Pre-print

A call for replications of addiction research: Which studies should we replicate & what constitutes a "successful" replication?

Robert M. Heirene (robert.heirene@sydney.edu.au)

Brain & Mind Centre, University of Sydney, Science Faculty, 94 Mallett St, Camperdown, NSW, 2050.

This is a pre-peer-review article. The final, peer reviewed version has now been published in full in Addiction Research and Theory:

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

A post-print, peer-reviewed version of this article can also be accessed via my personal website:

- https://robheirene.netlify.app

ORCID: orcid.org/0000-0002-5508-7102

Conflicts of interests: None

Funding sources: None

Abstract

Several prominent researchers in the problem gambling field have recently called for high-quality replications of existing gambling studies. This call should be extended to the entire field of addiction research: there is a need to focus on ensuring that the understanding of addiction and related phenomena gained through the extant literature is robust and replicable. This article discusses two important questions addictions researchers should consider before proceeding with replication studies: [1] which studies should we attempt to replicate? And: [2] how should we interpret the findings of a replication study in relation to the original study? In answering these questions, a focus is placed on experimental research, though the discussion may still serve as a useful introduction to the topic of replications for addictions researchers using any methodology.

Introduction

Perhaps more than ever, scientists are spending their time replicating existing studies in addition to conducting research that is novel or exploratory. The most well-known examples of this include the Open Science Collaboration's (2015) replication of 100 psychological studies and Camerer and colleagues' (2018) replication of 21 social science experiments published in Nature and Science. The failure to replicate many of the original findings in these projects has generated concerns regarding the reproducibility of scientific research and highlighted the importance of undertaking replication studies. In the addictions field, there has been little direct recognition of this importance until recently when, in a series editorials published in International Gambling Studies, the journal's Editors (Blaszczynski & Gainsbury, 2019) and others in the field (LaPlante, 2019; Wohl et al., 2019) called for good quality replications of existing gambling research. This call should be extended to the entire field of addiction studies: there is a need to focus on ensuring that the understanding of addiction and related phenomena gained through the existing literature is robust and replicable. The aim of this article is to discuss two important questions addictions

researchers should consider *before* proceeding on this replication crusade: [1] which studies should we attempt to replicate? And: [2] how should we interpret the findings of replication research?[1] In answering these questions, a focus is placed on experimental research, though the discussion may still serve as a useful introduction to the topic of replications for addictions researchers using any methodology.

Which studies should we replicate?

Addictions researchers, like all others, are unavoidably constrained by their available resources and therefore prioritising studies for replication attempts is desirable (Coles et al., 2018). Intuitively, we might consider the studies that have greatly influenced our thinking about addiction[2] as most worth replicating. This approach is logical and consistent with Makel and colleagues' (2012) cautiously proposed heuristic of replicating all studies that have received ≥100 citations to prevent flawed or fraudulent findings from going unchallenged for extended periods of time. Isager (2018) has found replication authors are often motivated by different types of impact, including theoretical, academic, and societal forms. Within the addictions field, theoretically impactful studies requiring replication, for example, might include those investigating nascent developments such as network theories of addictive disorders (e.g., Rhemtulla et al., 2016). The identification of so called "bridge symptoms" and symptoms with high "centrality" (see Fried et al., 2017) by such studies may help us better understand the relationship between addiction and comorbid psychopathologies and identify target symptoms for intervention.

A priority for addictions researchers considering replications may be clinical impact. Studies evaluating novel interventions or screening procedures, for example, should be candidates for replication whenever the original studies observe promising findings. This would streamline the testing process whilst still ensuring the efficacy of the procedures before their implementation in

[1] A focus is placed on answering these questions and not "how to replicate?" as this question has received significant attention in recent discussions (e.g., Wohl et al., 2019; Zwaan et al., 2018).
[2] I use "addiction" here as shorthand for all addiction-related phenomena of interest, including harms, interventions, associated cognitions and so on.

clinical practice. In context of intervention research, direct—as opposed to conceptual—replications (see Nosek & Errington, 2017 for an overview of this distinction) should be prioritised to ensure the reliability of the original finding under the same protocol for administration (Lilienfeld, 2018). In reality, however the process of evaluating a novel intervention or screening method does not always follow such clearly demarcated and linear phases in which a *first* study can be easily identified. In such cases, researchers may wish to first undertake meta-analysis of the existing research to obtain an understanding of the effect, before then—if the line of enquiry appears promising and/or conflicting findings are observed—considering which of the available studies is the most plausible and desirable (based on methodological rigour) to replicate, thereby further contributing to the estimate of the effect.

Isager (2019) provides a simple formula that can be used to quantify replication value (RV): $\text{RV} = \frac{impact}{corroboration}$. This, Isager suggests, could be operationalised in any number of different ways to suit the needs of researchers within a given field; for example, as: $\text{RV} = \frac{citations}{sample\ size}$. Addictions researchers could optimise this formula in several ways to increase its value in identifying and comparing the studies most worth replicating. First, if using the operationalisation above, using a more liberal citation count—such as that provided by Google Scholar—which includes an article's mentions in blogs and websites may be desirable if wishing to detect articles with the widest impact. Second, if considering societal influence, researchers could attempt to further quantify this by investigating the number of times an article is referred to on social media platforms[3]. For example, adding an indicator of societal impact in this way raises the RV of the now famous Rat Park study (Alexander et al., 1978) from $\frac{339}{140\ (n)} = 2.42$, to $\frac{339 + 101\ (mentions\ on\ Twitter\ \&\ YouTube^{4})}{140\ (n)} = 3.14$. The original Rat Park study and follow ups by

[3] This could be more formally achieved using the Altmetric service: https://www.altmetric.com.
[4] The exact details of these calculations are shared on Open Science Framework:

Alexander and colleagues (for an overview of this research line, see: Gage & Sumnall, 2019) showed reduced opiate consumption in rats living in enriched environments compared to those in standard environments. Subsequent research supports the value of environmental enrichment for addiction recovery (e.g., Imperio et al., 2018; Pooriamehr et al., 2017), though the only direct replication of the Rat Park experiment by an external researcher did not support the original findings (Petrie, 1996). Thus, despite subsequent replications lowering the RV of the study as the total participant number has increased, the widespread attention the study still receives online and in media cannot be ignored when considering its RV. However, whether societal impact (and any other form) can be accurately quantified remains uncertain and sound scientific and clinical judgment should also be used to determine RV.

While we should aim to replicate impactful addiction studies such as the Rat Park experiments, we should also be cautious to not exclude studies reporting null findings from those deemed worthy of replication that have—as a result of their negative findings—been less influential and received fewer citations (see Fanelli, 2010). Discussions of reproducibility to date have mostly focused on the concerning number of false positives or Type-I errors that are potentially published in the addiction (Wohl et al., 2019) and scientific literature more broadly (Simmons et al., 2011). These concerns are certainly legitimate (see Ioannidis, 2005) and underpin the need to replicate statistically significant, impactful findings in the addictions field to determine whether they represent false positives. Yet convincing evidence also exists to suggest a considerable number of published articles (~66.7% in psychology) contain at least one Type-II error (Hartgerink et al., 2017). This is unsurprising: in exploratory research where the chances of the null and alternative hypothesis being supported by the data are equally likely, false negatives are statistically more probable (by a ratio of 4:1)[5] than false positives when the conventional level of 80% statistical power is achieved and alpha (α) is set at 0.05. Even if we assume a true effect exists, there is still

[5] Assuming 50% power (the typical level achieved in psychological studies) and the same α, the rate of false negatives to false positives is 10:1! Calculations: false negatives: 1-power [β] x 50; false positives: α x 50.

a 36% probability that one or more non-significant results will be observed in any two studies where power = 80% and $\alpha = .05$[6]. Mixed results should be expected (Lakens & Etz, 2017).

It logically follows that there may exist a proportion of addiction-related studies which have asked important or influential research questions that report false negatives. Thus, it may be appropriate to ask: which studies had the *potential* to greatly influence our thinking about addiction, regardless of outcome? This distinction is only likely to become more pertinent given the increasing acceptance of publishing null findings—preliminary evidence indicates that 60.5% of preregistered and registered report studies are reporting null results, compared with ~5-20% of traditional studies (Allen & Mehler, 2019).

Finally, although the emphasis of this discussion has been placed on considering a study's (potential) impact when determining its RV, several other selection criteria can be used. For example, there may be concerns regarding the suitability or rigour of the original study's methodology (Isager, 2018; Mackey, 2012), or the existing evidence for an effect in a given line of research may be weak (for a details on how to calculate the relevant evidence for an effect using Bayes factors, see Field et al., 2019). Nonetheless, these concerns may be insufficient to motivate replication efforts if the research question(s) under study have limited scope to influence our understanding of addiction. Plausibility is also likely to substantially influence decisions regarding replication, with larger and more complex studies potentially being excluded based on implausibility alone. In such cases, addictions researchers could overcome the limitations of any single research teams by considering multi-lab collaborations, as has been done for replications of classic effects in psychology (Klein et al., 2014).

[6] Calculation: (100- [.80 x .80 x 100]). The probabilty of null effects can be easily calculated for different scenarios using the Shiny app provided by Lakens & Etz, 2017: http://shiny.ieis.tue.nl/mixed_results_likelihood/

How should we interpret the findings of a replication study?

Once we decide which addiction-related studies to replicate, it is important to clearly define what criteria will be used to determine the extent to which the findings of the replication corroborate those of the original study *before* commencing data analysis[7]. Four broad approaches available to addiction researchers for determining this are discussed below[8]. However, before considering which of these approaches to use, researchers may wish gage their confidence in the original outcomes—which may also influence the decision as to whether to replicate. For example, was the study preregistered and are data and materials shared (increasing confidence that results were not selectively reported or *p*-hacked)? Are the statistical procedures used clearly reported? Are outcomes reported accurately according to automated screening tools such as statcheck (Nuijten et al., 2016) and the GRIM test (Brown & Heathers, 2016)? Where data are shared, can the findings be independently reanalysed and reproduced? Asking these questions before choosing to replicate may prevent wasting time when the original finding(s) cannot be relied upon for comparison or when reanalysis suggests no effect exists (Nuijten et al., 2018).

Approach 1: p-values in NHST

The criterion for "successful" replication most consistent with common statistical thinking would be $p < \alpha$ when using traditional Null Hypothesis Significance Testing (NHST). That is, the data are surprising if the null hypothesis is true—that no difference between groups or association exists. Using *p*-values from traditional NHST to determine replication success is appealing as this is likely to be the approach used by the authors of the original study to interpret their findings, yet this approach has several limitations.

[7] Coles et al. (2018) recommend doing this in a pre-registered format in collaboration with the authors of the original study, wherever possible.

[8] As an extensive review of all approaches is beyond the scope of this article, the interested reader is referred to an annotated list of relevant articles and resources shared on this project's Open Science Framework (OSF) page: https://osf.io/5r7a9/.

First, as referred to above, *p*-values can be an unreliable indicator of outcome—particularly in the low powered studies typical of psychological research. Cumming (2008, 2014) has convincingly demonstrated this through statistical simulations of replication scenarios. Cumming shows that if we have with two populations whose normally distributed scores on an outcome differ by 10 points on average and we randomly sample 32 participants from each group 25 times, the resulting *p*s range from <.001 to .76 (12 of 25 are significant at $p < 0.05$), despite the *SD* being held at 20 for all randomly selected subsamples (Cumming, 2008). In these simulations, the 32 participants per group provides 52% power to detect the known population effect size of $d = 0.5$, which is consistent with the average level of power achieved in psychological studies historically (Fraley & Vazire, 2014). Thus, if we were to use $p < 0.05$ as the criterion for replication success in a typical addiction study where a true population effect size of exists, we would only expect ~50% of studies to replicate the original finding[9].

Second, *p*-values are not comparable. A *p* of 0.03 in both an original and replication study do not represent similar findings in terms of the magnitude of the effect studied (Sullivan & Feinn, 2012). Indeed, in large samples one could find an effect size one tenth the magnitude of an original study and still find a similar *p*. Whilst we would be reluctant to label this finding as corroborating the original, using *p*-values as our only method of inference could permit such a conclusion.

Third, and following on from the previous point, if we find a *p* less than our specified α (and are confident that this is not a Type-I error), this tells us little other than an effect is greater than 0 (Murphy et al., 2014). This is evident when samples are large enough to reach statistical significance despite negligible effects (e.g., $d = 0.001$; see Kramer et al., 2014). The limited value of *p* in this sense is problematic for evaluating replication outcomes as researchers are often interested in whether the *magnitude* of the effect is (approximately) replicable.

[9] Cumming (2008) has further calculated that only 12% of the variance in replication *p*s can be explained by the original value, with only smaller *p*s (i.e., $p < .001$) providing useful predictive information about future replications.

Approach 2: Effect sizes and their confidence intervals

A common way to overcome some of the aforementioned limitations of *p*-values is to report effect sizes and their confidence intervals alongside, or in place of, NHST outcomes. Effect sizes inform us of the magnitude of an effect and therefore convey practical information in an intuitive format (Cumming, 2014; Sullivan & Feinn, 2012), particularly when unstandardized versions are reported (Pek & Flora, 2018). By providing an indication of where a study's finding exists along a scale of possible outcomes, effect sizes avoid a dichotomous approach to thinking about study outcomes in which an effect is either true or false depending on whether a value crosses a threshold, as with *p*-values (Cumming & Fidler, 2009). In relation to replication, this allows us to determine the *degree* to which findings from a replication study mirror the original (Camerer et al., 2018; Piper et al., 2019) by calculating point and interval estimates for the difference. Effect sizes can also be compared in a significance testing paradigm using the *Q*-statistic used to study the heterogeneity of effect sizes in meta-analyses, although this approach may lack statistical power to detect meaningful levels of heterogeneity if comparing small numbers of effects (see Hedges & Schauer, 2019).

Effect sizes can also be combined meta-analytically to provide an estimate of the true effect based on the outcomes from both the replication(s) and the original study/studies (Cumming, 2008); although, a consideration of the study quality, as discussed above, should inform how meta-analytic techniques are used and interpreted. If using meta-analysis in this context, addiction researchers can choose between the fixed-effects model, which assumes the same common effect underlies all studies in the meta-analysis, or the random-effects model, which assumes each study's effect size is just one from a normal distribution of effect sizes. The choice between these models should be influenced by the specific context and aims of the meta-analysis. For example, while the random effects models is generally preferred (Borenstein et al., 2010), the fixed-effects model may be preferred if combining the results from only one original and one replication study as two

studies is insufficient for accurately estimating the variance of effect sizes in random-effects meta-analysis (van Aert & van Assen, 2018).

Confidence intervals for effect sizes and other parameters such a means, medians, probabilities (denoted by θ) provide additional information by illustrating the margin of error for the parameter (θ) estimate, offering a range of plausible values based on the *observed* data (Cumming & Fidler, 2009). Put simply, 95% confidence intervals will, on average and over time, contain the true population value of θ 95% of the time (Morey et al., 2016)[10]. Accordingly, seeing whether a replication study's θ fits within the confidence interval boundaries of the original study can be and is used as indicator of replication success (see Camerer et al., 2018; Open Science Collaboration, 2015). However, in replications terms, confidence intervals are not indicators that 95% of future θs will fall within the original interval boundaries[11]. Cumming and Maillardet, (2006) have calculated that 95% confidence intervals predict future θs with approximately 83.4% accuracy; although the predictive value of any individual parameter estimate and its confidence interval will depend on how well it reflects the true population value (which increases with sample size and reliable measurement).

While using effect sizes and their confidence intervals to interpret replication outcomes may be more informative than *p*-values, there are some important considerations and limitations associated with their use in this context. First, effect sizes in the existing literature appear to be inflated estimates. Camerer et al. (2018) found successful replication studies (as determined by $p < .05$) produced effect sizes 74.5% the size of the original, whilst the Open Science Collaboration (2015) found the average effect size across 99 replications (successful & unsuccessful) was half (48.9%) that of the originals. Publication and reporting biases are thought to explain the inflation

[10] See Morey et al. (2016) for an important discussion of the fallacies that surround confidence intervals.
[11] This is because the original parameter estimate and its confidence intervals may be a poor indication of the true population value, particularly if the sample size(s) studied were small.

of effect sizes (Open Science Collaboration, 2015), and thus it may be wise to anticipate smaller effects in replication studies. Researchers should also consider how these biases could lead to over-estimating effects in meta-analysis (see Kvarven et al., 2019).

It is also important to consider how sampling variation, contextual factors (e.g., experimenters, location, date), measurement error, and hidden moderators may have influenced the effect size of a study (Kenny & Judd, 2019; Stanley & Spence, 2014); as well as how these factors and deviations from the original design may influence a replication's effect size. There is increasing recognition that heterogeneity of effect sizes in the published literature is not only large, but larger than would be expected with sampling variation alone (Kenny & Judd, 2019; Stanley et al., 2018); particularly when larger and more consistent effects are studied (Klein et al., 2018). Based on simulated research scenarios, Kenny and Judd (2019) have found when heterogeneity of effect sizes exists, multiple smaller *N* studies (*N* = 100 x 5) may produce more precise estimates (i.e., narrower CIs) of an effect than one large *N* study (*N* = 500) even when only moderate power (69%) is achieved; though the importance of achieving sufficient sample sizes in replications should not be downplayed (Maxwell et al., 2015; Piper et al., 2019). Nonetheless, addiction researchers planning replications may want to consider the possibility of running multiple, multi-site studies to provide the most precise estimate of an effect and thereby test the robustness of the effect to minor variations in protocol and setting. Indeed, evidence of significant heterogeneity in effects has led to concerns regarding the utility of any single replication study (Kenny & Judd, 2019; Maxwell et al., 2015) and the proposal that researchers conducting replications should not be focused on *verifying* an existing finding, but rather on contributing to the overall *estimation* of the true underlying effect (Stanley & Spence, 2014). Thus, from this perspective the terms "successful" and "unsuccessful" replication are misleading (Gelman, 2018).

One approach to accounting for *some* of the heterogeneity in study outcomes when interpreting replication findings is to use the prediction interval method propounded by Patil et al. (2016). Prediction intervals have similar, although not identical, properties to confidence

intervals—given the original θ, a prediction interval can be calculated that is such that 95% of exact replications will produce θs that fall within the interval boundaries. Calculations for prediction intervals incorporate variation in both the original study and replication. For example, for prediction intervals around *r,* the following equation is used: $r_{orig} \pm z_{0.975} \sqrt{\frac{1}{n_{orig}-3} + \frac{1}{n_{rep}-3}}$ (additional information & calculations are provided by Patil and colleagues, 2016). Using prediction intervals to determine whether a replication's findings are consistent with the original can overcome the criticism of using confidence intervals for this purpose, namely that they do not account for sampling error (Kenny & Judd, 2019). A similar approach, also involving the calculations of "prediction intervals" (but using different calculations that estimate the variance of the sampling distribution for the original study), is offered by Spence and Stanley (2016). The authors suggest that these intervals can be used to see if the difference between an original study and replication study's outcomes is consistent with what would be expected due to sampling error alone. They provide open source software and R code for calculating prediction intervals for means, correlations, and Cohen's *d*.

Approach 3: Combining p-values and effect sizes

Significance testing may be more useful for interpreting replication outcomes when used in reference to a specific effect size. This approach encompasses several different methods including small telescopes, minimum effects testing, and equivalence testing. Each of these involve, in varying ways, testing against a Smallest Effect Size of Interest (SESOI), thereby overcoming some of the earlier-mentioned limitations of *p*-values in the traditional NHST context. Testing against a SESOI may always be preferential over testing against an effect of zero (i.e., H_0 in traditional NHST) as any effect under study is highly unlikely to be exactly zero given random variation. The corollary of this is that one can theoretically always find support for H_1 with sufficiently large sample sizes (Murphy et al., 2014), rendering significant outcomes in such circumstances uninformative. Thus, testing against a SESOI better enables falsifiability.

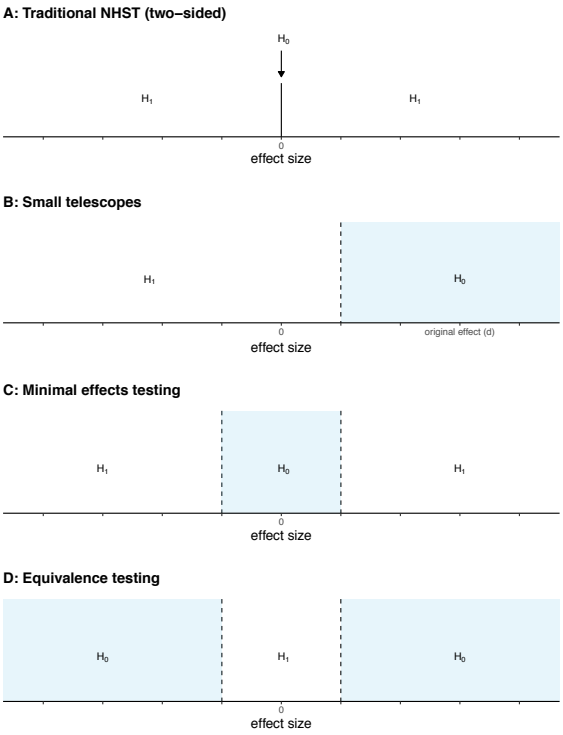

Figure 1. Different approaches to statistical testing: A) Traditional NHST, B) Small telescopes, C) Minimal effects testing, and D) Equivalence testing[12]. Δ_L and Δ_U are lower and upper equivalence boundaries, respectively.

Small telescopes: Proposed by Simonsohn (2015), the small telescopes approach labels a replication as a failure if it rejects an effect that the original study was 33% powered to detect ($d_{33\%}$). For example, if a study has a sample size of 50 participants for a paired-sample *t*-test, it has 33% power (or roughly a 1:2 chance) to detect an effect of $d_{33\%} = 0.22$ (for R code that can be used to calculate $d_{33\%}$, see projects OSF page: https://osf.io/tfmb8/). If a replication study's effect is statistically smaller than $d_{33\%}$, then we can conclude the original evidence to suggest a theoretically interesting effect existed is not convincing and was likely a false positive, *p*-hacked, or fraudulent (Simonsohn, 2015). To use the small telescopes approach, a typical one-sided significance test is used with the null hypothesis (H_0) equivalent to $\geq d_{33\%}$ and an alternative hypothesis (H_1) that $d < d_{33\%}$ (see Figure 1). One limitation of the Small telescopes approach is that it requires a sample size

[12] The R code provided by Lakens et al. (2018) was adapted (with permissions) to create this figure. The exact code used to create Figure 1 is shared on Open Science Framework: https://osf.io/5r7a9/.

roughly 2.5 times the original to detect $d_{33\%}$ and thus may not be applicable when the original study used a large sample. However, as Simonsohn notes, when the sample size of an original study was large, we may be less interested in knowing whether we could achieve $d_{33\%}$ as this could be negligibly small in such circumstances (e.g., $d_{33\%}$ for an independent *t*-test [*N* = 4,000] = 0.049).

Minimum effects testing: Using this method, the SESOI is a range that becomes H_0 (e.g., *d* = -.05 – .05), and effects larger than the SESOI range in either direction (e.g., *d* < -.05 or > .05) are set as H_1 (see Figure 1). Thus, we test whether an effect lays statistically within (H_0) or outside (H_1) a range of effects that would be so small so as to be of no interest if they were found, regardless of whether they reached significance using traditional NHST (Murphy et al., 2014). In replication terms, this option enables researchers to select a SESOI that, if achieved, would represent the minimal level of evidence required to support the original study. Addiction researchers should determine their SESOI based on a consideration of the theoretical or clinical importance of different effect sizes (Lakens, Scheel, & Isager, 2018).

Equivalence testing: This approach can be viewed as the inverse of minimum effects testing. It involves testing whether an effect is statistically within upper and lower equivalence boundaries that represent the SESOI (Lakens et al., 2018). In equivalence testing, H_0 represents a meaningful effect (i.e., larger than the SESOI range in either direction), while effects within the boundaries of the SESOI support H_1 (see Figure 1). Choosing between minimal effects and equivalence testing may depend on whether one wishes to place the burden of proof on replication (when H_1 is supported, findings support the original study) or non-replication (when H_1 is supported, findings do not support the original study; see Hedges & Schauer, 2019). Lakens and colleagues (2018) have developed the TOST (two one-sided significance tests) which can allow addiction researchers to easily include equivalence testing in their studies[13] (available for R, jamovi, excel).

[13] The TOST can also be adapted for use by those wanting to employ the small telescopes approach.

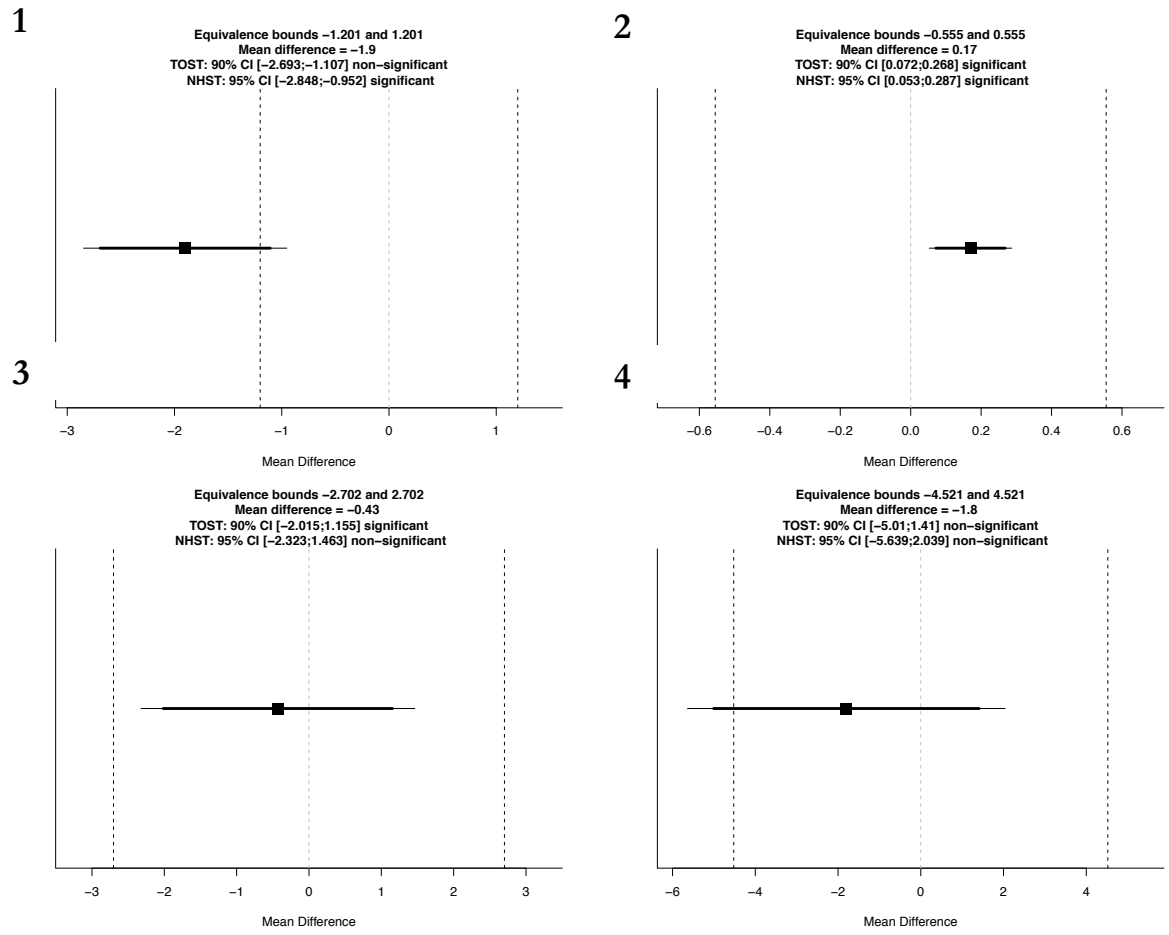

Figure 2. Examples of all possible equivalence testing outcomes from RCTs published in Addiction (panels 1, 3, & 4) & a recent study published in International Journal of Mental health & Addiction (panel 2; no example of this outcome could be identified in trials published in Addiction, likely because this finding typically requires very large sample sizes & small effects)

[1] Lintzeris et al. (2002), comparison of self-reported heroin use after 8 days between intervention (buprenorphine) & control groups; **[2]** Bener et al. (2019), difference between students with internet addiction disorder (1) vs. "normal" students (2) on question 4 of the 14-item Fatigue Scale ("Do you have problems starting things?"); **[3]** Petrakis et al. (2017), comparison of mean no. of drinks per day during 30-day follow-up between intervention (Mecamylamine) & placebo groups; **[4]** Grønbæk & Nielsen (2007), drinks per day comparison between Minnesota day clinic patients vs. standard public psychotherapy treatment patients in the 30-days post intervention.

Lakens et al. (2018) recommend combining equivalence testing with traditional NHST, resulting in four possible outcomes (see corresponding panels in Figure 2): [1] clear evidence of effect: not statistically equivalent and significantly different from zero (i.e., H_0 in NHST), [2] the effect is smaller than the SESOI, though is greater than zero: statistically equivalent and

significantly different, [3] no evidence of effect: statistically equivalent and not statistically different [4] the SESOI cannot be rejected, but the effect is not significant using NHST: not statistically equivalent and not significantly different. Figure 2 plots the findings from a reanalysis of the outcomes of three randomised control trials published in Addiction (panels 1, 3, & 4) and a study recently published in International Journal of Mental health & Addiction (panel 2). A medium effect size (d = 0.5)[14] was used for upper and lower boundaries for ease of illustration, though addictions researchers should give proper consideration to the equivalence boundaries that would be appropriate for their research (see Anvari & Lakens, 2019; Lakens et al., 2018). Each of the outcomes presented in Figure 2 is more informative than using NHST alone, and outcome 4 may be of particular interest in the context of replication research. This outcome indicates that although not statistically significant using NHST, the effect *may* be equal to or larger than the SESOI—which in this case is a noteworthy effect size of d = 0 .5—and therefore the original study may have lacked sufficient power to detect a meaningful effect. Going forward, addictions researchers can use the TOST, as done here, to determine whether other null findings reported in their subfields may require further investigation (as has been done for other fields: Quintana, 2018).

Maxwell et al. (2015) state that equivalence testing can be used to determine a range of values that, if a θ and its CI were observed within (they propose ranges for d of -0.10 to 0.10 or -0.05 to 0.05), would support the conclusion that the null hypothesis is "*for all intents and purposes essentially true*". Thus, if the range of parameter values observed in a replication study is within the equivalence boundaries, then strong evidence for the absence of an effect has been found and any effect(s) observed in previous investigations should be questioned. However, to classify an effect as equivalent to null based on such small ranges would require inordinately large sample sizes for addiction research. For example, to achieve 80% power for a between groups t-test using equivalence boundaries of -0.10 and 0.10 (d) and assuming the true effect is 0, a total of 3,426

[14] Cohen's d effect size was used to set boundaries here (hence the difference in scales along the x-axes in the plots), though raw difference scores can also be used for this purpose.

participants would be required (1,713 per group). For boundaries of -0.05 and 0.05 under the same conditions, 13,704 participants would be required (6,852 per group). The solution to this issue as recommended by Maxwell and colleagues (2015) echoes a recommendation made earlier in this article: addiction researchers must work collaboratively, undertaking multiple replication studies across multiple labs[15]. Only through such collaborations can we achieve the necessary sample sizes required to provide convincing evidence for the absence of an effect.

Approach 4: Bayesian statistics

A fourth approach is the use of Bayesian statistics and the calculation of Bayes Factors (BF). In NHST terms, BFs provide an indication for the relative evidence for H_0, and H_1 using likelihood ratios and therefore may be favoured over *p*-values. BFs also incorporate researchers' prior beliefs about the theory under study, which is included into the statistical model as a *prior* (e.g., one's prior belief about the probability of a coin landing on heads in a coin toss is likely to be 0.5). Once the data are observed, one's degree of belief in the theory is then updated accordingly into a *posterior belief* in the form of the resulting BF. According to convention, BFs > 3 represent substantial support for H_1 over H_0, BFs < 0.333 (or 1/3) represent substantial support for H_0 over H_1, and BFs between 0.333 and 3 reflect insufficient evidence for either hypothesis. While using a Bayesian approach to statistical inference was once seen as a complicated and unusual approach, such methods are becoming increasingly more common (e.g., Pisklak et al., 2019) and user-friendly software (e.g., JASP) has been developed that can now easily calculate BFs. Indeed, the flagship journal of the addictions field, Addiction, now recommends all authors report BFs in studies reporting null findings, with the suggestion that only BFs < 0.3 should be interpreted as representing no evidence of effect (i.e., support for H_0 over H_1).

[15] The upcoming SCORE (Systematizing Confidence in Open Research & Evidence) project by the Centre for Open Science and DARPA (Defence Advanced Research Projects Agency) provides another example of such multi-lab collaborations efforts in the wider social and behavioural sciences.

Bayesian methods and BFs possess a number of properties that make them useful for interpreting replication outcomes (Dienes, 2014, 2016), some of which mirror those provided by the methods outlined in Approach 3 above. For example, unlike in traditional NHST, researchers can clearly specify H_0 and H_1, potentially specifying H_1 as the original study's (or meta-analysis) finding for comparison. Such an approach has been advocated by Verhagen and Wagenmakers (2014) and then updated by Ly et al. (2018), who propose evaluating replication outcomes from both a sceptic's (i.e., no effect exists) and proponent's (i.e., the original study's finding reflects the true effect) perspective. The relevant evidence for the sceptic's perspective (H_o) over the proponent's (H_r) can then be calculated.

In addition to such model comparisons, Bayesian statistics can also be used for estimating parameters and their variance in a similar way to frequentist approaches (Kruschke, 2011). The 95% *credible* interval calculated for parameters in Bayesian statistics contain 95% of estimated parameter values that are most plausible based on the observed data. In an approach analogous to equivalence testing in frequentist statistics, researchers using Bayesian approach can identify a Region of Practical Equivalence (ROPE) that contains the range of values for θ we consider to be equivalent to null (Kruschke, 2011) and observe whether the credible interval within, outside, or partially within the ROPE. This approach therefore allows researcher to consider the *proportion* of θ values equivalent to null as well as making categorical interpretations, and again can provide convincing evidence for the absence (entire interval falls within the ROPE) or presence (entire interval falls outside of the ROPE) of an effect in the context of replication research.

Conclusions

There is a need for addiction researchers to conduct replication studies to ensure the veracity of our understanding surrounding addiction and related phenomena. Selecting appropriate studies for replication should involve taking into consideration each study's potential impact from theoretical, academic, societal, and clinical viewpoints. Consideration should also be given to how

the results of replication studies will be interpreted and compared with the original finding. A variety of approaches are available for this purpose, including effect sizes and their confidence intervals, interval predictions, small telescopes, and equivalence testing. In line with the Open Science Collaboration's (2015) view that "*No single indicator sufficiently describes replication success*", using several of these methods together will allow for the most comprehensive interpretation of findings. Finally, it is essential that addiction researchers recognise that multiple sources of variation can affect each study's outcome and therefore: [1] exact replication of an effect is unrealistic, and [2] one "failed" replication should not be cause to declare the original effect does not exist—the addictions field must adopt meta-scientific approach wherein the accumulation of evidence is prized and single findings are interpreted cautiously.

Supplemental materials

Supplemental materials, including an annotated list of relevant articles and resources relating to the interpretation of replication outcomes and the R code used to produce the figures presented here can be accessed via this project's Open Science Framework Project Page: https://osf.io/5r7a9/

Acknowledgements

I would like to thank Geoff Cumming, John Stapleton, Dylan Pickering, and Peder Isager for their helpful comments on earlier versions of this article. I would also like to thank Daniel Lakens and Peder Isager for kindly sharing the R code that was adapted to produce the figures presented here.

References

Alexander, B. K., Coambs, R. B., & Hadaway, P. F. (1978). The effect of housing and gender on morphine self-administration in rats. *Psychopharmacology*, *58*(2), 175–179. https://doi.org/10.1007/BF00426903

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

Anvari, F., & Lakens, D. (2019, pre-print). Using Anchor-Based Methods to Determine the Smallest Effect Size of Interest. https://doi.org/10.31234/osf.io/syp5a

Bener, A., Yildirim, E., Torun, P., Çatan, F., Bolat, E., Alıç, S., … Griffiths, M. D. (2019). Internet addiction, fatigue, and sleep problems among adolescent students: A large-scale study. *International Journal of Mental Health and Addiction*, *17*(4), 959–969. https://doi.org/10.1007/s11469-018-9937-1

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

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm.12

Brown, N. J. L., & Heathers, J. A. J. (2016). The GRIM Test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, *8*(4), 363–369. https://doi.org/10.1177/1948550616673876

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

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

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*(4), 286–300. https://doi.org/10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift Für Psychologie/Journal of Psychology*, *217*(1), 15–26. http://dx.doi.org/10.1027/0044-3409.217.1.15

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: where will the next mean fall? *Psychological Methods*, *11*(3), 217–227. https://doi.org/10.1037/1082-989X.11.3.217

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781–781. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2016). How Bayes factors change scientific practice. *Bayes Factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments*, *72*, 78–89. https://doi.org/10.1016/j.jmp.2015.10.003

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PloS One*, *5*(4), e10271–e10271. https://doi.org/10.1371/journal.pone.0010271

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

Fraley, R. C., & Vazire, S. (2014). The N-Pact Factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLOS ONE*, *9*(10), e109019. https://doi.org/10.1371/journal.pone.0109019

Fried, E. I., van Borkulo, C. D., Cramer, A. O. J., Boschloo, L., Schoevers, R. A., & Borsboom, D. (2017). Mental disorders as networks of problems: a review of recent insights. *Social Psychiatry and Psychiatric Epidemiology*, *52*(1), 1–10. https://doi.org/10.1007/s00127-016-1319-z

Gage, S. H., & Sumnall, H. R. (2019). Rat Park: How a rat paradise changed the narrative of addiction. *Addiction*, *114*(5), 917–922. https://doi.org/10.1111/add.14481

Gelman, A. (2018). Don't characterize replications as successes or failures. *Behavioral and Brain Sciences*, *41*, e128. https://doi.org/10.1017/S0140525X18000638

Grønbæk, M. & Nielsen, B. (2007). A randomized controlled trial of Minnesota day clinic treatment of alcoholics. *Addiction*, 102: 381-388. doi:10.1111/j.1360-0443.2006.01700.x

Hartgerink, C. H. J., Wicherts, J. M., & Van Assen, M. A. L. M. (2017). Too Good to be False: Nonsignificant Results Revisited. *Collabra: Psychology*, *3*(1). https://doi.org/10.1525/collabra.71

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

Imperio, C. G., McFalls, A. J., Hadad, N., Blanco-Berdugo, L., Masser, D. R., Colechio, E. M., … Freeman, W. M. (2018). Exposure to environmental enrichment attenuates addiction-like behavior and alters molecular effects of heroin self-administration in rats. *Neuropharmacology*, *139*, 26–40. https://doi.org/10.1016/j.neuropharm.2018.06.037

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Isager, P. M. (2018). What to replicate? Justifications of study choice from 85 replication studies. Zenodo. http://doi.org/10.5281/zenodo.1286715

Isager, P. M. (2019). Quantifying Replication Value: A formula-based approach to study selection in replication research. ZPID (Leibniz Institute for Psychology Information). https://doi.org/10.23668/psycharchives.2392

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, *24*(5), 578–589. https://doi.org/10.1037/met0000209

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating Variation in Replicability. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., … Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788. https://doi.org/10.1073/pnas.1320040111

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312. https://doi.org/10.1177/1745691611406925

Kvarven, A., Strømland, E., & Johannesson, M. (2019). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-019-0787-z

Lakens, D., & Etz, A. J. (2017). Too true to be bad: When sets of studies with significant and nonsignificant findings are probably true. *Social Psychological and Personality Science*, *8*(8), 875–881. https://doi.org/10.1177/1948550617693058

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

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*, 1–7. https://doi.org/10.1080/14459795.2019.1672768

Lilienfeld, S. O. (2018). Direct replication and clinical psychological science. *Behavioral and Brain Sciences*, *41*, e140. https://doi.org/10.1017/S0140525X18000754

Lintzeris, N., Bell, J., Bammer, G., Jolley, D. J. & Rushworth, L. (2002), A randomized controlled trial of buprenorphine in the management of short-term ambulatory heroin withdrawal. Addiction, 97: 1395-1404. doi:10.1046/j.1360-0443.2002.00215.x

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2018). Replication Bayes factors from evidence updating. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1092-x

Mackey, A. (2012). Why (or why not), when and how to replicate research. In: G. Porte (Ed.), Replication research in applied linguistics (pp. 34–69). Cambridge, UK: Cambridge University Press.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research: How Often Do They Really Occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *The American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

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

Murphy, K. R., Myors, B., & Wolach, A. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests (4th ed.)*. New York, USA: Routledge/Taylor & Francis Group.

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

Nuijten, M. B., Bakker, M., Maassen, E., & Wicherts, J. M. (2018). Verify original results through reanalysis before replicating. *Behavioral and Brain Sciences*, *41*, e143. https://doi.org/10.1017/S0140525X18000791

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

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

Patil, P., Peng, R. D., & Leek, J. T. (2016). What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *11*(4), 539–544. https://doi.org/10.1177/1745691616646366

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. - PubMed - NCBI. *Psychologial Methods*, *23*(2), 208–225. https://doi.org/10.1037/met0000126

Petrakis, I. L., Ralevski, E., Gueorguieva, R., O'Malley, S. S., Arias, A., Sevarino, K. A., Jane, J. S., O'Brien, E., & Krystal, J. H. (2018). Mecamylamine treatment for alcohol

dependence: a randomized controlled trial. *Addiction*, 113: 6– 14. doi: 10.1111/add.13943.

Petrie, B. F. (1996). Environment is not the Most Important Variable in Determining Oral Morphine Consumption in Wistar Rats. *Psychological Reports*, *78*(2), 391–400. https://doi.org/10.2466/pr0.1996.78.2.391

Piper, S. K., Grittner, U., Rex, A., Riedel, N., Fischer, F., Nadon, R., … Dirnagl, U. (2019). Exact replication: Foundation of science or game of chance? *PLoS Biology*, *17*(4), e3000188–e3000188. https://doi.org/10.1371/journal.pbio.3000188

Pisklak, J. M., Yong, J. J. H., & Spetch, M. L. (2019). The Near-Miss Effect in Slot Machines: A Review and Experimental Analysis Over Half a Century Later. *Journal of Gambling Studies*. https://doi.org/10.1007/s10899-019-09891-8

Pooriamehr, A., Sabahi, P., & Miladi-Gorji, H. (2017). Effects of environmental enrichment during abstinence in morphine dependent parents on anxiety, depressive-like behaviors and voluntary morphine consumption in rat offspring. *Neuroscience Letters*, *656*, 37–42. https://doi.org/10.1016/j.neulet.2017.07.024

Quintana, D. S. (2018). Revisiting non-significant effects of intranasal oxytocin using equivalence testing. *Psychoneuroendocrinology*, *87*, 127–130. https://doi.org/10.1016/j.psyneuen.2017.10.010

Rhemtulla, M., Fried, E. I., Aggen, S. H., Tuerlinckx, F., Kendler, K. S., & Borsboom, D. (2016). Network analysis of substance abuse and dependence symptoms. *Drug and Alcohol Dependence*, *161*, 230–237. https://doi.org/10.1016/j.drugalcdep.2016.02.005

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

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

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

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*(12), 1325–1346. https://doi.org/10.1037/bul0000169

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size-or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

van Aert, R. C. M., & van Assen, M. A. L. M. (2018). Examining reproducibility in psychology: A hybrid method for combining a statistically significant original study and a replication. *Behavior Research Methods*, *50*(4), 1515–1539. https://doi.org/10.3758/s13428-017-0967-6

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

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

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