Reasonable expectations for replications of psychedelic-assisted therapy: The case for prediction intervals rather than P-values

MITCH EARLEYWINE* and PHILIP KAMILAR-BRITT

Department of Psychology, University at Albany, State University of New York, NY, United States

ABSTRACT

Psychedelic compounds hold promise for alleviating human suffering. Initial trials of psychedelic-assisted treatments have established feasibility and safety, generating calls for replications. Meanwhile, social and medical sciences have drawn criticism due to perceptions of replication failures and varying public trust in empiricism. Data suggest that researchers and the public frequently misunderstand some of the statistical issues associated with replication, potentially leading to unrealistic expectations of treatment effects. Promoting discourse on what constitutes sufficient replication is especially warranted considering the ongoing progression of multi-site phase II and III clinical trials. Here, we review recent and classic work on prediction intervals and power analysis to reveal that trials of psychedelic-assisted therapy that emphasize statistical significance will likely include failures to replicate, especially if sample sizes do not increase dramatically. The field and the public should expect some failed replication attempts based on sampling variability alone. Continued emphasis on statistical significance will require markedly larger samples than those used in clinical trials to date, necessitating substantially greater resources. An alternative approach focused on prediction intervals has distinct advantages. We focus on a recent trial of MDMA-assisted therapy for PTSD to show that, based on prediction intervals, reasonable replications are well within reach. A lack of attention to these statistical issues could unnecessarily prompt widespread dismissal of these therapies before the intervention receives adequate investigation and a fair assessment. In contrast, realistic expectations and appropriate planning could help ensure that these treatments receive the opportunity to help those most in need.

KEYWORDS

prediction intervals, replication, power analysis, psychedelics, psychedelic-assisted therapy

Psychedelic-assisted treatments for psychiatric ailments look promising for appropriately screened samples (Ilingworth et al., 2021; Luoma, Chwyl, Bathje, Davis, & Lancelotta, 2020; Romeo, Karila, Martelli, & Benyamina, 2020; Zeifman et al., 2022). Most reviews of clinical trials with psychedelics end with the clarion call for replications. Tacit hopes for these replications are numerous but often include repeating the same therapy with comparable clients to confirm ameliorative effects. If a psychedelic-assisted treatment repeatedly demonstrates greater efficacy than an alternative intervention, the relevant data would serve as a first step toward empirical support, much like evidence-based approaches have benefited other forms of therapy and medicine (Sakulak et al., 2019). But even within a single trial, outcomes vary. For instance, the occasional member of a treatment group often ends up worse off than some members of the control group. The public purportedly trusts experts to conclude that, in the long run and on average, those who complete a treatment end up improving more than others who did not receive treatment. But traditions around replication and null hypothesis testing might lead any therapist, client, or public official to expect more than a good treatment could deliver. Expecting every trial to reveal that psychedelic-assisted treatment is superior to reasonable controls, let alone alternative treatments, might be more than simple sampling error can permit. Maintaining reasonable expectations as these lines of research continue will be essential for establishing true efficacy. Data suggest that even authors of published research do not have good intuitions about some of the difficulties
inherent in replication, leading them to overestimate the probability that one study’s statistically significant result will appear in a second experiment (Cumming, Williams, & Fidler, 2004).

Despite the promising reputation developing for psychedelic-assisted therapy during this new renaissance, many other publications lament a replication crisis in the social and medical sciences more broadly (See Amrhein, Trafimow, & Greenland, 2019; Ioannidis, 2005). Recent data also suggest a broader distrust of science in the public (Krause, Brossard, Scheufele, Xenos, & Franke, 2019). Meanwhile, many people with psychological ailments continue to suffer or find available treatments lacking (Earleywine & De Leo, 2020). Reasonable expectations for treatments and treatment outcome studies would help. A close look at available data and the natural variation in sampling suggests that a series of replications would prove informative. Nevertheless, expecting that some experiments will fail to replicate the initial promising, statistically significant results is actually quite reasonable. Tolerating these moments of uncertainty and occasional failures to replicate might prove difficult, but a complete absence of replication failures would likely suggest larger problems associated with incomplete reporting or publication bias. In fact, a series of results completely free of replication failures might suggest that some investigators decided not to write up experiments that showed no differences in treatment outcomes, or some editors did not accept those null findings for publication. Without infinite resources, we should expect some replication failures even when a treatment is effective, simply because of sampling error (Amrhein, et al., 2019).

Part of the problem with replication might arise because of the ritual surrounding P-values less than 0.05. Available randomized clinical trials of psychedelic-assisted therapy almost invariably examine the experimental treatment and a control, compare their average impacts, and report whether or not group differences reached statistical significance. Readers who are familiar with the spirited arguments about null hypothesis testing might relish this chance to repeat critiques of the practice that started several decades ago (Neyman & Pearson, 1928). But reconsidering what we do (and do not) call a replication might be in order. The question seems straightforward. Either the psychedelic-assisted treatment creates benefits or it does not. But small samples, including the trials already published, invariably show greater variation in their estimates of effect sizes in the population. Incredible statistical power could help, as enormous samples provide more trustworthy answers. But resources are limited, and these trials already require tremendous time, effort, and cash.

With these issues in mind, those eager to answer questions about the efficacy of psychedelic-assisted treatments are left in a quandary about what does and does not qualify as a replication. Strict adherence to null hypothesis testing could prove time consuming and expensive. But trusting in clinical lore and subjective impressions feels like something other than science. One helpful idea concerns the prediction interval—an anticipated effect size based on an initial estimate from an original experiment. An initial experiment’s effect and sample size can inform guesses about a subsequent replication’s effect for a specified sample size. Any effect that falls within this interval could qualify as a replication, regardless of statistical significance. This approach has proven helpful in the past when Psychology appeared to have a “replication crisis.” Although only 36% of findings appeared to replicate based on statistical significance, a close look at prediction intervals revealed that actually 77% of findings were consistent with the initial publications (Patil, Peng, & Leek, 2016). A focus on identifying effects within the interval, rather than reaching statistical significance, could have advantages for psychedelic-assisted treatments currently.

For example, a recent Phase III trial of MDMA-assisted treatment for post-traumatic stress disorder represents years of herculean work performed in multiple nations and reveals a promising effect size of $d = 0.91$ (Mitchell et al., 2021). The treatment itself makes sense. Alternative treatments lead to challenging drop-out rates and limited success. MDMA, when administered over multiple sessions as part of a manualized treatment lasting 18 weeks, outperformed the placebo control group. This original experiment had 46 people receive MDMA and 43 controls. The effect size, Cohen’s $d$, is a standardized difference between the means for the MDMA and control groups. (Literally the difference between the means divided by the pooled standard deviation).

Any team eager to replicate these results might initially plan to use the same sample sizes present in the original experiment. Understandably, they might wonder what sort of effect size they could anticipate if their results were from the same population as the original experiment, given expected variation due to sampling error. Note that anticipating the same statistical significance could be unrealistic. Ideally, the replicating team would change from expecting another $P$ value below 0.05 to a focus on identifying an effect within the prediction interval. The replication team (and consumers of their research) would essentially ask how different their $d$ could be from the original $d$ if their replication effect size really were identical except for the natural variation that we would expect based on sampling.

The result they would anticipate depends upon the original $d$, the sample size used to estimate it, and the sample size proposed in the replication. Computationally, the answer turns out to be the same as calculating a confidence interval for the difference between these two $d$ values (the original one and the replication). Tradition and desire might lead to a 95% prediction interval. An extension (Spence & Stanley, 2016) of a related proof (Zou, 2007) reveals that the sampling distribution of these $d$ values follows non-central $t$-distributions. The prediction interval ends up being asymmetric because non-central $t$-distributions are asymmetric, but the lower and upper limits are based on the original $d$ as well as the lower ($l_{original}$) and upper ($u_{original}$) limits of its confidence interval. We could assume that the replication’s $d$ is exactly the same as the original. The replication $d$ would also have the same confidence interval (since the replication
will have the same $N$, from $I_{\text{replication}}$ to $u_{\text{replication}}$. Thus, we can compute the 95% prediction interval from some straightforward equations.

**Prediction Interval Lower Limit**

$$d_{\text{original}}-\sqrt{\text{SQR}T\left((d_{\text{original}}-l_{\text{original}})^2\right) + \left(u_{\text{replication}}-d_{\text{original}}\right)^2}$$

**Prediction Interval Upper Limit**

$$d_{\text{original}}-\sqrt{\text{SQR}T\left((d_{\text{original}}-l_{\text{original}})^2\right) + \left(u_{\text{original}}-d_{\text{original}}\right)^2}$$

Thanks to a generous shiny app (https://replication.shinyapps.io/dvalue/) and simple calculations (Spence & Stanley, 2016), we learn that this interval includes all the values from $d = 0.29$ to 1.53. That is, if we assume a replication experiment with sample sizes identical to the original ones, the 95% prediction interval ranges from 0.29 to 1.53. The replication $d$ has a 95% chance of falling into this interval, assuming sampling error alone (note that 5% will still fall outside this range and could still stem from sampling error.) This range might appear vast. Cohen’s (1992) classic work on power and effect sizes suggests that 0.29 is between “small” and “medium” but 1.53 is much larger than the 0.8 considered “large.” The original sample size is fixed, but a team might imagine a bigger $N$ for the replication. With 50 per group, the 95% prediction interval would be a little smaller and range from 0.31 to 1.51. But the returns for a larger sample size diminish. With 500 participants per group, the new range is from $d = 0.45$ to 1.36. No research team is at fault here. These are the simple vagaries of sampling error. The variation and sample size within the original study leave the replication team with a potentially daunting task. In a sense, any replication that falls into this interval could be considered “a success” (Patil et al., 2016). That is, any result in this range is a reasonable expectation for a replication if this second sample differed from the original only based on sampling error.

In contrast, demands for identical statistical significance could be excessive. Any replicating team might dream of publishing the standard statistically significant result that appeared in the previous trial. They might sit down with power tables to see what they might need to reach the sacred $P < 0.05$ for the same dependent variable used in the original experiment. If they only were to accept a statistically significant result, even one-tailed (given that the therapy should do better than the control), they might want to specify power up front. Assuming a Type I error rate of 0.05, the knee-jerk response is often to set power to 0.80. But this plan literally means that 20% of the time, a true effect will go undetected. The team might prefer something more definite, like power of 0.99. But the results for this range of effect sizes is humbling.

For the high end of the prediction interval ($d = 1.53$), power of 0.99 and an alpha of 0.05 (one-tailed) requires a mere 15 participants per group. But trusting in an effect size estimate from such a small sample seems ill-advised given that the Phase III trial was almost three times as large. Should the effect exactly replicate the effect size of the Phase III trial at $d = 0.91$, an N of 39 per group would have a power of 0.99. But the lower bound of 0.29 would require 376 per group to reach power of 0.99, which is more than four times as large as the original experiment they want to replicate (Faul, Erdfelder, Buchner, & Lang, 2009). Demanding statistical significance could require dramatically more resources than focusing on the prediction interval. The time to recruit, screen, and complete the process for 758 people (376 for the treatment and 376 controls) could take years and years. Many people suffering from PTSD could miss out on an efficacious treatment waiting for these data to publish.

Alternatively, the attempts at replication could focus on sample sizes comparable to those used to demonstrate efficacy for other treatments of the same or comparable problems and focus on the prediction interval rather than statistical significance. For example, a meta-analytic review of psychological therapies for PTSD examined over 100 trials, with an average sample size much like the one used for this Phase III MDMA study (see Lewis, Roberts, Andrew, Starling, & Bisson, 2020). A replication using the same sample sizes as this original trial, and a hypothesis involving the prediction interval, might well be within reach. Nevertheless, a demand for a miniscule $P$-value risks abandoning a treatment simply because random error (related to sampling) got in the way. Resources are too valuable and the need for treatment is too dire for us to ignore prediction intervals as an alternative approach. A focus on prediction intervals, rather than $P$-values, would help establish psychedelic-assisted therapies as empirically validated treatments without risking dismissing their efficacy because of sampling error.

**REFERENCES**


