Psychological Medicine

cambridge.org/psm

Assessing expectancy and suggestibility in a trial of escitalopram v. psilocybin for depression

Balázs Szigeti[1,6], Brandon Weiss[1], Fernando E. Rosas[1,2,3,4], David Erritzoe[1], David Nutt[1] and Robin Carhart-Harris[5]

[1]Centre for Psychedelic Research, Imperial College London, UK; [2]Centre for Complexity Science, Imperial College London, UK; [3]Department of Informatics, University of Sussex, Brighton, UK; [4]Centre for Eudaimonia and Human Flourishing, University of Oxford, Oxford, UK; [5]Depts. of Neurology, Psychiatry and Behavioral Sciences, University of California San Francisco, San Francisco, USA and [6]Department of Psychiatry and Behavioral Sciences, University of California San Francisco, San Francisco, USA

Original Article

This work has not been presented previously elsewhere.

Cite this article: Szigeti B, Weiss B, Rosas FE, Erritzoe D, Nutt D, Carhart-Harris R (2024). Assessing expectancy and suggestibility in a trial of escitalopram v. psilocybin for depression. *Psychological Medicine* **54**, 1717–1724. https://doi.org/10.1017/S0033291723003653

Received: 2 March 2023
Revised: 29 September 2023
Accepted: 23 November 2023
First published online: 22 January 2024

Keywords:
clinical trial; comparative effectiveness trial; depression; expectancy; MDD; psilocybin; psychedelic; SSRI

Corresponding author:
Balázs Szigeti;
Email: balazs.szigeti07@imperial.ac.uk

Abstract

Background. To investigate the association between pre-trial expectancy, suggestibility, and response to treatment in a trial of escitalopram and investigational drug, COMP360, psilocybin, in the treatment of major depressive disorder (ClinicalTrials.gov registration: NCT03429075).
Methods. We used data ($n = 55$) from our recent double-blind, parallel-group, randomized head-to-head comparison trial of escitalopram and investigational drug, COMP360, psilocybin. Mixed linear models were used to investigate the association between pre-treatment efficacy-related expectations, as well as baseline trait suggestibility and absorption, and therapeutic response to both escitalopram and COMP360 psilocybin.
Results. Patients had significantly higher expectancy for psilocybin relative to escitalopram; however, expectancy for escitalopram was associated with improved therapeutic outcomes to escitalopram, expectancy for psilocybin was not predictive of response to psilocybin. Separately, we found that pre-treatment trait suggestibility was associated with therapeutic response in the psilocybin arm, but not in the escitalopram arm.
Conclusions. Overall, our results suggest that psychedelic therapy may be less vulnerable to expectancy biases than previously suspected. The relationship between baseline trait suggestibility and response to psilocybin therapy implies that highly suggestible individuals may be primed for response to this treatment.

Introduction

Depression affects approximately 400 million people globally and mental illness is forecasted to be the leading contributor to the global burden of disease by 2030 (Wellcome Global Monitor, 2021). The most prescribed medications for the treatment of depression are selective serotonin reuptake inhibitors (SSRIs) such as escitalopram (Lexapro), fluoxetine (Prozac), and sertraline (Zoloft). A recent comprehensive meta-analysis showed that these drugs are superior to placebo in the treatment of depression (Cipriani et al., 2018), although with small effect sizes relative to placebo (<0.3 standardized mean difference) (Hengartner & Plöderl, 2018). Considering the social and economic costs of depression and that many individuals reject or fail to comply with chronic medication strategies, alternative treatments are needed.

A promising new treatment for depression is psychedelic-assisted therapy (Nutt, Erritzoe, & Carhart-Harris, 2020). In this paradigm, psychedelics – such as psilocybin and lysergic acid diethylamide (LSD) – are assumed to interact positively with psychotherapeutic processes. During such 'psychedelic therapy', patients take a moderate-to-large dose of the psychedelic on one or two occasions with psychological supervision to guide the therapeutic process (Garcia-Romeu & Richards, 2018).

Recently, we conducted a head-to-head comparative trial of escitalopram, a highly selective SSRI and one of the most commonly prescribed antidepressants, *v.* investigational psilocybin-therapy, for the treatment of depression (Carhart-Harris et al., 2021). According to the pre-defined primary outcome, the mean of the self-rated 16-item *Quick Inventory of Depressive Symptomology* (Rush et al., 2003) scale (QIDS-SR-16), the between-treatment difference at the week 6 primary endpoint was not statistically significant. However, both response (70% *v.* 48%) and remission (57% *v.* 28%) rates, as scored on the QIDS-SR-16, favored psilocybin. Moreover, on all mental-health-related secondary outcomes, psilocybin therapy was superior by a greater than 95% confidence margin.

Patient expectations can influence therapeutic outcomes (Tambling, 2012). Researchers attempt to address expectancy effects in clinical trials by randomization and experimental 'blinding', i.e. by

concealing treatment allocation from patients and assessors. However, effective blinding is difficult to achieve in practice (Baethge, Assall, & Baldessarini, 2013), particularly in psychedelic trials (Muthukumaraswamy, Forsyth, & Lumley, 2021), due to the conspicuous subjective drug effects that enable most patients to deduce their treatment allocation. For example in another recent trial of psilocybin-assisted therapy, 94% of participants correctly guessed their treatment allocation, indicating that blinding was broken (Bogenschutz et al., 2022). Based on this consideration, a number of authors have expressed concerns over the methodological soundness of psychedelic trials, arguing that expectancy effects may be biasing the observed results despite the formal blinding procedures (Burke & Blumberger, 2021; Muthukumaraswamy et al., 2021; Szigeti, Nutt, Carhart-Harris, & Erritzoe, 2023). In the case of 'psychedelic microdosing', where users regularly take small doses of a psychedelic drug without clinical supervision (Polito & Liknaitzky, 2022), a number of studies, including some that were placebo-controlled (Cavanna et al., 2022; de Wit, Molla, Bershad, Bremmer, & Lee, 2022; Szigeti et al., 2021) suggest that positive expectancy may play an important role in driving positive responses highlighting a need to investigate expectancy and related effects in all psychedelic trials.

Methods

A trial of escitalopram v. psilocybin

This was a phase 2, investigator-initiated, double-blind, randomized trial in patients with moderate-to-severe major depressive disorder. The core treatment period was six weeks and the trial had two treatment arms. In the 'psilocybin arm', patients received two separate doses of 25 mg of investigational drug, COMP360, i.e. psilocybin, three weeks apart plus six weeks of daily placebo capsules. In the 'escitalopram arm' patients received two separate doses of 1 mg of psilocybin three weeks apart, which was viewed by the research team as a placebo, plus six weeks of daily oral escitalopram (10 mg/day the first 3 weeks, 20 mg/day for the final 3 weeks) – which was considered the main active component of this arm. Patients were randomly assigned to treatment groups in a 1:1 ratio. All patients received psychological support during the trial period. The trial was registered on ClinicalTrials.gov (NCT03429075), all the patients provided written informed consent and approval was obtained from all relevant regulator bodies, see (Carhart-Harris et al., 2021) for further details.

Baseline measures

To measure patients' expectations, the following two items were administered one day before both dosing days:

- 'Please rate the following with regards to the prospect of receiving 6 weeks of daily escitalopram. At the end of the trial, after receiving escitalopram every day for 6 weeks, how much improvement in your mental health do you think will occur?'
- 'Please rate the following with regard to the prospect of receiving two full strong doses of psilocybin, 3 weeks apart. At the end of the trial, 3 weeks after your second psilocybin dosing session, how much improvement in your mental health do you think will occur?'

Ratings of these items are referred to as '*escitalopram expectancy*' and '*psilocybin expectancy*', respectively. These items refer specifically to *efficacy-related* expectancy, e.g. as opposed to side effects, and will be referred to as *expectancy* from here on. Here, we use the expectancy measures obtained before the first dosing day, i.e. pre-treatment expectancy. Responses were collected on a 0–100 visual analog scale with anchor points at 0 ('0% improvement') and 100 ('100% improvement').

In this manuscript, we use '*received treatment expectancy*' as the expectancy measure, which is the expectancy associated with the actually received treatment for each patient (i.e. *escitalopram expectancy* when allocated to the escitalopram arm and *psilocybin expectancy* when allocated to the psilocybin arm). We choose to analyze the data this way because the 25 mg psilocybin used in the current trial induces strong psychological and physical effects that can be reliably recognized and attributed to psilocybin by most patients (Muthukumaraswamy et al., 2021), thus, blinding integrity is unlikely to have been maintained (Szigeti et al., 2023). Similarly, blinding integrity is also often violated in SSRI trials (Scott, Sharpe, & Colagiuri, 2022).

Suggestibility, which is the tendency to comply with suggestions from others (Wagstaff, 1991), was assessed with the *Short Suggestibility Scale* (SSS) (Kotov, Bellman, & Watson, 2004) at baseline. Absorption, which represents a predisposition to experience altered states of consciousness (Ott, Reuter, Hennig, & Vaitl, 2005), was assessed with the *Modified Tellegen Absorption Scale* (MODTAS) (Jamieson, 2005) also at baseline. These measures were included here because previous work has indicated that suggestibility and absorption are correlated (Milling, Kirsch, & Burgess, 2000), and that absorption may be predictive of the nature of psychedelic experiences (Aday, Davis, Mitzkovitz, Bloesch, & Davoli, 2021; Haijen et al., 2018).

Outcome measures

Pre-defined primary outcome was the change in the mean sum score of the self-rated *Quick Inventory of Depressive Symptomology Scale* (QIDS-SR-16) (Rush et al., 2003) at the six-week primary endpoint, while secondary outcomes included the clinician-rated *Beck Depression Inventory* (BDI) (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), the *Hamilton Depression Rating Scale* (HAM-D) (Hamilton, 1960) and the *Montgomery-Åsberg Depression Rating Scale* (MADRS) (Montgomery & Asberg, 1979). Here we re-analyzed these outcomes together with other mood-related secondary outcomes, specifically the self-rated *State-Trait Anxiety Inventory-Trait* (STAI-T) (Spielberger, 1983) and the *Warwick-Edinburgh Mental Well-being Scale* (WEMWBS) (Tennant et al., 2007).

Statistical models

We used linear mixed modeling to assess baseline differences. The first model had *expectancy* as the dependent variable, *patient ID* as random effect, and *expectancy type* (i.e. whether expectancy measure corresponds to escitalopram or psilocybin expectancy) as fixed effect. In a second model, we also added *treatment allocation* and its interaction with *expectancy type* as fixed effects to investigate potential between arm differences. Note that in these expectancy models each patient contributed two rows of data: one for escitalopram expectancy and one for psilocybin expectancy. Next, we constructed similar models for *suggestibility/absorption,* where *suggestibility/absorption* was the dependent variable and *treatment allocation* was the independent variable, see online Supplementary Table S1 for model formulas.

Linear mixed modeling was used to assess the 'within arm' association between outcomes and *expectation/suggestibility/*

https://doi.org/10.1017/S0033291723003653 Published online by Cambridge University Press

absorption, separately, for the two treatment arms. In these models, the dependent variable was *score* on one of six mood/well-being related outcomes (HAM-D, BDI, MADRS, QIDS-SR-16, STAI-T, or WEMWBS), four of which are widely used depression symptom severity scales (HAM-D, BDI, MADRS, QIDS-SR-16), *patient ID* as random effect, *timepoint*, one of the baseline covariates (*expectancy*/*suggestibility*/*absorption*) and it's interaction with *timepoint* as fixed effects, see online Supplementary Tables S2 and S3 for model formula.

Linear mixed modeling was also used to construct between-arms models adjusted for either *expectancy* or *suggestibility*. As before, the dependent variable was *score* on one of the mental-health-related outcomes (HAM-D, BDI, MADRS, QIDS-SR-16, STAI-T, or WEMWBS), with *patient ID* as a random effect, and *timepoint*, *treatment allocation*, *expectancy/suggestibility*, and their interactions as fixed effects, and see online Supplementary Table S4 and S5 for model formulas.

In all models, the pre-treatment covariates, i.e. *expectancy, suggestibility,* and *absorption*, were normalized by subtracting the median and then dividing by the standard deviation. Consequently, all results should be understood as representative at the median level of the covariate and estimates represent the change associated with an increase of 1 standard deviation. We choose to normalize the data at the median instead of the mean to protect against extreme values; however, we note here that normalizing at the mean yields the same qualitative results. In both the within- and between-arms models *expectancy* is defined as the 'received treatment expectancy', i.e., *escitalopram expectancy* for patients in the escitalopram arm and *psilocybin expectancy* for patients in the psilocybin arm, see Baseline measures for details. To control for multiple comparisons, we adjusted the p values with the Bonferroni method (Sedgwick, 2012). Throughout the manuscript, we report these adjusted p values, while all unadjusted p values can be found in the online Supplementary materials. For all models, the normality of the residuals was checked visually from QQ-plots. All models were constructed in R (v4.1.2) using the *lme4* (v1.1-27.1) and *lmerTest* (v3.1-3) packages.

Equivalence testing

In null hypothesis significance testing (NHST), the null hypothesis is either rejected or not, but the null hypothesis itself cannot be confirmed (Lakens, McLatchie, Isager, Scheel, & Dienes, 2020). Equivalence testing allows for an inference to be made on whether the null hypothesis can be accepted, i.e. results of this test can provide evidence to infer an absence of an effect. Specifically, equivalence testing can provide evidence that the true effect is smaller than a pre-specified equivalence bound, also known as 'smallest effect size of interest' or 'region of practical equivalence'. If an equivalence test yields significant results, it means that we can reject the hypothesis that the true effect is as extreme or more extreme than the chosen equivalence bound (Lakens et al., 2020). In contrast, a non-significant equivalence test means that effects as large as the equivalence bound cannot be ruled out.

We used the 'two one-sided t tests' (TOST) equivalence test procedure as implemented by the parameters package (https://rdrr.io/cran/parameters/) to further examine results where the null hypothesis was not rejected. We choose the equivalence bound to be 0.5 standardized mean difference (S.M.D.) because it corresponds to a suggested criterion for inferring *minimum clinically important difference* across a range of medical conditions (Norman, Sloan, & Wyrwich, 2003).

Data and code sharing

The manuscript's repository (https://github.com/szb37/psilodep2) contains a conda computational environment, the data used, and analysis scripts to reproduce all figures and major statistical findings presented here.

Results

Pre-treatment between-group differences

In the full sample, we found a significant difference between the pre-trial efficacy-related expectancy for *escitalopram v. psilocybin* (est ± S.E.: 25.8 ± 3.5; $p < 0.001$***), with estimated means of 54% (psilocybin) *v.* 28.2% (escitalopram) – on a scale of expecting 0–100% mental-health improvements, see Methods for details. There were no significant effects associated with *treatment* allocation (est ± S.E.: −3.2 ± 5.7; $p = 0.580$), nor with its interaction with *expectancy type* (est ± S.E.: 9.3 ± 6.9; $p = 0.183$), implying that, irrespective of group allocation, the sample had uniformly higher expectancy for psilocybin therapy. Changes in expectancy between the first and the second session are described in the online Supplementary materials. Briefly, no significant changes were observed for either *escitalopram* or *psilocybin expectancy* after the first and before the second psilocybin (1 or 25 mg) dosing session.

We found no significant between group differences with respect to baseline trait *suggestibility* (est ± S.E.: −2.5 ± 2.8; $p = 0.368$), *absorption* (est ± S.E.: 4.4 ± 4; $p = 0.272$), or any of the absorption related subscales, see online Supplementary Table S1 for details and Fig. 1 for boxplots.

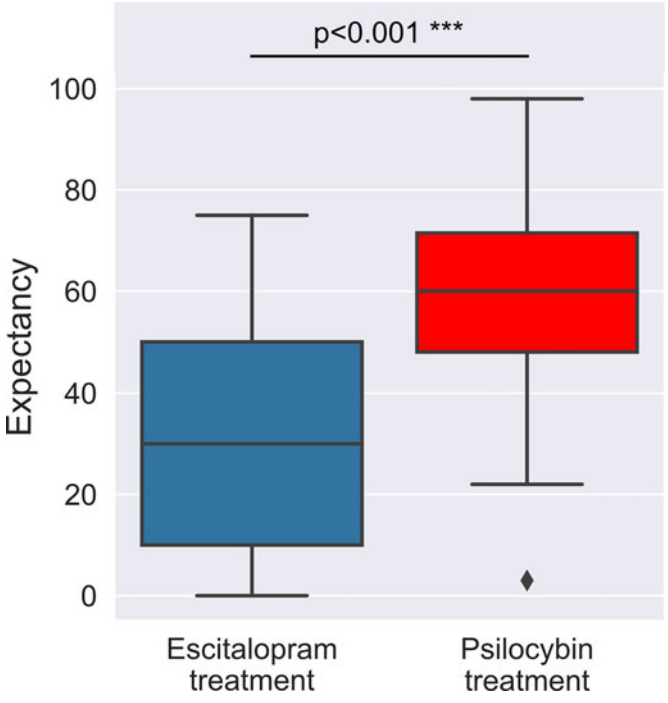

Figure 1. Boxplots of the observed *expectancy* scores at baseline. A substantial difference was found between *escitalopram* and *psilocybin expectancy* with estimated means of 54 (psilocybin) *v.* 28.2 (escitalopram). An expectancy score of 0 for either treatment implied an expectation of no improvement in mental health, whereas a score of 100 would have implied 100% improvement, see Methods for details. This expectancy imbalance was equally present in both treatment arms, see online Supplementary Table S1 for details.

https://doi.org/10.1017/S0033291723003653 Published online by Cambridge University Press

Within-treatment-arm association between expectancy *and therapeutic outcomes*

In the escitalopram arm, we found a significant interaction between *expectancy* and *timepoint* when predicting outcomes on the HAM-D (est. ± S.E.: −3.91 ± 0.9, adj. $p = 0.001$**), BDI (est. ± S.E.: −5.47 ± 1.61, adj. $p = 0.013$*), MADRS (est. ± S.E.: −4.87 ± 1.52, adj. $p = 0.022$*) and STAI-T (est. ± S.E.: −5.2 ± 1.68, adj. $p = 0.028$*) scales, but not on the QIDS-SR-16 (est. ± S.E.: −2.46 ± 0.98, adj. $p = 0.115$) and WEMWBS (est. ± S.E.: 2.95 ± 1.76, adj. $p = 0.641$) scales, see online Supplementary Table S2 for details. These findings suggest that on the HAM-D, BDI, MADRS, and STAI-T scales, there is a positive association between pre-treatment expectations for escitalopram and improved outcomes in the escitalopram arm. Specifically, on the HAM-D scale, each standard deviation (~22 points on the expectancy scale) increase in *expectancy* is associated with 3.91 points reduction in depression scores, etc.

Conversely, in the psilocybin arm, we found no significant interaction between *expectancy* and *timepoint* when predicting outcomes on any of the scales (HAM-D est. ± S.E.: 1.16 ± 1.05, adj. $p = 1$; BDI est. ± S.E.: 1.14 ± 2.4, adj. $p = 1$; MADRS est. ± S.E.: 1.69 ± 1.81, adj. $p = 1$; QIDS-SR-16 est. ± S.E.: 0.56 ± 1.25, adj. $p = 1$; STAI-T est. ± S.E.: 0.64 ± 2.66, adj. $p = 1$; WEMWBS est. ± S.E.: −2.96 ± 2.44, adj. $p = 1$), suggesting a lack of association between pre-treatment expectancy and therapeutic outcomes, see online Supplementary Table S3 for details. Equivalence testing the *expectancy* × *timepoint* interaction term yielded non-significant results on all scales, suggesting that we cannot rule out a true effect as large as the minimum important difference, see Equivalence testing and online Supplementary Table S6 for details. Figure 2 shows the *expectancy v.* outcomes regression lines for both treatment arms.

Within-treatment-arm association among suggestibility, absorption, *and therapeutic outcomes*

In the escitalopram arm, we found no significant interaction between baseline *suggestibility* and *timepoint* when predicting outcomes on any of the scales (HAM-D est. ± S.E.: 0.9 ± 1.07, adj. $p = 1$; BDI est. ± S.E.: 1.03 ± 1.83, adj. $p = 1$; MADRS est. ± S.E.: 2.78 ± 1.53, adj. $p = 0.490$; QIDS-SR-16 est. ± S.E.: 1.08 ± 1.01, adj. $p = 1$; STAI-T est. ± S.E.: 1.1 ± 1.71, adj. $p = 1$; WEMWBS est. ± S.E.: −2.78 ± 1.55, adj. $p = 0.516$), suggesting a lack of association between baseline *suggestibility* and therapeutic response to escitalopram, see online Supplementary Table S2 for details. Equivalence testing the *suggestibility* × *timepoint* interaction term yielded non-significant results on all scales except BDI and STAIT, see Equivalence testing and online Supplementary Table S6 for details; however, even on these two scales, the significance did not survive the Bonferroni correction.

In the psilocybin arm, we found a significant interaction between *suggestibility* and therapeutic response on all scales (HAM-D est. ± S.E.: −3.46 ± 0.92, adj. $p = 0.005$**; BDI est. ± S.E.: −7.16 ± 1.94, adj. $p = 0.006$**; MADRS est. ± S.E.: −6.36 ± 1.37,

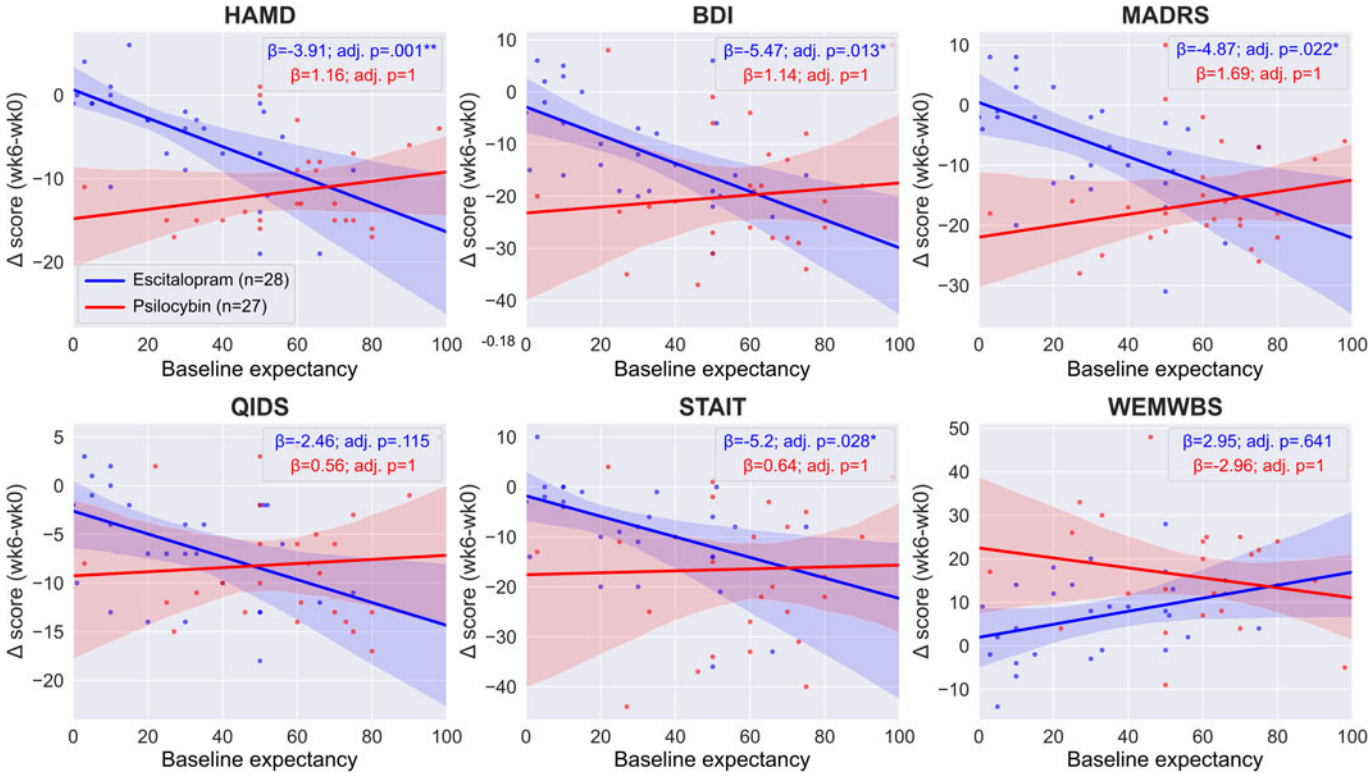

Figure 2. Regression lines of pre-treatment efficacy expectations *v.* clinical efficacy. Regression lines and coefficients were obtained from the two separate 'within arm' models, see Statistical models for details. There was a significant association between pre-treatment *expectancy* and response to escitalopram as assessed using the HAM-D, BDI, MADRS and STAI-T scales (blue lines), in contrast, there was no association for any outcome in the psilocybin arm (red lines). Note that the significance level is different between the two arms on four scales (HAMD, BDI, MADRS, STAI-T), but the difference reached significance only on two of them (HAMD, MADRS), see between-arm models for details. Boxes show the regression coefficients (β) associated with the *expectancy* × *timepoint* term in the two separate arms and the associated Bonferroni adjusted *p* values. Negative values indicate improved symptoms except for WEMWBS where positive values indicate improved well-being, see online Supplementary Tables S2 and S3 for details.

https://doi.org/10.1017/S0033291723003653 Published online by Cambridge University Press

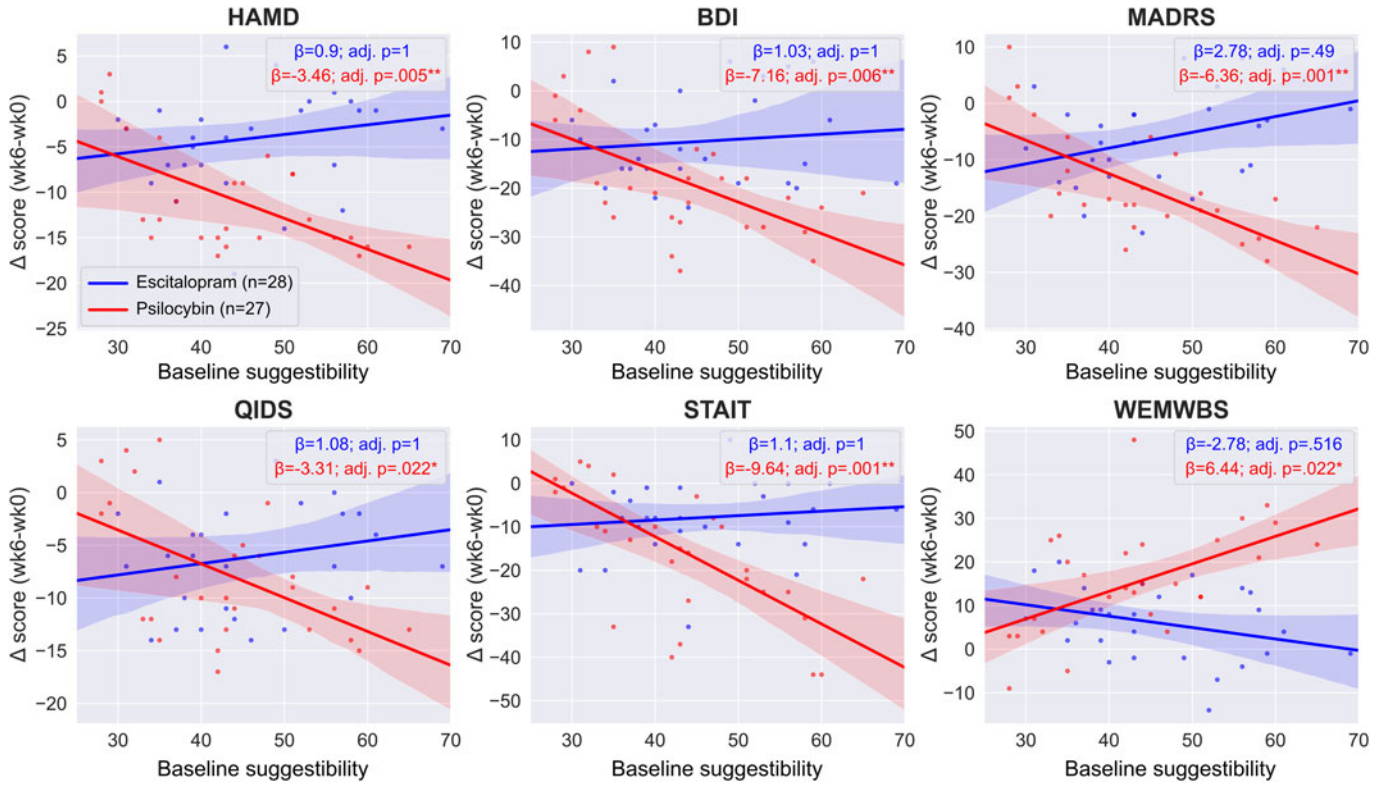

Figure 3. Regression lines of trait suggestibility *v.* clinical efficacy. Regression lines and coefficients were obtained from the two separate 'within arm' models, see Statistical models for details. There was a significant association between baseline *suggestibility* and outcomes on all scales in the psilocybin arm (red lines) that was maintained after adjusting for multiple comparisons. In contrast, there was no such significant relationship in the escitalopram arm (blue lines). Between-arm models indicate that not only the significance level was different between the treatment arms on all scales, but that the difference was also significant (Nieuwenhuis et al., 2011). Boxes show the regression coefficients (β) associated with the *suggestibility* × *timepoint* term in the two separate arms and the associated Bonferroni adjusted p values. Negative values indicate improved symptoms except for WEMWBS where positive values indicate improved well-being, see online Supplementary Tables S2 and S3 for details.

adj. $p = 0.001$***; QIDS-SR-16 est. ± s.e.: −3.31 ± 1.04, adj. $p = 0.022$*; STAI-T est. ± s.e.: −9.64 ± 2.1, adj. $p = 0.001$*; WEMWBS est. ± s.e.: 6.44 ± 2.02, adj. $p = 0.022$*), implying a robust association between baseline *suggestibility* and therapeutic response to psilocybin, see online Supplementary Table S3 for details. The findings suggest that, on the HAM-D scale, each standard deviation increase (~10 points on the *Short Suggestibility Scale*) of *suggestibility* is associated with 3.46 reduction in depression scores, etc. Figure 3 shows the *suggestibility v.* outcomes regression lines for both treatment arms.

We found no significant interaction between *absorption* and *timepoint* in either the escitalopram or the psilocybin arm on any of the scales, suggesting a lack of association between baseline *absorption* and response, see online Supplementary Tables S2 and S3 for details.

Between-treatment difference in models adjusted for expectancy *and* suggestibility

When adjusting the trial results for pre-trial expectancy, there was no significant interaction term between *timepoint* and *treatment* on any of the scales after adjusting for multiple comparisons (HAMD est. ± s.e.: −3.06 ± 1.67, adj. $p = 0.438$; BDI est. ± s.e.: −3.32 ± 3.49, adj. $p = 1$; MARDS est. ± s.e.: −4.52 ± 2.85, adj. $p = 0.711$; QIDS est. ± s.e.: 0.82 ± 1.93, adj. $p = 1$; STAI-T est. ± s.e.: −3.03 ± 3.92, adj. $p = 1$; WEMWBS est. ± s.e.: 7.82 ± 3.61, adj. $p = 0.214$), implying that there is no difference between the treatments after adjusting for expectancy. Equivalence testing yielded non-significant results on all scales, suggesting that we cannot rule out a true effect as large as the minimum important difference, see Equivalence testing and online Supplementary Table S6 for details.

The *treatment* × *timepoint* × *expectancy* interaction term was significant on the HAMD and MADRS scales (HAMD est. ± s.e.: 6.02 ± 1.63, adj. $p = 0.003$**; MARDS est. ± s.e.: 7.79 ± 2.78, adj. $p = 0.043$*), suggesting that, on these two scales, the difference between the treatment arms reached significance; however, the difference was not significant on the other 4 scales (BDI est. ± s.e.: 7.88 ± 3.39, adj. $p = 0.146$; QIDS est. ± s.e.: 3.6 ± 1.87, adj. $p = 0.361$; STAI-T est. ± s.e.: 6.96 ± 3.81, adj. $p = 0.441$; WEMWBS est. ± s.e.: −6.82 ± 3.54, adj. $p = 0.361$). When adjusting the trial results for suggestibility, the results qualitatively remained the same as for the unadjusted models (Carhart-Harris et al., 2021); specifically, there was a significant interaction term between *timepoint* and *treatment* on all scales except QIDS (HAMD est. ± s.e.: −5.88 ± 1.44, adj. $p < 0.001$***; BDI est. ± s.e.: −7.48 ± 2.7, adj. $p = 0.047$*; MARDS est. ± s.e.: −7.36 ± 2.09, adj. $p = 0.006$**; QIDS est. ± s.e.: −1.37 ± 1.46, adj. $p = 1$; STAI-T est. ± s.e.: −8.17 ± 2.75, adj. $p = 0.027$*; WEMWBS est. ± s.e.: 9.34 ± 2.59, adj. $p = 0.004$**). This finding suggests a between-treatment difference at the primary endpoint after adjusting for suggestibility, favoring the psilocybin condition on all scales, see online Supplementary Table S5 for details. The *treatment* × *timepoint* × *suggestibility* interaction term was significant on all scales (HAMD est. ± s.e.: −4.34 ± 1.42, adj. $p = 0.021$*; BDI est. ± s.e.: −8.12 ± 2.68, adj.

https://doi.org/10.1017/S0033291723003653 Published online by Cambridge University Press

$p = 0.023$*; MARDS est. ± S.E.: −9.12 ± 2.06, adj. $p < 0.001$***; QIDS est. ± S.E.: −4.37 ± 1.45, adj. $p = 0.024$*; STAI-T est. ± S.E.: −10.66 ± 2.73, adj. $p = 0.002$**; WEMWBS est. ± S.E.: 9.2 ± 2.57, adj. $p = 0.005$**), suggesting that not only was the significance level different between the treatment arms but that the difference was also significant (Nieuwenhuis, Forstmann, & Wagenmakers, 2011); see within-arm suggestibility models.

Discussion

Analyzing pre-treatment efficacy-related expectations in a trial of escitalopram *v.* psilocybin for the treatment of depression (Carhart-Harris et al., 2021), we found that patients had substantially higher expectancy for psilocybin therapy compared with escitalopram; however, when we assessed whether an association exists between pre-trial expectancy and therapeutic response, we found a significant association in the escitalopram arm, but not in the psilocybin arm.

The escitalopram results are consistent with previous findings pertaining to SSRIs (Bingel et al., 2011). However, the lack of association for the psilocybin arm is surprising given that expectancy effects are associated with improved outcomes across a wide range of medical diagnoses and therapeutic approaches (Tambling, 2012), including one naturalistic study of psychedelic use that assessed expectancy with a self-constructed binary (yes/no) questionnaire (Weiss, Miller, Carter, & Keith Campbell, 2021), rather than using a continuous scale. Suspicion has been expressed that in psychedelic trials the combination of the lack of effective blinding, strong demand characteristics, and related confirmation biases may positively bias trial outcomes (Burke & Blumberger, 2021; Muthukumaraswamy et al., 2021; Szigeti et al., 2023). Our results partially alleviate these suspicions, as we did not find a significant association between psilocybin-specific efficacy-related expectations and efficacy-related outcomes.

What explanations can be given for the lack of an expectancy effect in the psilocybin arm? Given that most of our trial patients were self-referred and it is reasonable to assume that many were seeking psilocybin treatment in particular, a ceiling effect on pre-trial expectancy for psilocybin was considered and examined; however, the average psilocybin expectancy score was 51% from a possible 100%, i.e. far from the ceiling. A second possibility is that the relationship is not linear in nature. For example, one could speculate that patients with unrealistically high expectations may be disappointed, leading to worse outcomes with higher expectations; indeed, the slopes of the expectancy *v.* outcome regressions are positive, see Fig. 2, although all of them are highly non-significant. Our sample was too small to investigate complex, non-linear models; however, this would be worth exploring via larger samples – achievable e.g. via pragmatic trials or real-world data collection (Carhart-Harris et al., 2022).

We failed to observe a significant expectancy effect in the psilocybin arm, but such a non-significant result should not be mistaken as evidence from which the absence of an effect can be inferred (Lakens et al., 2020). We performed equivalence testing to confirm the null hypothesis; however, this was non-significant as well. Therefore, from a strict inferential perspective, we cannot either rule out or confirm expectancy effects in the psilocybin arm, more data is needed to test and infer on this matter. We note that our data suggests that 'negative expectancy', i.e. higher expectancy associated with worse response, may be more likely than the generally assumed positive expectancy (Muthukumaraswamy et al., 2021), as indicated by the positive, although non-significant, slopes in Fig. 2. Thus, if there is a 'true' expectancy effect that we were underpowered to detect, it may be that higher expectancy for psilocybin could actually be associated with worse response to psilocybin.

If future research enabled us to accept the null hypothesis, i.e. that there is no association between expectancy and therapeutic response in psilocybin therapy, then this would imply that psilocybin therapy has a direct treatment effect that is independent of positive expectancy. More work is needed to determine what psilocybin's precise therapeutic action is, but some empirical clues and models are emerging (Carhart-Harris & Friston, 2019; Daws et al., 2022; Murphy et al., 2022; Zeifman, Wagner, Monson, & Carhart-Harris, 2023).

In this trial, response to psilocybin was not predicted by baseline expectancy, but the response to escitalopram was, therefore, the between-arm difference is also affected by expectancy. When we adjusted the models for baseline expectancy, there was no between-treatment difference in efficacy on any of the scales. In contrast, models unadjusted for expectancy produced a significant between-arm difference for all depression-related outcomes except on the QIDS-SR-16 scale, as originally reported (Carhart-Harris et al., 2021). This result implies that the observed expectancy imbalance biased the results in favor of psilocybin's superiority, see online Supplementary Table S7 for a direct comparison of the unadjusted and expectancy-adjusted between-arm models. Notably, this expectancy bias is not a result of the patients in the psilocybin arm benefitting from high expectations, as we found no expectancy effect in the psilocybin arm, but rather due to patients having low expectancy in the escitalopram arm, which can be interpreted as a nocebo effect.

Trait suggestibility was predictive of psilocybin efficacy here. Previous research indicates a link between verbal suggestibility and placebo responsiveness (Oakley, Walsh, Mehta, Halligan, & Deeley, 2021; Parsons, Bergmann, Wiech, & Terhune, 2021). Together, these findings could be interpreted as evidence for extra-pharmacological factors driving the response in the psilocybin arm, demand characteristics, and/or the Hawthorne effect, playing a role in psilocybin's efficacy, future trials may further examine this possibility. In a recent prospective naturalistic study on ayahuasca, suggestibility was associated with a greater reduction in trait neuroticism after ayahuasca (Weiss et al., 2021). One other naturalistic study failed to see a relationship between baseline trait suggestibility and either acute subjective experience or changes in well-being (Haijen et al., 2018); however, this latter null findings may have been a product of a multivariate regression approach and potential collinearity between model components. Baseline absorption has previously been found to be predictive of the acute subjective intensity of psychedelic effects (Aday et al., 2021; Haijen et al., 2018), which in turn may predict therapeutic outcomes (Murphy et al., 2022; Roseman, Nutt, & Carhart-Harris, 2017); however, here we did not find a direct link between absorption and response in either treatment arms. More work is needed to test the reliability with which trait suggestibility can predict response to psilocybin therapy, as well as what the mechanisms are for this apparent effect – e.g. is it more biologically grounded (Ott et al., 2005), or more psychologically based (De Pascalis, Chiaradia, & Carotenuto, 2002), or are the two inter-related and do they interact? High trait absorption could imply elevated sensitivity to direct drug effects (Ott et al., 2005), while high suggestibility could imply elevated attunement to acute insights, and influence from therapy personnel such

https://doi.org/10.1017/S0033291723003653 Published online by Cambridge University Press

as the therapist or clinical staff (Cherniak et al., 2021; Murphy et al., 2022).

Limitations and future work

The analysis presented here was not pre-registered; thus, our results should be understood as exploratory rather than confirmatory (Jaeger & Halliday, 1998). Furthermore, in the absence of any experimental manipulation of expectancy, all relationships reported here should be interpreted as correlational, not causal. Further studies are needed to assess causation, e.g. by seeking to manipulate expectations in a controlled and measured way.

The non-significant equivalence tests for the expectancy-outcome association in the psilocybin arm suggest that we cannot rule out an expectancy effect as large as 0.5 standardized mean difference (S.M.D.), corresponding to the *minimum important difference* (Norman et al., 2003). Our trial was not powered to reject effects as small as the *minimum important difference*, thus, the failed equivalence test may be a consequence of the small sample. Also, the expectancy measure used here was not a validated survey. It is possible that using validated expectancy measures would find different results from those presented in this paper.

No 'treatment allocation guess' data was collected either from patients or assessors in the current trial, meaning we could not evaluate blinding integrity or its interaction with expectancy. It is plausible that expectancy could interact with perceived treatment allocation – and the confidence level of this 'guess' – to influence response outcomes (e.g. disappointment at confidently realizing you have been allocated to the escitalopram arm). A new measure of blinding integrity that incorporates these features is introduced in another paper (Szigeti et al., 2023). We note that the expectancy measure used here was administered pre-trial for each arm when randomization had not yet determined treatment allocation. From the available data, we could derive a hypothetical treatment-agnostic expectancy measure, i.e., by averaging the expectancies for both treatments. However, this averaged or 'treatment agnostic' expectancy score did not qualitatively alter any of our conclusions, see online Supplementary materials for details.

We finally note that while the current paper has focused specifically on positive expectancy in relation to measures of therapeutic efficacy, i.e., mechanisms relevant to the so-called 'placebo effect', one could also examine expectancy regarding adverse effects - i.e., nocebo effects. The investigation of negative expectancy and negative outcomes in psychedelic trials is a worthwhile avenue for future investigations, as it could inform on risk type, prevalence, and mitigation.

Conclusions

We observed higher pre-trial positive expectancy for psilocybin *v.* escitalopram but found no evidence that efficacy-related expectations for psilocybin could predict therapeutic actual response to psilocybin. Conversely, pre-trial expectancy for escitalopram was reliably predictive of response to escitalopram across most of the efficacy-related outcome measures, in line with what is generally known about the influence of expectancy on response. Baseline trait suggestibility was predictive of response to psilocybin, but not to escitalopram.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291723003653

Funding statement. This work was funded by the Alexander Mosley Charitable Trust and Imperial College London's Centre for Psychedelic Research.

Competing interests. B.Sz., B.W., and F.E.R. declare no conflict. D.N. is an advisory to COMPASS Pathways, Neural Therapeutics, and Algernon Pharmaceuticals; received consulting fees from Algernon, H. Lundbeck, and Beckley Psytech; received lecture fees from Takeda and Otsuka and Janssen plus owns stock in Alcarelle, Awakn and Psyched Wellness. D.E. received consulting fees from Aya, Mindstate, Field Trip, and Clerkenwell Health. R.C.H. is an advisor to Mindstate, TRYP Therapeutics, Maya Health, Entheos Lab, and Journey Collab.

References

Aday, J. S., Davis, A. K., Mitzkovitz, C. M., Bloesch, E. K., & Davoli, C. C. (2021). Predicting reactions to psychedelic drugs: A systematic review of states and traits related to acute drug effects. *ACS Pharmacology & Translational Science*, *4*(2), 424–435. https://doi.org/10.1021/acsptsci.1c00014

Baethge, C., Assall, O. P., & Baldessarini, R. J. (2013). Systematic review of blinding assessment in randomized controlled trials in schizophrenia and affective disorders 2000–2010. *Psychotherapy and Psychosomatics*, *82*(3), 152–160. https://doi.org/10.1159/000346144

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571. https://doi.org/10.1001/archpsyc.1961.01710120031004

Bingel, U., Wanigasekera, V., Wiech, K., Ni Mhuircheartaigh, R., Lee, M. C., Ploner, M., & Tracey, I. (2011). The effect of treatment expectation on drug efficacy: Imaging the analgesic benefit of the opioid remifentanil. *Science Translational Medicine*, *3*(70), 70ra14. https://doi.org/10.1126/scitranslmed.3001244

Bogenschutz, M. P., Ross, S., Bhatt, S., Baron, T., Forcehimes, A. A., Laska, E., … Worth, L. (2022). Percentage of heavy drinking days following psilocybin-assisted psychotherapy vs placebo in the treatment of adult patients with alcohol use disorder: A randomized clinical trial. *JAMA Psychiatry*, *79*(10), 953–962. https://doi.org/10.1001/jamapsychiatry.2022.2096

Burke, M. J., & Blumberger, D. M. (2021). Caution at psychiatry's psychedelic frontier. *Nature Medicine*, *27*(10), 1687–1688. https://doi.org/10.1038/s41591-021-01524-1

Carhart-Harris, R., Giribaldi, B., Watts, R., Baker-Jones, M., Murphy-Beiner, A., Murphy, R., … Nutt, D. J. (2021). Trial of psilocybin versus escitalopram for depression. *New England Journal of Medicine*, *384*(15), 1402–1411. https://doi.org/10.1056/NEJMoa2032994

Carhart-Harris, R. L., & Friston, K. J. (2019). REBUS and the anarchic brain: Toward a unified model of the brain action of psychedelics. *Pharmacological Reviews*, *71*(3), 316–344. https://doi.org/10.1124/pr.118.017160

Carhart-Harris, R. L., Wagner, A. C., Agrawal, M., Kettner, H., Rosenbaum, J. F., Gazzaley, A., … Erritzoe, D. (2022). Can pragmatic research, real-world data and digital technologies aid the development of psychedelic medicine? *Journal of Psychopharmacology (Oxford, England)*, *36*(1), 6–11. https://doi.org/10.1177/02698811211008567

Cavanna, F., Muller, S., de la Fuente, L. A., Zamberlan, F., Palmucci, M., Janeckova, L., … Tagliazucchi, E. (2022). Microdosing with psilocybin mushrooms: A double-blind placebo-controlled study. *Translational Psychiatry*, *12*(1), 307. https://doi.org/10.1038/s41398-022-02039-0

Cherniak, A., Brulin, J. G., Mikulincer, M., Ostlind, S., Carhart-Harris, R. L., & Granqvist, P. (2021). Psychedelic science of spirituality and religion: An attachment-informed agenda proposal. *The International Journal for the Psychology of Religion*, *33*(4), 259–276. https://doi.org/10.31234/osf.io/x58m9

Cipriani, A., Furukawa, T. A., Salanti, G., Chaimani, A., Atkinson, L. Z., Ogawa, Y., … Geddes, J. R. (2018). Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: A systematic review and network meta-analysis. *FOCUS*, *16*(4), 420–429. https://doi.org/10.1176/appi.focus.16407

Daws, R. E., Timmermann, C., Giribaldi, B., Sexton, J. D., Wall, M. B., Erritzoe, D., … Carhart-Harris, R. (2022). Increased global integration in the brain

after psilocybin therapy for depression. *Nature Medicine*, *28*(4), 844–851. https://doi.org/10.1038/s41591-022-01744-z

De Pascalis, V., Chiaradia, C., & Carotenuto, E. (2002). The contribution of suggestibility and expectation to placebo analgesia phenomenon in an experimental setting. *Pain*, *96*(3), 393–402. https://doi.org/10.1016/S0304-3959(01)00485-7

de Wit, H., Molla, H. M., Bershad, A., Bremmer, M., & Lee, R. (2022). Repeated low doses of LSD in healthy adults: A placebo-controlled, dose–response study. *Addiction Biology*, *27*(2), e13143. https://doi.org/10.1111/adb.13143

Garcia-Romeu, A., & Richards, W. A. (2018). Current perspectives on psychedelic therapy: Use of serotonergic hallucinogens in clinical interventions. *International Review of Psychiatry* (*Abingdon, England*), *30*(4), 291–316. https://doi.org/10.1080/09540261.2018.1486289

Haijen, E. C. H. M., Kaelen, M., Roseman, L., Timmermann, C., Kettner, H., Russ, S., … Carhart-Harris, R. L. (2018). Predicting responses to psychedelics: A prospective study. *Frontiers in Pharmacology*, *9*, 897. https://doi.org/10.3389/fphar.2018.00897

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, *23*(1), 56–62.

Hengartner, M. P., & Plöderl, M. (2018). Statistically significant antidepressant-placebo differences on subjective symptom-rating scales do not prove that the drugs work: Effect size and method bias matter!. *Frontiers in Psychiatry*, *9*, 517. https://doi.org/10.3389/fpsyt.2018.00517

Jaeger, R. G., & Halliday, T. R. (1998). On confirmatory versus exploratory research. *Herpetologica*, *54*, S64–S66.

Jamieson, G. A. (2005). The modified tellegen absorption scale: A clearer window on the structure and meaning of absorption. *Australian Journal of Clinical & Experimental Hypnosis*, *33*(2), 119–139.

Kotov, R. I., Bellman, S. B., & Watson, D. B. (2004). Short Suggestibility Scale. Retrieved August 12, 2022, from https://renaissance.stonybrookmedicine.edu/sites/default/files/SSS_BLANK.pdf

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving inferences about null effects with bayes factors and equivalence tests. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *75*(1), 45–57. https://doi.org/10.1093/geronb/gby065

Milling, L. S., Kirsch, I., & Burgess, C. (2000). Hypnotic suggestibility and absorption: Revisiting the context effect. *Contemporary Hypnosis*, *17*(1), 32–41. https://doi.org/10.1002/ch.190

Montgomery, S. A., & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *The British Journal of Psychiatry: The Journal of Mental Science*, *134*, 382–389. https://doi.org/10.1192/bjp.134.4.382

Murphy, R., Kettner, H., Zeifman, R., Giribaldi, B., Kartner, L., Martell, J., … Carhart-Harris, R. (2022). Therapeutic alliance and rapport modulate responses to psilocybin assisted therapy for depression. *Frontiers in Pharmacology*, *12*. Retrieved from https://www.frontiersin.org/articles/10.3389/fphar.2021.788155

Muthukumaraswamy, S. D., Forsyth, A., & Lumley, T. (2021). Blinding and expectancy confounds in psychedelic randomized controlled trials. *Expert Review of Clinical Pharmacology*, *14*(9), 1133–1152. https://doi.org/10.1080/17512433.2021.1933434

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107. https://doi.org/10.1038/nn.2886

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care*, *41*(5), 582–592.

Nutt, D., Erritzoe, D., & Carhart-Harris, R. (2020). Psychedelic psychiatry's brave new world. *Cell*, *181*(1), 24–28. https://doi.org/10.1016/j.cell.2020.03.020

Oakley, D. A., Walsh, E., Mehta, M. A., Halligan, P. W., & Deeley, Q. (2021). Direct verbal suggestibility: Measurement and significance. *Consciousness and Cognition*, *89*, 103036. https://doi.org/10.1016/j.concog.2020.103036

Ott, U., Reuter, M., Hennig, J., & Vaitl, D. (2005). Evidence for a common biological basis of the absorption trait, hallucinogen effects, and positive symptoms: Epistasis between 5-HT2a and COMT polymorphisms. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *137B*(1), 29–32. https://doi.org/10.1002/ajmg.b.30197

Parsons, R. D., Bergmann, S., Wiech, K., & Terhune, D. B. (2021). Direct verbal suggestibility as a predictor of placebo hypoalgesia responsiveness. *Psychosomatic Medicine*, *83*(9), 1041. https://doi.org/10.1097/PSY.0000000000000977

Polito, V., & Liknaitzky, P. (2022). The emerging science of microdosing: A systematic review of research on low dose psychedelics (1955–2021) and recommendations for the field. *Neuroscience and Biobehavioral Reviews*, *139*, 104706. https://doi.org/10.1016/j.neubiorev.2022.104706

Roseman, L., Nutt, D. J., & Carhart-Harris, R. L. (2017). Quality of acute psychedelic experience predicts therapeutic efficacy of psilocybin for treatment-resistant depression. *Frontiers in Pharmacology*, *8*, 974. https://doi.org/10.3389/fphar.2017.00974

Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., … Keller, M. B. (2003). The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, *54*(5), 573–583. https://doi.org/10.1016/s0006-3223(02)01866-8

Scott, A. J., Sharpe, L., & Colagiuri, B. (2022). A systematic review and meta-analysis of the success of blinding in antidepressant RCTs. *Psychiatry Research*, *307*, 114297. https://doi.org/10.1016/j.psychres.2021.114297

Sedgwick, P. (2012). Multiple significance tests: The Bonferroni correction. *BMJ*, *344*, e509. https://doi.org/10.1136/bmj.e509

Spielberger, C. D. (1983). *State-trait anxiety inventory for adults*.

Szigeti, B., Kartner, L., Blemings, A., Rosas, F., Feilding, A., Nutt, D. J., … Erritzoe, D. (2021). Self-blinding citizen science to explore psychedelic microdosing. *eLife*, *10*, e62878. https://doi.org/10.7554/eLife.62878

Szigeti, B., Nutt, D., Carhart-Harris, R., & Erritzoe, D. (2023). The difference between 'placebo group' and 'placebo control': A case study in psychedelic microdosing. *Scientific Reports*, *13*(1), 12107. https://doi.org/10.1038/s41598-023-34938-7

Tambling, R. B. (2012). A literature review of therapeutic expectancy effects. *Contemporary Family Therapy*, *34*(3), 402–415. https://doi.org/10.1007/s10591-012-9201-y

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., … Stewart-Brown, S. (2007). The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): Development and UK validation. *Health and Quality of Life Outcomes*, *5*(1), 63. https://doi.org/10.1186/1477-7525-5-63

Wagstaff, G. F. (1991). Suggestibility: A social psychological approach. *Human suggestibility: Advances in theory, research, and application* (pp. 132–145). Florence, KY, USA: Taylor & Frances/Routledge.

Weiss, B., Miller, J. D., Carter, N. T., & Keith Campbell, W. (2021). Examining changes in personality following shamanic ceremonial use of ayahuasca. *Scientific Reports*, *11*(1), 6653. https://doi.org/10.1038/s41598-021-84746-0

Wellcome Global Monitor: Mental Health. (2021). Retrieved February 10, 2022, from Wellcome website https://wellcome.org/reports/wellcome-global-monitor-mental-health/2020

Zeifman, R. J., Wagner, A. C., Monson, C. M., & Carhart-Harris, R. L. (2023). How does psilocybin therapy work? An exploration of experiential avoidance as a putative mechanism of change. *Journal of Affective Disorders*, *334*, 100–112. https://doi.org/10.1016/j.jad.2023.04.105

https://doi.org/10.1017/S0033291723003653 Published online by Cambridge University Press