Published in final edited form as:
Psychol Addict Behav. 2022 May ; 36(3): 260–270. doi:10.1037/adb0000730.

Examining Replicability in Addictions Research: How to Assess, Ways Forward

Matthew R. Pearson[1],
Center on Alcohol, Substance use, and Addictions, University of New Mexico

Frank J. Schwebel,
Center on Alcohol, Substance use, and Addictions, University of New Mexico

Dylan K. Richards,
Center on Alcohol, Substance use, and Addictions, University of New Mexico

Katie Witkiewitz
Center on Alcohol, Substance use, and Addictions, Department of Psychology, University of New Mexico

Abstract

Objective: The high rate of statistically significant findings in the sciences that do not replicate in a new sample has been described as a "replication crisis." Few replication attempts have been conducted in studies of alcohol use disorder (AUD) and the best method for determining whether a finding replicates has not been explored. The goal of the current study was to conduct direct replications within a multisite AUD randomized controlled trial and to test a range of replication methods.

Method: We used data from a large AUD clinical trial (Project MATCH, n=1726) to simulate direct replication attempts. We examined associations between drinking intensity and negative alcohol-related consequences (Model 1), sex differences in drinking intensity (Model 2), and reductions in drinking following treatment (Model 3). We treated each of the 11 data collection sites as unique studies such that each subsample was treated as an "original" study, and the remaining 10 subsamples were viewed as "replication" studies. Replicability metrics included consistency of statistical significance, overlapping confidence intervals, and consistency of the direction of the effect. We also tested effect heterogeneity using meta-analysis.

Results: We observed between 0–100% replicability across the replicability metrics depending on which subsample was treated as the "original" study. Meta-analyses indicated results were more similar across subsamples with no significant heterogeneity for Models 1 and 2.

[1] Correspondence concerning this article should be addressed to Matthew R. Pearson, Center on Alcohol, Substance use, and Addictions, University of New Mexico, 2650 Yale SE MSC11-6280, Albuquerque, NM 87106, USA. mateo.pearson@gmail.com.
Author Note
Matthew R. Pearson, Frank J. Schwebel, and Dylan K. Richards, Center on Alcohol, Substance use, and Addictions, University of New Mexico; Katie Witkiewitz, Center on Alcohol, Substance use, and Addictions, Department of Psychology, University of New Mexico.

A portion of these results were presented at the virtual poster session at the 43rd Annual Scientific Meeting of the Research Society on Alcoholism (Schwebel, O'Sickey, Richards, Pearson, & Witkiewitz, 2020) and as part of a symposium presented at the 2021 Collaborative Perspectives on Addiction meeting (Richards, Schwebel, Pearson, & Witkiewitz, 2021).

Conclusions: We recommend researchers focus on effect sizes and use meta-analysis to evaluate level of replicability. We also encourage direct replication attempts and sharing of data and code to facilitate direct replication.

Keywords

Replicability; Addictions; Alcohol; Null Hypothesis Significance Testing; Meta-Analysis

Introduction

The sciences have been dealing with what many call a "replicability crisis," characterized by a high rate of replication failures. For example, the Open Science Collaboration (2015) conducted replication attempts of 100 experimental and correlational studies selected from three high impact psychology journals (*Psychological Science*, *Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition*) and found that, although 97% of the original studies had statistically significant effects, only 36% of the replications had significant results. Rate of replication ranged from 39% (subjective agreement on replication) to 47% (based on overlapping confidence intervals) depending on the method used, but either approach seems to support a stark and troubling conclusion: many findings do not replicate (i.e., "a replication crisis"). Regrettably, clinical psychology (including addiction science) has been slower to engage in reform efforts compared to other subfields of psychology, although there is substantial interest in these reform efforts among clinical psychologists (Tackett, Brandes, King, & Markon, 2019). Additionally, there has been a recent emphasis on replication seen in the field of addiction science (Heirene, 2020) driven, in large part, by a series of editorials published in the International Gambling Studies journal (Blaszczynski & Gainsbury, 2019; LaPlante, 2019; Wohl, Tabri, & Zelenski, 2019). Low replication rates have been attributed to questionable research practices used to obtain statistical significance (e.g., choosing among dependent variables; Simmons, Nelson, & Simonsohn, 2011) and systemic issues in psychological science such as valuing novel findings over replication (Open Science Collaboration, 2015). However, this issue remains contentious in that some have argued the level of crisis is overblown (Maxwell, Lau, & Howard, 2015; Pashler & Harris, 2012), based on false assumptions (Fiedler & Prager, 2018; Stroebe & Strack, 2014) and/or poor metrics of measuring replicability (Wilson, Harris, & Wixted, 2020).

Tackett and Miller (2019) generally define replicability as "a core principle of an objective, empirical science and reflects the extent to which scientific findings that are published hold up to independent replication" (p. 488). Tackett et al. (2019) argue that attempts to define and operationalize replication are necessary to improve clinical science. However, as with any statistical analysis, there are multiple ways to approach the analysis of replications (Zwaan et al., 2020), and thus there is currently no consensus of how best to define a successful replication. What is clear, however, is the concern among scientists about the reproducibility of research. A survey of more than 1,500 scientists across disciplines found that the vast majority (90%) believed that there is a reproducibility crisis (Baker, 2016). Further, irreproducibility was largely attributed to questionable research practices (e.g., selective reporting) and systemic issues with science (e.g., pressure to publish). In light of these concerns, the purpose of the present paper was to provide an overview of issues regarding replicability and reproducibility in addiction psychology and to conduct a case study to directly compare indices of replication that have been used in previous replication attempts (e.g., Open Science Collaboration, 2015).

Often, the terms "direct replication" and "conceptual replication" are used to distinguish replication attempts with the former referring to attempts to replicate a finding when following an original study's method as exact as possible, and the latter referring to attempts to replicate a finding when using a different method (see Nosek and Errington (2017) for further discussion). An important caveat, however, is that it is possible to conduct a secondary data analysis using combined data from an original study and a conceptual replication, reducing the discrepancies in methodology between the two studies, such as by applying the same statistical analysis, and thereby testing a direct replication. Such procedures are generally referred to as integrative data analysis, which we describe at length in the discussion section. Conceptual replications are more common than direct replications and can hold great promise in the face of consistent results. Stated differently, when diverse methods lead to the same or similar conclusions, then confidence in the findings should increase. Unfortunately, what is considered a conceptual replication depends on the theory driving the research, and all of its auxiliary assumptions (Trafimow, 2020). Thus, a conceptual replication is defined by the eye of the theorist. Beyond this definitional issue, when diverse methods lead to different conclusions, then confidence in the findings should decrease, and it is typically not decipherable what exactly accounts for the inconsistency, and whether it reflects random or systematic variation. This limitation of conceptual replications highlights the need for direct replication, which minimizes the number of factors that could account for failing to replicate findings.

In this manuscript, we use data from Project Matching Alcoholism Treatments and Client Heterogeneity (Project MATCH; Project MATCH Research Group, 1993), one of the largest clinical trials of alcohol use disorder (AUD) treatment ever conducted, to simulate direct replications attempts and assess replication using several different metrics described below (Heirene, 2020). Specifically, this multi-site trial involved 11 distinct data collection sites, and thus we treat each data collection site as an independent sample to evaluate replicability of specific research questions across these subsamples. Importantly, the aim of the present study addresses the *defining and operationalizing replication* domain of reform to improve clinical science proposed by Tackett et al. (2019).

How do we define replication? Subjective Assessment, Replicability Indices, and Meta-Analytic Approaches

Subjective assessment of replicability reflects the replication researchers' decision of whether results of an original study were replicated (cf. (Open Science Collaboration, 2015)). The subjective nature of this decision seems inherently problematic—it has the potential to introduce bias in the decision-making process. As such, researchers attempting to study replicability have developed a number of indices to assess the degree to which study findings replicate in a new study (or new sample). There is no single index for replicability, so we discuss the strengths and weaknesses of different indices that have been proposed and that are commonly used.

Pattern of statistical significance.—The most direct and obvious test of whether a particular finding replicates is to examine the pattern of statistical significance across studies, such that an original study and the replication study make the same decision in terms of rejecting or failing to reject the null hypothesis. With null hypothesis significance testing (NHST; (Trafimow & Earp, 2017)), the researcher sets the rate of Type I error that they are willing to accept (typically 5%, i.e., $p < .05$) and most commonly tests their data against a null hypothesis that there is no relationship, no differences among means, etc. The statistical test produces a *p*-value that represents the probability that the researcher would obtain a relationship/difference of the size observed or more extreme if the null hypothesis was actually true [P(data|null)]. Strengths of this approach relate to one of the (few) strengths of NHST, which is that the decision to reject/fail to reject the null hypothesis is set *a priori* ($p < .05$), and this produces a simple, binary decision.

Among the many limitations of NHST (Cohen 1994; Masson, 2011), NHST is designed to test the rejection of a null hypothesis and not the *acceptance* of a null hypothesis. Observing a result that does not reject the null hypothesis does not disconfirm a near infinite number of alternative explanations that could account for not rejecting the null hypothesis. In the context of NHST, non-significant results cannot truly “replicate” because if you fail to reject the null hypothesis (i.e., not statistically significant), you cannot conclude one way or the other whether the null hypothesis is valid (Trafimow, 2019). Using a metric of replication that only works for statistically significant results seems inherently problematic.

Confidence intervals.—Related to the *p*-value is the confidence interval. Confidence intervals can be used for NHST, but unlike a single *p*-value, confidence intervals provide a range of possible values and an estimate of precision (i.e., width of the confidence interval) (Cumming, 2012). Although often misinterpreted, 95% confidence intervals do not reflect any level of confidence that the actual population parameter falls within its lower and upper bounds. Rather, if one were to repeat an experiment indefinitely, we would expect to observe that 95% of the confidence intervals would contain the actual population parameter. Thus, the optimal way to use confidence intervals for replicability testing would be to conduct an experiment repeatedly (100+ times), and the true population parameter would be expected to fall within the bounds of 95% of the confidence intervals. Given this is exceedingly impractical, we doubt we will see this practice adopted by addiction science.

There are two other ways that confidence intervals can be used as a measure of replicability. One can test either whether estimates of effects in replication studies fall within the bounds of the original study’s confidence interval, or whether the estimates of effects in the original study falls within the bounds of the replication study’s confidence interval (as was done by Open Science Collaboration (Open Science Collaboration, 2015)). Unfortunately, both rely on an assumption that 95% confidence intervals provide us with something that they do not, that is, 95% confidence that the population parameter falls within the bounds of the confidence interval (Cumming & Maillardet, 2006). Although intuitively it makes sense to treat an original study as the referent to which a replication study is compared, the decision of making either the original study or the replication study the comparator is somewhat arbitrary. From the perspective that studies are sampling effect size estimates from a population, no one study should hold prominence over another. One suggestion to mitigate limitations of using confidence intervals is to conduct multiple, multi-site studies to provide a more precise estimate of effect by testing the robustness of the effect to variations in setting and protocol that would be expected when conducting multi-site studies (Heirene, 2020). It should be noted that Bayesian credible intervals do attempt to provide the bounds of the actual population parameter (Gokhale, Box, & Tiao, 1974). Although this manuscript focuses on frequentist methods that are most commonly used in addictions research, Bayesian statistics reflect a reasonable alternative approach, but is outside of the scope of the work addressed here (Gelman, 2008).

Type S.—One conceptually simpler indicator of replicability is whether an observed effect in a replication study is in the same sign (positive or negative) as the original study, which we refer to as Type S replicability (for a discussion of Type S errors, see (Gelman & Carlin, 2014)). Benefits of this index include avoiding some of the pitfalls of NHST. The greatest strength and greatest weakness of Type S replicability as a single index of replicability is that it ignores effect size. If we observe a positive effect in an original study, and a replication study shows a negative effect, intuitively, this finding seems sufficient to decide that the effect has not replicated, regardless of statistical significance or effect size. However, an effect has to be either positive or negative to some degree, as Meehl states, “the null hypothesis, taken literally, is always false” (p. 822, (Meehl, 2004)). We will rarely, if never, observe an exact zero in addiction science. Thus, similar to the problem of NHST methods, if the null hypothesis is practically true, then it is not possible for Type S to demonstrate that the null effect replicates as we would expect 50% of effects to sway positive and 50% of effects to sway negative.

Although many believe it is uncommon that scientists wish to prove a null hypothesis, it is not the case. Consider a standard randomized controlled trial. First, many studies inappropriately test for baseline equivalence on relevant variables, so a series of statistical tests are often used with the hope that none of these are significant (de Boer, Waterlander, Kuijper, Steenhuis, & Twisk, 2015; Hawkins et al., 2008). Next, researchers often wish to demonstrate that their retained sample does not significantly differ from the full sample (or those who dropped out), necessitating a series of additional statistical tests that the researcher again hopes to be non-significant. Thus, quite often researchers expect and hope that the null hypothesis is true for basic assumption checks to support their primary study aims. Further, in the case of multiple different treatments being delivered (i.e., comparative effectiveness designs, non-inferiority trials), knowing that treatment X is no more or less effective than treatment Y can be incredibly important (i.e., one treatment is new, one treatment is costlier, etc.). In these cases, equivalence testing like the two one-side tests (TOST) procedure are more appropriate (see Lakens et al., 2018).

Meta-analysis.—Using meta-analysis to determine level of replicability has several notable strengths compared to the previous indices. At a basic level, meta-analysis involves combining data across studies. This quantitative synthesis allows for directly examining and modeling the degree of consistency of effects across studies. In terms of determining replicability, there are several distinct ways in which meta-analysis can be used that need to be carefully considered. First, we must consider if a fixed effects or random effects meta-analysis is most appropriate. In a fixed effects model, we assume that there is a single population effect size, and that sampling bias is the only factor causing variability across studies. In a random effects model, we assume that there is a distribution of true population effect sizes, so sampling bias is only one factor that causes variability across studies (Hedges & Vevea, 1998; Hunter & Schmidt, 2000). The Open Science Collaboration (2015) used fixed effect meta-analysis, which is seldomly used given that there are few examples in behavioral sciences where our theoretical model is so strong that it predicts an exact effect size that would be expected to be invariant across samples/studies. Importantly, the way meta-analysis is conducted now, it still relies on NHST to determine statistical significance, though typically with a much higher level of precision than a primary study. The most common way of demonstrating effect sizes is through a forest plot showing confidence intervals across individual studies and constructing a confidence interval around the overall effect size estimate. Thus, many of the critiques of NHST apply to meta-analysis as well. For example, a meta-analytic approach (e.g., *Q*-statistic) may lack statistical power to detect meaningful levels of heterogeneity if comparing a small number of effects (Hedges & Schauer, 2018). When conducting a study, we can mistake the sign/direction of an effect (Type S error) and/or the size/magnitude of the effect (Type M error) (Gelman & Carlin, 2014). One strength of meta-analysis is that it can be used to test for both errors and can even investigate potential moderators that could account for these errors.

Replications Published in *Psychology of Addictive Behaviors*

Although many studies could be described to some extent as replication studies in the case that research questions are addressed that are similar to previous studies, we performed a cursory search of direct mentions of the word “replication” (without year restriction) appearing in the journal *Psychology of Addictive Behaviors*, which in July 2020 yielded 21 results. Of these papers, one study tested a within-subject replication using an animal model of alcohol self-administration (Cook et al., 2019), five studies reported measurement replications in new samples (Alterman et al., 1998; Isenhart & Silversmith, 1996; Jones, Spradlin, Joe Robinson, & Tragesser, 2014; Lee & Leeson, 2015; Pantalon, Nich, Frankforter, & Carroll, 2002), one used data from the Relapse Replication and Extension Project (Tonigan, McCallion, Frohe, & Pearson, 2017), five studies mentioned the importance of replication in future work (Hernández-López, Luciano, Bricker, Roales-Nieto, & Montesinos, 2009; Miller, Dibello, Lust, Meisel, & Carey, 2017; Schüz, Eid, Schüz, & Ferguson, 2016; Stanger et al., 2019; Von Sternberg, DiClemente, & Velasquez, 2018), and nine studies reported the results of replication efforts. Seven of these nine studies interpreted their results as replicating or partially replicating previous findings using either subjective assessment or consistency of statistical significance, though the method for determining replication was not always made explicit. By our assessment, three of the nine studies (Budney et al., 2015; Carney & Kivlahan, 1995; Hughes et al., 2016), replicated findings based on subjective assessment of replicating the prior findings. Four of the nine studies (Edens & Willoughby, 2000; Ham & Hope, 2006; Motschman & Tiffany, 2016; Rychtarik et al., 2017) more explicitly mentioned NHST to determine replication.

Two of the papers explicitly discussed the replication crisis and the importance of replication (Mun et al., 2015; Rothenberg et al., 2019). Rothenberg and colleagues (2019) reported a replication of the intergenerational transmission of cannabis use and a replication of the associations between coping, grades, parental AUD, and previous adolescent cannabis use in predicting cannabis use across intergenerational cohorts, where replication was defined as an original effect falling within the 95% confidence interval of the replication effect. Finally, Mun and colleagues (Mun et al., 2015) used an integrative data analysis meta-analytic approach to conduct a "built-in replication" (p. 35) of the efficacy of brief motivational interventions in reducing alcohol use among college student populations.

An Example: Clinical Trial of Treatments for Alcohol Use Disorder

We use data from one of the largest clinical trials conducted in the addictions field, to date, to serve as a case example of metrics for direct replications. By using data from Project MATCH (Project MATCH Research Group, 1993), we take advantage of the uncharacteristically large sample size of this multi-site randomized clinical trial (RCT). Given that participants were recruited by 11 clinical research units throughout the country with wide-ranging samples sizes, but these sample sizes are within the range of typical clinical trials ($57 < n < 226$), we reasoned that one sample could be treated as an "original study" and the other 10 sites as "replication studies" (similar to an approach used by Tonigan (2001) examining the effects of 12-step involvement). Given that in the real world the sample size of the original study may be small or large, we examined level of replicability across various indices treating each subsample as the original study, and each of the other subsamples as replication attempts.

One limitation of using real data to sort out the strengths and limitations of these distinct replicability indices is that failure to replicate could be either a problem with our method of testing replication or a problem that the effect does not exist in the population. To address this issue, we selected three models to test three different research questions that have been widely supported in previous research. This strategy allows us to consider failure to find these effects as a limitation of our statistical models rather than a limitation of the theory.

First, we tested whether drinking intensity is associated with alcohol-related consequences, hypothesizing that drinking intensity would be significantly positively correlated with consequences. Although there is not a one-to-one relationship between intensity of alcohol use and consequences (Prince, Pearson, Bravo, & Montes, 2018), it is extremely likely that intensity of alcohol use would be positively associated with experiencing alcohol-related negative consequences.

Second, we tested whether there are sex differences in drinking intensity, hypothesizing that drinking intensity would be significantly higher among men. On average, female drinkers consume less alcohol than male drinkers for many reasons, including sex differences in body weight and alcohol metabolism, among other factors. Parsing the distinct contributing factors would be more complicated (Becker, McClellon, & Glover Reed, 2017), but numerous population-level surveys have found females drink less than males, on average (Keyes, Li, & Hasin, 2011).

Third, we tested whether individuals who received treatment in Project MATCH would, on average, reduce their drinking from pre-treatment to post-treatment, hypothesizing significant reductions in drinking following treatment. Project MATCH was a comparative effectiveness trial for three treatments (i.e., motivational enhancement therapy, cognitive behavioral therapy, and 12-step facilitation) that largely failed to find differential response across treatments (Project MATCH Research Group, 1997). Although it could be considered contentious that one treatment is superior to another treatment, and considering the lack of a control group, one could even argue that changes from pre- to post-treatment reflect expectancy effects and are driven entirely from individuals' desire/motivation to change (DiClemente, Bellino, & Neavins, 1999). In the present example, we simply tested whether treatment-seeking individuals who obtain treatment tend to drink less over time, on average.

Method

Participants and Procedures

Secondary data analyses were conducted using data from Project MATCH (Project MATCH Research Group, 1993). In brief, 1,726 participants from 27 inpatient and outpatient treatment sites across nine research sites were recruited for Project MATCH. Two sites recruited clients for both outpatient and aftercare (i.e., post-inpatient treatment) samples, providing 11 distinct subsamples. Participants were mostly male (75.7%) and non-Hispanic white (80.0%) and were randomly assigned to receive one of three treatments for alcohol use disorder: cognitive behavioral therapy, motivational enhancement therapy, and 12-step facilitation. A detailed description of the participants and procedures for Project MATCH is reported elsewhere (e.g., Project MATCH Research Group, 1993).

Measures

Alcohol use.: Daily alcohol consumption during the past 90 days at baseline and 3-month follow-up was assessed using the Form 90 (Miller, 1996) delivered via an in-person interview. For the purpose of the present study, we used the Form 90 data to compute drinks per drinking day (DDD), or the average number of drinks per drinking occasion, as an indicator of drinking intensity/severity. Previous studies using these data have found high reliability for the Form 90 (e.g., Tonigan et al., 1997).

Alcohol-related consequences.: Alcohol-related consequences were assessed using the Drinker Inventory of Consequences (DrInC; Miller et al., 1995). The DrInC includes 50 items that represent a wide range of alcohol-related consequences (e.g., "My marriage or love relationship has been harmed by my drinking"). Participants indicate the frequency in which they experience each consequence using a 4-point scale (1 = never, 4 = daily or almost daily). Previous studies using these data have found adequate reliability for the DrInC (e.g., Miller et al., 1995). Reliability (Cronbach's alpha) of the DrInC at each site was excellent (α = .95 - .97).

Statistical Analysis

We estimated the replicability of three models. Across all three models, we examined each of the replication indices described above: a) pattern of statistical significance, b)

Type S, and c) 95% confidence intervals. We treated each of the 11 data collection sites in Project MATCH as an independent sample to test the functioning of distinct replicability indices. For the non-meta-analytic replicability indices, one sample needs to be treated as the original study in order to determine if the other studies replicate the original study findings. Given the arbitrariness of this designation, we evaluated the performance of these replicability indices treating each sample as the "original" study and the remaining subsamples as replication studies. For illustration purposes, we highlight the results using the largest subsample and the smallest subsample as the original study to simulate examining replicability in the case that the original study had high vs. low statistical power.

For the meta-analysis we used random-effects meta-analysis in Comprehensive Meta-Analysis V3 (Borenstein et al., 2010) calculating effect sizes for each of the 11 subsamples (Model 1: *r*, Models 2–3: Hedge's *g*). We used the Q statistic to determine if there was statistically significant heterogeneity in effect sizes, and we used I^2 as a measure of how much variation across studies was due to heterogeneity rather than chance (Higgins et al., 2003). An I^2 of around 20%, 50%, and 80% are considered small, medium, and large amounts of heterogeneity.

Model 1 Analyses

In Model 1, we examined the bivariate correlation (Pearson *r*) between an indicator of intensity of alcohol use (DDD) and the number of negative alcohol-related consequences (DrInC total score) during the same 3-month period by site. Specifically, we examined these associations at the 3-month follow-up, which included drinking and consequences experienced during treatment. Given that direct alcohol-related consequences presumably cannot occur when abstaining from drinking, this model only included those who reported drinking during this time period (total n = 895).

Model 2 Analyses

In Model 2, we examined sex/gender differences in average DDD at baseline using a series of independent-samples *t*-tests by site.

Model 3 Analyses

In Model 3, we examined the pre-post changes in drinking (DDD) following treatment using a series of paired-samples *t*-tests by site.

Results

Table 1 summarizes the results for all three models.

Model 1 Results

Among those who reported drinking between baseline and 3-month follow-up (n = 895), the positive association between drinks per drinking day (DDD) and negative alcohol-related consequences (DrInC total score) was significant in all but two subsamples. Given that it was non-significant in the smallest subsample, replicability based on pattern of significance was 0% using the smallest subsample as the referent group, and 90% using all other referent groups (median = 90%). Type S replicability was 100% using all referent groups (median = 100%). In terms of confidence intervals, 50% of correlation estimates fell within the bounds of the confidence interval of the largest subsample, and 70% of correlation estimates fell within the bounds of the confidence interval of the smallest subsample; replicability ranged from 40–90% using the other referent groups (median = 70%). Meta-analytic indices revealed that the moderate positive associations between DDD and DrInC total score (r_w = .449) did not demonstrate significant effect size heterogeneity across subsamples (see Figure 1 for a visual depiction), Q(1, 10)=12.368, p=.261, I^2=19.145, with only a small amount (less than 20%) of the heterogeneity attributable to true heterogeneity.

Model 2 Results

Model 2 showed baseline sex differences in DDD (i.e., males consumed more DDD compared to females) that were statistically significant in six subsamples, and not statistically significant in 5 samples. Thus, using the largest subsample as the referent, we observed 50% replicability and using the smallest subsample as the referent we observed 40% replicability; replicability ranged from 40–50% using the other referent groups (median=50%). Type S replicability was 100% using all referent groups (median 50%). In terms of confidence intervals, 90% of sex difference estimates fell within the bounds of the confidence intervals of either the largest subsample or the smallest subsample, ranging from 70–100% using the other referent groups (median 100%). Meta-analytic indices revealed that the moderate sex difference in DDD (g_w=.470) did not demonstrate significant effect size heterogeneity across subsamples (see Figure 2 for a visual depiction), Q(1, 10)=16.425, p=.088, I^2=39.118, with a moderate amount (~40%) of the heterogeneity attributed to true heterogeneity.

Model 3 Results

Model 3 indicated a reduction in DDD from baseline to 3-month follow-up, which was statistically significant across all subsamples; thus, pattern of significance and Type S replicability was 100% using all referent groups (median=100%). Only 30% of point estimates from other studies fell within the bounds of the confidence interval of the largest subsample (n=221), 70% fell within the bounds of the confidence interval of the smallest subsample (n=56), and 10–50% (median= 40%) fell within the bounds of the confidence intervals of the other referent groups. Meta-analytic indices revealed that the large reduction in DDD (g_w=−.968) did demonstrate significant effect size heterogeneity across subsamples (see Figure 3 for a visual depiction), Q(1,10)= 66.188, p<.001, I^2=84.89, with a large amount (~85%) of this heterogeneity attributable to true heterogeneity.

Discussion

The current study tested different replication indices and meta-analysis to examine three research questions relevant to the study of addictive behaviors: (1) is drinking intensity associated with alcohol-related consequences; (2) are there sex differences in drinking intensity; and (3) do we see reductions in drinking following alcohol treatment? For all three research questions, we estimated statistical models to test each question separately in each of 11 different treatment sites, which were treated as separate samples. The ultimate goal of this inquiry was to determine whether findings in any one of the samples replicated in the other 10 samples. Despite failing to find significant effect size heterogeneity using meta-analytic indices in Models 1 and 2, we observed between 0–100% replicability depending on the replication index examined, the subsample treated as the original sample, and the statistical model itself. Thus, we could have falsely concluded low levels of replication even when examining these large, reliable effects that are well supported by prior research and theory.

These findings demonstrate the fundamental weaknesses of individual replicability indices, and we believe the limitations of replicability metrics need to be carefully considered prior to declaring a “replicability crisis.” All else being equal, the larger the sample size of an original study, the narrower the confidence interval, and subsequent replications are more likely to find effect sizes that fall outside of this narrower range. Thus, by this definition, original studies with smaller sample sizes (i.e., wider confidence intervals) are more likely to be replicated, even though we know the effect size estimate is less reliable.

Meta-analyses avoid many of the pitfalls of these other replicability metrics, and focuses research on effect sizes, which we argue is a more fruitful direction in the field than the continued examination of whether *p*-values are less than .05 (Cohen, 1994). In fact, even when using a fixed effect meta-analysis, the Open Science Collaboration (2015) found that 68% of meta-analyzed effect sizes including the original and replication studies were significant, suggesting a substantially lower level of replication problems than other metrics. Using meta-analyses provides a direct test of the strength and heterogeneity of effect sizes across studies (or samples, in our case) and can be used at the study level or at the level of individual participant data (Mun et al., 2015).

Following the large-scale replication attempts cited throughout, multiple researchers have highlighted that in the face of NHST, one should expect regression to the mean, or regression shrinkage, of *p*-values and effect sizes as long as statistical significance has any relationship with whether a manuscript is published (Fiedler & Prager, 2018; Wilson et al., 2020). Thus, interpreting a reduction in effect size from an original study to a replication study should be viewed as a mathematical certainty (i.e., byproduct of regression to the mean), rather than a novel finding worth publishing. Rather, researchers should focus on obtaining large sample sizes, substantially larger than the norm, to protect from the inevitable Type S and Type M errors that occur with small sample sizes. For many reasons (e.g., limited resources, hard-to-reach populations), recruiting large samples in addiction science is impractical, if not impossible. However, an alternative method for obtaining large sample sizes is to combine the individual data from multiple studies for secondary data analysis, generally referred to as integrative data analysis (e.g., Curran & Hussong, 2009). Not only does integrative data analysis increase statistical power, but it is well-suited for tests of replicability due to the ability to model between-study heterogeneity within a single analysis of the combined data (Curran & Hussong, 2009). Integrative data analysis is becoming increasingly practical given greater expectations for data sharing and advances in statistical techniques for integrative data analysis, such as moderated nonlinear factor analysis for the development of commensurate measures across studies (e.g., Curran et al., 2014; Witkiewitz et al., 2016). When larger samples are obtained, the focus then moves away from statistical significance, which becomes virtually certain, toward effect sizes, which better helps to describe how impactful an effect might be in clinical practice.

Although conducting power analysis using an effect size observed in a published study is common practice, this practice will almost certainly result in underpowered research studies. Using the lower-bound estimate of effect size from a 95% confidence interval from a published study is viewed as a conservative way to conduct a power analysis; however, by not accounting for regression to the mean caused by publication bias and the likelihood that effect size estimates in the literature are inflated (Anderson & Maxwell, 2017; Ioannidis, 2008), this approach can still result in underpowered research (Anderson & Maxwell, 2017; Ioannidis, 2008). If researchers assume effect size inflation, they could directly reduce the effect size estimate being used to conduct power analysis adjusting for publication bias (Taylor & Muller, 1996). An alternative approach worth consideration is the a priori procedure (APP) introduced by Trafimow (Trafimow, 2019). Like the inferential statistics tests on which it is based, statistical power depends on the effect size of an association, which is to a varying degree, always unknown. With the APP, rather than focusing on statistical tests to be conducted *after* collecting data, one pivots to determining the sample size needed to assume that the sample estimates obtained are reasonably reflective of the underlying population estimates based on an *a priori* specified level of precision (Li, Trafimow, Wang, Wang, & Hu, 2020). Although this approach is advocated as a way to move away from relying on NHST, use of the APP demonstrates that 1) researchers should be obtaining substantially larger sample sizes than is currently the norm, and 2) researchers should focus on effect sizes.

It is important to highlight what we have and have not demonstrated in this example. We have demonstrated that some replicability indices suggest non-replication, even when the effect tested is quite robust and has been replicated in prior studies. What we have not done is to estimate the overall level of replicability in the field of addiction science. Large-scale replication attempts are needed to push the addictions field forward; however, the definitional issues of replication must be considered carefully, or this work will fail to achieve its primary objective. The field must also wrestle with what studies *should* be replicated (for further discussion on this topic, see Heirene, 2020). Our comparisons of relative strengths and benefits of these indices were conducted in the case where we have strong evidence that these effects are true (i.e., null=false). The relative strengths and weaknesses of these indices when there are no true effects (i.e., null=true) is an important topic that needs to be addressed in other ways (i.e., simulation studies). When we assert that direct replications are rare in clinical studies of addiction, we do not state this as an indictment of clinical researchers. We did find several excellent examples of replications published in *Psychology of Addictive Behaviors*, including novel replications that involved questions of intergenerational effects (Rothenberg et al., 2019) and integrative data analysis (Mun et al., 2015). The lack of more direct replication studies occurs for many reasons outside of investigators' control. First, there is a file drawer problem and it is unclear how many attempts at replication remain unpublished because findings did not replicate and were not statistically significant in the replication attempt (Ingre & Nilsonne, 2018). Second, randomized clinical trials (RCTs) are expensive and conducting a direct replication RCT is not likely to be feasible. Third, funding agencies prioritize “novel” research; thus, funding to conduct a costly study that uses the same methods as a previous trial is highly unlikely.

The National Institutes of Health (NIH) have demonstrated concerns of replicability over the past several years by requiring investigators to directly address “reproducibility” in grant applications. Further, institutes have been recommending or requiring increased transparency in the conduct of research. Two examples include the requirement that all clinical trials (broadly defined) be pre-registered at clinicaltrials.gov, and the requirement that investigators submit de-identified, individual-level data to an agency-sponsored data repository (e.g., National Institute on Alcohol Abuse and Alcoholism’s Data Archive, $\mathrm{NIAAA_{DA}}$). We believe that increasing use of such open science practices will help to reduce questionable research practices and will help science be more self-correcting.

Limitations and Future Directions

The current study has several limitations. First, the selection of a single study with the same investigators, same measures, and same training of assessors across sites is not the same as conducting independent replications by different investigators. Importantly, the consistency of investigators, measures, and training would likely increase the chances of replication and thus it is likely that replication would be potentially worse in an independent replication attempt. Second, we selected research questions that were well-supported in the literature and were based on bivariate analysis to provide straightforward comparisons across replication indices (e.g., only one effect per research question). The Type S replication was 100% across samples, which was not surprising given the straightforward and well-supported research questions examined. More complicated research questions that test multiple variables, which dominate the papers published in *Psychology of Addictive Behaviors,* would provide a more nuanced picture of the issues researchers encounter when attempting direct replication of prior findings and could potentially result in even less replication of findings (Tackett et al., 2019). Third, the Project MATCH study was conducted in the United States from 1991 to 1993 with a mostly non-Hispanic white and male sample, and it is unclear whether findings would replicate in other countries, in more diverse samples, and in a more contemporary sample. There is some evidence that the association between alcohol use and consequences (tested in Model 1) may be greater in racial and ethnic minority men, likely due to environmental factors and systemic racism (Witbrodt, Mulia, Zemore, & Kerr, 2014), and in certain countries (particularly in Eastern Europe; Shield et al., 2020). Given changes in drinking among women in more recent age cohorts (Keyes et al., 2011), we anticipate that sex differences in drinking intensity (tested in Model 2) would be less likely to replicate in more recently collected data. Similarly, Project MATCH was an abstinence-based treatment and many individuals had abstinence goals, it is unclear if drinking reductions would be as pronounced (tested in Model 3) in more recent alcohol treatment trials that have incorporated a broader focus on drinking reduction goals (Falk et al., 2019; Witkiewitz et al., 2019). Although accounting for effect size heterogeneity was not a focus of the present study, an important area for future research to explore the factors that may account for effect size heterogeneity.

With respect to future directions, we encourage open science practices to increase the potential to test the reproducibility of research findings and to aid in direct replication efforts. Open sharing of experimental protocols, analysis code, and data will increase the transparency and potential reproducibility of addiction science. Similarly, researchers could plan replication studies as part of ongoing research, such that initial findings can be replicated prior to publication, which has occurred in a number of papers we identified in our review of replications in *Psychology of Addictive Behaviors* (Alterman et al., 1998; Isenhart & Silversmith, 1996; Jones et al., 2014; Lee & Leeson, 2015; Pantalon et al., 2002). It is also imperative that NIH consider greater investment in replication via individual funding mechanisms, project grants and center grant mechanisms, as well as supplements to existing grants. Importantly, these studies should include both traditional meta-analysis to summarize the findings of a large number of unregistered studies as well as registered replication studies of targeted effects. This level of commitment would send a clear message to researchers that replication of results is an important and valued endeavor, particularly when novel and clinically meaningful effects are observed in individual studies.

Acknowledgments

This research was supported in part by a research grant (R01AA022328, PI: Witkiewitz) from the National Institute on Alcohol Abuse & Alcoholism (NIAAA). MRP is supported by a career development award (K01AA023233) from the NIAAA. FJS is supported by a training grant (T32AA018108) from the NIAAA. DKR is supported by an individual fellowship award (F32AA028712) from the NIAAA. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIAAA or National Institutes of Health.

References

Alterman AI, McDermott PA, Cook TG, Metzger D, Rutherford MJ, Cacciola JS, & Brown LS (1998). New scales to assess change in the addiction severity index for the opioid, cocaine, and alcohol dependent. Psychology of Addictive Behaviors, 12(4), 233–246. 10.1037/0893-164X.12.4.233

Anderson SF, & Maxwell SE (2017). Addressing the “Replication Crisis”: Using Original Studies to Design Replication Studies with Appropriate Statistical Power. Multivariate Behavioral Research, 52(3), 305–324. 10.1080/00273171.2017.1289361 [PubMed: 28266872]

Becker JB, McClellon ME, & Glover Reed B (2017). Sex differences, gender and addiction. J Neuropsych Res, 95(1–2), 136–147. 10.1002/jnr.23963

Blaszczynski A, & Gainsbury SM (2019, September). Editor’s note: replication crisis in the social sciences. International Gambling Studies, Vol. 19, pp. 359–361. 10.1080/14459795.2019.1673786

Budney AJ, Stanger C, Tilford JM, Scherer EB, Brown PC, & Walker DD (2015). Computer-Assisted Behavioral Therapy and Contingency Management for Cannabis Use Disorder. 29(3), 501–511.

Carney MM, & Kivlahan DR (1995). Motivational Subtypes Among Veterans Seeking Substance Abuse Treatment: Profiles Based on Stages of Change. Psychology of Addictive Behaviors, 9(2), 135–142. 10.1037/0893-164X.9.2.135

Cohen J (1994). The earth is round (p<. 05). American Psychologist, 49(12), 997–1003.

Cook JE, Chandler C, Rüedi-Bettschen D, Taylor I, Patterson S, & Platt DM (2019). Changes in the Elimination and Resurgence of Alcohol- Maintained Behavior in Rats and the Effects of Naltrexone. Psychology of Addictive Behaviors. 10.1037/adb0000525

Cumming G (2012). *Understanding the new* statistics: Effect sizes, confidence intervals, and meta-analysis. New York: Routledge.

Cumming G, & Maillardet R (2006). Confidence intervals and replication: Where will the next mean fall? Psychological Methods, 11(3), 217–227. 10.1037/1082-989X.11.3.217 [PubMed: 16953701]

Curran PJ, & Hussong AM (2009). Integrative data analysis: the simultaneous analysis of multiple data sets. Psychological Methods, 14(2), 81–100. 10.1037/a0015914 [PubMed: 19485623]

Curran PJ, McGinley JS, Bauer DJ, Hussong AM, Burns A, Chassin L, Sher K, & Zucker R (2014). A moderated nonlinear factor model for the development of commensurate measures in integrative data analysis. Multivariate Behavioral Research, 49(3), 214–231. 10.1080/00273171.2014.889594 [PubMed: 25960575]

de Boer MR, Waterlander WE, Kuijper LDJ, Steenhuis IHM, & Twisk JWR (2015). Testing for baseline differences in randomized controlled trials: An unhealthy research behavior that is hard to eradicate. International Journal of Behavioral Nutrition and Physical Activity, 12(1). 10.1186/s12966-015-0162-z

DiClemente CC, Bellino LE, & Neavins TM (1999). Motivation for change and alcoholism treatment. Alcohol Research and Health, 23(2), 86–92. [PubMed: 10890801]

Edens JF, & Willoughby FW (2000). Motivational patterns of alcohol dependent patients: A replication. Psychology of Addictive Behaviors, 14(4), 397–400. 10.1037/0893-164X.14.4.397 [PubMed: 11130158]

Falk DE, O’Malley SS, Witkiewitz K, Anton RF, Litten RZ, Slater M, … Alcohol Clinical Trials Initiative (ACTIVE) Workgroup. (2019). Evaluation of drinking risk levels as outcomes in alcohol pharmacotherapy trials. JAMA Psychiatry, 76(4), 374. 10.1001/jamapsychiatry.2018.3079 [PubMed: 30865232]

Fiedler K, & Prager J (2018). The Regression Trap and Other Pitfalls of Replication Science—Illustrated by the Report of the Open Science Collaboration. Basic and Applied Social Psychology, 40(3), 115–124. 10.1080/01973533.2017.1421953

Gelman A (2008). Objections to Bayesian statistics. Bayesian Analysis, 3(3), 445–450. 10.1214/08-BA318

Gelman A, & Carlin J (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspectives on Psychological Science, 9(6), 641–651. 10.1177/1745691614551642 [PubMed: 26186114]

Gokhale DV, Box GEP, & Tiao GC (1974). Bayesian Inference in Statistical Analysis. Biometrics, 30(1), 211. 10.2307/2529631

Group PMR (1993). Project MATCH: Rationale and Methods for a Multisite Clinical Trial Matching Patients to Alcoholism Treatment. Alcoholism: Clinical and Experimental Research, 17(6), 1130–1145. 10.1111/j.1530-0277.1993.tb05219.x [PubMed: 8116822]

Ham LS, & Hope DA (2006). Incorporating social anxiety into a model of college problem drinking: Replication and extension. Psychology of Addictive Behaviors, 20(3), 348–355. 10.1037/0893-164X.20.3.348 [PubMed: 16938075]

Hawkins JD, Catalano RF, Arthur MW, Egan E, Brown EC, Abbott RD, & Murray DM (2008). Testing communities that care: The tationale, design and behavioral baseline equivalence of the community youth development study. Prevention Science, 9(3), 178–190. 10.1007/s11121-008-0092-y [PubMed: 18516681]

Hedges LV, & Schauer JM (2018). Statistical Analyses for Studying Replication: Meta-Analytic Perspectives. Psychological Methods, 24(5). 10.1037/met0000189

Hedges LV, & Vevea JL (1998). Fixed- and random-effects models in meta-analysis. Psychological Methods, 3(4), 486–504. Retrieved from papers3://publication/uuid/87D1B83E-C6A2-474E-BD8B-00D92F21AAC9

Heirene RM (2020). A call for replications of addiction research: which studies should we replicate and what constitutes a ‘successful’ replication? Addiction Research and Theory, 0(0), 1–9. 10.1080/16066359.2020.1751130

Hernández-López M, Luciano MC, Bricker JB, Roales-Nieto JG, & Montesinos F (2009). Acceptance and Commitment Therapy for Smoking Cessation: A Preliminary Study of Its Effectiveness in Comparison With Cognitive Behavioral Therapy. Psychology of Addictive Behaviors, 23(4), 723–730. 10.1037/a0017632 [PubMed: 20025380]

Hughes JR, Naud S, Budney AJ, Fingar JR, & Callas PW (2016). Attempts to Stop or Reduce Daily Cannabis Use : An Intensive Natural History Study. 30(3), 389–397.

Hunter JE, & Schmidt FL (2000). Fixed Effects vs. Random Effects Meta-Analysis Models: Implications for Cumulative Research Knowledge. International Journal of Selection and Assessment, 8(4), 275–292. 10.1111/1468-2389.00156

Ingre M, & Nilsonne G (2018). Estimating statistical power, posterior probability and publication bias of psychological research using the observed replication rate. Royal Society Open Science, 5(9), 181190. 10.1098/rsos.181190 [PubMed: 30839704]

Ioannidis JPA (2008). Why most discovered true associations are inflated. Epidemiology, 19(5), 640–648. [PubMed: 18633328]

Isenhart CE, & Silversmith DJ (1996). MMPI-2 response styles: Generalization to alcoholism assessment. Psychology of Addictive Behaviors, 10(2), 115–123. 10.1037/0893-164X.10.2.115

Jones RE, Spradlin A, Joe Robinson R, & Tragesser SL (2014). Development and validation of the opioid prescription medication motives questionnaire: A four-factor model of reasons for use. Psychology of Addictive Behaviors, 28(4), 1290–1296. 10.1037/a0037783 [PubMed: 25180561]

Keyes KM, Li G, & Hasin DS (2011, December). Birth Cohort Effects and Gender Differences in Alcohol Epidemiology: A Review and Synthesis. Alcoholism: Clinical and Experimental Research, Vol. 35, pp. 2101–2112. 10.1111/j.1530-0277.2011.01562.x [PubMed: 21919918]

Lakens D, Scheel AM, & Isager PM (2018). Equivalence testing for psychological research: A tutorial. Advances in Methods and Practices in Psychological Science, 1(2), 259–269.

LaPlante DA (2019, September). Replication is fundamental, but is it common? A call for scientific self-reflection and contemporary research practices in gambling-related research. International Gambling Studies, Vol. 19, pp. 362–368. 10.1080/14459795.2019.1672768

Lee BW, & Leeson PRC (2015). Online Gaming in the Context of Social Anxiety. Psychology of Addictive Behaviors, 29(2), 473–482. 10.1037/adb0000070 [PubMed: 25938630]

Li H, Trafimow D, Wang T, Wang C, & Hu L (2020). User-friendly computer programs so econometricians can run the a priori procedure. Frontiers in Management and Business, 1(1), 2–6. 10.25082/fmb.2020.01.002

Masson MEJ (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. Behavior Research Methods, 43(3), 679–690. 10.3758/s13428-010-0049-5 [PubMed: 21302025]

Maxwell SE, Lau MY, & Howard GS (2015). Is psychology suffering from a replication crisis?: What does “failure to replicate” really mean? American Psychologist, 70(6), 487–498. 10.1037/a0039400 [PubMed: 26348332]

Meehl PE (2004). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Applied and Preventive Psychology, 11(1), 1. 10.1016/j.appsy.2004.02.001

Miller MB, Dibello AM, Lust SA, Meisel MK, & Carey KB (2017). Impulsive personality traits and alcohol use: Does sleeping help with thinking? Psychology of Addictive Behaviors, 31(1), 46–53. 10.1037/adb0000241 [PubMed: 28094998]

Motschman CA, & Tiffany ST (2016). Cognitive Regulation of Smoking Behavior Within a Cigarette: Automatic and Nonautomatic Processes. Psychology of Addictive Behaviors, 30(4), 494–499. 10.1037/adb0000157 [PubMed: 26866781]

Mun EY, De La Torre J, Atkins DC, White HR, Ray AE, Kim SY, … Huh D (2015). Project INTEGRATE: An integrative study of brief alcohol interventions for college students. Psychology of Addictive Behaviors, 29(1), 34–48. 10.1037/adb0000047 [PubMed: 25546144]

Nosek BA, & Errington TM (2017). Making sense of replications. ELife, 6. 10.7554/eLife.23383

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716. 10.1126/science.aac4716 [PubMed: 26315443]

Pantalon MV, Nich C, Frankforter T, & Carroll KM (2002). The URICA as a measure of motivation to change among treatment-seeking individuals with concurrent alcohol and cocaine problems. Psychology of Addictive Behaviors, 16(4), 299–307. 10.1037/0893-164X.16.4.299 [PubMed: 12503902]

Pashler H, & Harris CR (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. Perspectives on Psychological Science, 7(6), 531–536. 10.1177/1745691612463401 [PubMed: 26168109]

Prince MA, Pearson MR, Bravo AJ, & Montes KS (2018). A quantification of the alcohol use-consequences association in college student and clinical populations: A large, multi-sample study. American Journal on Addictions, 27(2). 10.1111/ajad.12686

Project MATCH Research Group. (1997). Title: Matching alcoholism treatments to client heterogeneity: Project MATCH posttreatment drinking outcomes. Journal of Studies on Alcohol, 58, 7–29. [PubMed: 8979210]

Rothenberg WA, Sternberg A, Blake A, Waddell J, Chassin L, & Hussong A (2019). Identifying Adolescent Protective Factors That Disrupt the Intergenerational Transmission of Cannabis Use and Disorder. Psychology of Addictive Behaviors. 10.1037/adb0000511

Rychtarik RG, McGillicuddy NB, Papandonatos GD, Whitney RB, & Connors GJ (2017). Randomized clinical trial of matching client alcohol use disorder severity and level of cognitive functioning to treatment setting: A partial replication and extension. Psychology of Addictive Behaviors, 31(5), 513–523. 10.1037/adb0000253 [PubMed: 28150955]

Schüz N, Eid M, Schüz B, & Ferguson SG (2016). Immediate effects of plain packaging health warnings on quitting intention and potential mediators: Results from two ecological momentary assessment studies. Psychology of Addictive Behaviors, 30(2), 220–228. 10.1037/adb0000146 [PubMed: 26999353]

Shield K, Manthey J, Rylett M, Probst C, Wettlaufer A, Parry CDH, & Rehm J (2020). National, regional, and global burdens of disease from 2000 to 2016 attributable to alcohol use: a comparative risk assessment study. The Lancet Public Health, 5(1), e51–e61. 10.1016/S2468-2667(19)30231-2 [PubMed: 31910980]

Simmons JP, Nelson LD, & Simonsohn U (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22(11), 1359–1366. 10.1177/0956797611417632 [PubMed: 22006061]

Stanger C, Scherer EA, Vo HT, Babbin SF, Knapp AA, McKay JR, & Budney AJ (2019). Working Memory Training and High Magnitude Incentives for Youth Cannabis Use: A SMART Pilot Trial. Psychology of Addictive Behaviors. 10.1037/adb0000480

Stroebe W, & Strack F (2014). The Alleged Crisis and the Illusion of Exact Replication. Perspectives on Psychological Science, 9(1), 59–71. 10.1177/1745691613514450 [PubMed: 26173241]

Tackett JL, Brandes CM, King KM, & Markon KE (2019). Psychology's Replication Crisis and Clinical Psychological Science. Annual Review of Clinical Psychology, 15(1), 579–604. 10.1146/annurev-clinpsy-050718-095710

Taylor DJ, & Muller KE (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. Communications in Statistics - Theory and Methods, 25(7), 1595–1610. 10.1080/03610929608831787

Tonigan JS (2001). Benefits of Alcoholics Anonymous attendance: Replication of findings between clinical research sites in Project MATCH. Alcoholism Treatment Quarterly, Vol. 19, pp. 67–77. 10.1300/J020v19n01_05

Tonigan JS, McCallion EA, Frohe T, & Pearson MR (2017). Lifetime Alcoholics Anonymous attendance as a predictor of spiritual gains in the Relapse Replication and Extension Project (RREP). Psychology of Addictive Behaviors, 31(1), 54–60. 10.1037/adb0000235 [PubMed: 28080094]

Tonigan JS, Miller WR, & Brown JM (1997). The reliability of Form 90: an instrument for assessing alcohol treatment outcome. Journal of Studies on Alcohol, 58(4), 358–364. 10.15288/jsa.1997.58.358 [PubMed: 9203116]

Trafimow D (2019). A Frequentist Alternative to Significance Testing, p-Values, and Confidence Intervals. Econometrics, 7(2), 26. 10.3390/econometrics7020026

Trafimow D, & Earp BD (2017). Null hypothesis significance testing and Type I error: The domain problem. New Ideas in Psychology, 45, 19–27. 10.1016/j.newideapsych.2017.01.002

Von Sternberg K, DiClemente CC, & Velasquez MM (2018). Profiles of behavior change constructs for reducing alcohol use in women at risk of an alcohol-exposed pregnancy. Psychology of Addictive Behaviors, 32(7), 749–758. 10.1037/adb0000417 [PubMed: 30451517]

Wilson BM, Harris CR, & Wixted JT (2020). Science is not a signal detection problem. Proceedings of the National Academy of Sciences of the United States of America, 117(11), 5559–5567. 10.1073/pnas.1914237117 [PubMed: 32127477]

Witkiewitz K, Hallgren KA, O'Sickey AJ, Roos CR, & Maisto SA (2016). Reproducibility and differential item functioning of the alcohol dependence syndrome construct across four alcohol treatment studies: An integrative data analysis. Drug and Alcohol Dependence, 158, 86–93. 10.1016/j.drugalcdep.2015.11.001 [PubMed: 26613839]

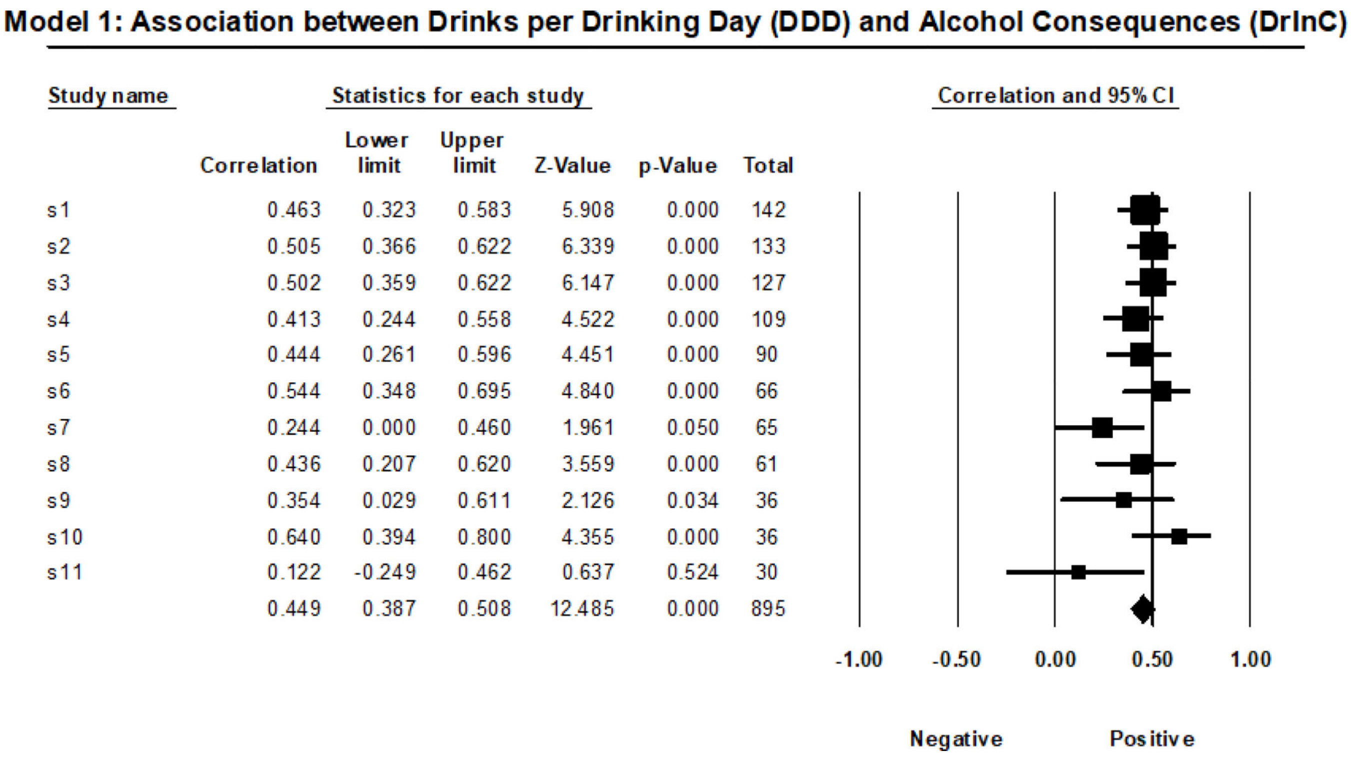
Witbrodt J, Mulia N, Zemore SE, & Kerr WC (2014). Racial/Ethnic disparities in alcohol-related problems: Differences by gender and level of heavy drinking. Alcoholism: Clinical and Experimental Research, 38(6), 1662–1670. 10.1111/acer.12398 [PubMed: 24730475]

Witkiewitz K, Stein ER, Votaw VR, Wilson AD, Roos CR, Gallegos SJ, … Claus ED (2019). Mindfulness-based relapse prevention and transcranial direct current stimulation to reduce heavy drinking: A double-blind sham-controlled randomized trial. Alcoholism: Clinical and Experimental Research, 43(6), 1296–1307. 10.1111/acer.14053 [PubMed: 30977904]

Wohl MJA, Tabri N, & Zelenski JM (2019, September). The need for open science practices and well-conducted replications in the field of gambling studies. International Gambling Studies, Vol. 19, pp. 369–376. 10.1080/14459795.2019.1672769

Public Health Significance Statement

This study highlights the importance of carefully considering how we define replication in the addictions field. Going forward, the field would benefit from an increased emphasis on effect sizes and more prevalent use of meta-analysis to evaluate replicability. We also encourage direct replication attempts and sharing data and code to facilitate reproducibility.

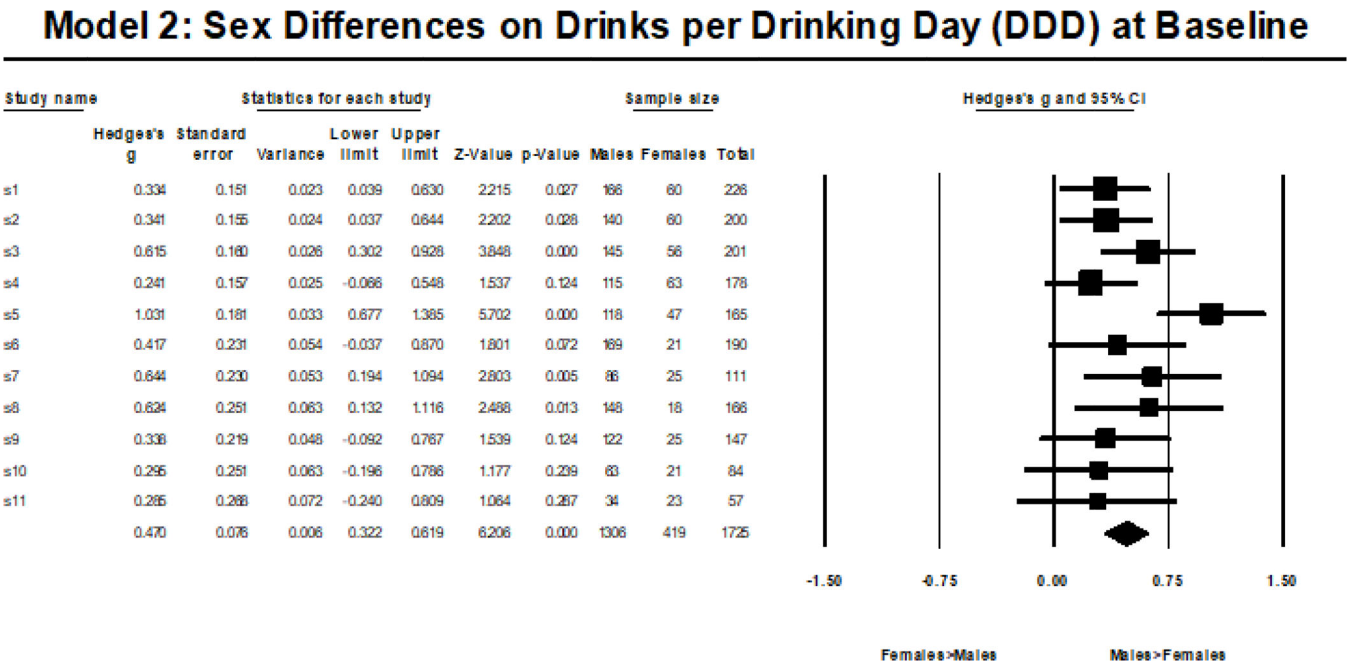


Model 1: Association between Drinks per Drinking Day (DDD) and Alcohol Consequences (DrInC)

Study name	Statistics for each study					
	Correlation	Lower limit	Upper limit	Z-Value	p-Value	Total
s1	0.463	0.323	0.583	5.908	0.000	142
s2	0.505	0.366	0.622	6.339	0.000	133
s3	0.502	0.359	0.622	6.147	0.000	127
s4	0.413	0.244	0.558	4.522	0.000	109
s5	0.444	0.261	0.596	4.451	0.000	90
s6	0.544	0.348	0.695	4.840	0.000	66
s7	0.244	0.000	0.460	1.961	0.050	65
s8	0.436	0.207	0.620	3.559	0.000	61
s9	0.354	0.029	0.611	2.126	0.034	36
s10	0.640	0.394	0.800	4.355	0.000	36
s11	0.122	-0.249	0.462	0.637	0.524	30
	0.449	0.387	0.508	12.485	0.000	895

Figure 1.

Random Effects Meta-Analysis Results from Model 1.

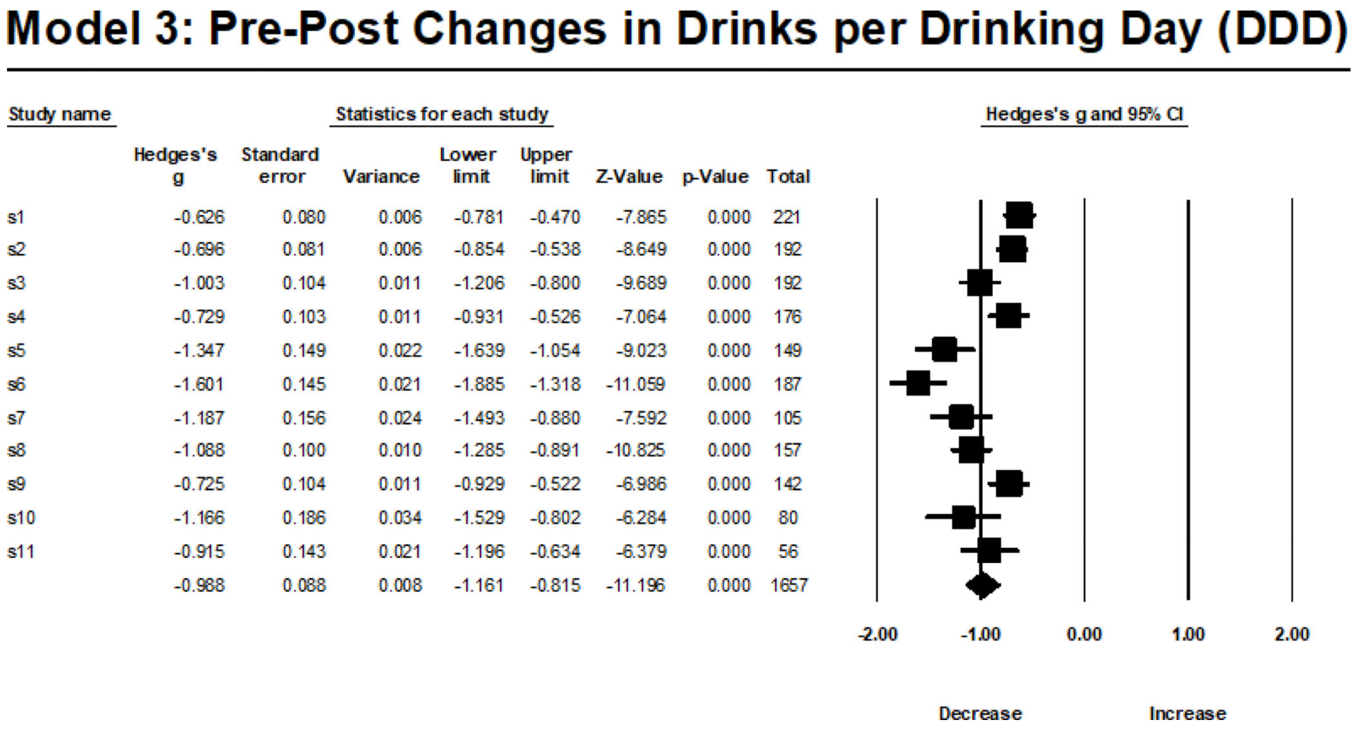


Model 2: Sex Differences on Drinks per Drinking Day (DDD) at Baseline

Study name	Statistics for each study							Sample size		
	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value	Males	Females	Total
s1	0.334	0.151	0.023	0.039	0.630	2.215	0.027	166	60	226
s2	0.341	0.155	0.024	0.037	0.644	2.202	0.028	140	60	200
s3	0.615	0.160	0.026	0.302	0.928	3.848	0.000	145	56	201
s4	0.241	0.157	0.025	-0.066	0.548	1.537	0.124	115	63	178
s5	1.031	0.181	0.033	0.677	1.385	5.702	0.000	118	47	165
s6	0.417	0.231	0.054	-0.037	0.870	1.801	0.072	169	21	190
s7	0.644	0.230	0.053	0.194	1.094	2.803	0.005	86	25	111
s8	0.624	0.251	0.063	0.132	1.116	2.488	0.013	148	18	166
s9	0.338	0.219	0.048	-0.092	0.767	1.539	0.124	122	25	147
s10	0.295	0.251	0.063	-0.196	0.786	1.177	0.239	63	21	84
s11	0.285	0.268	0.072	-0.240	0.809	1.064	0.287	34	23	57
	0.470	0.076	0.006	0.322	0.619	6.206	0.000	1306	419	1725

Figure 2.

Random Effects Meta-Analysis Results from Model 2.



Model 3: Pre-Post Changes in Drinks per Drinking Day (DDD)

Study name	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value	Total
s1	-0.626	0.080	0.006	-0.781	-0.470	-7.865	0.000	221
s2	-0.696	0.081	0.006	-0.854	-0.538	-8.649	0.000	192
s3	-1.003	0.104	0.011	-1.206	-0.800	-9.689	0.000	192
s4	-0.729	0.103	0.011	-0.931	-0.526	-7.064	0.000	176
s5	-1.347	0.149	0.022	-1.639	-1.054	-9.023	0.000	149
s6	-1.601	0.145	0.021	-1.885	-1.318	-11.059	0.000	187
s7	-1.187	0.156	0.024	-1.493	-0.880	-7.592	0.000	105
s8	-1.088	0.100	0.010	-1.285	-0.891	-10.825	0.000	157
s9	-0.725	0.104	0.011	-0.929	-0.522	-6.986	0.000	142
s10	-1.166	0.186	0.034	-1.529	-0.802	-6.284	0.000	80
s11	-0.915	0.143	0.021	-1.196	-0.634	-6.379	0.000	56
	-0.988	0.088	0.008	-1.161	-0.815	-11.196	0.000	1657

Figure 3.

Random Effects Meta-Analysis Results from Model 3

Table 1.

Summary of Rates of Replication across the 10 Subsamples versus an 11th Subsample as Referent and Meta-Analysis Results Combining all Subsamples

	Using Smallest n Site as Referent			**Using Largest n Site as Referent**			**Meta-Analytic Indices**			
	Significance	**CI overlap**	**Type S**	**Significance**	**CI overlap**	**Type S**	**ES**	***Q***	***p***	***I^2***
Model 1: Is alcohol use correlated with negative alcohol-related consequences?	10%	50%	100%	90%	70%	100%	r_w=.449	12.368	.261	19.145
Model 2: Are there sex differences in alcohol use?	40%	90%	100%	50%	90%	100%	g_w=.470	16.425	.088	39.118
Model 3: Does alcohol use decrease from pre- to post-treatment?	100%	30%	100%	100%	70%	100%	g_w=-.968	66.188	<.001	84.891

Note. DDD = Drinks per drinking day; DrInC=Drinker Inventory of Consequences; Model 1 was tested with a bivariate Pearson's correlation; Model 2 tested with an independent samples t-test; Model 3 tested with a paired samples t-test; "Significance" indicates the statistical test was significant at $p < 0.05$; "CI overlap" indicates whether the sample value in replication subsamples fall within the bounds of the referent sample's 95% confidence interval; "Type S" indicates the findings are in the same direction in the replication subsamples and the referent sample; ES = Effect size; r_W = weighted correlation coefficient from random-effects meta-analysis, g_W = weighted Hedges' g estimate from random-effects meta-analysis; Q = Cochran's Q, a chi-square distributed test of effect size heterogeneity ; *p* = represents the *p*-value associated with the Q statistic; I^2 = 100%*(Q-df)/Q, describes percentage of variation across studies that is due to heterogeneity rather than chance.