Do We Have a Replicability Crisis in Addictions Research? Depends How You Quantify It

Frank J. Schwebel[1], A. J. O'Sickey[1], Dylan K. Richards[2], Matthew R. Pearson[1], Katie Witkiewitz[1]

[1]Center on Alcoholism, Substance Abuse, and Addictions (CASAA), University of New Mexico

[2]Latino Alcohol and Health Disparities Research and Training (LAHDR) Center, Department of Psychology, University of Texas at El Paso

INTRODUCTION

- Science has struggled with findings failing to replicate, a problem termed the "replication crisis"
- Yet, some have challenged the extent to which there is a crisis
- Factors that can influence replication failures include differences across studies in: populations studied, measures/procedures used, sample size obtained (which strongly influences statistical power)
- It is unclear how to best conceptualize and assess replicability
- The present study examined whether the metrics used to determine if addictions study findings replicate influences level of replicability
- We also examined models with different levels of complexity to examine if more complex models are less replicable

METHOD

Participants

- The present study used data obtained from Project MATCH (n=1,726; 24% female)

Measures

- From the Timeline Followback (TLFB; Sobel et al., 1995), we used **drinks per drinking day (DDD)** in Models 1 and 2
- Total number of drinking consequences was evaluated via the **Drinkers Inventory of Consequences (DrInC**; Miller et al., 1995)

Analytic Plan

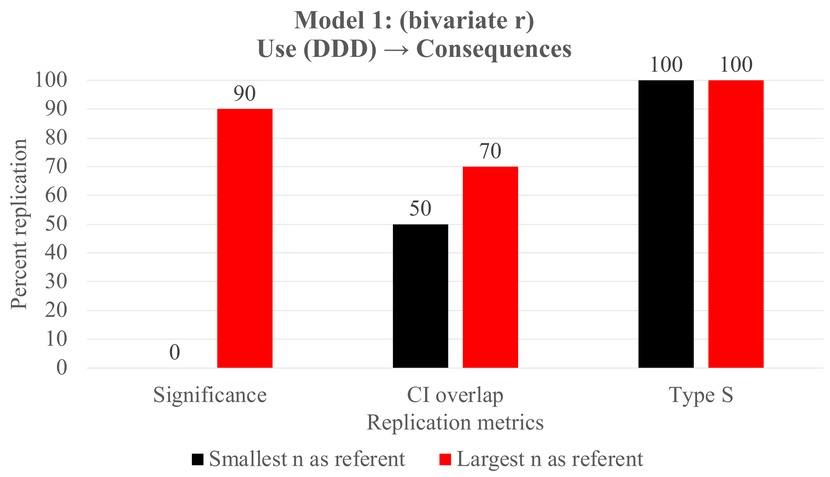
- Given participants were recruited from eleven research centers in the United States, we treated these subsamples as if they reflected one original study (using either the smallest or largest sample as the "original" study) and 10 "direct replication" attempts
- Thus, using different metrics of replicability, we could determine how many of the 10 subsamples replicated the original study
- **Model 1**: We examined the bivariate correlation between alcohol use (DDD) and consequences (DrInC) 3-months post baseline
- **Model 2**: We examined independent-samples *t*-tests to examine sex/gender differences in alcohol use quantity (DDD)

Software

- We used comprehensive meta-analysis (CMA) version 3 to conduct meta-analysis on effect sizes for each of the 4 models (r for Model 1, Hedge's g for Model 2, r calculated by square-rooting the R-square change from the step adding the highest-order interaction for Models 3 and 4)

This research is supported by a research grant from the National Institute on Alcohol Abuse and Alcoholism (R01-AA022328, PI: Witkiewitz). FJS and AJO are supported by a training grant (T32-AA018108) from the NIAAA. MRP is supported by a career development grant (K01-AA023233) from the NIAAA. Poster presented at the Virtual Poster session for the 43rd Annual Research Society on Alcoholism Scientific Meeting. For more information, contact mateo.pearson@gmail.com.

Figures 1 and 2. Level of Replicability by Model Type



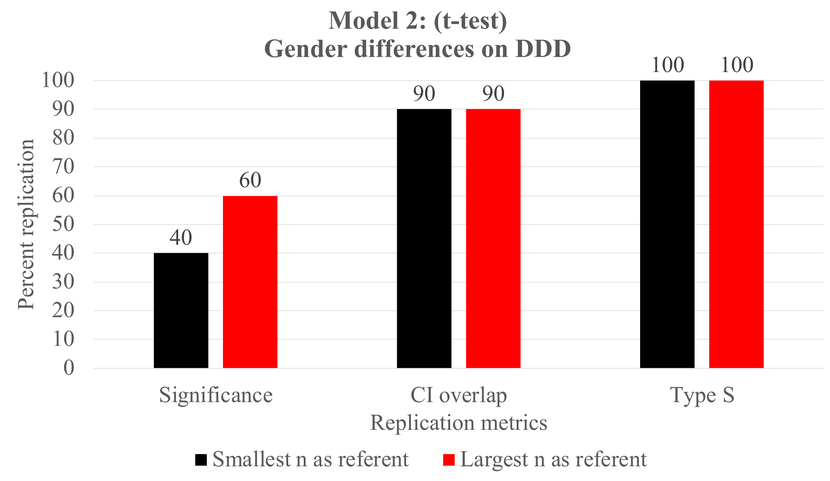


Table 1. Level of Replicability by Model Type

	Meta-Analytic Indices		
	Q	***p***	**I^2**
Model 1: (bivariate r) Use (DDD)→Consequences	12.368	.261	19.145
Model 2: (t-test) Gender Differences on DDD	17.789	.059	43.786

Definition of Replicability Metrics:

Significance – the pattern of significance (statistically significant vs. not statistically significant) is consistent across original sample and replication attempt(s)

Confidence Interval (CI) overlap – if replication attempts produce estimates that fall within the confidence interval of the original report. For example, the 95% CI of the referent group is 0.1 to 0.3. If the correlation found in a replication study was r=0.2, the finding would be considered replicated (within the referent CI). If the correlation was r=0.4, the finding would not be considered replicated (outside the referent CI).

Type S – the sign of the effect in the replication is consistent with the original sample (Gelman & Carlin, 2014)

Q statistic – tests if there is a significant level of heterogeneity in effect sizes (with p-value)

I^2 statistic – % of variation across studies that is due to heterogeneity rather than chance (Higgins & Thompson, 2002)

RESULTS

Despite failing to find significant effect size heterogeneity using meta-analytic indices (Table 1), we found varying levels of "replication" across models and replicability indices (Figures 1 and 2)

- With the least disputable model in question (model 1), we found either 0% replication (smallest site) or 90% (largest site) in terms of a consistent pattern of significance dependent on which sample was regarded as the original sample
- Using CI overlap as a replication metric, Model 1 ranged from 50% (smallest n) to 70% (largest n), whereas replication was equal for Model 2 (90%).

SUMMARY

- We have to carefully consider the limitations of our replicability metrics when declaring a "replicability crisis"
- The replicability metric strongly influences the level of replicability
- Not addressed here, there are significant limitations with using null hypothesis significance testing (NHST), particularly when assessing replication
- Meta-analyses avoid many of the pitfalls of traditional replicability metrics, and focuses research on effect sizes

We recommend focusing on effect size estimates, use of larger sample sizes (especially for more complex models), and conducting more replication attempts

For further discussion of more complex models and additional discussion of metric limitations please visit: mateolab.yolasite.com/openscience.php