Stage 1 Responses. Simulated Data



Results for Stage 2 and FOR

The mean response changes in the likelihood of submitting for publication were XX % (*SD* = XX, *Range* = XX-XX%), for reading were XX % (*SD* = XX, *Range* = XX-XX%), and for citing were XX % (*SD* = XX, *Range* = XX-XX%). Mean FOR judgements for initial submitting for publication were XX (SD = XX, Range = XX-XX), while mean FOR judgements for reading were XX (SD = XX, Range = XX-XX), and mean FOR judgements were XX (SD = XX, Range = XX-XX).

Independent t-tests revealed <no significant differences | significant differences> differences between statistically significant and non-significant abstracts for C-LoS in Experiment 1, t(XXXX) = X.XX, p = .XXX, as well as for C-LoR in Experiment 3, t(XXXX) = X.XX, p = .XXX, and for C-LoC in Experiment 3, t(XXXX) = X.XX, p = .XXX. Additionally, statistically significant differences were identified between hypothesis-consistent and hypothesis-inconsistent abstracts for C-LoS in Experiment 2, t(XXXX) = X.XX, p = .XXX, as well as for C-LoR in Experiment 4, t(XXXX) = X.XX, p = .XXX, and for C-LoC in Experiment 4, t(XXXX) = X.XX, p = .XXX..

Model Results

The results for the response change mediation models (Model X.2) for Experiment 1 (Model 1.2), 2 (Model 2.2), 3 (Model 3A.2 and Model 3B.2)) and 4 (Model 4A.2 and Model 4B.2)) are found in Table 5. Detailed model comparisons results for the response change models can be found in Appendix D.

H1.2: Statistical Significance and Response Changes in Publishability In the mediation model for hypothesis H1.2, Model 1.2, the presence of *statistically non-significant abstracts* exhibited a < significant | not significant> direct path to C-LoS (Beta = XX.XX, p=.XXX). Furthermore, the treatment variable demonstrated a < significant | not significant> mediated path to C-LoS (Beta = XX.XX, p=.XXX). < As hypothesized | Contrary to our hypothesis >, the effect of *statistically non-significant abstract* on FOR was < significant | not significant> and < negative | positive> (Beta = XX.XX, p=.XXX), and the effect of FOR on C-LoS was < significant | not significant> and < negative | positive> (Beta = XX.XX, p=.XXX).

H2.2: Hypothesis-Consistency and Response Changes in Publishability In the mediation model for hypothesis H2.2, Model 2.2, the presence of *hypothesis-inconsistent abstracts* showed a < significant | not significant> direct path to C-LoS (Beta = XX.XX, p=.XXX). Furthermore, the treatment variable exhibited a < significant | not significant> mediated path to C-LoS (Beta = XX.XX, p=.XXX). < As hypothesized | Contrary to our hypothesis >, the effect of *hypothesis-inconsistency* on FOR was < significant | not significant> and < negative | positive> (Beta = XX.XX, p=.XXX), and the effect of FOR on C-LoS was < significant | not significant> and < negative | positive> (Beta = XX.XX, p=.XXX).

H3A.2: Statistical Significance and Response Changes in Reception In the mediation model for hypothesis H3A.2, Model 3A.2, the treatment variable *statistically non-significant abstract* showed a < significant | not significant> direct path to C-LoR (Beta = XX.XX, p=.XXX). Furthermore, the treatment variable exhibited a < significant | not significant> mediated path to C-LoR (Beta = XX.XX, p=.XXX). < As hypothesized | Contrary to our hypothesis >, the effect of *statistically non-significant abstracts* on FOR was < significant | not significant> and < negative | positive> (Beta = XX.XX, p=.XXX), and the effect of FOR on C-LoR was < significant | not significant> and < negative | positive> (Beta = XX.XX, p=.XXX).

H3B.2: Statistical Significance and Response Changes in Citation In the mediation model for hypothesis H3B.2, Model 3B.2, the presence of an *statistically non-significant abstract* demonstrated a < significant | not significant> direct path to C-LoC (Beta = XX.XX, p=.XXX). Furthermore, the treatment variable exhibited a < significant | not significant> mediated path to C-LoC (Beta = XX.XX, p=.XXX). < As hypothesized | Contrary to our hypothesis >, the effect of *statistically non-significant abstracts* on FOR was < significant | not significant> and < negative |

positive> (Beta = XX.XX, p=.XXX), and the effect of FOR on C-LoC was < significant | not significant> and < negative | positive> (Beta = XX.XX, p=.XXX).

Table 5

Simulated Data: Model Parameters for All Response Change Random Slope Models

	Experiment 1 *RC: Publishing* Model 1.2		Experiment 3 *RC: Reading* Model 3A.2		Experiment 3 *RC: Citing* Model 3B.2	
Effects	*Beta*	*95%-CI*	*Beta*	*95%-CI*	*Beta*	*95%-CI*
ACME[a] N-Sig.[b]	0.03***	[0.01; 0.05]	0.03***	[0.01; 0.05]	0.03***	[0.01; 0.05]
N-Sig. → FOR[c]	-0.11***	[-0.14; -0.09]	-0.11***	[-0.14; -0.09]	-0.11***	[-0.14; -0.09]
FOR→ RC[d]	-0.27***	[-0.31; -0.22]	-0.27***	[-0.31; -0.22]	-0.27***	[-0.31; -0.22]
ADE[e] N-Sig.	0.26***	[0.18; 0.34]	0.26***	[0.18; 0.34]	0.26***	[0.18; 0.34]
Total Effect	0.29***	[0.21; 0.37]	0.29***	[0.21; 0.37]	0.29***	[0.21; 0.37]
Proportion Mediated	0.10***	[0.05; 0.18]	0.10***	[0.05; 0.18]	0.10***	[0.05; 0.18]
	Experiment 2 *RC: Publishing* Model 2.2		Experiment 4 *RC: Reading* Model 4A.2		Experiment 4 *RC: Citing* Model 4B.2	
Effects	*Beta*	*95%-CI*	*Beta*	*95%-CI*	*Beta*	*95%-CI*
ACME H-Inc.[f]	0.03***	[0.01; 0.05]	0.03***	[0.01; 0.05]	0.03***	[0.01; 0.05]
H-Inc → FOR	-0.11***	[-0.14; -0.09]	-0.11***	[-0.14; -0.09]	-0.11***	[-0.14; -0.09]
FOR→ RC	-0.27***	[-0.31; -0.22]	-0.27***	[-0.31; -0.22]	-0.27***	[-0.31; -0.22]
ADE H-Inc.	0.26***	[0.18; 0.34]	0.26***	[0.18; 0.34]	0.26***	[0.18; 0.34]
Total Effect	0.29***	[0.21; 0.37]	0.29***	[0.21; 0.37]	0.29***	[0.21; 0.37]
Proportion Mediated	0.10***	[0.05; 0.18]	0.10***	[0.05; 0.18]	0.10***	[0.05; 0.18]

Note. * $p \leq 0.05$ ** $p \leq 0.01$ and *** $p \leq 0.001$; [a] ACME = Average Causal Mediation Effect; [b] N-Sig. = Statistically Non-Significant Abstract; [c] FOR = Feeling of Rightness; [d] RC = Response Change; [e] ADE = Average Direct Effect; [f] H-Inc.= Hypothesis-Inconsistent Abstract.

H4A.2: Hypothesis-Consistency and Response Changes in Reception In the mediation model for hypothesis H4A.2, Model 4A.2, the presence of *hypothesis-inconsistent abstracts* showed a < significant | not significant> direct path to C-LoR (Beta = XX.XX, p=.XXX). Furthermore, the treatment variable exhibited a < significant | not significant> mediated path to C-LoR (Beta = XX.XX, p=.XXX). < As hypothesized | Contrary to our hypothesis >, the effect of *hypothesis-inconsistent abstracts* on FOR was < significant | not significant> and < negative | positive> (Beta = XX.XX, p=.XXX), and the effect of FOR on C-LoR was < significant | not significant> and < negative | positive> (Beta = XX.XX, p=.XXX).

H4B.2: Hypothesis-Consistency and Response Changes in Citation In the mediation model for hypothesis H4B.2, Model 4B.2, the treatment variable *hypothesis-inconsistent abstract* showed a < significant | not significant> direct path to C-LoC (Beta = XX.XX, p=.XXX). Furthermore, the treatment variable exhibited a < significant | not significant> mediated path to C-LoC (Beta = XX.XX, p=.XXX). < As hypothesized | Contrary to our hypothesis >, the effect of *hypothesis-inconsistent abstracts* on FOR was < significant | not significant> and < negative | positive> (Beta = XX.XX, p=.XXX), and the effect of FOR on C-LoC was < significant | not significant> and < negative | positive> (Beta = XX.XX, p=.XXX).

Exploratory Analyses

[Anything that makes sense to do after obtaining the data, and what has been described in the method section. Mainly: Exploratory models with the covariates familiarity with publication bias, pressure to publish, postdoc/professor; conscientiousness, and position. Only statistically significant exploratory results are reported. For example:]

In our exploratory analyses (summarized in Table 1), we investigated the influence of five variable categories on LoSRC, FOR, and C-LoSRC.

Familiarity with Publication Bias. MLM: A statistically < significant | non-significant > interaction was observed between familiarity with publication bias and the treatment. MLMM: The influence of familiarity with publication bias on FOR showed a statistically < significant | non-significant > effect.

Pressure to Publish. MLM: The pressure to publish presented a statistically < significant | non-significant > influence on LoSRC. MLMM: Additionally, a statistically < significant | non-significant > interaction between pressure to publish and treatment was found to affect FOR

Professional Status. MLM: Results indicated that both post-doctoral and professor categories < significantly | not significantly > influenced LoSRC when compared to the reference pre-doctoral category. MLMM: A statistically < significant | non-significant > effect of professional status was observed on FOR.

Conscientiousness (from BFI-2-S subscale): MLMM: Researchers' conscientiousness < significantly | not significantly > affected the likelihood to alter their initial abstract ratings. This indicates that those scoring higher on conscientiousness felt a stronger obligation to modify their responses when prompted to reconsider.

Stimulus Position. MLM & MLMM: Position, indexing the sequence in which abstracts were presented, revealed a statistically < significant | non-significant > effect on both LoSRC and FOR, suggesting a fatigue effect over time in evaluating research abstracts

Country: MLM & MLMM: Researchers from XX showed significantly lower/higher LoSRC rating compared to researchers from Germany. There was a/no significant difference in LoSRC ratings between researchers from recruitment track 1 (Germany) and track 2 (other countries).

Discussion

The results of the present study shed light on the decision-making processes that contribute to selective non-publication and non-reception. By fitting several within-subject models with significance and hypothesis-consistency as experimental treatment variables, we were able to isolate the effects of these factors using a nested linear mixed model approach. Our study also utilized a two-response decision-making procedure based on DPT to investigate the role of fast, automatic responses in stage 1 versus analytical decision-making processes in stage 2 and the relevance of FOR within a mediation framework.

Stage 1

Publishability

Previous reviews and meta-analyses illustrated that significant findings as well as hypothesis-consistent studies are more likely to be submitted for publication than non-significant findings

(Fanelli, 2011; Scheel et al., 2021), but evidence from experimental studies on publication bias is mixed (Atkinson et al., 1982; Augusteijn et al., 2023; Elson et al., 2020; Epstein, 1990; Mahoney, 1977). However, only few studies have experimentally investigated the influence of non-significant and hypothesis-inconsistent results on the evaluations of researchers on the quality and publishability of studies (Atkinson et al., 1982; Augusteijn et al., 2023; Elson et al., 2020; Epstein, 1990; Mahoney, 1977). Since studies in this field exclusively deployed between-subject findings, this mixed finding might result from the influence of single abstract-characteristics presented in these studies.

H1.1: First, we investigated whether the intuitive, fast choice for publication depended on statistical significance (H 1.1). *Positive Outcome*: [As predicted, statistically non-significant findings were less likely chosen to be submitted for publication than significant findings.] OR *Negative Outcome*: [Inconsistently with our initial hypothesis, statistically non-significant findings were not less likely chosen to be submitted for publication than significant findings]

H2.1 Second, we investigated the same for hypothesis-consistency (H2.1). *Positive Outcome:* [As predicted, hypothesis-inconsistent findings were less likely chosen to be submitted for publication than hypothesis-consistent findings. OR *Negative Outcome*: [Inconsistently with our initial hypothesis, hypothesis-inconsistent findings were not less likely chosen to be submitted for publication than hypothesis-consistent findings .]

Reception and Citation

Results regarding citation bias have been mixed and other reception-related factors, such as reading decision have not yet been investigated. While Callaham et al. (2002) found no bias in citation frequency, other studies found higher citation rates for articles with significant or hypothesis-consistent results (Jannot et al. 2013; Duyx et al. 2017). However, no experimental study has investigated reception or citation of non-significant or hypothesis-inconsistent findings.

Thus, we investigated whether immediate preferences for reading and citing also depend upon significance or hypothesis-consistency (H3A.1 – H4B.1).

H3A.1 *Positive Outcome*: [As predicted, statistically non-significant findings were less likely chosen to be read than significant findings.] OR *Negative Outcome*: [Inconsistently with our initial hypotheses our results indicate that there is no effect of statistical significance on the decision to read a full paper.]

H3B.1 *Positive Outcome*: [As predicted, hypothesis-inconsistent findings were less likely chosen to be read than hypothesis-consistent findings] OR *Negative Outcome*: [Inconsistently with our initial hypotheses our results indicate that there is no effect of hypothesis-consistency on the decision of a researcher to read an article]

H4A.1 *Positive Outcome*: [As predicted, non-significant findings were less likely chosen to be cited than significant findings. OR *Negative Outcome*: [Inconsistently with our initial hypotheses our results indicate that there is no effect of significance in the decision to cite an article.]

H4B.1 *Positive Outcome:* [As predicted, hypothesis-consistent findings were more likely chosen to be cited than hypothesis-inconsistent findings] OR *Negative Outcome*: [Inconsistently with our initial hypotheses our results indicate that there is no effect of hypothesis-consistency in the decision of a researcher to read an article.]

Stage 2 and FOR

Studies on publication bias have not investigated the dynamics in decision-making process related to submitting, citing or reading a paper. By applying a MLMM in our experimental design, we aimed to explore the relationship between fast, automatic, and more deliberate, analytic judgement processes as well as FOR-judgements as suggested by DPT.

Negative Outcome Stage 1 → *Positive Outcome Stage 2* [Even though we did not observe a significant effect for significance/hypothesis-consistency and submitting/reading/citing for stage 1, positive response changes in submitting/reading/citing were directly associated with non-significance/hypothesis-inconsistency. Furthermore, this association was also mediated by FOR judgements: Researchers were more likely to exhibit a positive response change regarding an abstract (a more positive considered evaluation than the initial intuitive evaluation in submitting to publish, citing, or reading) if the FOR of the initial evaluation was low, which itself was more likely if the experimental treatment was statistically significant or inconsistent with the hypothesis.]

Positive Outcome Stage 1 → *Negative Outcome Stage 2* [We observed a significant effect for significance/hypothesis-consistency and submitting/reading/citing for stage 1, but positive response changes in submitting/reading/citing were not directly associated with non-significance/hypothesis-inconsistency. Furthermore, this association was also not mediated by FOR judgements: Researchers were not more likely to exhibit a positive response change regarding an abstract (a more positive considered evaluation than the initial intuitive evaluation in submitting to publish, citing, or reading) if the FOR of the initial evaluation was low.]

Negative Outcome Stage 1 → *Negative Outcome Stage 2* [We did not find a significant effect for non-significance/hypothesis-inconsistency and submitting/reading/citing for stage 1, and positive response changes in submitting/reading/citing were not directly associated with non-significance/hypothesis-inconsistency as well. Furthermore, this association was also not mediated by FOR judgements: Researchers were not more likely to exhibit a positive response change regarding an abstract (a more positive considered evaluation than the initial intuitive evaluation in submitting to publish, citing, or reading) if the FOR of the initial evaluation was low.]]

Positive Outcome Stage 1 → *Positive Outcome Stage 2* [We observed a significant effect for significance/hypothesis-consistency and submitting/reading/citing for stage 1, but positive response changes in submitting/reading/citing were significantly associated with non-significance/hypothesis-consistency. Furthermore, this association was also mediated by FOR judgements: Researchers were more likely to exhibit a positive response change regarding an abstract (a more positive considered evaluation than the initial intuitive evaluation in submitting to publish, citing, or reading) if the FOR of the initial evaluation was low, which itself was more likely if the experimental treatment was statistically significant or inconsistent with the hypothesis.]

Exploratory Analyses

The exploratory analyses sought to illuminate possible confounding factors influencing the relationship between intuitive responses (LoSRC), subsequent changes in researchers' responses (C-LoSRC) and FOR.

Researchers who are more familiar with the detrimental effects of publication bias on research quality appear to evaluate non-significant results more positively/negatively/no effect regarding

submission to publication/ reading /citing. Also, response changes and FOR are/ are not influenced by familiarity with publication bias.

Similarly, the pressure to publish influenced all/none/some intuitive responses. Those who felt a heightened pressure were possibly leaning towards producing favorable outcomes, which underscores the intricate challenges of the current "publish or perish" culture in academia. Addressing this could have profound implications for the quality and reliability of published research.

Furthermore, academic professional status emerged as a potentially influential factor in determining FOR. This underscores the different professional dynamics and possibly inherent biases that researchers at various career stages might possess. Notably, senior/junior researchers like professors/predoctoral researchers, compared to their junior/senior counterparts, displayed heightened patterns/ no effect in FOR and in the likelihood to change responses.

Additionally, we noticed that abstracts presented later received, on average, more negative /positive/no effect evaluations in terms of both LoSRC and C-LoSRC. These differences could potentially indicate a fatigue effect, thus representing a confounding factor that influences our two-response framework:

Moreover, the personality trait of conscientiousness, as measured by the BFI-2-S subscale, was/was not indicative of a certain group of researchers' tendency to alter their initial decisions. This highlights a general limitation of DPT designs, as re-evaluations are pivotal in DPT experiments and appear to be influenced not only by analytical thinking but also by personality traits.

Lastly, researchers from XX exhibited a/no difference in their LoSRC ratings when contrasted with researchers from Germany. Regarding the recruitment strategies, we did/did not find differences between researchers from our two distinct recruitment tracks: manual collection of email addresses from clinical psychology departments and text scraping from PubMed clinical psychology articles.

Conclusion of Experiments

Taken together the data suggest that statistically nonsignificant and hypothesis-inconsistent results are <not significantly/ significantly> less likely to be chosen for submission to publication in intuitive responses. Furthermore, the experiments showed that statistically nonsignificant and hypothesis-inconsistent results are <not significantly/ significantly> more likely to be rejected in intuitive responses on reading the full paper. Moreover, the experiments revealed that intuitive positive decisions to cite a statistically nonsignificant and hypothesis-inconsistent abstract are <not significantly/ significantly> less likely compared to statistically significant and hypothesis-consistent abstracts.

However, drawing on a two-response paradigm we were able to show that considered positive responses indicative of deliberate and analytic Type 2 processing regarding decisions to submit to publish, cite and read are <not significantly/ significantly> more likely for statistically nonsignificant and hypothesis-inconsistent abstracts. Importantly, this association is influenced not only directly by these abstract characteristics but is also <not significantly/ significantly> mediated by FOR judgments. Additional factors, such as familiarity with the implications of publication bias, the prevalent pressure to publish in academia, professional status, and even fatigue, <not significantly/ significantly> modulate these decisions.

Scientific Implications

The present study makes significant contributions to the current literature on publication bias by addressing various gaps. Firstly, it fills the gap by using experimental within-subject design to examine publication bias in psychology, which has not been investigated in previous studies. Secondly, our study also addresses the need for experimental research on non-reception, which has been neglected in the literature. Finally, we investigated the dynamics of decision-making processes in non-publication and non-reception, and our findings provide insights into how such publications decisions are [not] biased by automatic and fast thinking styles, while deliberate and analytic thinking styles are [not] associated with publication, reception [and/or] citation bias.

Limitations

Replicating previous observations of publication bias (starting with observations as early as Sterling, 1959) within an experimental setting could provide insight on the degrees of biases and differential effects of biases regarding submitting for publication, reading and citing articles, while differentiating hypothesis consistency and statistical significance. Even though our design has some shortcomings that are discussed in the following, it offers the ability to test experimentally and control for confounding variables:

First, we investigated the effect of statistical significance and hypothesis-consistency. However, recent experimental studies on publication bias also investigated the effects for sample size and statistical reporting error (Augusteijn et al., 2023), and originality (Elson et al., 2020) and its interaction with statistical significance. Unfortunately, due to power restrictions we were not able to include further experimental treatment variables into our design.

Second, our abstracts were designed in a textbook fashion. In general, hypotheses are often not specified in the abstract of a publication, possibly indicating "HARKing" (hypothesis after results are known) or approaches that are not preregistered. For the present study, it was necessary for abstracts to be more simplified because sometimes several results - some significant, some non-significant - are presented in abstracts.

Third, the presence of social desirability in response patterns could compromise the study's validity. Future research should add a social desirability questionnaire or qualitative questions on how decisions were made when evaluating abstracts.

Fourth, p-values and statistical significance are often challenged in methodological discussions in psychology, instead, effect sizes, confidence intervals, or Bayesian statistics are possible alternatives for reporting results. Furthermore, clinical relevance is not directly reflected by statistical outcomes (Cuijpers et al., 2014), so biases and effects of clinical studies might be less clear than in other fields (e.g., in psychophysics with well-powered studies of multiple trials and well-defined outcome criteria). Hence, it is important to use experimental settings to clarify decision making in clinical research.

Fifth, while presenting abstracts may be a suitable format for decisions on reception (reading / citing), it is questionable in terms of making decisions about one's own publications, as researchers usually have much more information at hand when deciding whether to submit their own research for publication. Reading and citing an abstract also depends on contextual factors and on researchers' own research interests, which will likely influence the reception of the abstracts of this study. This should be taken into account in future studies.

Sixth, our study aimed at answering the question whether publication bias happens in the decision-making processes of researchers. Future studies could investigate if editors or journals are more biased towards significant or hypothesis-consistent findings.

Future Research

By analyzing and reporting on research practice and directly involving the research community in our experiments, we attempt to achieve a sustainable reduction of publication bias and a stronger reception of null findings in order to increase research quality in the field of clinical psychology. The notion of publication bias often refers specifically to the lack of publication of non-significant results and is therefore particularly applicable in quantitative and experimental disciplines such as the life sciences, psychology, and economics. However, the overarching tension between research quality and publication values (publishability and reception incentives) is also likely to exist in other disciplines that, due to their primarily qualitative or hermeneutic methods, do not have a focus on (non-)significant results (especially humanities and parts of the social sciences). Future studies should therefore examine further disciplines.

While the current study focused on significance and hypothesis-consistency, findings that are less striking or relate to issues that are considered to be of low recency in scientific discourse and not part of the zeitgeist, may also be more difficult to publish (Olson et al., 2002) or less likely to be cited (Jannot et al., 2013; Fanelli, 2013). Future studies should therefore also include these biases.

Author Contributions:

- Conceptualization: K. Eichel, L.J. Schiekiera, F. Hesselmann, J. Sachse, S. P. Müller, H. Niemeyer
- Methodology: L.J. Schiekiera, K. Eichel, F. Hesselmann, J. Sachse, S. P. Müller, H. Niemeyer
- Software: L.J. Schiekiera
- Validation: K. Eichel, L.J. Schiekiera
- Formal Analysis: K. Eichel, L.J. Schiekiera
- Investigation: L.J. Schiekiera, K. Eichel, H. Niemeyer
- Resources: H. Niemeyer
- Data Curation: K. Eichel, L.J. Schiekiera
- Writing – Original Draft: K. Eichel, H. Niemeyer
- Writing – Review & Editing: L.J. Schiekiera, K. Eichel, F. Hesselmann, J. Sachse, S. P. Müller, H. Niemeyer
- Visualization: L.J. Schiekiera, K. Eichel,
- Supervision: F. Hesselmann, K. Eichel, H. Niemeyer
- Project Administration: F. Hesselmann, K. Eichel, H. Niemeyer
- Funding Acquisition: F. Hesselmann, H. Niemeyer

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

Acknowledgments: Thanks to the Berlin University Alliance for funding this project (see Appendix F for a description of the overall project). Thanks to the Clinical-Psychological Interventions team at Freie Universität Berlin for participating in our pilot study and giving helpful feedback along the way. Thanks for the feedback to the students of the summer semester 2022 in "Good Science, Bad Science". Thanks to Olmo van den Akker for providing helpful comments. Thank you [in advance] for all the participants of the study. Thanks as well to Manuel Heinrich for important suggestions concerning the statistical modelling. Thanks to Sarah Mannion for language editing.

Funding: The project was funded by the Berlin University Alliance (312_OpenCall_3).

Appendix: See word document.

References

Ackerman, R., & Thompson, V. (2017). Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. Trends in Cognitive Sciences, 21(8), 607-617. doi:10.1016/j.tics.2017.05.004

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship?. Journal of Counseling Psychology, 29(2), 189.

Augusteijn, H. E., Wicherts, J., Sijtsma, K., & van Assen, M. A. (2023). Quality assessment of scientific manuscripts in peer review and education.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Borenstein, M. (2019). *Common mistakes in meta-analysis and how to avoid them.* Biostat. Inc., New Jersey.

Callaham, M. L., Wears, R. L., & Weber, E. (2002). Journal Prestige, Publication Bias, and Other Characteristics Associated With Citation of Published Studies in Peer-Reviewed Journals. *JAMA*, *287*(21), 2847–2850.

Carter, E.C., Schönbrodt, F.D., Gervais, W.M., & Hilgard J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. Advances in Methods and Practices in Psychological Science, 2(2), 115–44. https://doi.org/10.1177/2515245919847196.

Chopra, F., Haaland, I., Roth, C., & Stegmann, A. (2022). The null result penalty. CESifo Working Paper No. 9776.

Cuijpers, P., Turner, E.H., Koole, S.L., van Dijke, A., & Smit, F. (2014). What is the threshold for a clinically relevant effect? The case of major depressive disorders. *Depress Anxiety, 31*, 374-378. r

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods, 47(1),* 1-12. 10.3758/s13428-014- 0458-y.

Demian, A. F. (2019). Assess Three-Level-Cross-Classified Model in LME4 Package.

Dirnagl, U., & Lauritzen, M. (2010). Fighting publication bias: Introducing the Negative Results section. Journal of Cerebral Blood Flow and Metabolism, 30(7), 1263–1264. https://doi.org/10.1038/jcbfm.2010.51

Doran, C. M., & Kinchin, I. (2017). A review of the economic impact of mental illness. Australian Health Review, 43(1), 43-48.

Duyx, B., Urlings, M. J. E., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. (2017). Scientific citations favor positive results: A systematic review and meta-analysis. Journal of Clinical Epidemiology, 88, 92–101. https://doi.org/10.1016/j.jclinepi.2017.06.002

Elson, M., Huff, M., & Utz, S. (2020). Metascience on peer review: Testing the effects of a study's originality and statistical significance in a field experiment. Advances in Methods and Practices in Psychological Science, 3(1), 53-65.

Epstein, W. M. (1990). Confirmational response bias among social work journals. Science, Technology, & Human Values, 15(1), 9-38.

Evans, J. S. B. (1996). Deciding before you think: Relevance and reasoning in the selection task. British Journal of Psychology, 87(2), 223-240.

Evans, J. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. Psychonomic Bulletin & Review, 13(3), 378-395. doi:10.3758/BF03193858

Evans, J., & Stanovich, K. (2013). Dual-process theories of higher cognition: Advancing the debate. Perspectives on psychological science, 8(3), 223-241. doi:10.1177/1745691612460685

Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. Scientometrics, 90(3), 891-904.

Fanelli, D. (2013). Positive results receive more citations, but only in some disciplines. Scientometrics, 94(2), 701–709. https://doi.org/10.1007/s11192-012-0757-y

Fleming, N. (2019). Top US institutes still aren't reporting clinical-trial results on time. Nature, Epub ahead of print. 10.1038/d41586-019-00994-1

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. Science, 345(6203), 1502-1505. 10.1126/science.1255484

Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in Psychology Experiments: Evidence from a Study Registry. Social Psychological and Personality Science, 7(1), 8–12. https://doi.org/10.1177/1948550615598377

Gabbard, G. O., Lazar, S. G., Hornberger, J., & Spiegel, D. (1997). The economic impact of psychotherapy: A review. american Journal of psychiatry, 154(2), 147-155.

Garcia-Martinez, A. T., Guerrero-Bote, V. P., & Moya-Anegon, F. de. (2012). World scientific production in Psychology. Universitas Psychologica, 11(3), 699-717.
Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*(1), 1-20.
Haustein, S. Grand challenges in altmetrics: heterogeneity, data quality and dependencies. Scientometrics 108, 413–423 (2016). https://doi.org/10.1007/s11192-016-1910-9
Hox, J. J. (2002). Multilevel Analysis: Techniques and Applications. Mahwah, NJ: Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9781410604118
Ioannidis, J. P. (2005). Why most published research findings are false. PLoS medicine, 2(8), e124.
Jannot, A.-S., Agoritsas, T., Gayet-Ageron, A., & Perneger, T. V. (2013). Citation bias favoring statistically significant studies was present in medical research. Journal of Clinical Epidemiology, 66(3), 296–301. https://doi.org/10.1016/j.jclinepi.2012.09.015
Kahneman, D. (2003). A Perspective on Judgment and Choice: Mapping Bounded Rationality. American Psychologist, 58(9), 697-720.
Knapp, M., & Wong, G. (2020). Economics and mental health: the current scenario. World Psychiatry, 19(1), 3-14.
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. Journal of statistical software, 82, 1-26.
Lee, K.P., Boyd, E.A., Holroyd-Leduc, J.M., Bacchetti, P., & Bero, L.A. (2006). Predictors of publication: characteristics of submitted manuscripts associated with acceptance at major biomedical journals. *The Medical Journal of Australia, 184(12),* 621-6. doi: 10.5694/j.1326-5377.2006.tb00418.x.
Lorah, J. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. Large-Scale Assessments in Education, 6(1), 1-11.
Lortie, C. J., Aarssen, L. W., Budden, A. E., Koricheva, J. K., Leimu, R., & Tregenza, T. (2007). Publication bias and merit in ecology. Oikos, 116(7), 1247–1253. https://doi.org/10.1111/j.0030-1299.2007.15686.x
Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. Cognitive therapy and research, 1, 161-175.
McDaid, D., Park, A. L., & Wahlbeck, K. (2019). The economic case for the prevention of mental illness. Annual review of public health, 40, 373-389.
Niemeyer, H., Musch, J., & Pietrowsky, R. (2012). Publication bias in meta-analyses of the efficacy of psychotherapeutic interventions for schizophrenia. *Schizophr Res, 138*(2-3), 103-112. doi:10.1016/j.schres.2012.03.023
Niemeyer, H., van Aert, R.C.M., Schmidt, S., Uelsmann, D., Knaevelsrud, C. & Schulte-Herbrueggen, O. (2020). Publication Bias in Meta-Analyses of Posttraumatic Stress Disorder Interventions. Meta-Psychology, 4. https://doi.org/10.15626/MP.2018.884
O'Gorman, J., Shum, D. H., Halford, W. K., & Ogilvie, J. (2012). World trends in psychological research output and impact. International Perspectives in Psychology, 1(4), 268-283.
Okike, K., Kocher, M.S., Mehlman, C.T., Heckma,n J.D., & Bhandari, M. (2008). Publication bias in orthopaedic research: an analysis of scientific factors associated with publication in the Journal of Bone and Joint Surgery (American Volume). The *Journal of Bone and Joint Surgery,* 90(3), 595-601. doi: 10.2106/JBJS.G.00279.
Olson, C. M., Rennie, D., Cook, D., Dickersin, K., Flanagin, A., Hogan, J. W., Zhu, Q., Reiling, J., & Pace, B. (2002). Publication Bias in Editorial Decision Making. Jama-Journal of the American Medical Association, 287(21), 2825–2828. https://doi.org/10.1001/jama.287.21.2825
Open Science Collaboration, Nosek, Brian A., Aarts, Alexander A., Anderson, Christopher J., Anderson, Joanna E. and Kappes, Heather Barry,... (2015) Estimating the reproducibility of psychological science. *Science, 349*(6251). 10.1126/science.aac4716.
Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. Cognitive psychology, 80, 34-72.
Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2011). Altmetrics: A manifesto, 26 October 2010. http://altmetrics.org/manifesto
R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http:// www.R-project.org
Revilla, M., & Höhne, J. K. (2020). How long do respondents think online surveys should be? New evidence from two online panels in Germany. *International Journal of Market Research, 62*(5), 538–545. doi:10.1177/1470785320943049
Scheel, A. M., Schijen, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the standard Psychology literature with Registered Reports. Advances in Methods and Practices in Psychological Science, 4(2), 25152459211007467.

Shapiro, D. (2002). Renewing the scientist-practitioner model. The Psychologist, 15, 232–234.
Shaw, D. M., & Penders, B. (2018). Gatekeepers of Reward: A Pilot Study on the Ethics of Editing and Competing Evaluations of Value. Journal of Academic Ethics, 16(3), 211–223. https://doi.org/10/gg5m37
Smulders, Y. M. (2013). A Two-Step Manuscript Submission Process Can Reduce Publication Bias. Journal of Clinical Epidemiology, 66(9), 946–947. https://doi.org/10.1016/j.jclinepi.2013.03.023
Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., . . . Harvey, I. (2010). Dissemination and publication of research findings: an updated review of related biases. *Health Technol Assess, 14*(8), iii, ix-xi, 1-193. doi:10.3310/hta14080
Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. Journal of Research in Personality, 68, 69-81. https://doi.org/10.1016/j.jrp.2017.02.004
Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance--or vice versa. *Journal of the American Statistical Association*, 30-34.
Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's Replication Crisis and Clinical Psychological Science. *Annual review of clinical psychology, 15*, 579–604.
Tenopir, C., & King, D. W. (2002). Reading behaviour and electronic journals. Learned Publishing, 15(4), 259-265.
Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. Thinking & Reasoning, 20(2), 215-244.
Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. Cogn Psychol, 63(3), 107-140. doi:10.1016/j.cogpsych.2011.06.001
Timmer, A., Hilsden, R.J., Cole, J., Hailey, D., & Sutherland, L.R. (2002). Publication bias in gastroenterological research - a retrosp,ective cohort study based on abstracts submitted to a scientific meeting. *BMC Medical Research Methodology, 2(7).* doi: 10.1186/1471-2288-2-7.
Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis.
Wang, S., & Thompson, V. (2019). Fluency and feeling of rightness: The effect of anchoring and models. Psychological Topics, 28(1), 37-72.
Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. Journal of Experimental Psychology: General, 143(5), 2020–2045. https://doi.org/10.1037/xge0000014
Wieschowski, S., Riedel, N., Wollmann, K., Kahrass, H., Müller-Ohlraun, S., Schürmann, C., ... & Strech, D. (2019). Result dissemination from clinical trials conducted at German university medical centers was delayed and incomplete. *Journal of Clinical Epidemiology, 115,* 37-45. 10.1016/j.jclinepi.2019.06.002.