Preregistration specificity & adherence: A review of preregistered gambling studies & cross-disciplinary comparison

Robert M. Heirene[1], Debi A. LaPlante[2], Eric R. Louderback[2], Brittany Keen[1], Marjan Bakker[3], Anastasia Serafimovska[1], and & Sally M. Gainsbury[1]

[1] School of Psychology
Brain & Mind Centre
University of Sydney
Australia
[2] Division on Addiction
Harvard Medical School
USA
[3] Methods and Statistics Department
Tilburg University
Netherlands

Submitted to Meta-Psychology. Participate in open peer review by commenting through hypothes.is directly on this preprint. The full editorial process of all articles under review at Meta-Psychology can be found following this link: https://tinyurl.com/mp-submissions

You will find this preprint by searching for the first author's name.

All materials associated with this study can be accessed on our Open Science Framework page, including the preregistration, scoring protocols, raw datasets, analysis scripts, and the RMarkdown script used to create this manuscript: https://osf.io/n8rw3/

The authors made the following contributions. Robert M. Heirene: Conceptualization (equal), Project administration (lead), Methodology (equal), Investigation (equal), Formal analysis (lead), Visualization, Writing - Original Draft Preparation, Writing - Review & Editing (equal); Debi A. LaPlante: Conceptualization (equal), Funding acquisition, Methodology (equal), Writing - Review & Editing (equal), Supervision (equal); Eric R. Louderback: Conceptualization (equal), Methodology (equal), Formal analysis (support), Writing - Review & Editing (equal); Brittany Keen: Project administration (support), Data curation (lead), Investigation (equal), Writing - Review & Editing; Marjan Bakker: Methodology (equal), Data curation (support), Writing - Review & Editing (equal); Anastasia Serafimovska: Project administration (support), Data curation (support), Investigation (equal), Writing - Review & Editing (equal); Sally M. Gainsbury: Conceptualization (equal), Methodology (equal), Writing - Review & Editing (equal), Supervision (equal).

Correspondence concerning this article should be addressed to Robert M. Heirene, Gambling Treatment & Research Clinic, Brain & Mind Centre, University of Sydney, 94 Mallett St, Camperdown, Sydney, NSW, Australia, 2050. E-mail: robert.heirene@sydney.edu.au; robheirene@gmail.com

Abstract

Study preregistration is one of several "open science" practices (e.g., open data, preprints) that researchers use to improve the transparency and rigour of their research. As more researchers adopt preregistration as a regular research practice, examining the nature and content of preregistrations can help identify strengths and weaknesses of current practices. The value of preregistration, in part, relates to the specificity of the study plan and the extent to which investigators adhere to this plan. We identified 53 preregistrations from the gambling studies field meeting our predefined eligibility criteria and scored their level of specificity using a 23-item protocol developed to measure the extent to which a clear and exhaustive preregistration plan restricts various researcher degrees of freedom (RDoF; i.e., the many methodological choices available to researchers when collecting and analysing data, and when reporting their findings). We also scored studies on a 32-item protocol that measured adherence to the preregistered plan in the study manuscript. We found that gambling preregistrations had low specificity levels on most RDoF. However, a comparison with a sample of cross-disciplinary preregistrations (N = 52; Bakker et al., 2020) indicated that gambling preregistrations scored higher on 12 (of 29) items. Thirteen (65%) of the 20 associated published articles or preprints deviated from the protocol without declaring as much (the mean number of undeclared deviations per article was 2.25, $SD = 2.34$). Overall, while we found improvements in specificity and adherence over time (2017-2020), our findings suggest the purported benefits of preregistration—including increasing transparency and reducing RDoF—are not fully achieved by current practices. Using our findings, we provide 10 practical recommendations that can be used to support and refine preregistration practices.

Note: *This is a pre-printed manuscript that has not yet undergone peer-review.*

Introduction

A preregistration is a time-stamped, immutable document posted on an online repository that outlines the details of a proposed research study, including the hypotheses, methods, outcomes of interest, and data analysis plan. Historically, preregistration has been used primarily for randomised control trials (RCTs) (Dickersin & Rennie, 2003) and later for systematic reviews and meta-analyses (Stewart et al., 2012). More recently, researchers performing other forms of quantitative and qualitative studies (Haven & Van Grootel, 2019) have begun to adopt this practice, and the number of researchers preregistering these types of studies is increasing year-on-year (Kuperschmidt, 2018), with 17,000 new preregistrations posted on the online repository Open Science Framework (OSF) in 2020 alone (Centre for Open Science, 2020). This trend has been largely prompted by concerns regarding the replicability and reproducibility of the extant literature (Allen & Mehler, 2019; Simmons et al., 2011), and preregistration is one of several practices (e.g., open data, preprints) that researchers are using to improve the transparency and rigour of their research as a part of the *open science movement*.

Proponents of study preregistration have advanced three overlapping and mutually compatible perspectives regarding its value. First, preregistration increases transparency (Nosek et al., 2018). Transparency in the research process has multiple benefits, such as improving the ability to detect questionable research practices (QRPs; e.g., *hypothesizing after the results are known* ["HARKing"] and selective outcome reporting; Kerr (1998), Norris et al. (2012)) and publication bias (Munafò et al., 2017), and enabling the differentiation of planned, *a priori* analyses from unplanned, *post hoc* analyses (Nosek et al., 2019).

Second, preregistration assists with reducing Researcher Degrees of Freedom (RDoF)—that is, the many methodological choices available to researchers when collecting, analysing, and reporting their findings (Bakker et al., 2020; Wicherts et al., 2016). Reducing RDoF can be important as the freedom to make data-contingent decisions during the research process (e.g., when deciding which inference criteria to use or how to deal with outliers) can inflate the risk of finding false-positive results or Type-I errors (Wicherts et al., 2016), which when done strategically is known as *p-hacking* or *asterisk hunting* (Head et al., 2015).

Third, Lakens (2019; pg. 1) argues that preregistration is valuable as it allows for "*others to transparently evaluate the capacity of a test to falsify a prediction.*" The degree to which a test is capable of falsifying a prediction is termed its "severity" and, as Lakens discusses, more severe tests are more impressive and indicative of a solid theoretical underpinning.[1] Several QRPs can reduce the severity of tests by reducing the likelihood of researchers being able to falsify their hypothesis, including optional stopping (i.e., continuously checking & analysing data during the collection phase & only stopping when a statistically significant

[1]For example, if a researcher studying behavioural addictions predicts that a sample of problem gamblers will differ from non-problem gamblers on one personality index of a multidimensional measure, without specifying which specific index will differ or the direction or magnitude of the effect, then the test of this claim will lack severity as it is highly unlikely that the difference between the two samples will be exactly zero on all indices. If, by contrast, the researcher predicts that the samples will only differ in extroversion levels, with the problem gambling sample displaying a mean score of 2-4 points higher than non-problem gamblers, the test of this claim will be high in severity as it is highly capable of being falsified.

result is observed) and HARKing. Thus, readers can better evaluate the severity of tests reported in preregistered compared to non-preregistered studies as these QRPs can be more easily detected in the former (Lakens, 2019).

Available research is limited but supports the value of study preregistration. Preregistering RCTs has been shown to reduce the likelihood of finding statistical false-positives (Kaplan & Irvin, 2015) and help detect outcome switching (Chen et al., 2019; Vassar et al., 2020). Preliminary evidence shows that the effect sizes reported in preregistered studies in the psychology literature are considerably smaller than non-registered studies, suggesting the latter contain effects that are inflated by QRPs and publication bias (Schäfer & Schwarz, 2019). Yet, the value of preregistration is limited by at least two factors. First is the degree to which preregistrations specifically describe all aspects of the planned study. If key study details like hypotheses, primary outcomes, sampling procedures, and analysis plans are not clearly and comprehensively specified, then the many benefits of preregistration listed above fail to materialize. Second is the extent to which researchers actually follow (i.e., adhere to) their pre-specified plans and declare any deviations (Nature Human Behaviour, 2020). The benefits of the practice are again lost if post-preregistration, for example, a researcher changes their criterion for outlier removal or the cut-off score used to divide groups and fails to declare such deviations.

To date, three studies have evaluated modern study preregistration practices according to the specificity of the research plans and the degree to which the researchers adhered to them. Bakker et al. (2020) examined the specificity of preregistrations registered on OSF (osf.io) during 2016 that used either structured or unstructured templates. The authors adapted the list of RDoFs developed by Wicherts et al. (2016) to create a scoring protocol that assessed the extent to which the preregistration restricted each RDoF (e.g., "Deciding on how to deal with outliers in an *ad hoc* manner") by being "*specific*" (i.e., all phases of research process are described), "*precise*" (i.e., descriptions of the research plan can only be interpreted in one way), and "*exhaustive*" (i.e., explicit acknowledgment that the plan will not be deviated from). We use the term specificity here as shorthand for these three principles. Bakker and colleagues found that specificity was higher in the sample using a structured template but was relatively low for both samples, particularly regarding the selection of measured variables and covariates.

Claesen et al. (2019) investigated 16 articles published in *Psychological Science* between 2015 and 2016 with 27 corresponding preregistrations (some articles contained multiple, separately preregistered studies). They assessed whether the authors of these publications adhered to their preregistrations in eight areas (e.g., exclusion criteria, statistical model), finding that 26 articles (96%) included at least one deviation that was not declared. Only one study disclosed all deviations, and all studies deviated from their preregistration in one of the eight areas. Ofosu and Posner (2019) evaluated 195 preregistrations from the economics and political science fields registered between 2011 and 2016. Only 49.7% of the sample were judged to contain sufficiently detailed descriptions of the four key areas they deemed necessary for a complete preregistration (i.e., hypotheses, primary dependent variables, treatments or independent variables, & the statistical model[s] to be tested). Of the 95 preregistrations with a corresponding published article, more than a third failed to include at least one preregistered hypothesis and 18% presented tests of unregistered

hypotheses.

Collectively, the findings from Bakker et al. (2020), Claesen et al. (2019), and Ofosu and Posner (2019) highlight a need to continue to examine preregistration practices and how they can be improved in order to maximise the potential scientific benefits of preregistration. In the present study, we aimed to better understand the strengths and weaknesses of more recent preregistration practices (i.e., for 2017 onwards). We did this by simultaneously determining the degree to which researchers studying gambling sufficiently specify all aspects of their studies in preregistrations and the extent to which they adhere to their pre-specified plans. Preregistered studies from the gambling field were selected as the sample for two reasons. First, most of our research team are experienced gambling researchers, which uniquely positioned us to determine whether all relevant details were specified when describing the use of field-specific measures and datasets (e.g., online gambling account data). Second, the gambling field is fraught with concerns regarding impartiality and QRPs due to the frequent involvement of the gambling industry in funding and supporting research (Livingstone & Cassidy, 2014), and open science practices such as preregistration have been proposed as a strategy to combat the risk of bias when undertaking industry-supported research (Louderback et al., 2020). Accordingly, we aimed to understand how effectively gambling researchers are currently preregistering their studies by comparing their preregistration specificity scores (according to Bakker et al.'s [2020] scoring protocol) with the specificity scores recorded for the randomly selected, cross-disciplinary preregistrations in Bakker and colleagues' study. As the discussion of open science principles and practices in the gambling field has been limited until recently (Blaszczynski & Gainsbury, 2019; Heirene, 2020; Heirene & Gainsbury, 2020; LaPlante, 2019; Louderback et al., 2020; Wohl et al., 2019), we hypothesized that preregistrations of gambling-focused research studies would have lower specificity levels (i.e., have lower scores on the RDoF scoring protocol) than the cross-disciplinary sample.

Methods

The hypotheses, methods, and analysis plan for this study were preregistered on OSF (https://osf.io/3jy6q). Unless otherwise stated, we adhered to the methods outlined in our preregistration. We have included a "Deviations from preregistration" subsection later in the methods section outlining any major deviations from our preregistration and minor deviations are presented in footnotes. The study data, analysis scripts, and materials, including details of transparent changes, can all be accessed on our *OSF page* (our project's *OSF Wiki* lists and describes all documents related to this study).

Search & selection process

Our complete process of searching for and selecting registrations is presented in Figure 1. We searched the OSF repository (www.osf.io) on three occasions throughout 2020 for preregistrations of gambling studies by searching the terms "gambling," "wagering," and "betting" separately. To be included, a preregistration had to:

- outline the plan for a study that was primarily focused on a gambling-related concept or concepts;
- be written in English;
- report at least one hypothesis;
- not be for a review and/or meta-analytic study as these studies involve unique forms of RDoF and risks of bias that would require a separate scoring system [e.g., PRISMA-P; Moher et al. (2015)];
- not be for a clinical trial as these also involve unique forms of RDoF and risks of bias that would require a separate scoring system[2] [e.g., CONSORT; Schulz et al. (2010)].

OSF searches and the selection of preregistrations were performed by BK. RH checked 20% of included and excluded registrations for the accuracy of the selection process according to the above eligibility criteria and agreed with all original selection decisions[3].

Sample size determination

To compare our sample with the 52 cross-disciplinary preregistrations analysed by Bakker et al. (2020) and thereby test our hypothesis, we aimed to include a minimum of 53 gambling study preregistrations. This was based on an a priori power analysis conducted using G*Power V3.1.9.4 for a Wilcoxon-Mann-Whitney test with power of 0.80, an effect size of 0.5 (d), and alpha (α) at 0.05, which estimated that 53 preregistrations per group (N = 106 overall; 53 gambling and 53 cross-disciplinary studies) would be required (Bakker et al. originally selected 53 preregistrations for evaluation but had to remove one as it was withdrawn from OSF). Our effect size of was based on Bakker and colleagues' suggestion that a medium effect ($d = 0.5$) is indicative of a practically meaningful difference between two samples of preregistrations.

We needed to conduct three separate searches of the OSF repository between March and October 2020 in order to identify 53 preregistrations meeting our criteria (see Figure 1). We did not summarise or analyse the data until all 53 preregistrations were identified and coded by two researchers. Although there were 55 gambling preregistrations meeting our eligibility criteria available on OSF at the time of our third and final search (see Figure 1), we restricted our sample size to the number provided by our *a priori* power analysis.

Sample description

The characteristics of our sample are presented in Table 1 alongside the characteristics of the 52 cross-disciplinary preregistrations evaluated by Bakker et al. (2020) for comparison. The data for the cross-disciplinary sample studied by Bakker et al. were accessed from the authors' *OSF page*[4]. All of these preregistrations were posted on OSF as part of the

[2]We decided this at the preregistration stage and, in retrospect, we believe the specificity scoring protocol used would be suitable for evaluating the specificity of clinical trial preregistrations as well.

[3]In our preregistration we stated that a second researcher would only check 10% of included & excluded registrations but we decided to review a larger sample of selections to ensure the accuracy of the process.

[4]Bakker et al.'s OSF page: https://osf.io/fgc9k/. The sample we extracted & studied here are labelled as group "1" in Bakker et al.'s R data file.

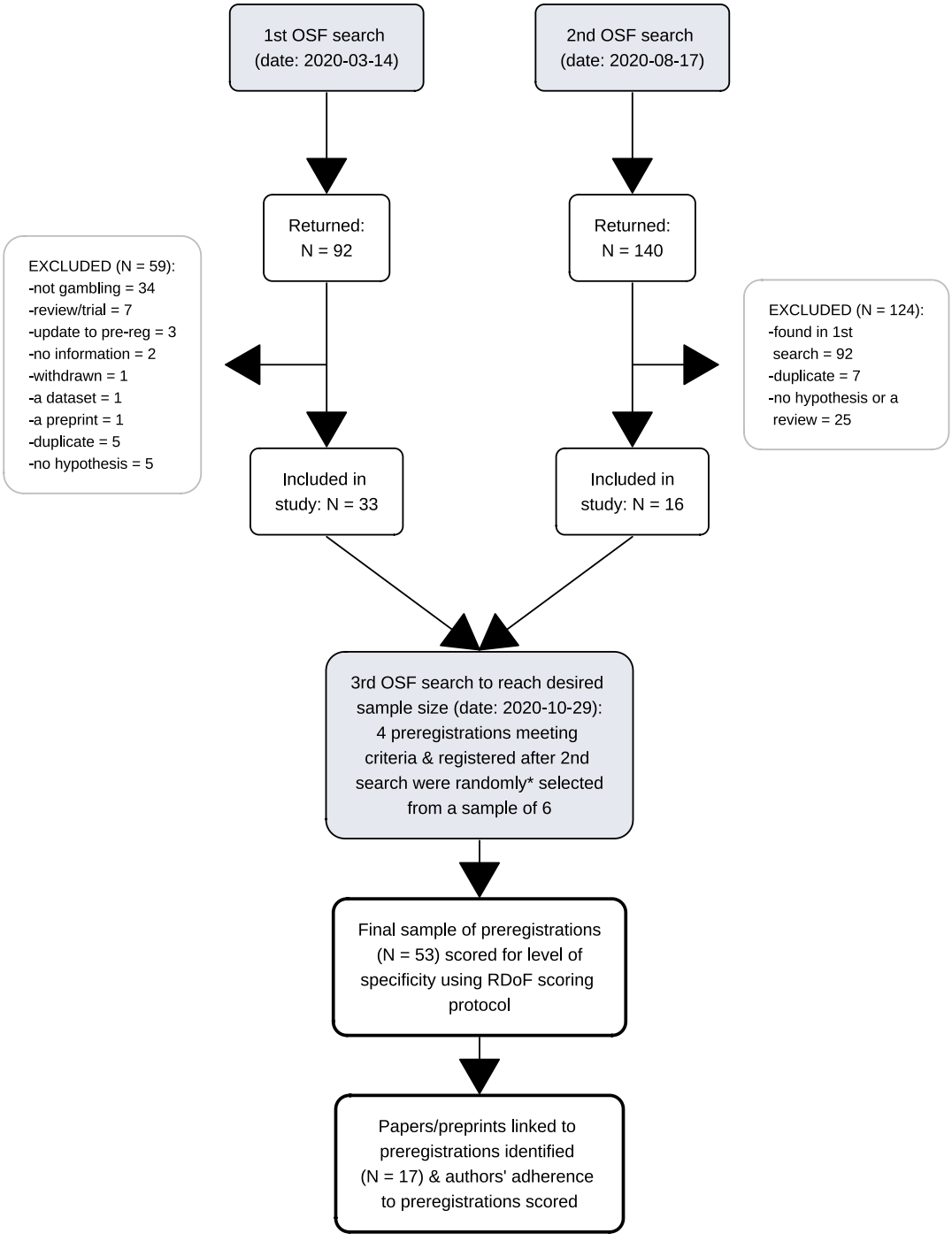

Figure 1

Flow-chart: Identification & selection of preregistrations. Figure legend: *This PRISMA-style flowchart presents the process of identifying and selecting our sample of preregistrations. Eligibility criteria are presented in order in which they were applied during the selection process. *Random selection performed using a random number generator (the R script used for this is shared on OSF).*

Table 1

Sample characteristics

Variable	Preregistration sample	
	Cross-disciplinary, N = 52	Gambling, N = 53
Template		
Open-Ended Registration	0 (0.0%)	11 (20.8%)
OSF Preregistration (formerly 'Prereg Challenge')	52 (100.0%)	32 (60.4%)
Preregistration Template from AsPredicted.org	0 (0.0%)	5 (9.4%)
OSF-Standard Pre-Data Collection Registration	0 (0.0%)	5 (9.4%)
Year		
2016	52 (100.0%)	0 (0.0%)
2017	0 (0.0%)	4 (7.5%)
2018	0 (0.0%)	9 (17.0%)
2019	0 (0.0%)	17 (32.1%)
2020	0 (0.0%)	23 (43.4%)

Note:

Statistics presented: n(%)

"Preregistration Challenge" (or "Prereg Challenge"), a competition held between 2015 and 2018 by the Centre for Open Science. The competition aimed to increase researchers' experience with preregistration and required participants to use a highly structured template to preregister their studies (a cash prize of $1,000 was awarded to all researchers who preregistered their studies using this template and published their findings in an eligible journal). The template asked researchers 26 questions about their planned study, including the research questions, hypotheses, sampling plan, variables, design, and analysis plan. This template remains available on OSF as the "OSF Preregistration" format (the form can be accessed *here*).

Bakker et al. (2020) labelled the Prereg Challenge template as a "structured format," compared to the "*Standard Pre-Data Collection*" template which they labelled an "unstructured format" as it only contains two questions that ask authors whether they have began data collection and whether they have looked at data. We compared our sample with Bakker et al.'s structured format preregistrations instead of their unstructured sample as our preliminary scans of OSF indicated that the OSF Preregistration format was most commonly used by gambling researchers. This template was the most frequently used format in our final sample (Table 1). There was no overlap between the two samples.

Scoring preregistration specificity

Table 2

Researcher degrees of freedom & associated preregistration specificity scoring protocol

Code	Researcher Degrees of Freedom (RDoF)	Associated preregistration specificity question
T1	Conducting exploratory research without any hypothesis	1: Is at least one hypothesis specified such that it is clear what are the IV(s) and DV(s)?
T2	Studying a vague hypothesis that fails to specify the direction of the effect	2: Is the direction of the hypothesis specified?
D1	Creating multiple manipulated independent variables and conditions	3: Does the text exclude the possibility that at least one of the manipulated variables will be omitted in the test of the hypothesis?
		4: Does it specify exactly how the manipulated variable will be used in the analysis to test the hypothesis?
D2	Measuring additional variables that can later be selected as covariates, independent variables, mediators, or moderators	5: Does it exclude the possibility that at least one other variable (e.g., covariate) is included in the analysis?
D3	Measuring the same dependent variable in several alternative ways	6: Does it specify which measurement instrument will be used as the main outcome variable?
D4	Measuring additional constructs that could potentially act as primary outcomes	7: Does it specify that the confirmatory analysis section of the paper will not include another DV than the ones specified in all hypotheses?
D5	Measuring additional variables that enable later exclusion of participants from the analysis (e.g., awareness or manipulation checks)	8: Does the pre-registration indicate inclusion and exclusion criteria in selecting data points?
D6	Failing to conduct a well-founded power analysis	9: Is a power analysis reported?
D7	Failing to specify the sampling plan and allowing for running (multiple) small studies	10: Is the sampling protocol outlined, including the exact number of participants, recruitment strategy, eligibility criteria, and stopping rules?
C1	Failing to randomly assign participants to conditions	11: Is it specified how randomization is implemented?
C2	Insufficient blinding of the participants and/or experiments	12: Does it describe procedures to blind participants to and/or experimenters to conditions?
C3	Correcting, coding, or discarding data during data collection in non-blinded manner	13: Does it include protocols concerning coding of data, discarding of cases, or correction of scores during data collection?
C4	Determining the data collection stopping rule on the basis of desired results or intermediate significance testing	Same as RDoF D7 (Question 10)
A1	Choosing between different options of dealing with incomplete or missing data on ad hoc grounds	14: Does it indicate how the study deals with incomplete or missing data?
A2	Specifying pre-processing of data (e.g., cleaning, normalization, smoothing, and motion correction) in an ad hoc manner	15: Does it offer a protocol for pre-processing the data when required (e.g., corrected for motion and other artifacts)?
A3	Deciding how to deal with violations of statistical assumptions in an ad hoc manner	16: Does it indicate how to test for and deal with violations of statistical assumptions ?
A4	Deciding on how to deal with outliers in an ad hoc manner	17: Does it indicate how to detect outliers and how they should be dealt with?
A5	Selecting the dependent variable at several alternative measures of the same construct	Same as RDoF D4 (Question 6)
A6	Trying out different ways to score the chosen primary dependent variable	18: Is the method used to measure the primary outcome variable(s) fully described?
A7	Selecting another construct as the primary outcome	Same as RDoF D4 (Question 7)
A8	Selecting independent variables out of the set of manipulated independent variables	Same as RDoF D1 (Question 3)
A9	Operationalising manipulated independent variables in different ways (e.g., by discarding or combining levels of factors)	Same as RDoF D1 (Question 4)
A10	Choosing to include different measured variables as covariates, independent variables, mediators, or moderators	Same as RDoF D2 (Question 5)
A11	Operationalising nonmanipulated independent variables in different ways	19: Are the methods to measure non-manipulated IV(s) fully described?
A12	Using alternative inclusion and exclusion criteria for selecting participants in analyses	Same as RDoF D5 (Question 8)
A13	Choosing between different statistical models	20: Does it specify the statistical model(s) that will be used to test the hypothesis (e.g., logistic regression)?
A14	Choosing the estimation method, software package, and computation of SEs	21a: Does it indicate details of the estimation technique used to estimate the statistical model and compute standard errors?
		21b: Does it specify which statistical software package and version is used for running the analyses?
A15	Choosing inference criteria (e.g., Bayes factors, alpha level)	22: Does it indicate the inference criteria (e.g., Bayes factors, Alpha level)?
R6	Presenting exploratory analyses as confirmatory (HARKing)	Same as RDoFs T1 (Question 1) and D4 (Question 7)

Note:

Specificity questions are summarised here for space purposes. The scoring protocol containing all full questions can be found on our OSF page

We used Bakker et al.'s (2020) scoring protocol to evaluate the specificity of preregistrations. This contains 23 questions[5] which provide scores for 29 RDoFs from Wicherts et al.'s (2016) checklist (all 29 RDoFs and the associated preregistration specificity scoring question are presented in Table 2). Thus, scores (i.e., specificity scores) represent the extent to which the preregistration restricts potential RDoFs arising during the research process. Specificity scores range between 0 and 3:

- 0 = not specified: opportunistic use of RDoF not restricted at all.
- 1 = some specification but lacking details: opportunistic use of RDoF is restricted to some extent[6].
- 2 = detailed specification: opportunistic use of RDoF is completely restricted, but no explicit statement confirming that authors will not deviate from this plan by adding additional methods/processes.
- 3 = detailed specification and statement that authors will not deviate from their plan by adding additional methods/processes: opportunistic use of RDoF is completely restricted. For example, in a *recent preregistration* written by two of the present authors, we outlined the reasons why a participant's data may be excluded from analyses before stating: "Individuals will not be excluded from analyses for reasons other than those stated here."
- N.A. = RDoF item not relevant to preregistration.

Like Bakker et al. (2020) and Ofosu and Posner (2019), we counted the number of hypotheses proposed in each preregistration. Further, given concerns regarding many gambling researchers' potential conflicts of interest due to their connections with industry and/or government, we also scored preregistrations on whether relevant disclosures were reported. We used the journal *International Gambling Studies'* (IGS) three-factor disclosure framework to structure our assessment. IGS' framework requires authors to disclose [1] funding sources for the work, [2] any competing interests, and [3] any constraints on publishing the findings made by funders or stakeholders. We scored preregistrations on each of the three factors as either 0 (no mention) or 3 (relevant disclosure reported).

During the scoring process we found it necessary to add our own "decision rules" to Bakker and colleagues' (2020) protocol that helped inform how we scored particular questions and enhanced our consistency across preregistrations. For example, in order to obtain a score of two or higher on question 10 (corresponding to RDoFs D7 and C4), researchers need to specify various details of the sampling plan, including "*how many and how additional participants or data points are sampled when pre-set sample size is not reached?*"; however, many of the studies preregistered in our sample involved online convenience sampling with minimal criteria for eligibility and, as a result, these researchers had almost total control over the number of participants they recruited. Therefore, not reaching their pre-set sample

[5]The original scoring protocol lists 22 questions, but one of these (Q21) has two questions (21a & 21b) with clearly distinct responses.

[6]For some RDoFs, there are fewer gradations of specificity possible & so scores of 1 are not possible for the RDoFs T1, T2, D1, D3, A2, A5, A8, & A9. For the same reason, scores of 1 & 2 are not possible for RDoFs D2, D4, A7, & A10.

size was not a concern for them and an associated plan did not need to be pre-specified. As such, we developed a decision rule which stipulated that preregistrations of these studies could score ≥ 2 for question 10, provided they had specified all other required details of their sampling plan. Our *full scoring protocol*, including these decisions rules, is shared on OSF and the original protocol by Bakker et al. can be accessed on their *OSF page*.

Scoring preregistration adherence

We developed a protocol for scoring gambling researchers' adherence to their preregistrations]with 32 questions—29 corresponding to the 29 RDoFs and three corresponding to disclosures (i.e., funding, conflicts of interest, & constraints on publishing). For example, for RDoF A1 ("*Choosing between different options of dealing with incomplete or missing data on ad hoc grounds*") we asked: "*Are the procedures used to deal with missing data consistent with those reported in the preregistration?*" Our *full adherence scoring protocol* is available on OSF and summarised versions of the questions are outlined in Table 5.

There were eight possible responses to each question:

- 0 = Yes, consistent with preregistration—no deviation.
- 1 = No, deviation from preregistration made and declared by the authors and a justification for change is provided.
- 2 = No, deviation from preregistration made and declared, but no justification for deviation is provided.
- 3 = No, deviation made and not declared or justified by the authors.
- U = unable to determine due to lack of detail reported in: [1] the preregistration [scored as U_P] (e.g., randomisation procedure not reported in preregistration but used in study), [2] the article [U_A] (e.g., randomisation procedure described in preregistration but not in the article), or [3] both [U_B] (e.g., randomisation is used but is not specified in either the preregistration or article).
- NA = Not applicable.

Scoring risk of bias in reporting

As we scored all articles for adherence according to the 29 RDoFs proposed by Wicherts et al. (2016), we decided (post-preregistration of the present study) to provide further information about the quality of the preregistered study articles by assessing them according to the remaining six RDoFs proposed by Wicherts et al. relating to the risk of bias in reporting. For example, for RDoF R3 ("*Failing to mention, misrepresenting, or misidentifying the study preregistration*") we asked: "*Is the preregistration clearly mentioned and linked/signposted to in the article and easily accessible (e.g., not embargoed)?*" We developed seven questions to cover these RDoF (see Table 6) and appended them to our adherence scoring protocol; all were scored as "1" (yes) or "2" (no).

Scoring procedure

Two researchers (RH + BK or AS) independently coded each preregistration and associated article[7] using the scoring protocols outline above, before convening to discuss any inconsistencies and to agree on final scores. Coders documented their scores in two separate "scoring frameworks" (i.e., Microsoft Excel files). All disagreements were resolved by the two coding pairs without the need to consult a third team member. No researcher was involved in coding their own preregistered study and the scores of preregistrations authored by one or more of our research team (N = 17), were checked by an external researcher for accuracy.

In our preregistration we stated that we would pilot code 10% of our sample. There were 33 preregistrations in our sample after the first OSF search and so we selected four of those with associated articles for pilot coding. After independently coding these, the level of inter-coder reliability achieved for specificity and adherence scores was quantified using Krippendorff's alpha ($_k\alpha$). We used the R package "irr" (Gamer & Lemon, 2012) to calculate $_k\alpha$ (analysis script shared on OSF). We achieved an level of inter-coder consistency of $_k\alpha = 0.859$ (2 raters, 104 items) for specificity scores and $_k\alpha = 0.809$ (2 raters, 156 items)[8] for adherence scores. As we achieved our pre-specified minimum level of consistency (i.e., $\geq$ 0.7), we proceeded to score the remainder of the sample. The master scoring framework containing the final, agreed-upon scores used to compute the findings presented here can be accessed on OSF. The time required to score preregistrations and associated articles was considerable—approximately 1 hour for specificity scoring, 1.5 hours for adherence scoring and 15 minutes for scoring risk of bias in study reporting per researcher.

Data analysis

All data analyses were performed using R version 4.0.2 (R Core Team, 2020). We have shared all of the *analysis scripts* used for this study on OSF, along with a *html document* presenting the annotated analysis code (and associated outputs) used to pre-process the data and compute all of the results presented here.

We summarised specificity scores by computing the arithmetic mean, standard deviation (*SD*), and median values for each RDoF and overall (i.e., mean scores on all items were summed and divided by the total number of items [N = 29]). For adherence and risk of bias in reporting scores, we simply tallied the number of each type of response for every question.

To compare gambling and cross-disciplinary preregistration specificity scores, we employed 30 Wilcoxon-Mann-Whitney (Wilcoxon) tests (29 RDoF specificity scores & 1 overall score). The decision to use non-parametric Wilcoxon tests is consistent with the strategy used by Bakker et al. (2020) and did not require data to be normally distributed (scores were right skewed; see Figure 2). As NA scores were common, particularly for some

[7] We use the term "article" to refer to published reports on findings, including journal articles and preprints.

[8] We did not include the question relating to the number of hypotheses in the inter-coder analysis of specificity scores, but we did include the three questions relating to disclosures; hence: (23+3)*4 = 104 items. For the analysis of adherence scores, we included questions related to disclosures & risk of bias in reporting, making a total of 37 items per article; hence: (29+3+7)*4=156.

items (i.e., RDoFs D1, C1, C2, A2, A8, A9, & A11; see Table 3), we used the same method of dealing with missing values employed by Bakker and colleagues. That is, we used a two-way imputation procedure based on corresponding row and column means. To minimise the false discovery rate (FDR), we used the Benjamini-Hochberg correction technique (Benjamini & Hochberg, 1995). This process involved multiplying all 30 p-values returned from our Wilcoxon tests by their rank after ordering them from largest to smallest (e.g., if our fifth largest p-value was 0.006 this would become: 0.006*5 = 0.03)[9]. To determine the magnitude of differences in specificity scores between the samples we calculated Cliff's Delta (D) effect sizes (Cliff, 1993).[10]

Deviations from our preregistration

We made a small number of deviations from our preregistered plan to best address the aims of the present study. We recorded all deviations and our reasoning for each in Transparent Changes Documents (hereafter "changes documents") that were uploaded to OSF.

All major deviations are also reported here. First, as described in our *changes document 1*, we decided to score specificity by providing a response for each of the 23 questions in Bakker et al.'s (2020) protocol and then later use these question responses to impute a score for each of the 29 RDoFs. This enabled us to provide a more detailed overview of preregistration specificity because of the dependencies present when scoring according to RDoFs. For example, RDoF A14 is "*Choosing the estimation method, software package, and computation of SEs [standard errors]*" and—when using Bakker et al.'s original protocol—a single specificity score is assigned to this RDoF based on two questions with unique answers: 21a and 21b (see Table 2). Thus, we prevented the loss of granular information provided by paired questions like 21a and 21b. The outcomes for each question (as opposed to RDoF) are shared on OSF.

Second, in our preregistration we stated that we would perform a maximum of two search and selection processes and stop sampling after the second, regardless of whether we had identified 53 preregistrations (our pre-specified target). However, after the second search we had identified 49 relevant preregistrations (see Figure 1) and, as we were still coding these several months later (thus sufficient time had lapsed to ensure more gambling studies had been preregistered), we decided to undertake a third search to try and reach our desired sample size (see *changes document 2*).

[9] In our preregistration, we stated that we would compare all original p-values to their corresponding Benjamini-Hochberg "critical value"—calculated as: (i/m)Q, where i = the rank of the p-value (ordered from smallest to largest), m = the total number of tests undertaken, & Q = the FDR we select (i.e., 0.05). However, instead we multiplied p-values by their rank to produce "corrected p-values" & make for easier interpretations of our findings in our summary table (Table 4).

[10] When used as an effect size, D represents the the extent to which two distributions of ordinal values overlap (Romano et al., 2006). D values range between -1 (all scores in Group 2 > all scores in Group 1) and 1 (all scores in Group 2 < all scores in Group 1), with 0 representing total overlap between samples. Romano and colleagues have compared D values to benchmark values for effect sizes when using Cohen's d (Cohen, 1988), finding a d of 0.2 (small effect) corresponds to a D of approximately 0.147, a d of 0.5 (medium effect) corresponds to a D of approximately 0.33, and a d of 0.8 (large effect) corresponds to a D of approximately 0.474.

Table 3

Preregistration specificity: Summary of specificity scores for gambling & cross-disciplinary preregistrations

	Gambling preregistrations				Cross-disciplinary preregistrations			
RDoF	Mean	*SD*	Median	NA (n)	Mean	*SD*	Median	NA (n)
Hypotheses								
T1: Hypothesis	2.32	0.47	2	0	2.02	0.14	2.0	0
T2: Direction of hypothesis	2.26	0.88	2	0	1.54	1.20	2.0	0
Study design								
D1: Multiple manipulated IVs	0.12	0.61	0	29	1.03	1.42	0.0	15
D2: Additional IVs	0.06	0.41	0	0	0.12	0.58	0.0	0
D3: Multiple DV measures	1.75	0.70	2	0	1.62	0.80	2.0	0
D4: Additional constructs	0.62	1.23	0	0	0.00	0.00	0.0	0
D5: Adding exclusion variables	1.55	0.91	2	0	1.23	0.70	1.0	0
D6: Power analysis	0.79	1.01	0	0	0.96	0.99	0.5	0
D7: Sampling plan	1.42	0.75	2	0	0.71	0.58	1.0	1
Data collection								
C1: Random assignment	0.45	0.83	0	33	0.86	0.92	1.0	15
C2: Blinding	0.75	0.50	1	49	0.02	0.14	0.0	3
C3: Data handling/collection	0.14	0.35	0	11	0.04	0.19	0.0	0
C4: Stopping rule	1.42	0.75	2	0	0.71	0.58	1.0	1
Analysis								
A1: Missing data	0.55	0.50	1	0	0.76	0.55	1.0	1
A2: Data pre-processing	1.33	1.15	2	50	0.50	0.93	0.0	44
A3: Statistical assumptions	0.45	0.67	0	0	0.18	0.48	0.0	1
A4: Outliers	0.38	0.77	0	0	0.69	0.92	0.0	0
A5: Selected DV measured	1.75	0.70	2	0	1.62	0.80	2.0	0
A6: DV scoring	1.21	0.93	2	0	0.65	0.65	1.0	0
A7: Primary outcome selection	0.62	1.23	0	0	0.00	0.00	0.0	0
A8: IV selection	0.12	0.61	0	29	1.14	1.48	0.0	15
A9: Defining manipulated IVs	1.96	0.46	2	29	1.92	1.19	2.0	15
A10: Adding additonal IVs	0.06	0.41	0	0	0.12	0.58	0.0	0
A11: Defining non-manipulated IVs	1.31	0.87	2	18	0.63	0.67	1.0	22
A12: Eligbility criteria	1.55	0.91	2	0	1.21	0.72	1.0	0
A13: Statisical model selection	1.36	0.56	1	0	1.31	0.51	1.0	0
A14: Method and package	0.13	0.44	0	0	0.13	0.44	0.0	0
A15: Inference criteria	1.02	0.77	1	0	1.08	0.33	1.0	0
Reporting hypotheses								
R6: HARKing	0.62	1.23	0	0	0.00	0.00	0.0	0

Note:

Specificity scores range between 0 & 3 (higher scores indicating greater specificity). See subsection 'Scoring preregistration specificity' for more details on the scoring protocol. All figures reported here were calculated using non-imputed specificity scores.

Third, and as stated in our *changes document 3*, we planned to calculate summary descriptive values (i.e., arithmetic mean & median) for adherence scores but we agreed that the scores 1-3 represented qualitative categories that described whether and how authors deviated from their preregistration and not an ordinal scale from "no deviation" to "major deviation." Additionally, we added the option to assign "U" (unable to determine) scores (see changes document 1) and these were common, meaning any summary values (e.g., means) would have not accounted for these categorical scores. Finally, we initially hypothesised that gambling registrations would have consistently lower specificity scores than the cross-disciplinary sample and chose to use one-tailed Wilcoxon tests; however, after performing the one-tailed tests as preregistered it became clear that the direction of differences was not consistent and therefore two-tailed tests were warranted to detect all differences between the samples. As such, we have recorded the outcomes from the one-tailed tests and report these on OSF, but report two-tailed test outcomes here (see *changes document 3*).

Results

Preregistration specificity

RDoF specificity scores

Table 3 presents a summary of the specificity scores for gambling preregistrations and for the cross-disciplinary sample studied by Bakker et al. (2020) for comparison. To allow further comparisons between gambling and cross disciplinary registrations, the frequency of specificity scores given to each RDoF for both samples is presented in Figure 2.

Confirmatory analyses.

Outcomes from the Wilcoxon tests comparing the groups' specificity scores are presented in Table 4. Gambling studies preregistrations were significantly more likely to include hypotheses that clearly described the variables of interest (RDoF H1: medium effect size) and stated the direction of the hypothesised effect (RDoF H2: medium effect), potentially reducing the risk of HARKing (RDoF R6: small effect).

In relation to study design, gambling preregistrations contained significantly more specification of sampling plans (D7: large effect) than cross-disciplinary preregistrations and were more likely to explicitly exclude the possibility of studying additional dependent variables other than those preregistered (D4: small effect). Conversely, descriptions of manipulated variables were significantly more specific in cross-disciplinary preregistrations (D1: medium effect).

In relation to data collection procedures, gambling preregistrations were significantly more specific in their descriptions of blinding (C2: very large effect), data handling during collection (C3: small-medium effect), and when data collection will end (i.e., "stopping rules"; C4: large effect).

Gambling preregistrations were also significantly more specific in their descriptions of four (of 15) RDoFs relating to the analysis process, including data preparation when working with complex datasets requiring pre-processing (A2: very large effect), the process of

Table 4

Preregistration specificity: Comparisons between gambling & cross-disciplinary registrations' specificity scores

	Wilcoxon test			Cliff's D effect size		
RDoF	W	p	Corrected p*	Effect	95% CIs	
Hypotheses						
T1: Hypothesis	962.5	0.0000	**0.0011**	-0.301	-0.428, -0.163	
T2: Direction of hypothesis	911.0	0.0013	**0.0261**	-0.339	-0.515, -0.135	
Study design						
D1: Multiple manipulated IVs	1852.0	0.0015	**0.0262**	0.344	0.13, 0.528	
D2: Additional IVs	1405.0	0.5562	2.7808	0.020	-0.046, 0.085	
D3: Multiple DV measures	1274.0	0.3100	2.4801	-0.075	-0.219, 0.071	
D4: Additional constructs	1092.0	0.0006	**0.0118**	-0.207	-0.316, -0.094	
D5: Adding exclusion variables	1046.5	0.0230	0.3226	-0.241	-0.435, -0.025	
D6: Power analysis	1501.0	0.3687	2.5808	0.089	-0.107, 0.279	
D7: Sampling plan	665.5	0.0000	**0.0000**	-0.517	-0.668, -0.325	
Data collection						
C1: Random assignment	1630.5	0.1008	1.0077	0.183	-0.043, 0.392	
C2: Blinding	69.5	0.0000	**0.0000**	-0.950	-0.987, -0.821	
C3: Data handling/collection	1048.0	0.0016	**0.0280**	-0.239	-0.378, -0.09	
C4: Stopping rule	665.5	0.0000	**0.0000**	-0.517	-0.668, -0.325	
Analysis						
A1: Missing data	1633.5	0.0579	0.6374	0.185	-0.007, 0.365	
A2: Data pre-processing	216.0	0.0000	**0.0000**	-0.843	-0.931, -0.663	
A3: Statistical assumptions	1068.5	0.0103	0.1654	-0.225	-0.384, -0.052	
A4: Outliers	1651.0	0.0329	0.3948	0.198	0.013, 0.37	
A5: Selected DV measured	1274.0	0.3100	2.7901	-0.075	-0.219, 0.071	
A6: DV scoring	906.5	0.0014	**0.0258**	-0.342	-0.526, -0.127	
A7: Primary outcome selection	1092.0	0.0006	**0.0124**	-0.207	-0.316, -0.094	
A8: IV selection	2000.5	0.0000	**0.0008**	0.452	0.248, 0.617	
A9: Defining manipulated IVs	1316.0	0.6873	1.3745	-0.045	-0.268, 0.182	
A10: Adding additonal IVs	1405.0	0.5562	3.3369	0.020	-0.046, 0.085	
A11: Defining non-manipulated IVs	641.0	0.0000	**0.0000**	-0.535	-0.698, -0.319	
A12: Eligbility criteria	1037.0	0.0196	0.2934	-0.247	-0.441, -0.032	
A13: Statisical model selection	1301.5	0.5676	1.7028	-0.056	-0.242, 0.135	
A14: Method and package	1380.5	0.9799	0.9799	0.002	-0.112, 0.116	
A15: Inference criteria	1453.5	0.5608	2.2432	0.055	-0.144, 0.249	
Reporting hypotheses						
R6: HARKing	1092.0	0.0006	**0.0129**	-0.207	-0.316, -0.094	
Overall						
Overall mean score	1024.0	0.0235	0.3051	-0.257	-0.456, -0.034	

Note:

*Corrected using Benjamini-Hochberg method (i.e., ranked from largest to smallest & then multiplied by rank); Bold p-values were statistically significant after the Benjamini-Hochberg correction; CIs = 95% confidence intervals; Plots show Cliff's Delta (D) & effect sizes 95% CIs—lower D values indicate higher specificity levels among gambling registrations.

measuring or scoring of the primary dependent variable (A6: medium effect), excluding the possibility of studying additional dependent variables (A7: small effect), and the process of measuring or scoring non-manipulated independent variables (A11: large effect). Descriptions of how manipulated variables will be used in analyses, however, were significantly more specific in cross-disciplinary preregistrations (A8: medium-large effect).

Overall, the mean specificity score for the gambling sample was greater than for the cross-disciplinary sample (medium-large effect), although this difference was not statistically significant after correcting for multiple testing with the Benjamini-Hochberg procedure.

Exploratory analyses.

We calculated the mean overall score per gambling study preregistration, grouped them by year of registration, and plotted them in Figure 3A. The mean specificity score of preregistrations increased year on year from 2017 (median = 0.73), through 2018 (median = 0.78) and 2019 (median = 0.98), and then dropped slightly in 2020 (median = 0.86).

We also grouped the mean overall score per preregistration by the template used and plotted this in Figure 3B. Open-ended preregistrations had the highest specificity score (median = 1.46), followed by those using the OSF preregistration template (formerly "Prereg Challenge"; median = 0.82), the template from AsPredicted.org (median = 0.83), and finally the OSF standard pre-data collection template (median = 0.59). However, 10 (91%) Open-ended preregistrations actually used the OSF preregistration template in a Word document format. Combining all preregistrations that used the OSF template in some form results in a median specificity score of 0.90. The conspicuous outlier in both panels of Figure 3 (mean score = 2.64) was a preregistration written by the first and last authors before we conceived of this study and was developed specifically to achieve high scores on the RDoF scoring protocol developed by Bakker et al. (2020). Overall, the mean specificity score was higher for the 17 preregistrations written by one of the present authors ($M = 1.27$, $SD = 0.47$) compared to the rest of the sample ($M = 0.83$, $SD = 0.26$).

We performed Spearman's rank-order correlations between the aggregated scores for all RDoF categories (e.g., Data collection, analysis). Specificity scores in every domain were strongly and positively correlated with every other (see Figure 4).

Number of hypotheses

Many hypotheses reported in preregistrations could be interpreted as single predictions or multiple independent but related predictions. For example, one hypothesis was: "*We predict that participants will report a higher likelihood of winning, excitement, and urge to gamble as well as hypothetically purchase more scratch cards when scratch cards are presented with unclaimed prize information compared to when scratch cards are presented without unclaimed prize information (i.e., ticket remaining information and game number conditions)*" which, while reported as a single hypotheses (no. 2 in a list of 4), contains four predictions that could be tested separately. The number of hypotheses therefore varied depending on whether all predictions reported as one hypothesis were assumed to be one hypothesis ($M = 3.96$, $SD = 3.51$, min = 1, max = 22) or multiple independent hypotheses

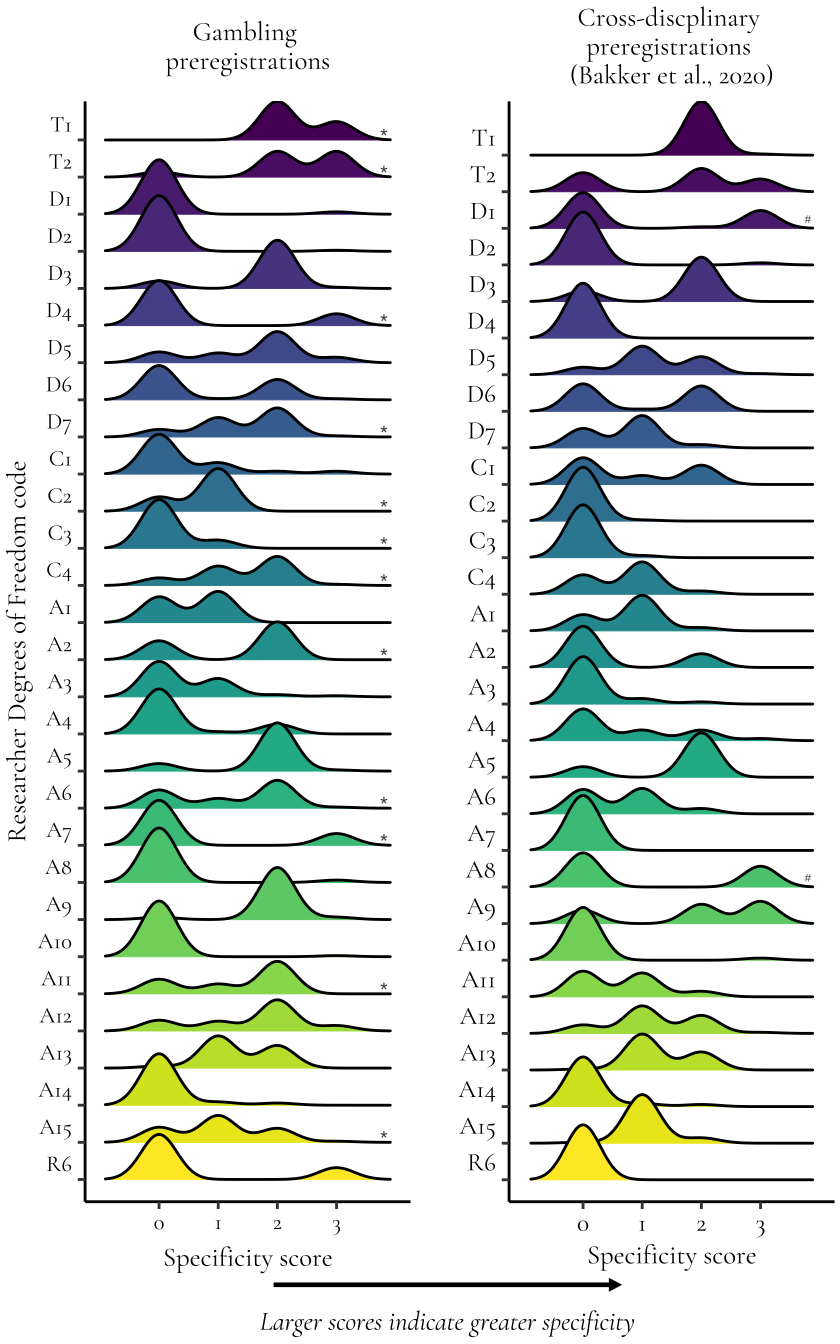

Figure 2

Distribution of specificity scores for gambling & cross-disciplinary preregistrations. Figure legend: *These density plots show the relative distribution of specificity scores given for each RDoF item for both samples of preregistrations (non-imputed scores used). * & # indicate statistically significant difference between samples: * = gambling preregistrations > cross-disciplinary; # = cross-disciplinary > gambling preregistrations (see Table 4 for test outcomes). Note: Scores of 1 were not possible for the following RDoFs: T1, T2, D1, D3, A2, A5, A8, and A9. Scores of 1 and 2 were not possible for the following RDoFs: D2, D4, A7, and A10. Also, while this figure displays the relative distribution of scores for each RDoF rather than exact frequency counts, the number of scores contributing to each density plot varies between the samples due to differences in the number of NA scores (see Table 3).*

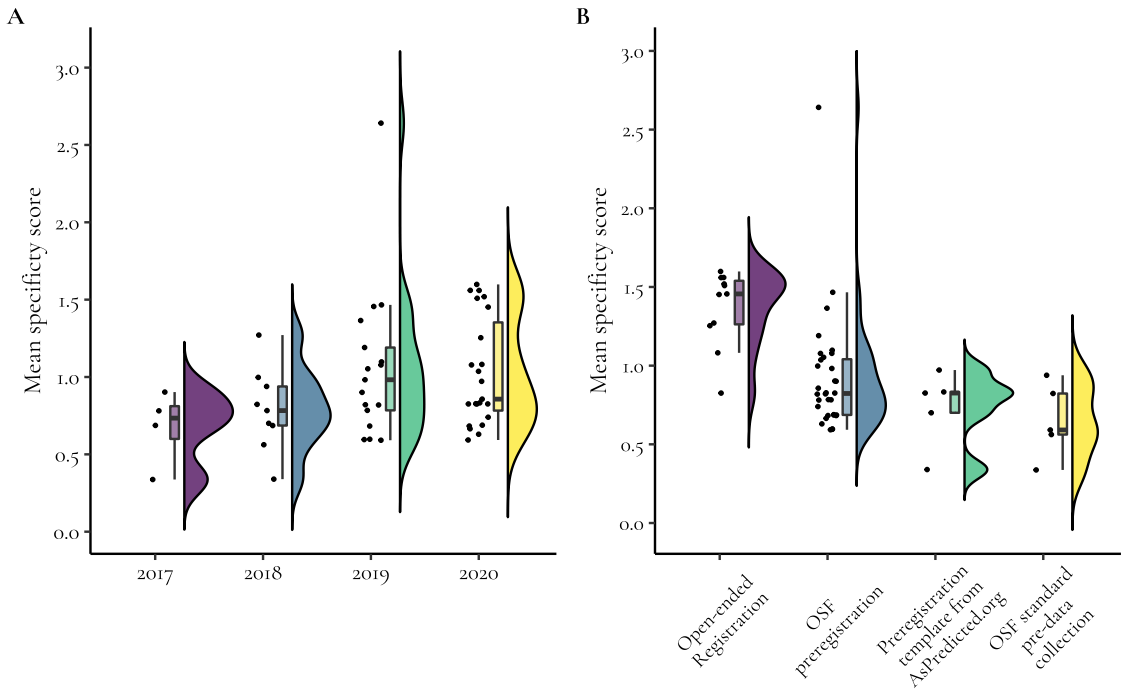

Figure 3

Preregistration specificity scores over time (A) & for different templates (B). Figure legend: *Figure 3A shows each preregistration's mean overall specificity score, grouped by the year of registration. Figure 3B shows the same values but grouped by the template used to structure the preregistration. Both use non-imputed, original scores*

(M = 6.4, SD = 7.54, min = 1, max = 44). Eleven (20.75%) articles presented their hypotheses in this way.

Reporting of disclosures

Sixteen (30.2%) preregistrations included a funding disclosure, 10 (18.9%) reported a conflicts of interest statement, and 9 (17.0%) reported whether there were any restrictions on publishing. However, almost every preregistration that included a disclosure was authored by one or more of the present team. After removing our preregistrations, only 1 (2.8%) of the remaining 36 included a funding disclosure, and none reported conflicts of interest statements or restrictions on publishing.

Adherence to preregistrations

We found 17 articles associated with 20 preregistrations. Scoring was done at the level of the preregistered study and thus scores for 20 articles are presented. We found 13 (65%) articles included at least one undeclared deviation (i.e., a score of 3). The number of undeclared deviations per study ranged from 0 to 8 (M = 2.25, SD = 2.34). The number of articles containing at least one undeclared deviation was 3 (100.0%) in 2017, 4 (66.7%) in

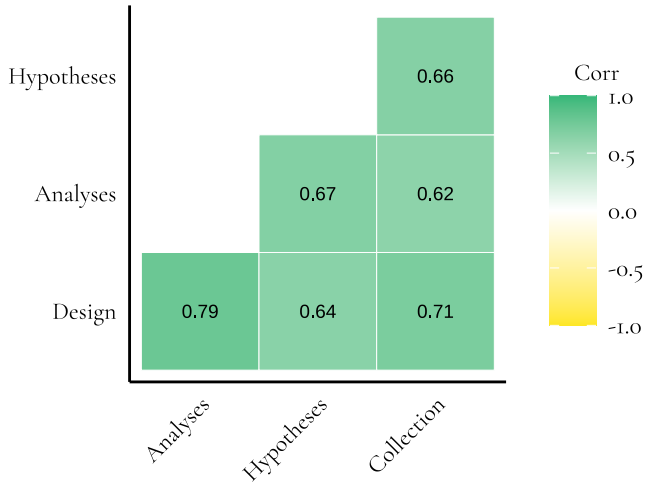

Figure 4

Correlation matrix for the relationships between aggregated specificity scores. Figure legend: *All correlations were significant at the $p < 0.05$ level.*

2018, 4 (50.0%) in 2019, and 2 (66.7%) in 2020. Only 4 articles declared a deviation from the preregistration and provided a rationale for the change (i.e., a score of 1; the range of this score [per article] was 0-8, $M = 0.85$, $SD = 2.3$).

Figure 5 presents the proportion of each adherence scores as given across all questions and articles. A score of 0 was most common, indicating no deviation from the preregistration. The different "U" scores were also common, indicating that it was frequently difficult to determine whether authors had deviated from their preregistrations. Combined, U scores made up 40.6% of the total responses given, with most (36.5%) made up by U_P (unable to determine due to a lack of information in preregistration) and U_B (unable to determine due to a lack of information in both the preregistration and article) scores. A score of 2 was not awarded to any article, indicating that all reported deviations were accompanied with rationale.

Table 5 presents the distribution of adherence scores for each question. Undeclared deviations most commonly related to the hypotheses tested, the variables included in tests, and the statistical analyses selected to test hypotheses. U_P scores, which indicate that there was poor specificity of an item in the preregistration despite being relevant to the study, were common in relation to the operationalisation of independent variables, the estimation techniques used to estimate the statistical model(s), the statistical software used to conduct analyses, inference criteria, research funding, and competing interests. U_B scores, which indicate a lack of specificity in both the preregistration and article despite being relevant to the study, were common in relation to the procedures used to randomly allocate participants

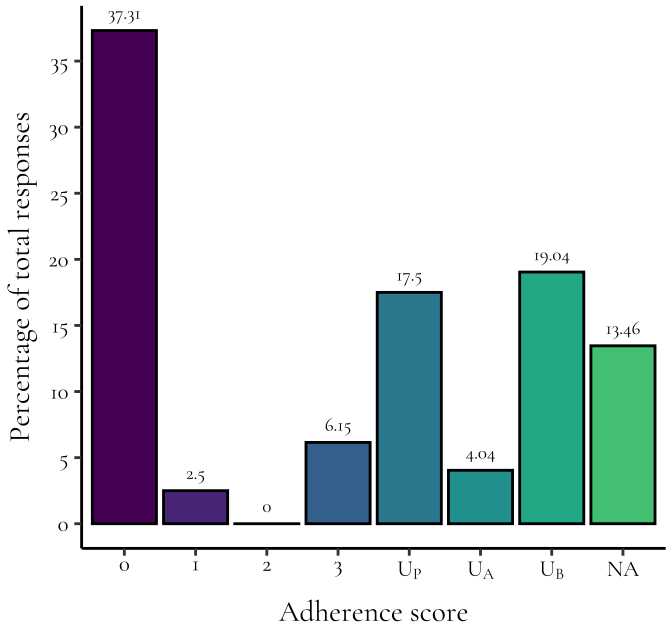

Figure 5

Distribution of adherence scores. Figure legend: *The proportion of each type of adherence score for the entire set of responses across all questions & articles. There were 520 total responses (26 questions * 20 articles). Scoring: 0 = Yes, consistent with preregistration—no deviation; 1 = No, deviation from preregistration made and declared by the authors and a justification for change is provided; 2 = No, deviation from preregistration made and declared, but no justification for deviation is provided; 3 = No, deviation made and not declared or justified by the authors; U = unable to determine due to lack of detail reported in the preregistration [U_P], the article [U_A], or both; [U_B]; NA = Not applicable.*

to conditions, coding and handling data during data collection (e.g., dealing with mistakes made by participants or equipment), dealing with missing data, handling outliers, testing statistical assumptions, the software used to perform analyses, and possible constraints on publishing findings.

Risk of bias in reporting

The outcomes from scoring the risk of bias in study reporting are presented in Table 6. We operationalised RDoF R5 (misreporting results and p-values) as failing the online tool 'statcheck' (http://statcheck.io), which uses the test statistic and degrees of freedom from reported outcomes to recalculate *p*-values and highlight any discrepancies between reported and recalculated values. Statcheck was able to identify all of the components required to recompute 60 *p*-values in seven articles (the tool may have been unable to find the information required to compute p-values in some articles for several reasons, including because none were reported, results were not reported in APA style, or difficulty reading PDF files). We found six (10.0%) statistical reporting errors, one one (1.67%) of which was a decision error (i.e., a *p*-value misreported in a way that may affect whether it is interpreted as statistically significant [it crosses the 0.05 threshold]), spread across two articles (which

Table 5

Gambling researchers' adherence to their preregistrations: Frequency of adherence scores by question

	Adherence scores (n)							
Abbreviated question	0	1	2	3	U_P	U_A	U_B	NA
Hypotheses								
Are the hypotheses reported the same as in the preregistration?	12	0	0	7	1	0	0	0
Is the direction of each hypothesis the same?	14	0	0	1	1	4	0	0
Study design								
Are the manipulated independent variables operationalised in the same way, stated in the protocol?	10	0	0	0	1	0	0	9
Are all variables included in analyses testing hypotheses, consistent with the preregistered analysis plan?	14	0	0	5	1	0	0	0
Are dependent variables measured in the same way as stated in the preregistration?	17	1	0	0	2	0	0	0
Are all dependent variables included in analyses reported in the preregistration?	17	1	0	2	0	0	0	0
Data collection								
Are the criteria for including datapoints in analyses consistent?	8	1	0	2	2	6	1	0
Is the sampling protocol stated in the preregistration followed?	9	2	0	2	3	3	1	0
Is the randomisation procedure used consistent with that reported in the preregistration?	2	0	0	0	0	1	7	10
Is the blinding procedure used consistent with that reported in the preregistration?	0	0	0	0	0	1	1	18
Are the procedures used to code and manage data during the data collection process consistent?	0	0	0	0	1	0	14	5
Analysis								
Are the procedures used to deal with missing data consistent with those reported in the preregistration?	4	0	0	0	2	3	11	0
Are the procedures used to preprocess data consistent?	0	0	0	0	1	0	0	19
Are the procedures used to test for statistical assumptions consistent?	3	0	0	1	4	0	12	0
Are the procedures used to identify and deal with outliers consistent?	1	1	0	0	3	1	14	0
Are the dependent variables scored in a way that is consistent?	11	1	0	1	4	0	3	0
Are the dependent variables used in primary analyses all the same as reported in the preregistration?	18	1	0	1	0	0	0	0
Are the independent variables used in primary analyses all the same?	17	0	0	2	0	0	0	1
Are non-manipulated IVs operationalised in a way consistent with the preregistration?	6	0	0	1	5	0	0	8
Are the statistical tests used to test hypotheses consistent?	10	3	0	5	1	1	0	0
Are the estimation techniques used to estimate the statistical model(s) consistent?	1	1	0	0	15	0	3	0
Was the statistical software used to conduct analyses consistent with the preregistered plan?	7	0	0	0	6	0	7	0
Are the inference criteria used consistent?	8	1	0	2	5	1	3	0
Disclosures								
Are the funding sources reported the same as stated in the preregistration?	2	0	0	0	15	0	3	0
Are the competing interests reported the same?	1	0	0	0	16	0	3	0
Are the constraints on publishing reported the same?	2	0	0	0	2	0	16	0
Overall								
Summation	194	13	0	32	91	21	99	70

Note:

We answered all questions in relation to the confirmatory, hypothesis tests. Undeclared deviations (i.e., scores of 3) are coloured red for ease of detection. While we developed 29 questions for each of the 29 RDoFs (and 3 related to disclosures), due to dependencies in the RDoFs the same question was asked for 6 pairs of items (e.g., RDoFs D4 and C4) and so we removed all responses to duplicated questions before performing calculations to prevent weighting some questions more than others.

U_P = Unable to determine to due lack of specificity in preregistration

U_A = Unable to determine to due lack of specificity in article.

U_B = Unable to determine to due lack of specificity in both the preregistration and article.

Table 6

Summary of risk of bias in reporting scores

			Scores (N)		
Code	Researcher Degrees of Freedom	Question	Yes	No	NA
R1	Failing to assure reproducibility (verifying the data collection and data analysis)	Are data shared and accessible to all?	12	8	0
		Are the data analysis scripts shared and accessible?	6	14	0
R2	Failing to enable replication (re-running of the study)	Are the methods reported sufficiently, to allow replication? Including all study materials used?	17	3	0
R3	Failing to mention, misrepresenting, or misidentifying the study preregistration	Is the preregistration clearly mentioned and linked/signposted in the article and easily accessible?	16	4	0
R4	Failing to report so-called "failed studies" that were originally deemed relevant to the research question	Are any experiments that were preregistered not reported?	2	18	0
R5	Misreporting results and p-values	Does running the paper through statcheck highlight any potential statistical errors?	4	6	10
R6	Presenting exploratory analyses as confirmatory (HARKing)	Are any hypotheses reported that weren't stated in the preregistration?	3	17	0

Note:
Scores were assigned for each preregistered study reported as opposed to each article, other than for RDoF 5 which had to be scored at the article level and therefore scores for two of the 17 articles are represented five times in the frequency counts presented as these reported results from five of the preregistrations in our sample.

reported four preregistered studies between them). However, we decided to manually inspect all errors and found that one non-decision error and the one decision error were mistakes made by statcheck misidentifying outcome values. After removing these two incorrectly identified reporting errors, the remaining four errors were still reported in two articles.

Discussion

The aim of this study was to better understand modern preregistration practices and how these can be improved to maximise their potential scientific benefits. We assessed the degree to which gambling studies researchers sufficiently specified all aspects of their studies in preregistrations (N = 53), the extent to which they adhered to their plans, and the risk of bias in the reporting of preregistered studies in the field. We also compared the results for our sample with the results from a similar study that analysed a cross-disciplinary sample of 52 preregistrations (Baker et al., 2020). In the following subsections we discuss the results from each of these assessments, the implications and limitations of our findings, and recommendations for improving preregistration practices.

Preregistration specificity

Similar to Bakker et al. (2020), we found that gambling researchers' level of specificity was low for many RDoFs, indicating that RDoF in these particular areas was not restricted by preregistrations. Mean specificity scores were less than 1 (which represents the minimal possible specificity, and 0 represents 'not specified') for 15 RDoFs, including descriptions

of: the independent variables and how they will be measured (D1 & A8); all variables (e.g., covariates, moderators) included in analyses (D2 & A10); the primary dependent variable(s) (D4 & A7), power analyses (D6), participant randomisation (C1); blinding procedures (C2); coding and handling data during collection (C3); handling missing data (A1); dealing with statistical assumptions testing (A3); handling outliers (A4); the estimation method software package and computation of standard errors (A14); and the hypotheses, sufficiently so as to prevent HARKing (R6). These findings suggest the intended benefits of preregistration—such as restricting and enabling an evaluation of test severity—are not fully achieved by current levels of reporting within preregistrations. One area where specificity levels were relatively high (mean >2) was in the description of study hypotheses. While some hypotheses were vaguely specified (see Number of hypotheses subsection of results), most researchers presented hypotheses that enabled us to discern the key variables under study as well as the direction of the predicted effect(s). This is positive given the centrality of hypotheses to preregistrations, and represents an area of good practice.

Despite generally low specificity levels and contrary to our hypothesis, 12 RDoF specificity scores from our gambling studies sample were significantly higher than those from the cross-disciplinary sample in Bakker et al (2020). There are a number of possible reasons for this. First, all studies in the cross-disciplinary sample were registered in 2016 and mean specificity scores appear to have improved over time (70.5% of articles in our sample were published in 2020, 23.5% in 2019, & 6% in 2018). Second, there may have been differences in scoring between our study and that of Bakker and colleagues. As stated in the *Scoring preregistration specificity* subsection, we developed multiple decision rules to guide our scoring and these often focused on how we could award *more* scores in circumstances where the proposed methods were not aligned with the scoring system so as not to unfairly disadvantage these preregistrations. For example, question two in the scoring protocol asks, "*Is the direction of the hypothesis specified?*" and in order to obtain a high score of 3, a preregistration must also state the sidedness of the statistical test of the hypothesis; however, some of the preregistrations used Chi-Squared tests and/or analysis of variance (ANOVA) and the sidedness of these tests cannot be specified. As such, we awarded a score of 3 in these cases so long as the direction of all predicted differences were clearly specified. Third, 17 (32.1 %) of the gambling preregistrations were authored by one or more of the present study's team, all of whom are dedicated to improving the transparency of their work through preregistration. The mean overall specificity score for registrations authored by one of the present team was considerably higher than the remaining sample of registrations (1.27 and 0.83, respectively).

Adherence to preregistrations

Researchers may deviate from their preregistration for a number of reasons—due to requests from referees or editors during the peer review process; after finding a more appropriate statistical test of their hypothesis or unexpected, but logical, reasons to exclude particular participants; or more concerningly, in order to increase the chance of observing statistically significant findings and/or to inflate effect sizes. Thus, deviations can be positive, resulting in more informative and/or scientifically rigorous outcomes, or negative, resulting

in misleading or inaccurate findings. Either way, it is essential that researchers transparently report any deviations so that others can judge their appropriateness and potential impact on the validity of the findings reported.

Our findings support existing research on clinical trial registration (Goldacre et al., 2019; Vassar et al., 2020) and general study preregistration (Claesen et al., 2019; Ofosu & Posner, 2019) in suggesting that many researchers do not transparently declare deviations from their pre-specified plans. We found a lower proportion of articles included undeclared deviations (65%) than Claesen et al. found in their sample of preregistered studies published in *Psychological Science* (96%). This could be explained by the outlet of publication (none of our sample were published in *Psychological Science*) or, perhaps more likely, improved reporting standards since the 2015-2017 period studied by Claesen and colleagues. Unreported deviations in our sample were most common in relation to hypotheses (35% of articles), the variables included in hypothesis tests (25%), and the statistical models used to test hypotheses (25%). These results are consistent with Ofosu and Posner's (2019) observations in the economics and political science literature, who found the median article failed to report 25% of registered hypotheses, 18% included tests of non-registered hypotheses, and 19% articles deviated in the statistical models used (only one of which declared this deviation). Breaking down the types of hypothesis deviations in our study, four articles (20%) failed to report preregistered hypotheses, two (10%) reported non-registered hypotheses, and one (5%) altered preregistered hypotheses (e.g., by changing non-directional to directional predictions). These findings suggest changes to hypotheses post-registration are more diverse than simply developing *post-hoc* hypotheses most consistent with the outcomes (i.e., what Kerr [1998] termed "pure HARKing").

Our findings expand on previous fidelity studies (Claesen et al., 2019; Ofosu & Posner, 2019) by also reporting the number of instances when we were unable to tell whether authors deviated from their preregistrations due to insufficient detail in their preregistration (U_P), article (U_A), or both (U_B). These outcomes are essential for understanding the value of current preregistration practices. If, as was frequently the case in our study, one cannot determine whether the methods reported in an article are consonant with the allied preregistration, then the value of the practice is seriously diminished. Our breakdown into U_P, U_A, and U_B scores revealed that ambiguous and/or incomplete reporting in both preregistrations and resulting articles often precludes efforts to cross-check pre-planned methods with those actually used. Preregistrations often included insufficient details of statistical estimation methods to enable comparisons with published articles, and both preregistrations and articles frequently failed to provide any detail regarding procedures used to handle outliers, data handling during collection, testing of statistical assumptions, dealing with missing data, the software used to perform analysis, and randomisation procedures. Claesen and colleagues (2019) also reported that they found it difficult to assess whether authors had deviated from their preregistrations because neither "*preregistrations nor the published studies were written in sufficient detail*" (p. 9).

Risk of bias in reporting preregistered studies

Our evaluation of the risk of reporting bias is, to our knowledge, the first study to use Wicherts et al.'s (2016) checklist for this purpose and provides further insights into preregistration and reporting practices. Of 20 preregistered studies, data were shared for 12 and analysis scripts were available for six. These rates are substantially higher than those found in the wider gambling literature for sharing data and analysis scripts, which were both found in less than 4% of studies in a random sample of 500 gambling research studies for the 2016-2019 period (Louderback et al., *In preparation*). The higher rates found in our study might be because researchers who preregister their studies are more likely to engage in other open science practices. We found four articles (of 17) that did not mention the study preregistration or link to it, hampering attempts by readers to compare the article with the preregistration. One article (for two preregistrations) did not report a third study that was preregistered. When we contacted the author to enquire about this, they stated that they had originally submitted the preregistered study to a journal and reviewer comments led them to perform two additional experiments, but they did not explain why the outcomes from the original study were not reported anywhere. Further, three articles were not reported sufficiently to enable replication and two (for four preregistrations) contained statistical reporting errors, obfuscating interpretations of findings and replication attempts. These instances represent opportunities for additional education about transparency in reporting that will help advance the gambling field and beyond.

Limitations

There are three limitations that are important to note. First, our sample of preregistrations and articles was restricted to the gambling studies field. Although this conferred the benefits discussed in our introduction (i.e., subject expertise aided evaluation of reporting; concerns of bias in the field), gambling research is multidisciplinary and researchers typically come from the fields of psychology, neuroscience, and public health. Therefore, our outcomes might not generalise beyond these disciplines, despite the similarities between our findings and evaluations of preregistered studies in economics and political science (Ofosu & Posner, 2019). Second, our exploration of preregistration adherence was limited because articles were only available for 20 of the preregistrations in our sample. Third, there might also be limitations to the specificity scoring protocol we used to evaluate preregistrations. To achieve a maximum score of 3 on most RDoF items requires preregistration authors to explicitly state that they will not deviate from their pre-specified method by, for example, using additional eligibility criteria or reasons for excluding data points. Although such statements may add value in restricting RDoF, this approach is unconventional in scientific research and therefore scores of 2 and 3 could be viewed as equivalent until the value of making explicit promises not to deviate from preregistrations has been empirically evaluated.

Implications of findings

Our findings have several important implications for understanding and advancing the value of preregistration in scientific research. At present, study plans presented in

preregistrations would benefit from additional specificity so as to prevent researchers needing to make data-contingent decisions (e.g., when to cease data collection) that could potentially bias findings (Wicherts et al., 2016). Further, the majority of articles reporting preregistered studies contain at least one undeclared deviation from the preregistration and a notable proportion do not mention that the studies were preregistered, precluding evaluations of test severity (Lakens, 2019) and preregistration fidelity. What is more, the failure to clearly describe methods in both preregistrations and corresponding articles was problematic and obfuscated evaluations of consistency. In one case, it took two researchers six hours each to score one preregistration for specificity and adherence due to ambiguity and a lack of clarity in the preregistration and inconsistencies with the article. There are a number of factors that likely contribute to these difficulties beyond the control of researchers. For example, strict journal word counts can prevent authors from fully explaining their methods and requests from reviewers and editors made during the review process can lead to changes in the terms used or the analyses conducted that make comparisons with preregistrations difficult. While these issues are not present when writing preregistrations, preregistration remains a relatively new component of the research process. To date, research institutions have provided little formal training and guidance for preparing preregistrations. Additionally, the time and resources required to undertake preregistration has not been factored into existing funding structures.

Overall, our findings suggest that the purported benefits of preregistration—increasing transparency, restricting RDoF, enabling evaluations of test severity—are not fully achieved by gambling studies researchers' current preregistration practices. This conclusion is concerning as the time required to preregister studies is not insubstantial. Ofosu and Posner (2019) found 88% of economics and political science researchers surveyed spent, on average, at least a week writing their preregistration, 32% spent 2-4 weeks, and 26% spent more than a month; although the majority of those surveyed agreed that the time dedicated to preregistration was worthwhile and that it allowed them to receive useful pre-study feedback and/or it saved time downstream. Still, the time investment has been raised as an objection to preregistration (Ofosu & Posner, 2020) and preregistering one's study with sufficient detail is challenging (Nosek et al., 2019) . Evaluations of how preregistering studies impacts the reporting quality, reproducibility, and replicability of published research are needed to confirm whether the benefits justify the additional effort required to review preregistrations.

Preregistration practices appear to be improving. We observed increases in specificity and decreases in the proportion of articles containing undeclared deviations from 2017 to 2020. We provided further evidence that more structured templates like the OSF preregistration[11] and AsPredicted.org formats typically result in higher levels of specificity than less structured templates like the OSF standard pre-data collection format. Finally, undertaking this study has provided unique insights into the difficulties faced when trying to interpret preregistrations and evaluate researchers' adherence to them, which we have used to proffer suggestions for improving the value of preregistration for researchers and organisations involved in the scientific enterprise (journals, research institutions, and funding bodies) below.

[11]Recall that while preregistrations listed as using the "Open-ended" format had the highest specificity scores (Figure 3B), 91% of these actually used the OSF preregistration template in a Word document.

Five reccomendations for researchers preregistering their studies

1. **State what it takes to falsify your hypothesis:** Lakens (2019) recommended that authors of preregistrations do this, and this strategy would overcome many of the issues we observed in gambling study preregistrations. As described, several authors presented multiple predictions as a single hypothesis without specifying whether one or all needed to be supported in order to view the hypothesis as being supported by their data (and possibly increasing the likelihood of authors being able to state that their hypothesis was at least "*partially supported*"). Further, some hypotheses were so vague as to be almost impossible to falsify (e.g., "*The removal of opportunities to bet on live sporting events [due to COVID-19 shutdowns] will lead some sports bettors to engage in other forms of gambling.*"[12])—they lacked severity (Lakens, 2019). Both of these issues can be avoided by stating what outcome(s) would falsify one's hypotheses.
2. **Use a structured preregistration template:** Structured templates like the OSF preregistration format are associated with better specificity and can help researchers to understand what information they need to include in their preregistrations to ensure their study plan is sufficiently specified. Authors can further enhance the specificity of their preregistrations by using Bakker et al.'s (2020) scoring protocol as a guide, as we did when preregistering this study.
3. **Ensure consistency between preregistration and article:** Researchers should make it as easy as possible for others to compare their pre-specified study plan with the resulting article. This can be achieved by using consistent terminology between the two (e.g., for variables & statistical models); by providing each hypothesis with the same, consistent label (e.g., H^1); and, if using OSF to post preregistrations, by (re)naming their overarching project page (or relevant subcomponent) with the title of the final article. We found many OSF pages contained multiple preregistrations with similar names and overlapping content, making it difficult to discern which preregistration belonged to which article.
4. **Clearly and directly link to your preregistration:** Difficulties in connecting preregistrations and articles were also found by Claesen et al. (2019) and, as they recommended, could be further avoided by including a clear link directly to the allied preregistration(s) in articles and not simply a link to the overall project page.
5. **Report all deviations from your preregistration:** We recommend that authors report all protocol deviations within their study article under a clear heading like "Deviations from preregistration," as we have done here. However, space constraints may make it difficult to fully report each deviation, the rationale for the change, and the likely effect on study outcomes. Claesen et al. (2019) have developed a document for recording all of this information (https://osf.io/xv5rp/) and we have used similar "Transparent changes documents" for this study (https://osf.io/qep2a/) and others (https://osf.io/j6tud/). Whichever format chosen, researchers should share these documents on an accessible repository (e.g., OSF) and/or alongside their article

[12]This particular preregistration also contained the hypothesis "some sports bettors stop gambling because they are primarily interested in sports, not other things." Thus, one (but likely both) of these two preregistered hypotheses literally has to be true (i.e., sports bettors must either stop gambling or gamble on other activities in the absence of opportunities to bet on sports).

as supplemental material.

Five recommendations for journals, research institutions, and funding bodies to improve the value of preregistration

1. **Support transparency, not a clean narrative:** Echoing the arguments made by the Nature Human Behaviour editorial team (2020), journals should encourage researchers to transparently report all aspects of their studies, including deviations, regardless of whether this makes the findings appear less conclusive or compelling. Others (e.g., Frankenhuis & Nettle, 2018) have suggested that a fully transparent presentation of results, including clear labelling of confirmatory and exploratory analyses, can actually foster creativity and knowledge sharing because all results are presented instead of only significant or "interesting" findings.
2. **Remove word count restrictions on methods sections:** The ability to understand exactly how research data were obtained, analysed, and interpreted is fundamental to scientific understanding. Yet, many journals' word limit policies leave researchers with too little space to fully describe these processes. Word restrictions, if required at all, should be reserved for the introduction and discussion sections of articles so that researchers can freely describe all aspects of their methods and results.
3. **Review preregistrations alongside articles:** As highlighted by Claesen et al. (2019), existing systems (e.g., open science badges) reward authors for simply performing the act of preregistration, regardless of what information is included. Reviewing preregistrations alongside submitted manuscripts could determine whether authors have preregistered a minimum set of study details (e.g., hypotheses, sample size rationale, measurements, analyses) and any deviations. However, this would likely require incentivising reviewers, whether monetarily or via increased recognition of peer-reviewing contributions when considering candidates for jobs, promotions, and funding opportunities (see Moher et al., 2020).
4. **Provide training and guidance on preregistration:** Teaching researchers about the scientific benefits conferred by study preregistration and providing training courses and guidance on how to write preregistrations will help to ensure that we maximize the benefits of this practice and avoid wasting resources on insufficiently detailed and poorly followed preregistrations.
5. **Make preregistration mainstream:** Research institutions and funding bodies should consider study preregistration a normal component of conducting hypothesis-testing research. The time and resources required to preregister studies should be factored into funding programmes and workloads so that researchers have sufficient time to write their preregistrations in a way that will achieve the intended benefits. Journals can also support this effort by including links to preregistrations alongside their articles' key information (e.g., DOI, author list), by considering the development of novel direct integration strategies within methods sections, and by requiring manuscript sections dedicated to highlighting deviations.

Conclusions

A preregistered study is not necessarily better, more rigorous, or more impactful than a non-preregistered one. Preregistrations allow others to better evaluate studies by being able to detect deviations from pre-specified plans and to differentiate confirmatory from exploratory analyses. They may also reduce the number of data-contingent decisions researchers need to make when performing their studies and thereby reduce the effects of (conscious or unconscious) bias on study outcomes. Our evaluation of preregistration practices in gambling studies suggests that preregistration activity is increasing in the field and improvements in specificity are occurring. Still, improvements in writing preregistrations and reporting the associated studies are necessary if we want to maximise the value of this process and improve the quality of the scientific literature, and we hope the recommendations provided here will be useful for all researchers in achieving these goals, both in gambling-focused research and in science more generally.

Acknowledgements

The research team would like to thank Dr Dylan Pickering and Thomas Swanton for their assistance with reviewing our preregistration specificity scores. We thank Dr Nicola Black for providing constructive feedback on an earlier version of this article and Su Jeong Cho for her assistance with administration tasks.

Conflict of Interest and Funding

Funding for this project was provided by the Division on Addiction to the University of Sydney via a research contract between the Division on Addiction and GVC Holdings, PLC. GVC Holdings is a large international gambling and online gambling operator. GVC had no involvement with the development of our research questions or protocol or development of this preregistration. The authors declare no conflicts of interest in relation to this study.

Disclosures

The authors declare that there are no competing interests associated with this work. There were no constraints on publishing the findings of this study and we did not require approval from Entain PLC (formerly GVC Holdings), PLC to disseminate any of the findings.

Analysis code & *R* packages used

We used R [Version 4.0.2; R Core Team (2020)] and the R-packages *apa* [Version 0.3.3; Gromer (2020); Aust and Barth (2020)], *english* [Version 1.2.5; Fox et al. (2020)], *forcats* [Version 0.5.1; Wickham (2020a)], *ggplot2* [Version 3.3.3; Wickham (2016)], *papaja* [Version 0.1.0.9997; Aust and Barth (2020)], *purrr* [Version 0.3.4; Henry and Wickham (2020)], *readr* [Version 1.4.0; Wickham and Hester (2020)], *stringr* [Version 1.4.0; Wickham (2019)], *tibble* [Version 3.1.2; Müller and Wickham (2020)], *tidyr* [Version 1.1.3; Wickham (2020b)], and *tidyverse* [Version 1.3.0; Wickham et al. (2019)] for all our analyses.

All analysis scripts and datasets can be accessed on our OSF project page (https://osf.io/n8rw3/). We used an RMarkdown script to perform all our analyses and develop the figures presented in this manuscript. A HTML document was generated from this script that presents the analysis code alongside the outcomes generated and annotation detailing process of analysing the data and the decisions made throughout. This manuscript and all tables included were also developed entirely in R using an RMarkdown script and the package *papaja*. We have developed a guide on how to independently reproduce our results and this manuscript using the data and analysis scripts available online (see: https://osf.io/ns32y/; please contact the corresponding author with any questions about this process: robert.heirene@sydney.edu.au; robheirene@gmail.com).

References

Allen, C., & Mehler, D. M. A. (2019). Open science challenges, benefits and tips in early career and beyond. *PLOS Biology*, *17*(5), e3000246. https://doi.org/10.1371/journal.pbio.3000246

Aust, F., & Barth, M. (2020). *Papaja" Create APA manuscripts with R Markdown* [Manual].

Bakker, M., Veldkamp, C. L. S., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, *18*(12), e3000937. https://doi.org/10.1371/journal.pbio.3000937

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Blaszczynski, A., & Gainsbury, S. M. (2019). Editor's note: Replication crisis in the social sciences. *International Gambling Studies*, *19*(3), 359–361. https://doi.org/10.1080/14459795.2019.1673786

Centre for Open Science. (2020). *Impact report 2020; maximizing the impact of science together*.

Chen, T., Li, C., Qin, R., Wang, Y., Yu, D., Dodd, J., Wang, D., & Cornelius, V. (2019). Comparison of Clinical Trial Changes in Primary Outcome and Reported Intervention Effect Size Between Trial Registration and Publication. *JAMA Network Open*, *2*(7), e197242–e197242. https://doi.org/10.1001/jamanetworkopen.2019.7242

Claesen, A., Gomes, S. L. B. T., Tuerlinck, F., Vanpaemel, W., & Leuven, K. (2019). *Preregistration: Comparing Dream to Reality (pre-print)*. https://doi.org/10.31234/osf.io/d8wex

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494–509. https://doi.org/10.1037/0033-2909.114.3.494

Dickersin, K., & Rennie, D. (2003). Registering clinical trials. *JAMA*, *290*(4), 516–523. https://doi.org/10.1001/jama.290.4.516

Fox, J., Venables, B., Damico, A., & Salverda, A. P. (2020). *English: Translate integers into english* [Manual].

Frankenhuis, W. E., & Nettle, D. (2018). Open Science Is Liberating and Can Foster Creativity. *Perspectives on Psychological Science*, *13*(4), 439–447. https://doi.org/10.1177/1745691618767878

Goldacre, B., Drysdale, H., Dale, A., Milosevic, I., Slade, E., Hartley, P., Marston, C., Powell-Smith, A., Heneghan, C., & Mahtani, K. R. (2019). COMPare: A prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*, *20*(1), 118. https://doi.org/10.1186/s13063-019-3173-2

Gromer, D. (2020). *Apa: Format outputs of statistical tests according to APA guidelines* [Manual].

Haven, T. L., & Van Grootel, L. (2019). Preregistering qualitative research. *Accountability in Research*, *26*(3), 229–244. https://doi.org/10.1080/08989621.2019.1580147

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLOS Biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Heirene, R. M. (2020). A call for replications of addiction research: Which studies should we replicate and what constitutes a 'successful' replication? *Addiction Research & Theory*, *0*(0), 1–9. https://doi.org/10.1080/16066359.2020.1751130

Heirene, R. M., & Gainsbury, S. M. (2020). Can the open science revolution revolutionise gambling research? *The BASIS*.

Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools* [Manual].

Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. *PLOS ONE*, *10*(8), e0132382. https://doi.org/10.1371/journal.pone.0132382

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kuperschmidt, K. (2018). More and more scientists are preregistering their studies. Should you? *Science*. https://doi.org/10.1126/science.aav4786

Lakens, D. (2019). *The Value of Preregistration for Psychological Science: A Conceptual Analysis*.

LaPlante, D. A. (2019). Replication is fundamental, but is it common? A call for scientific self-reflection and contemporary research practices in gambling-related research. *International Gambling Studies*, *19*(3), 362–368. https://doi.org/10.1080/14459795.2019.1672768

Livingstone, C., & Cassidy, R. (2014). The problem with gambling research. In *The Conversation*. http://theconversation.com/the-problem-with-gambling-research-31934.

Louderback, E. R., Wohl, M. J. A., & LaPlante, D. A. (2020). Integrating open science practices into recommendations for accepting gambling industry research funding. *Addiction Research & Theory*, 1–9. https://doi.org/10.1080/16066359.2020.1767774

Moher, D., Bouter, L., Kleinert, S., Glasziou, P., Sham, M. H., Barbour, V., Coriat, A.-M., Foeger, N., & Dirnagl, U. (2020). The Hong Kong Principles for assessing researchers: Fostering research integrity. *PLOS Biology*, *18*(7), e3000737. https://doi.org/10.1371/journal.pbio.3000737

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

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Sert, N. P. du, Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A.

(2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9. https://doi.org/10.1038/s41562-016-0021

Müller, K., & Wickham, H. (2020). *Tibble: Simple data frames* [Manual].

Nature Human Behaviour. (2020). *Tell it like it is.* *4*(1), 1–1. https://doi.org/10.1038/s41562-020-0818-9

Norris, S. L., Holmer, H. K., Ogden, L. A., Fu, R., Abou-Setta, A. M., Viswanathan, M. S., & McPheeters, M. L. (2012). *Selective Outcome Reporting as a Source of Bias in Reviews of Comparative Effectiveness [Internet].* Agency for Healthcare Research and Quality.

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration Is Hard, And Worthwhile. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2019.07.009

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

Ofosu, G., & Posner, D. N. (2019). Pre-analysis Plans: A Stocktaking (Pre-Print). *MetaArXiv*. https://doi.org/10.31222/osf.io/e4pum

Ofosu, G., & Posner, D. N. (2020). Do Pre-analysis Plans Hamper Publication? *AEA Papers and Proceedings*, *110*, 70–74. https://doi.org/10.1257/pandp.20201079

R Core Team. (2020). *R: A language and environment for statistical computing* [Manual]. R Foundation for Statistical Computing.

Romano, J., Kromrey, J. D., Coraggio, J., Skowronek, J., & Devine, L. (2006). Exploring methods for evaluating group differences on the NSSE and other surveys: Are the t-test and Cohen's d indices the most appropriate choices? *Nnual Meeting of the Southern Association for Institutional Research*.

Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology*, *10*, 813. https://doi.org/10.3389/fpsyg.2019.00813

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *BMJ (Clinical Research Ed.)*, *340*, c332. https://doi.org/10.1136/bmj.c332

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Stewart, L., Moher, D., & Shekelle, P. (2012). Why prospective registration of systematic reviews makes sense. *Systematic Reviews*, *1*(1), 7. https://doi.org/10.1186/2046-4053-1-7

Vassar, M., Roberts, W., Cooper, C. M., Wayant, C., & Bibens, M. (2020). Evaluation of selective outcome reporting and trial registration practices among addiction clinical trials. *Addiction*, *115*(6), 1172–1179. https://doi.org/10.1111/add.14902

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-Hacking. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01832

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string operations* [Manual].

Wickham, H. (2020a). *Forcats: Tools for working with categorical variables (factors)* [Manual].

Wickham, H. (2020b). *Tidyr: Tidy messy data* [Manual].

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Hester, J. (2020). *Readr: Read rectangular text data* [Manual].

Wohl, M. J. A., Tabri, N., & Zelenski, J. M. (2019). The need for open science practices and well-conducted replications in the field of gambling studies. *International Gambling Studies*, 1–8. https://doi.org/10.1080/14459795.2019.1672769