UNIVERSITY OF CALIFORNIA SAN DIEGO

Cognitive Development as a Model for the Effects of Psychedelics: Do Changes in Cognitive Flexibility Underly the Clinical Benefits of Psychedelic Therapy?

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Experimental Psychology

by

Ethan S. Hurwitz

Committee in charge:

Professor Adena Schachner, Co-Chair
Professor Caren Walker, Co-Chair
Professor Timothy Brady
Professor Mark Geyer
Professor Adam Halberstadt
Professor Fadel Zeidan

2024

The Dissertation of Ethan S. Hurwitz is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

To Roland Griffiths, who instilled in me that science should be equal parts rigor, curiosity, and fun, and to always remain aware of awareness.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This dissertation would not be possible without the mentorship, guidance, and support, of the following:

Roland Griffiths, Frederick Barret, and Theresa Carbonaro, for helping instill the foundation of my philosophy and perspective on science early in my career development.

Meredith Berry, for perfectly balancing mentorship and friendship.

Elizabeth Lapidow, my fellow cohort of two, and Tanushree Agrawal for their unwavering support, professionally and personally.

Jon Dean, for being with me in the trenches for years and always having my back.

Adena Schachner and Caren Walker, for their willingness to allow me to pursue my passions and provide mentorship and guidance.

My committee members, for supporting my ideas and allowing this work to occur.

Emma Geller, for her mentorship, friendship, support, and providing the best example of what an educator should be.

My family, who may not have understood what I was doing but were unconditionally supportive.

All of the Indigenous Peoples who developed the technology and laid the foundation to make this work possible.

All of my research assistants, mentees, and students, for helping make my work possible and fueling my passion for pedagogy.

Chapter 2 is currently being prepared for submission for publication of the material. Hurwitz, Ethan; Brockbank, Erik; and Walker, Caren. The dissertation author was the primary investigator and author of this material.

Chapter 3 is currently being prepared for submission for publication of the material. Hurwitz, Ethan; Dean, Jon; Brockbank, Erik; McKinty, Arwynn; Farrell, Briana; Gopnik, Alison; and Walker, Caren. The dissertation author was the primary investigator and author of this material.

Chapter 4 is currently being prepared for submission for publication of the material. Hurwitz, Ethan; Brockbank, Erik; Walker, Caren. The dissertation author was the primary investigator and author of this material.

VITA

2013 Bachelor of Science in Psychology, Towson University

2014-2017 Research Coordinator, Johns Hopkins University School of Medicine

2017-2024 Teaching Assistant, University of California San Diego

2019 Master of Arts in Experimental Psychology, University of California San Diego

2022-2024 Associate in Teaching, University of California San Diego

2024 Doctor of Philosophy in Experimental Psychology, University of California San Diego

PUBLICATIONS

[see Google Scholar or www.ethanhurwitz.com]

ABSTRACT OF THE DISSERTATION

Cognitive Development as a Model for the Effects of Psychedelics:

Do Changes in Cognitive Flexibility Underly the Clinical Benefits of Psychedelic Therapy?

by

Ethan S. Hurwitz

Doctor of Philosophy in Experimental Psychology

University of California San Diego, 2024

Professor Adena Schachner, Co-Chair
Professor Caren Walker, Co-Chair

Psychedelics have shown promise as a novel therapeutic, providing clinical benefits for a variety of conditions. Emerging evidence provides a promising cognitive-behavioral account for their therapeutic efficacy, specifically that psychedelics increase cognitive flexibility. However, work to date has yielded inconsistent results. In this dissertation, I overview several experiments utilizing modern cognitive science and computational methodology which test predictions for specific behaviors symptomatic of increased cognitive flexibility. These behaviors are gleaned

from an unlikely source: children. In Chapter 2, I present a novel measure of problem solving under dynamic constraints which can be used to quantify search and sampling strategies. In Chapter 3, I present a pilot test of whether psychedelic treatment will result in (1) decreased influence of prior knowledge on inferences, (2) employment of broader, more exploratory, search strategies when sampling hypotheses, and (3) more diffuse exogenous attention. Finally, in Chapter 4, I examine whether these cognitive flexibility tasks are susceptible to practice effects. If differences in these behaviors can be identified between a control and psychedelic treated group, it would provide initial evidence for what the specific cognitive mechanisms underlying psychedelic therapy may be. With this, it would be possible to better identify who, and what clinical conditions, may be receptive to psychedelic therapy.

Chapter 1 General Introduction

In recent years there has been a resurging interest in studying psychedelic drugs. Through both classic and this more recent work, psychedelics have shown promise as a novel therapeutic, providing clinical benefits for a variety of conditions. The question of **how** remains unanswered. **What are the mechanisms underlying these clinical benefits?** Emerging evidence from clinical interviews and survey studies provides a promising cognitive-behavioral account, specifically that psychedelics increase cognitive flexibility. However, experimental studies implementing cognitive tasks have yielded inconsistent results. In this dissertation, I overview several experiments applying recently developed cognitive tasks and computational methods to test predictions for specific behaviors known to be symptomatic of increased cognitive flexibility. These behaviors are gleaned from an unlikely source: children. If differences in these behaviors can be identified between a control and psychedelic treated group, it would provide initial evidence for what the specific cognitive mechanisms underlying psychedelic therapy may be. With this, it would be possible to better identify who, and what clinical conditions, may be receptive to psychedelic therapy.

Introduction and History of Psychedelics

The term "psychedelics" has conventionally been used to refer to a broad swath of psychoactive drugs characterized by their consciousness altering properties. Psychedelic compounds have been utilized for religious and medicinal purposes throughout history. Some posit that their use extends back as far as the Paleolithic era (Stamets, 1996), and have even suggested they were a contributing factor for the development of modern day humans (an evolutionary account known as the "Stoned Ape Theory"; McKenna, 1999). Wasson and Schultes were among the first Westerners to document current ceremonial use of psilocybin

mushrooms in indigenous Mesoamerican people and tie this ceremonial use to many artifacts with mushroom motifs excavated from Mayan and Aztec temples. Psilocybin was known to the Aztecs as *teonanácatl*, which is widely translated as, "flesh of the gods" (Schultes, 1940; Wasson, 1980). A less popular translation for teonanácatl is, "bread of the gods," which have led some to postulate that the mana referenced in the Hebrew Bible was actually psilocybin mushrooms (Schultes et al., 2001; Strassman, 2014).

Psilocybin, along with *N*,*N*-dimethyltryptamine (DMT), mescaline, and lysergic acid diethylamide (LSD), comprise the group of "classic psychedelics", which have a similar primary mechanism of action as agonists at 5-HT_{2A} receptors and elicit profound acute alterations in a variety of mental functions (Nichols, 2004; Vollenweider & Preller, 2020). These include changes in cognition, perception, emotional processing and empathy, sense of self and volition, as well as persisting psychological and behavioral changes. Acute effects include visual changes (e.g., brighter and more distinct color perception and complex kaleidoscopic visions), distorted perceptions of time and space, increased heart rate and blood pressure, feelings of fear or anxiety, and decreased psychomotor performance (e.g., Carbonaro et al., 2018; Griffiths et al., 2011). A majority of psychedelic users endorse benefiting from their experiences, even when used in less controlled and supportive settings that result in challenging experiences, those characterized by high degrees of anxiety, fear, and difficulty (Carbonaro et al., 2016). This may be due to the mystical-type experiences (MTEs) that psychedelics have been shown to uniquely and reliably elicit (e.g., Griffiths et al., 2006; Ott, 1996; Smith, 1964). MTEs can also be naturally occurring and are characterized by a deeply felt positive mood (peace, joy, love, tranquility, etc.), noesis (attaining intuitive knowledge), sacredness, ineffability (a difficulty adequately describing experience in words), altered perception of time and space (feeling

"outside of" time and space), internal unity (pure awareness), and external unity (connection of all people and things) (Barrett et al., 2015; Hood, 1975; Hood, 2005; Pahnke, 1969).

Psychedelics as Treatment

Decreases in mood, well-being, quality of life, and social relations, are symptoms of many pathologies and associated with decreased treatment adherence (Arrieta et al., 2013). Considering most treatments require prolonged use to be effective, this results in a lack of symptom alleviation for patients. Since studies have shown that even single discrete experiences with psychedelics can elicit enduring positive outcomes, many researchers began to wonder whether psychedelics could be used as an effective treatment option. The lack of understanding about the powerful influence of set and setting (Johnson et al., 2008) resulted in high frequencies of adverse events in early clinical studies of psychedelics (Isbell, 1959; Malitz et al., 1960). Subsequent research involved preparation before, and support throughout, experiencing the acute effects of the drugs. This resulted in greater reports of positive experiences and decreases in psychological harm (Leary et al., 1963; Pahnke, 1969).

Despite emerging evidence of safety and therapeutic potential, the Controlled Substances Act, effective as law on May 1st, 1971, classified psilocybin and other psychedelics as illegal Schedule I substances (Controlled Substances Act, 1971). As a result, clinical research with these substances was largely non-existent until the early 2000's, first in Europe characterizing the acute psychological and physiological effects (Gouzoulis-Mayfrank et al., 1999; Vollenweider et al., 1998) and then in the US establishing basic safety (Griffiths et al., 2006; Johnson et al., 2008). Since then, a so-called "psychedelic renaissance" has begun (Sessa, 2020), with clinical trials investigating the therapeutic potential of psychedelic therapy for a variety of conditions. The emphasis of this more recent research has focused largely on asking *if* psychedelics work for

some clinical indication, rather than *how*. However, the indications that psychedelic treatment has been shown effective in treating (outlined below), are almost all characterized by pathologically rigid cognition (e.g., Chamberlain et al., 2006; Lee & Orsillo, 2014; Palm & Follette, 2011; Todd et al., 2015), a point I will revisit later in this section.

Depression and Anxiety Disorders

The most extensive research on the clinical benefits of psychedelics has been on their capacity to elicit antidepressant and anxiolytic effects. First investigated in patients with advanced-stage cancer, LSD (Grof et al., 1973) and N,N-Dipropyltryptamine (Richards et al., 1977) were both shown to significantly reduce existential depression and anxiety after treatment. Psilocybin was also shown to significantly reduce depression scores compared to baseline only at 6 months post-treatment following 2 treatment sessions in an initial pilot study (Grob et al., 2011). In two follow-up placebo controlled clinical trials, a single administration of psilocybin treatment resulted in significant increases in subjective quality-of-life and optimism ratings, as well as clinically significant reductions in anxiety and depression in about 80% of patients (Griffiths et al., 2016; Ross et al., 2016). These antidepressant and anxiolytic effects of a single administration of psilocybin, observed in two samples of patients, endured outwards of at least 6 months. In light of these findings, the antidepressant and anxiolytic effects of psilocybin have also been investigated in more general depression populations, including in treatment-resistant depression (TRD) and major depressive disorder (MDD).

In an open-label pilot study where participants received a fixed order of two doses of psilocybin (10mg then 25mg) 7 days apart, depression scores (as measured by the Quick Inventory of Depressive Symptoms [QIDS]) were significantly reduced compared to baseline after both treatment sessions. These effects remained consistent at all post-treatment time points

outwards of 3 months (Carhart-Harris et al., 2016). In a follow-up, randomized and double-blind study, psilocybin was directly compared to conventional depression treatment (the selective serotonin reuptake inhibitor [SSRI] escitalopram). At 6 weeks post-treatment, while both groups showed significant reductions from baseline on QIDS scores, there were no between-group differences when comparing the 2 doses of 25mg psilocybin group to the daily oral escitalopram group (Carhart-Harris et al., 2021). However, it is noteworthy that two administrations of psilocybin were equally effective in reducing depression compared to daily administrations of escitalopram, which can often have non-trivial negative side-effects. These effects were purported to be correlated with decreases in brain network modularity (Daws et al., 2022), though this account has been criticized for discrepancies in reports compared to the pre-registered study designs and statistical flaws (including the use of one-tailed tests and interpreting nonsignificant interactions; Doss et al., 2022). When directly comparing a single treatment with low (10mg) vs high (25mg) doses of psilocybin and placebo (1mg psilocybin), both low and high doses significantly reduced depression scores (measured with the Montgomery–Åsberg Depression Rating Scale [MADRS]) one day following treatment compared to placebo. However, while there were no differences between the placebo and low dose psilocybin groups at one week following treatment, the high dose group had significantly lower MADRS scores at all post-treatment time points outwards of 12 weeks (Goodwin et al., 2022). This same pattern of results was found for other measures of clinical efficacy, including anxiety, affect, and quality of life (Goodwin et al., 2023).

In participants with MDD, after administration of 20mg and 30mg of psilocybin (one week apart), participant scores on the GRID Hamilton Depression Rating Scale (GRID-HAMD) were significantly reduced outwards of one month compared to a waitlist control group (Davis et

al., 2020). This clinically significant reduction in GRID-HAMD scores persisted at 3, 6, and 12 months post-treatment (Gukasyan et al., 2022). Using the MADRS as the primary outcome measure, consistent with these findings above, participants treated with a single administration of 25mg of psilocybin had significantly reduced depression the day following treatment and lasting outwards of 43 days. This reduction was significantly greater than that observed in the niacin control group (Raison et al., 2023).

Since the effects of psilocybin are serotonergically mediated, the concomitant use of SSRIs is usually an exclusion criterion in most clinical trials. Participants taking SSRIs are usually required to taper off under the guidance of their prescribing healthcare provider, and remain abstinent for several months, before being eligible to participate. Though recent evidence suggests that there may not be any contraindications for concomitant SSRI use. In a double-blind placebo-controlled crossover study, participants were pretreated with 14 days of placebo or escitalopram prior to the administration of 25mg of psilocybin. The acute subjective effects of psilocybin were largely the same between both groups, with the main differences being escitalopram significantly reducing “bad drug effects” (including anxiety and cardiovascular changes). Further, there were no differences in blood concentrations of psilocybin metabolites between groups (Becker et al., 2022). Concomitant SSRI use also does not attenuate the antidepressant effects of psilocybin. In a sample of individuals with TRD taking concomitant SSRIs, a single administration of 25mg of psilocybin resulted in significant decreased post-treatment MADRS depression scores at all time points outwards of 12 weeks (Goodwin et al., 2023). When comparing the treatment effects of a single dose of 25mg psilocybin in patients with TRD and either taking or not taking concomitant SSRIs, the average reductions in MADRS scores from baseline to 12 weeks post-treatment were similar in both groups (-14.9 and -12.0

respectively). This suggests that the concomitant use of SSRIs does not attenuate the antidepressant effects of psilocybin treatment.

While there is relative homogeneity among the antidepressant effects observed across these studies in various depressive populations, it is notable that some studies have found no treatment effects. For example, microdosing (consumption of sub-hallucinogenic doses of psychedelics – typically at 10% or less than a standard dose) psilocybin resulted in no antidepressant or anxiolytic effects (Marschall et al., 2022). More interestingly, a within-subjects placebo-controlled fixed-order (placebo then psilocybin) study found significant decreases in GRID-HAMD scores after both placebo and psilocybin. There were no significant differences in the degree of change between the two treatments (Sloshower et al., 2023), highlighting the potential impact of expectancy, therapy, and placebo effects in the results from previous studies.

Addiction

Addiction has also shown to be responsive to treatment with psychedelics. In an open-label pilot study of individuals with alcohol dependence, the intensity of psilocybin effects during treatment sessions were strongly predictive of subsequent reductions in drinking behavior and cravings, as well as increases in abstinence self-efficacy (Bogenschutz et al., 2015). In a follow-up double-blind placebo-controlled study, the percentage of heavy drinking days and average drinks consumed per day was significantly decreased in both groups at all post-treatment assessments outwards of 36 weeks, but was lower for the psilocybin treated group (Bogenschutz et al., 2022; O'Donnell et al., 2022). Retrospective survey studies suggest that these effects on reducing or eliminating drinking behaviors are more pronounced for individuals with more severe a priori alcohol use disorder (Kervadec et al., 2023).

Psilocybin has also been shown to be highly effective in improving outcomes for long-term smokers with multiple failed quit attempts. Following psilocybin treatment, 80% of these individuals successfully quit and were abstinent six months later. 67% were still abstinent a full year later (Johnson et al., 2017). Qualitative analysis of interviews with these participants revealed that their psilocybin sessions elicited vivid insights into their self-identity and motivations for smoking. Furthermore, that any short-term withdrawal symptoms were overshadowed by the acute and longitudinal effects of their psilocybin experiences (Noorani et al., 2018).

Given the prevalence and impact of the opioid crisis in the US (Vadivelu et al., 2018) finding non-opioid pain management therapies and treatments for substance use disorder has been a recent funding priority (*NIH Launches HEAL Initiative, Doubles Funding to Accelerate Scientific Solutions to Stem National Opioid Epidemic*, 2018). Survey studies have found that lifetime psychedelic use was associated with lower odds of opioid use disorder (Jones et al., 2022; Pisano et al., 2017). Naturalistic use of psychedelics was associated with significant decreases in substance use disorder (SUD) in general, with 96% meeting criteria for SUD before the experience, and 26% meeting criteria after (Garcia-Romeu et al., 2020). Clinical trials are currently underway to experimentally test the efficacy of psilocybin to treat opioid use disorder (University of Wisconsin, Madison, 2024).

Pain

Chronic pain and pain disorders are particularly difficult to treat given the complex nature of these indications. In addition to the pain itself, chronic pain often results in comorbidities including depression, anxiety, and adjustment disorders (Bryl et al., 2021). Early work demonstrated LSD engendering pronounced analgesia among individuals with metastatic

cancer. Pain relief was more effective than high potency opiates and endured past the acute subjective effects of the drug (Kast, 1966, 1967; Kast & Collins, 1964). Survey studies have found psilocybin and LSD may reduce the severity and frequency of cluster migraine headaches (Andersson et al., 2017; Schindler et al., 2015; Sewell et al., 2006). Randomized, double-blind, placebo-controlled studies have replicated these effects with psilocybin, showing that cluster headache attack frequency was significantly lower after 3 doses (5 days apart) of psilocybin than placebo (Schindler et al., 2021, 2022). While most clinical trials utilize moderate to high doses of psychedelics, which result in acute subjective effects, recent work has demonstrated analgesia in doses below the threshold to bring about such effects. Low doses of LSD decreased subjective levels of acute pain experience and unpleasantness in response to cold water exposure (Ramaekers et al., 2021).

Psychedelics may effectively alleviate phantom limb pain as well, one of the most mysterious and intractable chronic pain conditions. Evidence for this began with pilot studies showing that low doses of LSD temporarily attenuated phantom limb pain. While pain returned at a diminished intensity, a second LSD administration cured pain in half of participants (Kuromaru et al., 1967). A follow-up further probed the effects of repeated doses of LSD and found reductions in pain intensity and reliance on analgesic medication in a majority of participants (Fanciullacci et al., 1977). A case report published recently by Ramachandran and colleagues provides additional evidence by presenting an instance of psilocybin, in combination with mirror visual feedback, resulting in elimination of chronic phantom limb pain (Ramachandran et al., 2018).

Other Conditions and Considerations

Psychedelics have also recently shown clinical promise for several other conditions. While there is a high degree of variance in how specifically, psilocybin experiences are related to acute alterations in individuals' self-perception in ways that are generally positive (Elias et al., 2023). Consistent with this correlational finding, several pilot studies have demonstrated psilocybin's efficacy in treating conditions related to self-perception. For example, at one month post-treatment, psilocybin reduced body weight and shape concerns, trait body image anxiety, and overall trait anxiety, compared to baseline scores. Body weight concerns remained reduced at three months post-treatment (Peck et al., 2023). In another study looking at body dysmorphic disorder, body dysmorphia-related obsessive compulsions were compared at baseline and various time points outwards of 12 weeks following treatment with psilocybin. Scores were significantly reduced at all follow-up time points (Schneier et al., 2023). Psilocybin may also be beneficial for conditions related to altered immune function and inflammation. While little research has been done to date, psilocybin, compared to placebo, was found to reduce the pro-inflammatory cytokine tumor necrosis factor-α acutely and reduce the inflammatory markers interleukin-6 and C-reactive protein at 7 days post-treatment (Mason et al., 2023). One case study has reported clinical benefits in microdosing psilocybin for the treatment of Lyme disease (Kinderlehrer, 2023), with clinical trials currently underway to experimentally test this (Johns Hopkins University, 2023).

Possible Mechanisms

Despite mounting evidence of the therapeutic potential of psychedelic treatment for a variety of clinical indications, the mechanisms underlying these effects have received comparatively less investigation. While there have been few hypothesis-driven studies about possible causal mechanisms underlying the positive outcomes and clinical benefits documented

by the work reviewed above, post-hoc and exploratory follow-up analyses have offered possible neural and psychological accounts of these effects. These accounts seem to converge on the same conclusion: that psychedelic use results in increased flexibility, and this flexibility is associated with clinical benefits. Flexibility can be considered at different levels of analysis: neurally, cognitively, or psychologically. While each higher level is thought to afford changes in the lower, little research to date has attempted to empirically establish this link. Further, cognitive accounts of the effects of psychedelics have received comparatively less attention in general. Without this, it is difficult to make any causal claims for this theory.

Psychological accounts

Several psychological accounts for the clinical benefits of psychedelics have been proposed. One such idea relates to the experiences of "ego dissolution," or loss of the sense of self, afforded by psychedelics (Carhart-Harris et al., 2018; Carhart-Harris et al., 2014; Lebedev et al., 2015; Nour & Carhart-Harris, 2017; Tagliazucchi et al., 2016). This ego-dissolution experience is positively correlated with improvements in well-being attributed to psychedelic use (Nour et al., 2016). Another is related to the MTEs that psychedelics reliably elicit. In their foundational study, Griffiths and colleagues (2006) found that psilocybin significantly increased scores on mysticism scales when administered in controlled settings. Over 60% of participants fulfilled criteria for having had a classically defined MTE. Participants reported their experiences as extremely impactful, with almost 75% rating it among the top 5 or single most personally meaningful and spiritually significant experience of their lives. Participants also attributed enduring positive changes in attitudes and behaviors to this experience, which were affirmed by 3rd party ratings made by friends and family. Such changes included improved social relationships with family and others, as well as increased physical and psychological well-being

and care (Griffiths et al., 2006). These interpretations and enduring effects were found to be persistent up to 14 months after the experience (Griffiths et al., 2008). Psychedelic-induced MTEs have also been found to mediate their resulting positive outcomes and clinical efficacy (Griffiths et al., 2008, 2016).

Another is related to the spontaneous experiences of insight that psilocybin has been found to dose-dependently elicit (Carbonaro et al., 2018, 2020; Davis et al., 2021). These insight experiences are correlated with improvements in depression symptoms after treatment (Roseman et al., 2018). Further evidence for their clinical relevance comes from interviews with patients who quit smoking after psilocybin treatment. Qualitative analyses of these interviews revealed patients attributed their smoking cessation to the vivid insights gained from their experiences with psilocybin (Noorani et al., 2018).

While this seems to suggest a role of MTEs or insight experiences in the clinical benefits of psychedelics, these effects are mediated by survey measures of psychological flexibility (Davis et al., 2020, 2021). In other words, changes in psychological flexibility may afford these MTE and insight experiences which are correlated with subsequent clinical benefits. Further, the personality domain of openness, which psychedelics increase (MacLean et al., 2011), is associated with greater flexibility (Silvia et al., 2009). These psychological accounts, focusing on changes in the mind, thus seem to suggest the role of increased flexibility underlying the effects of psychedelic therapy.

Neural Accounts

While the neural effects of psychedelic drugs are widespread, neural accounts of the mechanisms underlying their clinical benefits have largely focused on changes to the Default Mode Network (DMN) and Salience Network (SN). This focus on the DMN and SN are

informed by studies of MTEs elicited by other altered states of consciousness. For example, deep meditative practices can also elicit MTEs (d'Aquili & Newberg, 1998; Newberg & d'Aquili, 2000), and the DMN and SN are altered acutely during meditation practices (Barrett & Griffiths, 2018; Brewer et al., 2011).

Both the DMN and SN are large connector hubs, comprising among the highest number of cortico-cortical connections in the brain (Hagmann et al., 2008). The DMN is comprised mainly of the medial prefrontal cortex (mPFC), posterior cingulate cortex (PCC), parahippocampal cortex (PHC), and parts of the inferior parietal lobe (IPL) (Buckner et al., 2008; Fox et al., 2005). These areas have dense expression of 5-HT_{2A} receptors. Also known as the "task-negative network," the DMN is associated with several cognitive processes including self-referential and ruminative processing, self-other distinctions, inner-speech, mentalizing, and theory of mind (Barrett & Griffiths, 2018). The SN, also known as part of the "task-positive network," is primarily comprised of the anterior cingulate cortex (ACC) and anterior insula. These areas are associated with externally directed attention, detecting and maintaining goal-relevant stimuli, and coordination between other brain networks (Bonnelle et al., 2012).

Both functional Magnetic Resonance Imaging (fMRI; e.g., blood oxygenation level dependent, or BOLD, response) and molecular imaging methods (e.g., Positron Emission Tomography) have demonstrated the impact of psychedelics on activity in the DMN and SN. Early work, utilizing molecular imaging techniques, indicated acute increases in mPFC glucose metabolism after administration of psilocybin (Gouzoulis-Mayfrank et al., 1999; Vollenweider et al., 1997). Conversely, more recent work has found that psilocybin acutely decreases BOLD activity in the mPFC, ACC, and PCC. These decreases in mPFC and ACC predicted the intensity of subjective drug effects (Carhart-Harris et al., 2012, 2017). Additionally, mescaline was found

to decrease IPL activity (Hermle et al., 1992). These discrepancies are likely due to methodological differences in both measurement of neural processes and experimental design (e.g., differences in drug metabolism and CNS absorption due to oral vs. intravenous administration; Barrett & Griffiths, 2018). As a result of such inconsistencies, across a small number of studies, the nature of neural activity changes is still somewhat ambiguous. Changes in insula activity are more variable, with psilocybin acutely increasing activity in right insula but decreasing in left insula (Lewis et al., 2017).

In contrast to neural activity, changes to neural connectivity have yielded more consistent results across different designs and methods. Psychedelics have been found to cause changes in *structural* connectivity by way of increased neural plasticity. Increases in neuritogenesis (formation of new neurites), spinogenesis (development of new dendritic spines), and synaptogenesis (formation of new synapses between neurons) have been reported after administration of serotonergic psychedelics both in vivo and in vitro (Ly et al., 2018). Psychedelics have also been shown to increase brain-derived neurotrophic factor expression, a protein integral in neurogenesis and synaptogenesis (Vaidya et al., 1997).

In addition to these *structural* connectivity changes, the effects of psychedelics on *functional* connectivity changes have been more thoroughly investigated. Functional connectivity measures the change in linear coupling (the influence of one region on another by way of temporal correlations in activity) between brain regions. In other words, predicting activity in one brain area based on the activity in another (Carhart-Harris et al., 2012). fMRI scans during acute drug effects have shown that psilocybin decreases functional connectivity within the DMN, specifically between mPFC to PCC (Carhart-Harris et al., 2012; Tagliazucchi et al., 2014) and PCC to IPL (Carhart-Harris et al., 2012). Global Brain Connectivity analyses have shown

psilocybin decreases mPFC, ACC, and insula connectivity (Preller et al., 2020). Outside the DMN though, there are generally overall increases in functional connectivity, including between the DMN and SN (Roseman et al., 2014), and in sensory areas (Preller et al., 2020).

Curiously, fMRI scans taken post-acutely reveal increases in within-network DMN functional connectivity. There are increases in PCC to ACC and ACC to IPL 24 hours after administration of the DMT containing brew ayahuasca (Sampedro et al., 2017). Increases in connectivity between mPFC to ACC are observed 24 hours after psilocybin administration (Carhart-Harris et al., 2017) and between mPFC to PCC 48 hours after administration (Smigielski et al., 2019). Within-network DMN connectivity increases, as well as global connectivity increases, have been found to persist at 1-week and 1-month post-psilocybin administration (Barrett et al., 2020). These overall increases in global connectivity, with acute decreases followed by post-acute increases in within-network DMN connectivity, have been likened to a 'reset' mechanism whereby acute disintegration affords post-acute re-integration leading to overall improvements in normal functioning (Carhart-Harris et al., 2017).

Interestingly, these brain regions most affected by psychedelic use (mPFC, ACC, and IPL) are associated with behaviors symptomatic of increased flexibility (Filipowicz et al., 2016; Kerns et al., 2004; Kim et al., 2012). Further, increases in whole brain (global) connectivity are thought to be characteristic of increased neural flexibility, as there is increased "communication" between brain regions that are normally more distinct (Pang et al., 2016; Siegel et al., 2024). Like the psychological accounts outlined above, which focus on changes in the mind, these neural accounts focusing on changes in the brain strongly suggest the role of increased flexibility in the effects of psychedelics.

Cognitive Flexibility

The neural accounts reviewed above provide a compelling descriptive, implementation level, account. Though by focusing on brain-specific changes they can provide little insight into how or in what ways one's subjective experience is changed. This is particularly important as many of the clinical benefits of psychedelics have been documented in conditions that are primarily psychological and cognitive in nature. In contrast, by focusing on the mind rather than the brain, the psychological accounts can begin to address this. However, as these are assessed through correlational study designs, they cannot afford any causal explanations. Thus, while both the neural and psychological accounts reviewed above seem to implicate increases in flexibility as a potential mechanism underlying the effects of psychedelics, there exists a mind-brain gap in the evidence supporting this theory. To address this gap, help link the neural changes with psychological survey outcomes, and afford more definitive causal claims, it is imperative to more rigorously investigate flexibility at the cognitive level.

Cognitive flexibility is a broad concept that was originally posited to characterize one's ability to selectively change their concept systems in response to exogenous input. This was described in terms of switching behaviors, e.g., learning/adopting new rules in a task paradigm, simultaneously maintaining internal representations for multiple concepts and shifting attention between them, etc. (Scott, 1962). Early attempts to measure cognitive flexibility were done with task switching paradigms like the Dimensional Change Card Sorting Task (Zelazo, 2006), Multiple Classification Card Sorting Task (Bigler & Liben, 1992), and Wisconsin Card Sorting Task (Berg, 1948), which all require participants to learn and switch between rules when sorting or classifying test objects. For example, first matching objects by color and then by shape, quantity, or category. However, these behaviors are all related to executive function, which psychedelics have been found to impair (Barrett et al., 2018).

More recently, cognitive flexibility has been described more broadly in terms of one's ability to adjust their thinking from old situations to new ones (Moore & Malinowski, 2009), adapt to new situations by overcoming habitual thought patterns or responses (Deák, 2004), or the simultaneous awareness of multiple possibilities in a given situation (Martin & Rubin, 1995). These broadened conceptualizations of cognitive flexibility, which may in fact be afforded by less developed executive function (Gopnik, 2024), have led to more varied and nuanced predictions of behaviors reflective of being more cognitively flexible. For example, being less biased by prior knowledge (Gopnik et al., 2017), considering a wider range of possible solutions to a problem (Bonawitz et al., 2014), being more exploratory even when aware of the costs of doing so (Liquin & Gopnik, 2022), and having better memory for information outside the scope of one's goal in directed attention (Deng & Sloutsky, 2016). Here, I refer to this broader framing of cognitive flexibility that is afforded by less developed executive function (Gopnik, 2024).

As cognitive flexibility is a broad cognitive process which encompasses several cognitive functions (Miyake et al., 2000), it is instantiated through distributed brain areas and networks. For example, switching behaviors generally involve prefrontal cortex (PFC) activity, with specific types of switching recruiting different areas of the PFC (Kim et al., 2012). The IPL is associated with the determination of saliency of information relative to prior beliefs, i.e., when new information is consistent with expectations of an existing mental model or new models should be considered (Filipowicz et al., 2016). And the ACC is associated with conflict monitoring and cognitive control (Kerns et al., 2004). These regions are among those with the greatest changes in activity and functional connectivity resulting from the use of psychedelics. Notably, they are associated with *decreased* activity, which may afford increases in flexibility.

Cognitive flexibility has historically been associated with creativity and creative problem solving, and early work investigating the effects of psychedelics reported improved creative problem-solving after use (Harman et al., 1966; Spitzer et al., 1996). As noted above, psychedelics also increase the personality domain of openness, which is associated with both creativity and cognitive flexibility (Silvia et al., 2009). Additionally, the DMN has been found to play a central role in flexible thinking (Beaty et al., 2016, 2020). Consistent with these earlier findings, evidence from both animal models (Torrado Pacheco et al., 2023) and humans (Doss et al., 2021) suggests psilocybin therapy increases cognitive flexibility and these changes persist outwards of one month post-treatment, and formal models of cognitive flexibility have been proposed to underly the effects of psychedelics (Carhart-Harris & Friston, 2019; Kuypers, 2018).

In line with these models, a few studies have tested whether psychedelics alter scores on tasks measuring divergent thinking (DT), an instantiation of cognitive flexibility. DT is a method of exploring possible solutions to generate creative ideas. It is contrasted with convergent thinking (CT), which refers to the ability to quickly produce a single correct response (Guilford, 1950). However, the results from studies on the effects of psychedelics on DT are inconsistent. Kuypers and colleagues (2016) found that while CT was decreased, DT was increased during the acute effects of ayahuasca. However, DT increases were found in only one of two DT tasks administered (Kuypers et al., 2016). On this same task, the Picture Concept Task (PCT), Mason and colleagues (2019) observed DT score increases the day after psilocybin administration. Scores returned to baseline at 7 days post administration. In contrast, CT scores were unaffected the day after and significantly increased 7 days post administration (Mason et al., 2019). In a follow-up study, Mason and colleagues (2021) found that both DT and CT scores (as assessed by the PCT and alternative use task) decreased during the acute effects of psilocybin. At 7 days post

psilocybin administration, there were no differences in DT. However, CT was decreased compared to placebo and was no different from baseline (Mason et al., 2021). Further, the placebo group showed evidence of practice effects, making the effects from the psilocybin group difficult to interpret and raising questions about the results from previous studies.

The inconsistencies between Kuypers et al., (2016) and Mason et al., (2019, 2021) may be due to the way in which DT was operationalized. Specifically, both employed summary score metrics like fluency (number of responses generated) and originality (overall frequency of generated responses). This approach quantifies responses, rather than the processes underlying their production, and has been criticized for its inability to provide information about the cognitive processes underlying DT (Hass, 2017). While data from both psychological and neuroimaging research seems to implicate increased cognitive flexibility as a potential mechanism underlying the clinical benefits of psychedelics, a hypothesis-driven approach that moves beyond summary score metrics is needed in order to make conclusive statements. Specifically, predictions for specific behaviors that would be symptomatic of more flexible cognition must be tested directly.

Cognitive Development as a Model

How could such behaviors be identified? One approach is to look to populations known to be more cognitively flexible than baseline adults. If such a population could be identified, it would allow for the generation of testable predictions for the specific ways in which psychedelics may impact cognition that would result in increased flexibility. Fortunately, such a population exists, is widespread, and (relatively) easy to access. Those people are children. As described above, the most salient neural changes that psychedelics engender are overall increases in global functional connectivity, with specific local decreases within the DMN and SN,

increases in neural plasticity, and decreased activity in DMN and prefrontal regions. Children (ages 7-9) have lower within-network DMN functional connectivity and fiber density between the mPFC and PCC than adults, as well as decreased gray matter volume (Supekar et al., 2010). Across a number of studies, these mPFC and PCC connectivity differences show a distinct linear trend such that connectivity increases with age between ages 7 to 35 (Fair et al., 2007; Sherman et al., 2014; Uddin et al., 2011). Infancy and early childhood are periods of development characterized by high degrees of neuroplasticity, including increased synaptogenesis and spinogenesis, which plateaus around 12 months and steadily declines across development (Huttenlocher, 1990). These plasticity changes are heterogeneous, with sensory cortices reaching peak shortly after birth but the PFC taking much longer (Huttenlocher & Dabholkar, 1997). Thus, children also have less developed frontal regions and corresponding executive function (Thompson-Schill et al., 2009), which may lead to their increased cognitive flexibility.

In addition to these neural similarities between children and psychedelic treated adults, behavioral research also reports similarities in performance on creativity tasks. While DT performance decreases into middle to older adulthood (Jaquish & Ripple, 1985), adults with PFC damage exhibit increased performance compared to controls (Reverberi et al., 2005). Further, adults who receive transcranial direct current stimulation over their PFC also exhibit increased performance (Chrysikou et al., 2013). Thus, in instances where the adult brain is made more child-like, through injury or temporary change, they perform better on tasks requiring cognitive flexibility.

It is worth noting that one of the consistent findings reported in cognitive development work is that children's responses tend to be "noisier" than adults. Children have immature PFCs which results in less developed executive function, inhibition, and constraint. The variability in

their behavior has traditionally been viewed as sub-optimal, irrational, or even inflexible. Critically, however, these conclusions are only reasonable if children have the same goals and utilities as adults. Recent work suggests this is not the case. Whereas adults tend to prioritize reward maximization (Sumner et al., 2019), children tend to prioritize learning and exploration (Sumner et al., 2019). This protracted period of learning and exploration may facilitate humans' unique cognitive capabilities (Gopnik, 2020), as converging evidence from animal research suggests that a species' length of immaturity is correlated with their intelligence and relative brain size (Bennett & Harvey, 1985; Snell-Rood, 2013; Weisbecker & Goswami, 2010). As a result, children have been shown to outperform adults on tasks requiring flexible thinking (e.g., Gopnik et al., 2017; Plebanek & Sloutsky, 2017; Sumner et al., 2019). The variability observed between children and adults may therefore be reconstrued as both systematic and rational given differences in exploratory strategies and cognitive flexibility more broadly, afforded by these differences in executive function (Denison et al., 2013; Gopnik, 2020; Gelpi, 2021).

In sum, both neural and behavioral data suggest that, when compared to baseline adults, adults treated with psychedelics and children exhibit some of the same differences. Indeed, I am not the first to draw this comparison between childhood and psychedelic experiences (Gopnik, 2018). In this dissertation, I therefore treat children as a proxy psychedelic group. In doing so, leveraging what is known about the cognitive differences between children and adults to guide predictions for the specific ways psychedelics could be impacting cognition that would result in increased flexibility. What, then, are the specific cognitive features that allow children to be more flexible and exploratory? Much of what is known about these features comes from computational approaches to cognition, specifically those utilizing probabilistic models.

Probabilistic Models of Cognition

It has been proposed that our causal knowledge of the world shares the same structure and function as scientific theories (e.g., Gerstenberg & Tenenbaum, 2017). Like scientific theories, our "intuitive theories" are structured representations of abstract causal relationships which can be used to explain and interpret events, and also support predictions and counterfactual inferences. Further, both scientific and intuitive theories are dynamic and consistently revised in light of new information (Gopnik & Wellman, 2012). This account, known as the rational constructivist account, views the process of cognitive development and learning as analogous to scientific theory change, where beliefs and information are analogous to hypotheses and evidence (e.g., Gopnik, 1996; Gopnik & Wellman, 2012).

From birth humans are equipped with a powerful domain general learning mechanism which allows us to be remarkably prolific learners, using probabilistic information to infer the abstract causal structure of the world from observed events. We generate and revise our causal theories by selecting hypotheses to test from the space of all possibilities and inferring which provides the best explanation for some observed evidence. While several hypotheses could generate the observed evidence, some will be more probable than others. Thus, instead of having right and wrong answers, our theories are formed probabilistically by assessing the probability that a candidate hypothesis is true, compared to other alternatives, and identifying which has the highest probability. Probabilistic models of cognition use the principles of probability theory to formally model this process. The statistical relationship between some observed evidence to be explained and a candidate explanatory hypothesis can then be computed via Bayesian inference. In Bayesian inference, prior knowledge is formalized as a probability distribution over possible hypotheses. These "priors" represent the probability that a given hypothesis (h) is true prior to

observing any evidence. The posterior probability of a hypothesis, the probability of the hypotheses being true given the observed evidence (d), can thus be computed using Bayes' Rule:

$$p(h|d) \approx p(d|h) * p(h)$$

Specifically, the posterior is proportional to the probability of seeing the observed evidence if a given hypothesis were true (the likelihood) weighted by the prior probability of that hypothesis (Perfors et al., 2011).

With this mathematical formalization, these models can make specific quantitative predictions about which hypothesis best explains some evidence. Across a variety of learning paradigms, both adults and children choose the hypothesis that is most probable according to Bayesian inference (e.g., Bonawitz et al., 2014; Denison et al., 2013; Gopnik & Wellman, 2012; Tenenbaum et al., 2011). Through this computational account which characterizes theory formation and revision (i.e., learning) across development, three key cognitive features which are consistent with the recent broadened conceptualizations of cognitive flexibility have been proposed to afford children relatively greater flexibility in forming and revising their theories.

Cognitive Features Affording Flexibility

First, **children and adults differ in how much their prior knowledge guides their inferences.** Theories must simultaneously be flexible enough for belief revision to occur, but conservative enough to prevent strong beliefs from being too easily overturned. These inferences are therefore based on the integration of current evidence with one's prior knowledge. Since children have comparatively less prior knowledge than adults, this results in more weakly held beliefs. As a consequence, children's inferences are more sensitive to the current evidence, often leading to more flexible belief revision (Gopnik et al., 2017; Lucas et al., 2014; Seiver et al., 2013).

Second, **children and adults differ in how they choose from the space of possible hypotheses.** The process of Bayesian inference comes with computational cost, and the space of possible hypotheses to evaluate is intractably large (Gittins & Jones, 1979; Griffiths et al., 2012; Sanborn et al., 2010). Much empirical work suggests that people across development instead *approximate* Bayesian inference by sampling a subset of hypotheses to evaluate (e.g., Denison et al., 2013; Sanborn, 2017; Sanborn & Chater, 2016; Thomas et al., 2008; Vul et al., 2014; Vul & Pashler, 2008). This process of searching and sampling from hypothesis space can unfold in different ways. It has been proposed that while adults tend to search narrowly, only sampling hypotheses that incrementally differ, children tend to search broadly, allowing them to arrive at solutions that may ultimately be better than local alternatives (Gopnik, 2020).

Third, **children and adults differ in how they search their exogenous environment for evidence.** Given their lack of inhibition and executive control, children have less selective attention, allocating exogenous attentional resources more diffusely and exploring their environments more broadly than adults. This broader exploration allows them to pick up on various relevant cues in their environment that adults may potentially miss (e.g., Blanco & Sloutsky, 2020; Plebanek & Sloutsky, 2017; Rich & Gureckis, 2015).

In this dissertation, I hypothesize that these three key features which discriminate adult and child cognition, allowing children to be more flexible, will also discriminate between adults in psychedelic-treated and control groups. In Chapter 2, I present a novel measure of problem solving under dynamic constraints which can be used to quantify search and sampling strategies. In Chapter 3, I present a pilot test of whether psychedelic treatment will result in (1) decreased influence of prior knowledge on inferences, (2) employment of broader, more exploratory, search strategies when sampling hypotheses, and (3) more diffuse exogenous attention. Finally,

in Chapter 4, I examine whether these cognitive flexibility tasks are susceptible to practice effects. Collectively, this work aims to provide initial support for a more holistic and comprehensive account of the benefits of psychedelic therapy, improving the research on cognitive flexibility in psychedelic studies and more generally.

Chapter 2 Contexto: A Novel Measure of Flexible Reasoning Under Dynamic Constraints

Human reasoners are routinely faced with situations where they must think flexibly and creatively under dynamic constraints, and seem to enjoy doing so (e.g., in the case of many strategy and word games). Longstanding empirical interest in these capabilities have spurred an entire field of research on creativity and cognitive flexibility. However, to date, the tools available to quantify and better understand this behavior are limited. Below, I review the existing research from this field, paying special attention to the limitations of existing methods for measuring flexible thinking, before proposing a novel approach.

Creativity is typically studied using one of four approaches based on Rhodes' 4P's model of creativity (Rhodes, 1961), each of which focuses on a different dimension of creativity: 1. Person – the contributions of individual traits and characteristics (Do successful artists share unique personality traits?), 2. Process – how one generates, develops, and refines creative ideas (How does the process of idea generation vs evaluation differ?), 3. Place/Press – the effect of environmental factors and contexts (Do people generate different ideas in laboratory settings vs in the real world?), 4. Product – evaluating the results and outcomes of the creative process (How "good" is a particular creative solution?). The Person, Place, and Product approaches are all contingent on first having a creative output: One needs an output to evaluate its goodness, identify creative individuals to investigate their individual characteristics, and compare outputs generated in different contexts. Thus, much of the creativity literature has focused on Process approaches, and the primary method to assess creativity is through measures which quantify Process (Long, 2014).

Torrance (1966) characterized the creative process as first requiring the development of hypotheses or proposed solutions, then testing and evaluating these candidate hypotheses, and

ultimately sharing the created product. Guilford's Structure-of-Intellect model (1956), which Torrance expanded, frames the creative process as a form of problem solving which recruits two distinct operations: divergent and convergent thinking. Divergent thinking (DT) characterizes the process of broadly searching for and generating candidate solutions, whereas convergent thinking (CT) characterizes the selection of the best option among candidates. DT has conventionally been viewed as more relevant to the success of creative problem solving (Guilford, 1956), and as a result is the most widely used operational definition of creativity (Hocevar, 1981).

Despite the widespread use of DT tasks, it has been a challenge for researchers to find appropriate measures to quantify this construct. Guilford (1957) proposed several scoring measures of DT, including *fluency* (number of responses produced), *flexibility* (category variation amongst responses), *originality* (uniqueness or novelty of responses), and *elaboration* (depth or detail included in responses), and applied these scoring measures to the data produced by the Alternative Use Task. In this task participants must generate as many "creative and unusual" uses as possible for a common household object in a specific timeframe (Guilford, 1967). Torrance (1966) built on Guilford's DT framework and developed the Torrance Test for Creative Thinking (TTCT), operationalizing these same DT scoring measures across several tasks (Torrance, 1966). These tasks are amongst the earliest, and still the most widely used measures of creativity (see. for example, Said-Metwaly et al., 2017). However, both have received widespread criticism.

The most common criticisms of DT tasks, and creativity measures more generally, regard their psychometric shortcomings. The validity of DT tasks has been routinely questioned (Baer, 2016; Cropley, 2000; Hennessey & Amabile, 2010; Lemons, 2011; Said-Metwaly et al., 2017),

and indeed, empirical evidence suggests their validity varies depending on the conditions in which the test is administered (Harrington, 1975; Katz & Poag, 1979; McCrae, 1987; Wallach & Kogan, 1965). Even more troubling is the extensive evidence of a lack of convergent validity between DT and other creativity tasks (Taylor, et al., 1963, Davis & Belcher 1971, Getzels & Csikszentlmihalyi, 1973; Ellison, 1973, Andrews, 1975; Barron, 1969; Beittel, 1964; Brittain and Beittel, 1961; Dillehunt, 1973; Goolsby & Helwig, 1975; Gough, 1976; Hadden & Lytton, 1971; Jordan, 1975; Karlins et al., 1969; Kogan & Pankove, 1974; Popperova, 1972; Roweton et al., 1975; Skager et al., 1967; Hoecevar, 1983). Strikingly, several researchers have found that performance on creativity tasks does not correlate with actual creative ability, as measured by an individual's success in various creative fields (Feldhusen & Goh 1995; Kanter, 1984; Martin et al., 1981).

While concerning, these inconsistencies should not necessarily come as a surprise, as creativity is a broad construct. It is reasonable to think that the skills and characteristics that afford one to be a prized chef may differ from those of a musician, painter, or writer. Nevertheless, creativity is largely treated as a single unitary construct, without appreciation for its different dimensions (Hocevar, 1981). This is perhaps the most salient issue with DT tasks, they neglect an entire component of creativity: convergent thinking.

To account for this, Mednick (1962) developed an alternative measure of creativity called the Remote Associates Task (RAT). In each task trial, participants are given three target cue words and must generate a fourth word which individually pairs with all three cues. For example, for the cue words "Comb, Dew, Moon," the correct answer is "Honey." To solve RAT trials, participants must generate candidate answers, evaluate their fit with each of the target cue words, and then select which of their candidate answers is best. This process is thought to involve an

initial stage of DT, followed by a subsequent stage of CT, thereby recruiting both operations involved in the creative process (Smith et al., 2013). Further, unlike many other measures of creativity, it enforces constraints on the generated ideas, which is more akin to real-world creative scenarios in which situational demands or goals must be met. As a result of its improvements over pure DT tasks and other measures, the RAT has become a commonly used measure of creativity (Wu et al., 2020). However, the RAT has also received its own criticisms. Despite recruiting both DT and CT processes, scores on the RAT are generally not correlated with scores on other measures of DT or creativity (Lee Bae et al., 2014). Studies which have found correlations are generally weak, with correlation coefficients usually around $r = 0.10$ (Taft & Rossiter, 1966, Akbari Chermahini et al., 2012, Lee Bae et al. 2014). On the other hand, RAT scores have been found to have stronger correlations with CT tasks (Akbari Chermahini et al., 2012; Laughlin et al., 1968; Taft & Rossiter, 1966), particularly on insight problem solving tasks (Huang et al., 2012; Chang et al., 2016). While insight problem solving (overcoming mental blocks to see problems in a new way) requires creative and flexible thinking, it is *also* a measure of intelligence. Indeed, the strongest correlations between RAT scores and other constructs has been to general measures of intelligence (Lee & Therriault, 2013; Lee Bae et al., 2014).

Beyond the issues raised above, the RAT has two other significant limitations. First, though it does purport to capture the DT process, it does not provide concrete data on this portion of the process. Conventionally, only participant's proposed answer on each trial is recorded and scored. Recently, researchers have taken a more computational approach, asking participants to record all the responses that came to mind while trying to arrive at their answer. This allows the semantic distance between responses to be computed as a more direct quantification of DT

(Smith et al., 2013). Despite being asked to report everything that comes to mind, participants usually give only a few responses per trial, substantially undermining the rigor of the analyses.

Second, while the RAT strives for more realistic problem solving by imposing constraints, these constraints are *static*. In many cases of real-world problem-solving, people get some form of feedback on their proposed solutions, which serves to further constrain the generation of subsequent solutions. In this way, they face problem-solving scenarios with *dynamic* constraints. For example, in the scientific research process, one has some initial hypothesis, tests it, and observes the results. Using this feedback, one then re-evaluates their hypothesis, revises it, and then tests the revised hypothesis, steadily making incremental changes as this process is repeated. A similar process also occurs in many strategy and word games. When one is playing a crossword puzzle, the clues stay the same, but the quality of proposed solutions (the words that answer the clues) is modified as the player fills out other words on the game board. The RAT falls short of capturing this continual updating inherent in many problem-solving scenarios.

Creative foraging models have come closest to assessing problem solving under dynamic constraints (Hart et al., 2017). However, it is unclear if this framework, which was initially designed to characterize animal foraging behavior via navigation of physical spaces, could appropriately capture hypothesis search in semantic space. The present work proposes to address this gap by developing a novel task, based on the popular internet game *Contexto*, to measure how people search over a hypothesis space with dynamically updating constraints. While in the original game participants have access to one trial per day and can make an unlimited number of guesses, here they complete multiple trials in which they attempt to guess a secret target word within a finite number of guesses. When they submit a guess, participants receive feedback about

how similar (in semantic distance) their guess is to the target. To be successful, participants must utilize this feedback to continuously update the constraints on their subsequent guesses. Across two experiments, we first assess whether this novel task can be used to measure creative problem solving. Further, whether performance is susceptible to practice effects to assess its suitability for use in pre- and post-test designs (for intervention studies, tracking cognitive decline, etc.). Broadly speaking, we hypothesized that participants will be sensitive to the semantic distance feedback and use it to inform their subsequent guesses. Although intuitively plausible, this is an aspect of the process of creative thinking that prior tasks have thus far been unable to capture.

Experiment 1

In Experiment 1, we assess whether participants are sensitive to the feedback they receive. If people are indeed sensitive to feedback and use it to inform their subsequent guesses, they should behave differently depending on the proximity of their previous guess to the trial's target word. However, since this process is contingent on the particular guesses a participant produces, we provide them with a start word on each trial. This design allows us to anchor participants at various distances from the target word.

We hypothesized that: (1) The average similarity between sequentially generated responses, and between responses and the target word, will be higher when the starting word is closer to the target. (2) Participants should produce better guesses (i.e., a subsequent guess closer to the target word) when the starting word is closer to the target. (3) Participants should successfully guess the target word more often when the starting word is closer to the target.

Method

Participants. 61 participants were recruited from the University of California, San Diego, undergraduate research pool. Participants answered basic demographic questions

including age (M = 21.50 years, SD = 3.67, 13.11% declined to answer), gender identity (72.1% female, 16.4% male, 3.3% other gender identity, 8.2% declined to answer), and race (24.6% East Asian, 19.7% Hispanic, 16.4% White, 11.5% multiracial, 4.92% South Asian, 4.92% Black, 1.64% Native Hawaiian or Pacific Islander, 1.64% Middle Eastern, 1.59% other racial identity, 13.1% declined to answer).

Materials. The target words were chosen based on a pilot study. We selected target words that were correctly identified by a majority of participants, but took an average of about twenty guesses. To generate the rankings of potential guesses to the target words, a dictionary of approximately 80,000 words was rank-ordered by their similarity to the target word for every trial. Similarity was computed using the Global Vectors for Word Representation (GloVe; Pennington et al., 2014). GloVe is an unsupervised learning algorithm that produces vector representations for words. A set of pre-trained word vectors was used which contained 840 billion tokens gathered via Common Crawl. The original set of word vectors was filtered to include only a subset of words originally used by Contexto, then further filtered to remove words with two or fewer letters, all stop words, words flagged as inappropriate (e.g., curse words and slurs), words containing numbers or punctuation, and words with multiple accepted spellings (e.g., keeping barbecue and removing barbeque). Finally, all words were lemmatized and any resulting duplicates were removed. This process resulted in a final dictionary containing 80,224 words. Similarity was defined by the cosine between any two word vectors. Compared to Latent Semantic Analysis (LSA), which captures semantic structure by focusing on singular value decomposition (SVD) to reduce the dimensionality of a term-document matrix, GloVe uses global word co-occurrence to capture meaning from the entire corpus. This approach has been shown to produce higher-quality word embeddings and perform better on semantic measures like

word analogy tasks (Pennington et al., 2014). After computing the cosine similarity between each word in the dictionary and the target word for a given trial, the dictionary was arranged by the resulting values in descending order. A given word's rank was defined by its numerical position in the arranged dictionary (where the first word was the target), and that ranking was given to participants as feedback for their guesses in a given trial. To account for idiosyncrasies that occurred when creating a sorted dictionary for a word in its singular vs plural form (i.e., differences between the ranked dictionary for the target word "frog" vs. "frogs"), the associated cosine similarities for each word in the dictionaries were averaged and then rank-ordered. The resulting "denoised" dictionary was used in all applicable cases.

Procedure. Participants first saw an instructions screen that gave them the rules of the game. They were informed that each trial contains a randomly selected secret target word, and they had to figure out what that word was. They were also told that the secret word will always be a noun, that they have a maximum of ten guesses per trial, and that they will receive a randomly chosen word to start with at the beginning of each trial (see Appendix A for full instructions text). The start word on each trial varied in closeness to the target words. The proximity of the start words to the target were either close (approximately rank 10), medium (approximately rank 100), or far (approximately rank 1000). Each participant saw three trials from each start word proximity condition, for a total of nine trials. The same nine target words (cookie, flower, barbecue, tree, moose, pancake, camera, car, and pencil) were used for all participants, but each target word was randomly assigned to one of the start word proximity conditions. Thus, while all participants had a trial where, for example, "cookie" was the target word, some received a start word on this trial that was close and others receive a start word that was medium or far (see Appendix A for the full list of start words per start proximity condition

for each target word). Participants submitted their guesses one at a time and were shown the guess' rank after submission. If the target word was correctly guessed, a congratulatory screen was displayed, and the participant moved on to the next trial. If a participant exhausted all ten guesses without ultimately guessing the trial's target word, they were shown what the target word was and then began their next trial.

Results

We first assessed the difference between the average similarity of sequentially generated responses in different start word proximity conditions. We fit a linear mixed-effects model including start word proximity (Close, Medium, and Far) as a fixed effect. Participant, target word, trial, and the interaction of participant and start word proximity, were included as random effects. The resulting full model was as follows: similarity ~ start_word_proximity + (1 | participant_id) + (1 | target_word) + (1 | trial) + (1 | participant_id:start_word_proximity). Significance of the main effect was assessed using a Type III Analysis of Variance with Satterthwaite's method. This revealed a significant main effect of start proximity, $F(2, 114.2) = 15.09$, $p < 0.001$. Post-hoc pairwise comparisons using the Tukey method for multiple comparisons indicated that the mean similarity score for sequentially generated responses in the Close condition ($M = 0.435$) was significantly higher than in the Medium ($M = 0.410$), $t(153) = 3.25$, $p = 0.004$, $d = 0.16$, 95% CI [0.06, 0.27]. The mean similarity score in the Close condition was significantly higher than in the Far ($M = 0.393$), $t(146) = 5.46$, $p < 0.001$, $d = 0.27$, 95% CI [0.16, 0.37]. Also, the mean similarity score in the Medium condition was significantly higher than in the Far, $t(93) = 2.49$, $p = 0.038$, $d = 0.11$, 95% CI [0.02, 0.20].

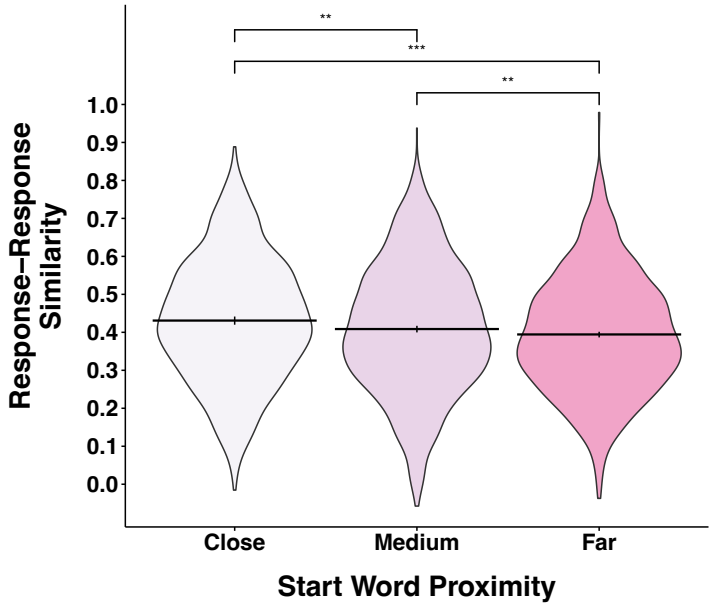


Figure 2.1 Comparison of similarity between sequentially generated responses across trials with different start word proximities.

The same model specification and analysis was applied to comparing the difference in similarity between responses and the target word. This revealed a significant main effect of start proximity, $F(2, 105) = 258.57$, $p < 0.001$. Post-hoc pairwise comparisons using the Tukey method for multiple comparisons indicated that the mean similarity score between responses and the target word in the Close condition ($M = 0.496$) was significantly higher than in the Medium ($M = 0.373$), $t(135) = 13.00$, $p < 0.001$, $d = 0.81$, 95% CI [0.67, 0.96]. The mean similarity score in the Close condition was also significantly higher than in the Far ($M = 0.283$), $t(131) = 22.70$, $p < 0.001$, $d = 1.42$, 95% CI [1.27, 1.56]. Further, the mean similarity score in the Medium condition was significantly higher than in the Far, $t(104) = 10.22$, $p < 0.001$, $d = 0.60$, 95% CI [0.47, 0.73].

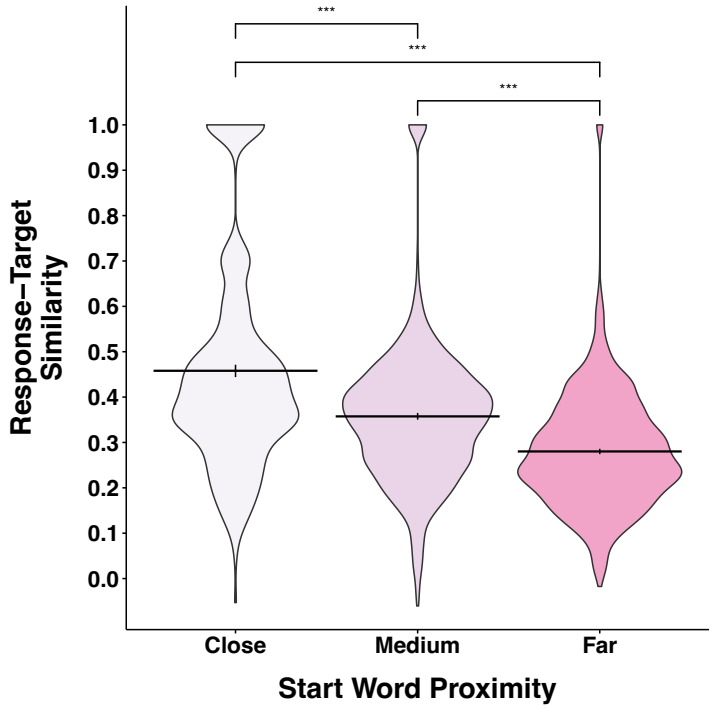


Figure 2.2 Comparison of similarity between responses and target word across trials with different start word proximities.

We next assessed whether there was a difference in guess improvement between conditions. That is, did the extent to which participants produced subsequent guesses closer to the target differ between start word proximity conditions? For all *n* guesses a participant produced within a trial after their first guess, we coded whether the response to target similarity in guess *n* was greater than that in guess *n*-1. We then fit a mixed-effects logistic regression to examine the effect of start word proximity (Close, Medium, and Far) on the likelihood of making a better subsequent guess. The model including both participant and trial as random effects resulted in a singular fit model. However, the inclusion of participant did not significantly improve model fit ($p > 0.05$), and the results from both models were identical. Thus, we dropped the random effect of participant, and the resulting full model was as follows: better_guess ~ start_word_proximity + (1 | trial). Significance of the main effect was assessed with nested model comparisons. Results revealed no significant main effect of start word proximity $\chi^2(2) =$

4.24, p = 0.12. Participants were equally likely to make better guesses regardless of start word proximity to the target.

Finally, to assess whether participants ultimately guessed the target word correctly more often when they received a starting word that was closer, we compared the average number of correctly guessed words between start proximity conditions. Trials where the target word was correctly guessed were coded 1, and those where the target word was not guessed were coded 0. We then fit a mixed-effects logistic regression to examine the effect of start word proximity (Close, Medium, Far) on the likelihood of correctly guessing the target word. Participant, target word, and trial were included as random effects. Including the interaction of participant and start word proximity resulted in a singular fit model, but this term did not result in an improved model fit ($p > 0.05$). The resulting full model was as follows: correct ~ start_word_proximity + (1 | participant_id) + (1 | target_word) + (1 | trial). Significance of the main effect was assessed with nested model comparisons. Results revealed a significant main effect of start word proximity, $\chi^2(2) = 179.7$, $p < 0.001$. Post-hoc pairwise comparisons using the Tukey method for multiple comparisons indicated that participants guessed the word correctly more often in the Close condition (M = 0.61) than in the Medium (M = 0.22), $z = 7.77$, $p < 0.001$, $d = 2.96$, 95% CI [2.21, 3.71]. They also did so more often in the Close condition compared to the Far (M = 0.08), $z = 9.01$, $p < 0.001$, $d = 4.55$, 95% CI [3.54, 5.55], and the Medium condition compared to the Far, $z = 3.84$, $p < 0.001$, $d = 1.58$, 95% CI [0.77, 2.40].

Discussion

The results from Experiment 1 provide initial evidence that participants do in fact engage with this task in a meaningful way. They integrate the feedback they receive from each guess to constrain their subsequent guesses. Further, they exhibit different search behaviors depending on

how close to the target word their guesses were. (1) They guessed more similarly to other guesses and the target when close and less similarly when further away, and (2) they were more likely to correctly guess the target when closer. Additionally, the task yielded more data per participant than comparable tasks. Across all trials, participants guessed an average of 8.11 words out of a maximum of 10.

Experiment 2

People often develop strategies when engaging in this type of reasoning. In fact, several academic papers (Anderson & Meyer, 2022) and online resources (Benveniste & Frere, 2022) have been dedicated to investigating optimal strategies when playing similar games. The ability to administer a task multiple times provides several advantages, including being able to assess for pre- and post-intervention changes in experimental studies. In Experiment 2, we tested whether participants' performance exhibited any practice effects across multiple administrations.

Method

Participants. 41 participants were recruited from the University of California, San Diego, undergraduate research pool. 19 were excluded due to providing nonsense answers which compromised the similarity measure between responses, resulting in a final dataset of 22 people. Participants answered basic demographic questions including age (M = 21.45 years, SD = 2.84), gender identity (77.3% female, 18.2% male, 4.55% other gender identity), and race (50.0% East Asian, 27.3% White, 13.6% multiracial, 4.55% Black, 4.55% Native American).

Procedure. The procedure for Experiment 2 was identical to Experiment 1, with the following exceptions. First, participants were not given a start word at the beginning of each trial, and this was removed from the instructions text. Second, participants completed the task three separate times, each one week apart. The sets of target words were randomized between

time points (Appendix A). That is, while the same five target words were always used at a given time point, which time point they occurred in varied between participants. Finally, participants had 20 guesses max per trial instead of 10.

Results

We first assessed the difference between the average similarity of sequentially generated responses at different time points. We fit a linear mixed-effects model with time point (1, 2, and 3) as a fixed effect. Time point was dummy coded and mean-centered. Random effects were included for participant, target word, and the interaction of participant and time point. The resulting full model was as follows: similarity ~ time_point + (1 | participant_id) + (1 | target_word) + (1 | participant_id: time_point). Significance of the main effect was assessed with nested model comparisons. This revealed no significant main effect of time point, $\chi^2(1) = 1.35$, $p = 0.24$. Participants' sampling strategies did not change at different time points (Figure 2.3).

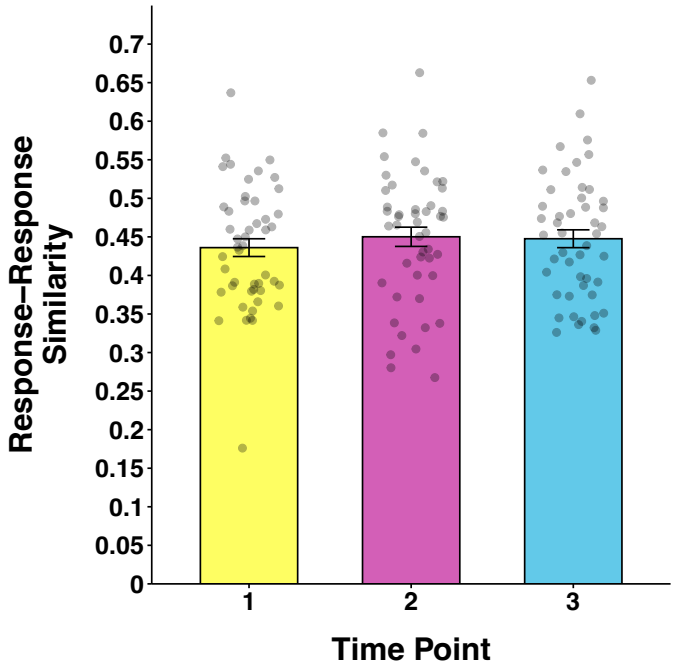


Figure 2.3 Comparison of similarity between sequentially generated responses across time points.

We next assessed whether there were any differences in the extent to which participants ultimately guessed the target word correctly between time points. To do so, we compared the average number of correctly guessed words between time points. Trials where the target word was correctly guessed were coded 1, and those where the target word was not guessed were coded 0. We then conducted a mixed-effects logistic regression to examine the effect of time point on the likelihood of correctly guessing the target word, including participant, target word, and the interaction of participant and time point as random effects. The full model fit was: correct ~ time_point + (1 | participant_id) + (1 | target_word) + (1 | participant_id:time_point. Significance of the main effect was assessed with nested model comparisons. This revealed a significant main effect of time point, $\chi^2(1) = 5.55$, $p = 0.019$. Post-hoc pairwise comparisons using the Tukey method for multiple comparisons indicated participants were more likely to correctly guess the secret word at time point 1 (M = 0.13%) than at time point 3 (M = 0.03%), $z = 2.54$, $p = 0.03$, OR = 4.4. They were also more likely to correctly guess the secret word at time point 2 (M = 0.19%) than at time point 3, $z = 3.41$, $p = 0.0019$, OR = 6.84. However, there was no difference between time point 1 and 2, $z = -1.20$, $p = 0.45$, OR = 0.64. This suggests that participant performance did not improve with repeated administrations of the task.

Discussion

In Experiment 2, we found no evidence for practice effects on this task. If participants were improving on the task or developing effective strategies, we would expect them to (1) change their search strategies such that sequentially generated responses would be more or less similar at different time points, and (2) correctly guess the target word more often in later time points. Despite completing the task on three separate occasions, participants' overall strategies did not seem to change. There were no differences in the similarity of sequentially generated

responses across time points. Moreover, their performance did not improve. They did not correctly identify the target words more often in subsequent time points. In fact, they performed worse at time point 3 than the first two. This suggests that the trials at time point 3 were more difficult than the first two, rather than participants exhibiting practice effects. Overall, we found no evidence of practice effects for this task, suggesting it is amenable to being administered multiple times in a particular study.

General Discussion

Humans routinely face a unique problem-solving challenge where they must think flexibly and conduct a hypothesis search under dynamic constraints. This occurs in many professional contexts, for example when engaging in the scientific process. It is also something we enjoy and seek out, in the case of strategy and word games. To date, the tools available to quantify and better understand this behavior are limited. As noted above, both general creativity measures and DT tasks focus on this idea of the search problem, but do not explicitly capture the filtering and constraint satisfaction aspects of these problems, and have been long criticized for psychometric shortcomings. The RAT, a measure that involves both DT and CT, is an improvement, but only incorporates static constraints and typically yields smaller amounts of data. Here, we introduce a novel task based on the popular internet game Contexto that better captures how people solve this problem and yields more repeated observations per subject. Thus, this novel task addresses several shortcomings of extant measures and allows us to more robustly quantify how people approach the search problem.

In Experiment 1, we tested several intuitive hypotheses about the ways in which people engage in this type of problem solving. We found that people are sensitive to the feedback they receive and use it to constrain their hypothesis search. They produced guesses more similar to

each other and to the target when it is close and less similar when it is further away. They were also more likely to correctly guess the target when they were closer to it. In Experiment 2, we observed no practice effects from repeated administrations of this task. Participants' performance did not improve, and their search strategies did not differ, even after completing the task multiple times. Together, the results of these initial experiments provide support for the use of this task as a novel measure of reasoning under dynamic constraints.

One additional feature of this task is its inherent flexibility. There are many parameters that researchers can modify to fit the needs of a particular experimental paradigm, making it potentially useful in a variety of contexts: Participants can be given start words to artificially situate them at different distances from the target word to assess their behaviors at these different points. The number of trials and guess limit per trial (should one want to include them) can vary. Different success criteria can be set – for example a requirement that they must guess the target correctly, or end within the top *N* words. The semantic content of the target words can also vary to assess whether participants have specific lexical biases (e.g., more likely to generate nouns). Further, the data generated by this task can be analyzed in different ways to examine different aspects of this reasoning process: One can look at the number of words correctly guessed, the proportion of responses that were better than previous ones, average similarity between sequentially generated responses, average similarity between responses and the target, maximum pairwise similarity between any two responses, etc.

Finally, this task addresses recent calls for increasing the use of games as experimental paradigms in psychology research (Allen et al., 2023). Many, if not most, experiments involve tedious and onerous procedures requiring researchers to incentivize participation. Playing games has been an activity enjoyed by people across different cultures (Suchow et al., 2020), ages

(Brändle et al., 2021), and throughout human history (Depaulis, 2020). Tasks that are intrinsically rewarding can result in increased quality and quantity of data (Hartshorne et al., 2019), and the results from the current work are consistent with this.

The results from Experiments 1 and 2 demonstrate that this task successfully quantifies this unique form of reasoning under dynamic constraints, presenting an improvement over the currently available methods for assessing this and affording researchers sufficient flexibility to be applied across contexts. Future research can apply computational approaches to better characterize the specific search and sampling strategies people employ. For example, performance can be compared to creative foraging tasks (e.g., Hart et al., 2017) and the predictions from different foraging algorithms can assess the extent to which people deviate from optimal search. We can additionally test whether and how people's performance and search strategies change in different clinical conditions, or in response to different clinical treatments (see Chapter 3). Finally, given that there is no evidence for practice effects on this task, it can be used to validate interventions intended to affect cognitive flexibility. Our results provide initial support for intuitive hypotheses about how people engage in reasoning under dynamic constraints and provide a paradigm for future research to quantify the specific search and sampling strategies employed by humans in a variety of contexts and clinical conditions. This work will collectively help us better understand a unique and ubiquitous form of reasoning.

Acknowledgements

Chapter 2 is currently being prepared for submission for publication of the material. Hurwitz, Ethan; Brockbank, Erik; and Walker, Caren. The dissertation author was the primary investigator and author of this material.

Thank you to Nildo Junior from the Contexto team for sharing the word list resources.

Chapter 3 The Effects of Psilocybin on Cognitive Flexibility: A Pilot Study

As described in Chapter 1, psilocybin therapy has recently shown promise as an effective treatment option for a variety of clinical indications (Andersen et al., 2020; Johnson & Griffiths, 2017; Van Amsterdam & Van Den Brink, 2022), but its mechanism of action remains unclear. The current work proposes a novel explanation for these therapeutic benefits: increasing cognitive flexibility. The conditions most responsive to psychedelic treatment are all characterized by pathologically rigid cognition (e.g., Chamberlain et al., 2006; Lee & Orsillo, 2014; Palm & Follette, 2011; Todd et al., 2015). For example, high trait anxious individuals display an increased usage of their priors, failing to update predictions in response to changes in evidence (Browning et al., 2015; Kraus et al., 2021). Individuals with depression and anxiety exhibit exogenous attentional biases, such that they tend to narrowly fixate on negative or seemingly threatening stimuli (Abend et al., 2018; Mennen et al., 2019). They also exhibit increased negative and decreased positive interpretation biases, which are resistant to disconfirming information (Everaert et al., 2018). These attentional and interpretation biases often create negative feedback loops whereby people inappropriately generalize from negative experiences and interpret novel stimuli as negative or threatening. This results in increased attention to them at the cost of other things in their environment. In this way, people limit their ability to be exposed to new evidence that might otherwise result in theory revision.

These so called “learning traps” (Rich & Gureckis, 2015) are thought to underly many pathologies. In fact, cognitive behavioral therapy and exposure therapy, the conventional non-pharmacological treatments for depression and anxiety, specifically target cognitive rigidity and learning traps by forcing people to revise their theories by exposing them to evidence they otherwise would not get themselves. Further, aberrant DMN connectivity has routinely been

found among individuals with major depressive disorder (e.g., Leibenluft & Pine, 2013; Seminowicz et al., 2004; Sheline et al., 2009).

Chronic pain patients also display similar hallmarks of cognitive rigidity (Ellis, 1987; Heapy et al., 2006). Pain is a complex experience. There is both a sensory and an affective component to pain. It is felt in terms of intensity as well as unpleasantness (Price et al., 1987). How one evaluates the context surrounding their pain impacts their experience of that pain (Grant & Zeidan, 2019). By affording theory revision, possibly through functional connectivity and plasticity changes (Castellanos et al., 2020), psychedelics may not only alleviate the anxiety and depression that are often chronic pain comorbidities but provide analgesia directly through changing a patient's relationship with their pain. Specifically, changing the contextual evaluation of their pain which may attenuate affective-driven pain. Such perspective shifts were reported as crucial to the phantom limb pain relief that psilocybin engendered (Ramachandran et al., 2018).

The work outlined below seeks to clarify whether and how psilocybin may result in changes to cognitive flexibility. Specifically, three key features symptomatic of more flexible cognition are tested. These features are inspired by developmental differences from the proposed cognitive development model in Chapter 1. Specifically, that children and adults differ in how much their prior knowledge guides their inferences, how they sample from the space of possible hypotheses and interpretations, and how they search their exogenous environment for evidence. They are also inspired by reported differences in clinical populations that psychedelics are effective in treating. The interpretation biases expressed in individuals with anxiety and depression could result from hyper local sampling of some high prior negative hypotheses. Their deficits in theory revision could result from not sampling any disconfirming hypotheses or restricting their attention and not exposing themselves to disconfirming evidence. Together the

results of these experiments represent an important step toward understanding how this novel and unprecedently effective therapeutic treatment affords its clinical benefit.

Study Design

Data for the cognitive tasks outlined below were collected from a much larger double-blind placebo-controlled study examining the therapeutic efficacy of psilocybin as a treatment for chronic phantom limb pain. Importantly, the Results sections below present preliminary data from the pilot phase of this study and the reported analyses serve to communicate the analysis plan for the follow-up work with a larger sample size. Thus, the statistical analyses here are solely to preview what they look like in the current sample and are not intended to be taken as representative of the final sample. In this pilot study, nine total participants with chronic phantom limb pain were enrolled and randomized to receive either psilocybin (25 mg p.o.; n = 5) or a niacin control (100mg p.o.; n = 4). Due to limb amputation, finding weight-adjusted drug doses is complex. Considering that recent evidence suggests there are no differences in the subjective effects of psilocybin as a result of weight (Garcia-Romeu et al., 2021), fixed doses were administered. fMRI scans were conducted at baseline and the day after the experimental drug session. Clinical outcomes (e.g., pain, depression, anxiety, etc.) were collected at multiple intervals throughout the study and longitudinally (1-, 2-, and 4-weeks following the experimental session). All cognitive tasks were conducted the day following the experimental drug session. For the purposes of this dissertation, only data from the cognitive tasks will be reported.

Participants

Participants with a single extremity amputation were recruited through flyers, the internet, physician referral, and direct contact (if they had a MyChart account and record of amputation). Of the 159 individuals who completed the initial telephone screening, nine were

randomized and completed the study protocol. Demographics for all participants are shown in Table 3.1. All participants reported experiencing phantom limb pain with an average pain rating of at least 3/10 and occurring at least once per week. All participants in this sample had either a leg (seven) or finger (two) amputated. The side of the amputated extremity varied, with four having a right-side amputation, and five having a left-side amputation. This study was approved by the UC San Diego IRB, and written informed consent was obtained from all participants.

Table 3.1 Participant Demographics

Measure	Placebo (n=4)	Psilocybin (n=5)	All Participants (n=9)
Gender (% female)	50%	40%	44%
Age in years (mean, SEM)	31.2 (4.33)	41.0 (6.55)	36.7 (4.24)
Race / Ethnicity			
White	100%	60%	77.8%
Non-White	0%	40%	22.2%
Years since Amputation (mean, SEM)	9.22 (5.38)	3.82 (0.69)	6.22 (2.42)
Phantom Limb Pain Duration in years (mean, SEM)	1.22 (0.61)	6.94 (3.10)	4.40 (1.47)
Lifetime Use of Psychedelics			
Percent reporting any past use	75%	80%	77.8%

Procedure

Participants were first interviewed via a preliminary phone screening to assess for study eligibility by checking whether they met all inclusion criteria and did not meet any exclusion criteria. Eligible participants were invited for an in-person screening visit at the UCSD Altman Clinical and Translational Research Institute (ACTRI) which consisted of five components: 1. Informed Consent, 2. Questionnaires and pain testing, 3. A physical examination, 4. A psychological examination, and 5. A blood draw and urine testing. The physical and psychological examinations were to verify the information collected during the phone screening to assess for study eligibility. The blood test was to confirm study inclusion/exclusion criteria that could not be assessed during the phone screening. During the course of their study

participation, participants were required to refrain from using illicit drugs. Additionally, individuals who were pregnant or nursing were ineligible for the study. Thus, the urinalysis consisted of a drug test and a pregnancy test (for participants of child-bearing potential). Participants who remained eligible after the in-person screening next attended a baseline assessment visit at the UCSD Keck Center for fMRI where they completed several survey measures[1], psychophysical testing, and an fMRI scan. The data collected at this visit served as the baseline, pre-intervention, time point for which all post-intervention and longitudinal data were compared. After this visit, participants were randomized to either the psilocybin or niacin condition and began their preparation for the experimental drug sessions.

In line with the established protocols and guidelines implemented in clinical research with psilocybin to reduce anxiety, fear, and other adverse reactions (Johnson et al., 2008), each study participant was assigned two monitors (or guides) who supervised the experimental drug sessions. Prior to the experimental drug sessions, the participants had three preparatory meetings with their monitors. The purpose of these preparatory meetings is three-fold (Cosimano, 2021). First, to establish trust and rapport between the participant and their monitors. Second, so that the monitors have all relevant points of reference to provide therapeutic support as needed during the experimental drug sessions. In service of these first two aims, the monitors would review the participant's life history, major life events including those leading up to and immediately following the amputation, social and familial relations, etc. They also discussed the participant's expectations and intentions for their study participation. The third and final aim is to directly prepare the participants for the experimental session day. This is accomplished by undergoing a "mock" session conducted in the session room, establishing participant preferences regarding assisting with navigation and physical movement during the session, reviewing the range of

possible experiences elicited by psilocybin and niacin, as well as how to handle any potential disappointment that may arise if participants suspect they did not receive psilocybin.

Following the preparatory visits, within approximately one week, participants underwent their experimental drug session. They arrived at the ACTRI around 8:30am, completed presession questionnaires assessing their expectations and noting what they ate for breakfast, provided a urine sample for drug and (if relevant) pregnancy tests, and had a baseline set of vitals (heart rate and blood pressure) taken. While the sessions took place in an overnight room at the ACTRI clinic (a spacious room with two beds positioned adjacently and an adjoined bathroom), the room itself was furnished to be "living room-like", adorned with hung tapestries, salt lamps, blackout curtains, and other miscellaneous items. Since many of the participants experienced medical-related trauma or PTSD as a result of their amputations, great care was taken to minimize as best as possible the hospital aesthetic of the room. For example, all medical and research equipment were either removed or covered. Once the urinalysis and all preparation was completed, participants were administered an opaque gelatin capsule containing either 25mg psilocybin or 100mg niacin. After administration of the blinded study drug, participants were instructed to lay on the bed with eyeshades and headphones on and were encouraged to "go into the experience." A standardized playlist implemented in many clinical trials administering psilocybin (Strickland et al., 2021) was played through the headphones and synchronously throughout the room through speakers, so as to not "break" the experience if the participant needed to use the restroom. The playlist consisted mostly of classical music and exclusively songs without English words, to prevent the linguistic content of the music from influencing the participants' experiences. At regular intervals throughout the day, the monitors took measurements of the participant's vitals and completed assessments noting participant behaviors

and their inferred intensity of drug effects. Support would be provided as needed and requested by the participant, but otherwise the monitors played and ancillary role, intervening only if deemed appropriate in their clinical judgement. A variety of "rescue" medications were on standby in the event that the participant experienced a significant adverse event. The administration of such rescue medication is very rare, and was not implemented in any participants in this study. A study team physician and ACTRI nursing staff remained on call for the duration of expected drug effects (approximately 7 hours), at which point the participants completed a battery of post-session questionnaires[1] assessing their subjective drug effects before being allowed to depart the ACTRI.

The day after their experimental session, participants returned to the Keck Center for fMRI for their first follow-up visit. This visit largely mirrored the baseline visit, completing the same battery of questionnaires, psychophysical testing, and fMRI protocol. However, participants additionally completed a battery of cognitive flexibility tests. Upon completion of all data collection, participants met with at least one of their study monitors for an integration meeting. In this meeting, participants would discuss a narrative description of their experiences the previous day, their reflections and feelings about the drug session, and any difficulties that arose during the day or evening. The purpose of this meeting is to help participants safely integrate their experiences into their lives, and has been shown to help reduce incidences of post-acute anxiety, depression, or ontological shock (Cosimano, 2021). While for most participants this integration meeting took place the day after their experimental drug session, for some participants this meeting took place within 1-3 days after due to scheduling constraints. Participants subsequently and remotely completed the same battery of questionnaires[1] at 1-, 2-,

[1]Because these participants were involved in this larger study, several additional measures were given but are outside of the scope of this dissertation. The full list of assessments can be found in Appendix B.

and 4-weeks following their experimental drug session, at which point their study participation concluded.

Experiment 1: Does Psychedelic Treatment Change the Influence of Prior Knowledge on Inferences?

According to Bayes' Rule, the stronger one's priors, the more their inferences are guided by prior knowledge and the less sensitive they become to current evidence, even if it contrasts with their existing theories. As noted in Chapter 1, we observe differences in children and adults' prior knowledge. Adults have strong priors about many things. As we grow and learn, our confidence in our theories strengthens, and we become less likely to change them. Several studies have shown that while adults tend to rely on their stronger priors and update their theories much less, children readily and flexibly update their theories in response to counterevidence (e.g., Gopnik et al., 2015; Kimura et al. in prep; Lucas et al., 2014; Seiver et al., 2013). Here, I explore whether these differences are also observed when looking at adults after taking psilocybin. Changes in the strength of priors underlies a prominent theory for the effects of psychedelics known as the Relaxed Beliefs Under Psychedelics (REBUS) account (Carhart-Harris & Friston, 2019). However, the REBUS account is founded almost exclusively on neuroimaging data with no direct behavioral evidence.

To formally test this, as well as the predictions of the REBUS account, I adapted an experimental paradigm originally developed by Lucas and colleagues (2014) and replicated in a variety of follow up studies (Gopnik et al., 2015, 2017; Lucas et al., 2014; Walker et al., 2020). Participants were first introduced to a novel machine and told that is activated only by objects that are 'blickets'. To determine whether an object is a blicket or not, participants must observe whether the machine activates (lights up) when the object is placed on top of it. After being oriented to the machine, participants were shown some unambiguous training data which implied

that the machine operated according to a general rule. The causal reasoning literature suggests that disjunction (in which individual objects are causal) is the default assumption for adults, associated with the highest prior (e.g., Cheng, 1997; Griffiths & Tenenbaum, 2005). Lucas et al. (2014) originally tested three conditions. In one (disjunctive condition), the training data indicated that the machine operated by such a disjunctive rule: each object either does or does not activate the machine. The hypothesis implied by this evidence is therefore **consistent** with their high prior. In the second (conjunctive condition), the training data indicated that the machine instead operated by a conjunctive rule: two objects must be placed on the machine at the same time to make it activate. The hypothesis implied by this evidence therefore **contrasts** with their high prior. In the third (baseline condition), no training data was presented, and participants could only rely on their prior beliefs. After seeing the unambiguous training data (or no data in the baseline condition), participants were shown some ambiguous test data with new objects that was consistent with *both* their default high prior disjunctive theory that only one object was needed to activate the machine *and* the less likely conjunctive theory (see Figure 3.1).

Participants were then asked to identify which objects from their test data were blickets. For one critical object, it is ambiguous from the test data alone as to whether or not it is a blicket. This judgement is thus informed by the rule participants infer for how the machine operates. They were also asked to try to activate the machine themselves using the available objects. This second test question is needed to disambiguate whether the results would be solely a consequence of linguistic nuance, i.e., whether participants had a general reluctance to ascribe an object the "blicket" label. Thus, rather than being explicitly reported, participant behavior and responses to follow-up test items demonstrates which conclusion they drew. The use of this

experimental paradigm allows the likelihood of the evidence to be kept constant, and any observed differences can be attributed to changes in the weight of the priors.

Across several experiments, the authors found that in the conjunctive condition, the proportion of participants who correctly labeled the critical object as a blicket and chose multiple objects when prompted to activate the machine themselves generally decreased with age. Importantly, there were no differences between children and adults in the baseline condition: They both inferred that the machine operated according to a disjunctive rule. This implies that the developmental differences are not the result of differences in prior beliefs, but rather the relative *strength* of those prior beliefs: Children are more willing to switch to an initially unlikely hypothesis when presented with counterevidence (Lucas et al., 2014).

In the current work, only the conjunctive condition is tested. This is because it is the critical condition that assesses whether there are changes to the strength of one's priors, our sample size would not allow for additional conditions, and, importantly, previous work has demonstrated the robustness of adults' priors for disjunction (Lucas et al., 2014; Gopnik et al., 2015; Gopnik et al., 2017). If psilocybin indeed decreases the strength of adults' priors, then participants in the psilocybin condition should be more sensitive to the observed evidence and perform similarly to children in previous work. Specifically, they should correctly label the critical object as a blicket and select multiple objects when asked to activate the machine. In contrast, participants in the control condition should remain biased by their priors and be more likely to ignore the counterevidence, performing like adults in previous work. Specifically, they should fail to label the critical object as a blicket and select a single object when asked to activate the machine. Alternatively, if psilocybin does not lower the strength of people's priors, the

proportions between the two conditions should be similar and in line with results from adults in previous studies.

Method

Materials. This experiment was conducted electronically, but several physical stimuli were constructed and implemented in the creation of the images and videos used in the task. A "blicket machine" was constructed with a small wooden box painted black that had a frosted glass window on all sides. Inside the box was a light bulb that was activated via a hidden remote control. The objects were eleven different shapes made of standard grey clay. Clear plastic containers with masking tape labels were used in the instructions video to organize the shapes (See Appendix C for all stimuli). The videos were recorded in front of a plain white background and all objects (shapes and the machine) were placed on a brown mat on top of a white table.

Procedure. Participants first watched an instructions video where they were introduced to a novel machine and told that is activated only by objects that are 'blickets', which cannot be readily identified by any visual properties. Thus, to determine whether an object is a blicket or not, participants must observe whether the machine activates (lights up) when the object is placed on top of it. They are told their task is to figure out which objects are blickets. Though blickets should be a novel concept for which participants do not have any specific prior beliefs, the following was done in an attempt to control for any potential differences in beliefs about the prevalence of blickets that participants may have held. First, the experimenter states that only a few of the objects are blickets and most are not. Second, the experimenter pulls out two clear containers with the labels "Blickets" and "Not Blickets". The "Blickets" container had a single object in it, whereas the "Not Blickets" container had 4 objects. None of these objects were used again in the experiment. The experimenter then pulled out and counted the number of objects

from each container, and reiterated that only a few objects are blickets. The instructions video ends by noting that the participant will next be shown unsorted objects, for which it is unclear whether they are a blicket or not.

The training trials video begins with the experimenter seated at the location with the blicket machine and 3 objects next to it. The experimenter begins by picking up each object individually and verbally labeling them by their shape. Then, the experimenter goes through a series of training events where a single object or a combination of objects were picked up and placed on the machine, which either does or does not activate. Specifically, the experimenter states, "Let's see what happens when we put X (and Y) onto the machine," places the single object or combination of objects onto the machine, and then verbally communicates the result ("It did not turn on" / "It turned on!"). There are six training events in total, where each of the three objects are placed on the machine alone and all pairwise combinations of two objects are placed on the machine. Importantly, the evidence provided by these training events is unambiguous, and indicates that the machine operates according to a conjunctive rule. The only instances where the machine activates occurs when multiple objects are placed on the machine in conjunction. The full list of training events is displayed in Figure 3.1. When the training video ends, the participant is shown a series of follow-up questions where a picture of each object is shown individually and they are asked a two-alternative forced choice question denoting whether they think that object is a blicket or not. Unlike in previous work, we also ask them to rate how confident they are in their assessment about whether the object is a blicket or not (on a scale of 0 = Not Confident at All to 100 = Extremely Confident). These questions are repeated for each of the three objects shown in the training trials video. Upon completion, participants are then told they will next see three new objects and try to determine whether they are blickets or not.

The test trials video follows a similar format to the training trials video. Three novel objects are shown and the experimenter starts by picking them up individually and verbally labeling them according to their shape. The experimenter then proceeds to go through a series of test events where a single object or a combination of objects will be picked up and placed on the machine, which either does or does not activate. During these test events, one critical object is placed on the machine alone, which does not activate, but the machine does activate each time it is placed with another object. Critically, in all events where multiple objects are used, the combinations include one object that is never placed on the machine independently. Thus, these test events are ambiguous: either multiple objects are needed to activate the machine, or the machine is activated by the single object that was never tested individually. The full list of test events is displayed in Figure 3.1. Participants are then shown the same follow-up questions about each object individually. However, upon completing these questions, they are then prompted to try to activate the machine themselves by selecting only the objects necessary to activate it. They are shown a screen with pictures of each of the three objects and select which single or combination of objects they would use. They are then asked to rate how confident they are about whether their choice will activate the machine. Full video scripts and question text are displayed in Appendix C.

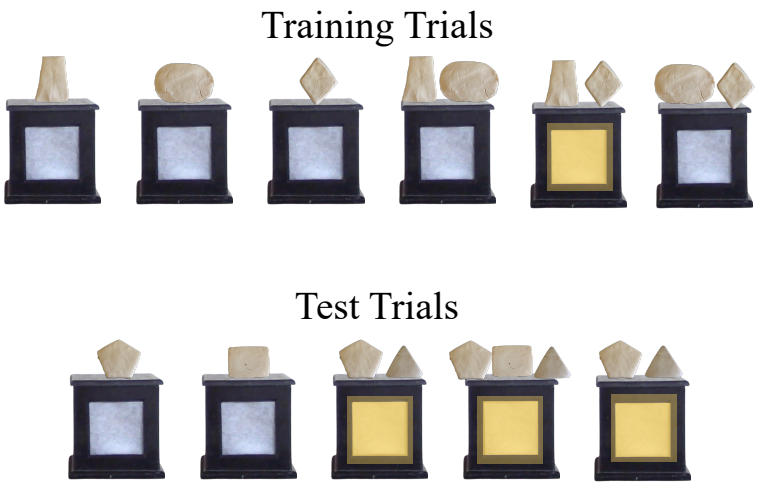


Figure 3.1 Unambiguous training data implying the machine operates according to a conjunctive rule (top row), and ambiguous test data that can be explained by conjunction or disjunction (bottom row).

Results

Blicket Judgements. Of particular interest is whether participants labeled the critical test object as a blicket, as this informs whether they were sensitive to the evidence shown in the training events and revised their prior belief that the machine would operate according to a disjunctive rule. If psilocybin treatment reduces the strength of one's priors and makes them more sensitive to current evidence and likely to update their beliefs, participants in the psilocybin condition should be more likely to identify the critical object as a blicket than those in the niacin condition. Given the small sample size in the current work, a Fisher's Exact Test was used to test for differences in the rate of labeling the critical object as a blicket between conditions. This revealed that, although participants in the psilocybin condition (2 / 5) were more likely to label the critical object as a blicket than those in the niacin condition (0 / 4), this difference was not significant ($p = 0.44$). Effect sizes could not be computed given that one cell had a count of 0 (resulting in an odds-ratio of infinity). A one-way ANOVA was conducted to compare the difference in confidence ratings for this judgement between conditions. Likewise, this revealed that although participants in the psilocybin condition ($M = 60.8$) tended to be more confident in

their judgements than those in the niacin condition (M = 51.8), this difference was not significant, $F(1, 7) = 0.08$, $p = 0.79$, $d = 0.19$, $BF = 0.52$.

Intervention Choices. As mentioned above, there may be some linguistic nuance driving participants' responses, whereby they have a general reluctance to ascribe an object the unusual and abstract "blicket" label. Thus, participants were also asked to intervene on the machine and try to activate it themselves. The critical test here is whether they chose a single object (regardless of which) or multiple objects to try and activate the machine themselves. If psilocybin treatment reduces the strength of one's priors and makes them more sensitive to current evidence and likely to update their beliefs, participants in the psilocybin condition should be more likely to choose multiple objects than those in the niacin condition. Given the small sample size in the current work, a Fisher's Exact Test was used to test for differences in the rate of choosing multiple objects when prompted to try and activate the machine between conditions. This revealed that participants in the psilocybin condition (5 / 5) were more likely to choose multiple objects than those in the niacin condition (1 / 4), and despite the small sample size, this difference was significant ($p = 0.048$). Effect sizes again could not be computed given that one cell had a count of 0 (resulting in an odds-ratio of infinity). A one-way ANOVA was conducted to compare the difference in confidence ratings that this choice would activate the machine between conditions. Participants in both the psilocybin (M = 85.8) and niacin (M = 88.2) conditions had similar ratings of confidence, $F(1, 7) = 0.04$, $p = 0.84$, $d = 0.14$, $BF = 0.52$.

Exploratory Analyses. Given that Fisher's Exact Tests provide very conservative estimates (Fisher, 1935), the intervention choice data was also modeled using a logistic regression, the approach one would take with a larger sample size. This revealed a significant main effect of condition, $\chi^2(1) = 6.96$, $p = 0.008$. In considering the hypothesized linguistic

nuance that may impact participants' likelihood of labeling an object as a blicket in general, we compared the confidence ratings across all participants on their blicket judgements and intervention choices using a one-way ANOVA predicting confidence rating by trial type (blicket judgement vs intervention). While not a significant difference with the current sample size, $F(1, 16) = 3.43$, $p = 0.08$, $d = 0.87$, $BF = 1.26$, participants overall were more confident in their intervention choices ($M = 86.9$) than their blicket judgements ($M = 56.8$).

Discussion

In this experiment, we tested whether psilocybin treatment affects the strength of people's priors. We hypothesized that, compared to niacin, psilocybin would reduce the strength of people's priors, making them more sensitive to current evidence and consequently more likely to revise their beliefs. The results from this pilot data provide initial support for this hypothesis. While the difference was not significant, participants in the psilocybin condition were more likely to label the critical object as a blicket than participants in the niacin condition. Additionally, while also not a significant difference, participants in the psilocybin condition were more confident in their judgements than those in the niacin condition, though both were only moderately confident. When asked to intervene and try to activate the machine themselves, all participants in psilocybin condition choose multiple objects compared to one in the niacin condition, and both were similarly highly confident. While this difference just crossed the threshold for significance, it is likely that this would be a more significant effect with a larger sample size, as was shown in the exploratory logistic regression analysis of the same data. While participants' confidence in their blicket judgements and intervention choices did not differ between conditions, surprisingly, their collective confidence ratings were numerically much higher for the intervention choice than the blicket judgement. This suggests the hypothesized

linguistic nuance may have impacted participants' blicket judgements, making them more reluctant to label an object as a blicket. Future research might consider weighting intervention choices more heavily than blicket judgements when using this paradigm to test adults.

Together, these results provide initial behavioral evidence in favor of the REBUS model (Carhart-Harris et al., 2014; Carhart-Harris & Friston, 2019). They also provide a mechanism which could underlie previous findings that psychedelic use is associated with increased suggestibility (Carhart-Harris et al., 2015; Lebedev et al., 2023; McGovern et al., 2023), which has long a storied history dating back to the CIA's MKULTRA project (Linville, 2016). It is unclear how long the period of decreased priors lasts. Recent work has likened the effects of psychedelics to a 'reset' mechanism, whereby acute disintegration of psychological processes affords a post-acute re-integration which leads to overall improvements in normal functioning (Carhart-Harris et al., 2017). The decreased strength of prior beliefs may thus characterize a therapeutic window of opportunity during which critical therapy can be administered and lead to the widespread benefits psilocybin has demonstrated eliciting.

Experiment 2: Does Psychedelic Treatment Change Hypothesis Search Strategy?

As noted in Chapter 1, people are thought to approximate Bayesian inference by stochastically sampling candidate hypotheses to evaluate from a probability distribution over possible hypotheses (e.g., Sanborn, 2017; Sanborn & Chater, 2016; Thomas et al., 2008; Vul & Pashler, 2008). This sampling process is not done randomly, hypotheses are instead sampled with frequencies proportional to their probability. After sampling and evaluating a hypothesis, its probability will be updated, the probability distribution over all hypotheses is adjusted, and a new hypothesis is sampled from this updated distribution. Given that this process is costly, there is a tradeoff between exploration for new hypotheses and exploiting current ones, known as the

exploration-exploitation dilemma (Mehlhorn et al., 2015). The tension produced by this dilemma is driven by opposing goals: 1. Gaining useful information by exploring unfamiliar options/hypotheses despite the potential for poorer immediate rewards, and 2. Maximizing rewards by exploiting options/hypotheses with high reward expectations.

This exploration-exploitation dilemma has been thoroughly investigated in the context of reinforcement learning paradigms (Sutton & Barto, 1998). Here, there is some unknown reward structure, and agents (human or computer models / algorithms), through their chosen actions, must learn which are most rewarding and which are less so or even costly. Critically, only chosen actions will provide feedback to help agents learn. These paradigms thus force the agents to resolve an explore-exploit dilemma by balancing exploration to learn the underlying reward structure with exploitation to maximize reward. Computational approaches to better understanding how this dilemma is resolved have proposed several potential algorithmic accounts, and despite their differences most have a property in common: Effective exploitation of an environment's reward structure requires first understanding that reward structure. Thus, in these algorithms, exploration is prioritized initially and then its use fades over time as information is accumulated, a process referred to as simulated annealing. Simulated annealing is used analogously, referencing how a piece of metal is maximally flexible initially when it is very hot and becomes less malleable as it cools over time, but more formally is a stochastic optimization algorithm. These algorithms are maximally likely to arrive at the optimal solution in a given environment (Kirkpatrick et al., 1983). Developmental differences in search strategies have been posited to be analogous to simulated annealing (Gopnik, 2020; 2024) and resemble a process of stochastic optimization, framing development as a process of parameter optimization (Giron et al., 2023).

Adults tend to privilege efficiency and seek to maximize their utility (Dennett, 1989; Gergely et al., 1995; Jara-Ettinger et al., 2016). As a consequence, they are very exploitive and employ local, "low-temperature," search strategies when sampling hypotheses (e.g., Herbst et al., 2017; Smith et al., 2013). Guided by all the information they have accumulated throughout their life, a high prior hypothesis is initially sampled and strategically incremented such that successive samples are close in hypothesis space. Adults thus tend to base their decisions on just a few samples (Goodman et al., 2008), and often times taking a single sample is the optimal strategy (Vul et al., 2014). These exploitive strategies lead to quick, "good enough" solutions. However, they leave one susceptible to being stuck in a local optima, whereby their current hypothesis may be better than all local alternatives but much worse than alternatives that are further away (Gopnik, 2020).

Children, on the other hand, engage in directed and systematic exploration (Schulz et al., 2019), even when this is associated with a greater cost (Liquin & Gopnik, 2022). This behavioral pattern of adults being more exploitative and children being more exploratory has been documented in causal inference tasks (e.g., Gopnik, 1996, 2012; Lapidow & Walker, 2020) and in general reinforcement learning paradigms where they freely act in an environment and choose whether to continue exploring or exploit known rewards (Blanco & Sloutsky, 2020, 2021; Liquin & Gopnik, 2020; Schulz et al., 2018; Sumner et al., 2019; Sumner et al., 2019). Thus, these differences in explore-exploit behaviors may reflect a developmental change in the *process* of theory revision via changes in search strategies, rather than in the strength of their priors alone.

Here, I test whether these differences are also observed when looking at adults after psilocybin treatment. Exploiting and exploring map onto more rigid and flexible cognitive strategies, respectively. To test for possible changes in search strategy when sampling, two

different types of paradigms were implemented: serial production tasks (SPTs; both single and multiple cue) and general exploration paradigms.

Serial Production Tasks (SPTs)

In SPTs participants are prompted to generate candidate solutions based on a single or multiple cues. These cues serve to constrain the responses that are generated. Search strategies can then be quantified in several ways, for example by the average similarity between sequentially generated responses. If the average sequential response similarity is high, then local search strategies are being employed, as each response is close to each other in hypothesis space (low distance). If the average sequential response similarity is low, then more global strategies are being employed, as responses are further away in hypothesis space (high distance).

Exploration Paradigms

While SPTs produce data amenable to investigating hypothesis search and sampling strategies, they have several limitations compared to other more general exploration and search paradigms. When generating responses in an SPT, individuals explore word or concept associations one at a time, characterizing sequential exploration rather than the simultaneous exploration that is more representative of what people do in the real world. SPTs often do not provide feedback, and when they do it is usually after a trial has concluded. Participants are thus unable to update their strategies based on this feedback, again unlike in the real world. Further, and critically, given that responses are generated by the participant rather than chosen amongst a set of provided options, there is not the option for participants to be exploitive. In this way, they cannot provide insight into how respondents are resolving the exploration-exploitation dilemma, because it does not exist. In contrast, general exploration paradigms do require participants to choose among a provided set of options, and thus create an exploration-exploitation dilemma.

A common general exploration paradigm is the multi-armed bandit task, a specific type of reinforcement learning paradigm. In these, participants choose between a fixed number of options which have varying but initially unknown reward amounts. Participants complete several rounds but have a limit on the number of actions they can take in each, and are free to explore options which may potentially be more rewarding or exploit the options with a known reward. Through this controlled environment, various factors such as the limit on number of actions, number of available options, reward distribution, and fixed vs dynamic nature of that distribution, can be manipulated. They therefore present more direct assessments of how agents resolve an exploration-exploitation dilemma.

Experiment 2.1: Single Cue SPT

To assess search strategies with a single cue SPT, I adapted an experimental paradigm used by Bonawitz and colleagues (in prep). In this task, participants were introduced to a novel machine that is activated by a particular member of a category and generated 5 hypotheses about what they thought would activate it, without receiving feedback. Using a novel machine with no information or feedback regarding what would be a likely cause controlled for any potential differences in priors between participants, helping to disambiguate changes in search strategy from changes in priors. After providing their responses, participants completed a conceptual sorting task where they organized their guesses with proximity corresponding to their similarity. They found that adults had a significantly lower average distance between their sequentially generated responses than children. That is, their sequentially generated responses were more similar. Additionally, of the total responses provided by both age groups, children provided a significantly greater proportion of unique responses than adults. Though adults had greater knowledge of possible responses to generate, they still generated a less diverse set. Further, there

was a general correspondence between children and adults' responses such that the average distance between pairs of items was correlated between age groups, suggesting that adults and children agreed on the similarity between pairs of specific items.

It was hypothesized that psilocybin will change people's search strategies, making them more global. Thus, participants in the psilocybin condition should perform like children in previous work, having greater average sequential response distance (less similarity) and covering more hypothesis space with their guesses. Participants in the niacin condition should employ local search strategies, performing like adults in previous work, having less average sequential responses distance (greater similarity) and covering less hypothesis space. Alternatively, if psilocybin does not change people's search strategies, the average sequential response distance and amount of hypothesis space covered between the two conditions should be similar and in line with results from adults in previous studies.

Method.

Materials. The original task was completed in person, which used a 36-inch x 36-inch cardboard sheet broken into 10 x 10 grids. Guess items were recorded on 2-inch x 2-inch cards and placed along the grid by participants. Here, the task was adapted to be completed electronically, and thus there were no physical materials.

Procedure. Participants were told there was a novel machine that is activated by a particular member of a category (animals). They were given no information about which animal was likely to activate the machine and were asked to generate five successive guesses (hypotheses) about which animal they thought would activate it, receiving no feedback between responses. Specifically, they were asked, "What animal do you think made the machine light up?" After providing their responses, participants completed a conceptual sorting task. They

were told they were looking at the plans for a new zoo that is under construction, which has the exhibits planned but not their location, and asked to help the designers arrange the layout of the exhibits such that animals that are most similar are close together and animals that are dissimilar are further apart. The page following these instructions displayed an orange square outline with 12 items along the edges. These 12 items contained their five responses along with seven additional items selected from a predetermined anchor set (Appendix D). Specifically, items not present in the participant's set of guesses would be included. This provided an individualized measure of subjective item similarity, as measured by the Euclidian distance between any two items on the sorting grid. Once all twelve items were arranged inside the sorting grid, the square would turn green and a "continue" button would appear, which ended the experiment upon clicking.

Results. We first examined for differences between the average sequential response similarity between conditions. If participants were employing local search strategies, they might generate candidate response chains like, "Dog, Cat, Bunny." These responses are all common house pets and would be grouped relatively closely (less distance between). If they were employing more global search strategies, they might generate candidate response chains like, "Beetle, Horse, Owl." These responses have less readily apparent connections and would be grouped further apart (more distances between). The average Euclidian distance between all sequentially generated response pairs within each participant was calculated. Since participants may use the sorting space differently, these distances were normalized by dividing by the average similarity between all pairwise responses, or the overall distance expected if participants were sampling randomly. Differences in normalized sequential response distance between conditions (psilocybin vs niacin) were compared using independent samples t-tests. Participants

in the niacin condition (M = 0.92) tended to have less similarity than those in the psilocybin condition (M = 0.78), but this difference was not significant, $t(7) = 1.48$, $p = 0.18$, $d = 0.99$, BF = 0.92. This suggests participants in both conditions sampled with similar search strategies.

We next examined the overall amount of space covered by participants' guesses, defined by the number of unique responses generated. Participants' guess lists were collated between conditions, and repeat responses were removed, to generate the total number of unique guesses reported by each condition. The proportion of unique responses between conditions was compared using a Chi-Squared test. Following a similar pattern, while participants in the niacin condition (18/20) generated a greater proportion of unique responses than those in the psilocybin condition (17/25), this difference was not significant, $\chi^2(1) = 3.11$, $p = 0.15$.

Discussion. In this experiment a single-constrained serial production task was used to investigate whether psilocybin treatment will impact search and sampling behaviors. It was hypothesized that participants treated with psilocybin will exhibit more global, exploratory, search and sampling behaviors. We found that, participants in both conditions produced sequential responses that were similarly distanced. Additionally, they did not differ in their proportion of unique responses. Together, these suggest that participants were employing similar search and sampling strategies, and that psilocybin did not affect these behaviors.

Experiment 2.2: Multiple Cue SPT

In single cue SPTs, close associates of a response also tend to satisfy the constraints of the problem. For example, when generating responses for what animal may activate a machine, both "cat" and "dog" satisfy the constraints of the problem. However, "cat" and "shark," equally satisfy these constraints, despite being very dissimilar. Further, in the real world there are often multiple constraints on an inference. Responses, and their close associates, that satisfy one

constraint might not satisfy the others. Multiple cue SPTs thus provide a more robust examination of search strategies.

The Remote Associates Task (RAT) is among the most commonly used SPT (Mednick, 1962). In each trial of the RAT, participants are given 3 target cue words and their task is to generate a 4th word which pairs with all three. For example, in the cue set "Comb, Dew, Moon," the correct answer is "Honey." The RAT is conventionally used to assess creativity and problem solving abilities by summing the number of correctly answered trials (Mednick, 1962), but the data it provides affords much richer analysis. Smith and colleagues demonstrated that the data generated by the RAT can be used to assess search strategies (Smith et al., 2013). They did so by using a 300-factor Latent Semantic Analysis (LSA) model to compute similarity metrics between responses. How similar or dissimilar the responses are in a given trial reflects what type of search strategy participants are exhibiting. If responses are similar, then local search strategies are being employed. If responses are less similar, then global search strategies are being employed.

However, as outlined in Chapter 2, the RAT has two major limitations. First, participants often produce a small number of responses. This may be because they are able to both sample a candidate solution and evaluate its quality internally, prior to actually reporting the response. Following the example above, a participant may come up with the candidate solution "hair", which at first might appear to match with the cues "dew" and "comb". They might then realize that it does not match with the cue "moon", and that "hairdo" is a single word and not associated with the word "dew". In this case, they may discard "hair" as a candidate response and not actually report it, despite being encouraged to do so by the experiment's instructions. Through a pilot experiment, this was found to be the case even in a modified version of the RAT where

participants' audio responses were recorded and they were encouraged to "just think out loud" without needing to directly write down their candidate responses. Second, while the RAT does impose constraints, making it more like the real-world problem-solving humans face, the constraints are static. In many cases, people get some form of feedback on their proposed solutions, and this feedback serves to further constrain the generation of proposed solutions. Thus, the RAT additionally falls short in capturing this continual updating inherent in many problem-solving scenarios.

To account for these issues, a novel experimental paradigm was used based on the popular internet game "Contexto" (see Chapter 2). In this task, participants complete multiple trials where they have to guess a secret target word. They are initially given no information as to what the target word might be and must simply generate a series of guesses to try and identify it. Participants receive feedback after submitting each guess on how similar it was to the secret target word, and thus must continuously update the constraints (imposed by their own previously generated guesses) on their subsequent guesses. It was hypothesized that psilocybin will change people's search strategies, making them more global. Thus, participants in the psilocybin condition should have *lower* average sequential response similarity and cover more hypothesis space in their response sets. Participants in the niacin condition should continue to employ local search strategies, performing like adults in previous work and have *greater* average sequential responses similarity and cover less hypothesis space in their response sets. Alternatively, if psilocybin does not change people's search strategies, the two groups should be similar and in line with results from adults in previous studies on all outcomes.

Method.

Materials. The target words implemented here were chosen based on a pilot study, where they were correctly identified by a majority of participants but took an average of about twenty guesses. Thus, these words were thought to yield sufficient data. To generate the rankings of potential guesses to the target words, a dictionary of approximately 80,000 words was rank-ordered by their similarity to the target word for every trial. Similarity was computed using the Global Vectors for Word Representation (GloVe; Pennington et al., 2014). GloVe is an unsupervised learning algorithm that produces vector representations for words. A set of pre-trained word vectors was used which contained 840 billion tokens gathered via Common Crawl. The original set of word vectors was filtered to include only a subset of words originally used by Contexto, then further filtered to remove words with two or less letters, all stop words, words flagged as inappropriate (e.g., curse words and slurs), words containing numbers or punctuation, and words with multiple accepted spellings (e.g., keeping barbecue and removing barbeque). Finally, all words were lemmatized and any resulting duplicates were removed. This process resulted in a final dictionary containing 80,224 words. Similarity was defined by the cosine between any two word vectors. Compared to Latent Semantic Analysis (LSA), which captures semantic structure by focusing on singular value decomposition (SVD) to reduce the dimensionality of a term-document matrix, GloVe uses global word co-occurrence to capture meaning from the entire corpus. This approach has been shown to produce higher-quality word embeddings and perform better on semantic measures like word analogy tasks (Pennington et al., 2014). After computing the cosine similarity between each word in the dictionary and the target word for a given trial, the dictionary was arranged by the resulting values in descending order. A given word's rank was defined by its numerical position in the arranged dictionary (where the first word was the target), and that ranking is what was given to participants as feedback for their

guesses in a given trial. To account for idiosyncrasies that occurred when creating a sorted dictionary for a word in its singular vs plural form (i.e., differences between the ranked dictionary for the target word "frog" vs "frogs"), the associated cosine similarities for each word in the dictionaries were averaged and then rank-ordered. The resulting "denoised" dictionary was used in all applicable cases.

Procedure. Participants first saw an instructions screen that gave them the rules of the game. They were informed that each trial contains a randomly selected secret target word, and they had to figure out what that word was. They were also told that the secret word will always be a noun, and they have a maximum of thirty guesses per trial (see Appendix A for full instructions text). Participants submitted their guesses one at a time and were shown the guess' rank after submission. If the target word was correctly guessed, a congratulatory screen was displayed and the participant moved on to the next trial. If all thirty guesses were exhausted without identifying the trial's target word, the participant was shown what the target word was and then began their next trial. Each participant saw five trials. The same five target words ("moose", "cookie", "pencil", "flower", "car") were used for all participants, and the order of the trials was randomized.

Results. We first assessed for differences in the average similarity of sequentially generated responses between conditions. A linear mixed-effects model was fit with condition (psilocybin vs niacin) as a fixed effect. Participant and target word were included as random effects. The resulting full model was as follows: similarity ~ condition + (1 | participant_id) + (1 | target_word). Significance of the main effect was assessed with nested model comparisons. If participants were exhibiting local, exploitive, search strategies, their sequentially generated responses would be *more* similar. If they were exhibiting global, exploratory, search strategies,

their sequentially generated responses would be *less* similar. This analysis revealed a significant main effect of condition, $\chi^2(1) = 4.0$, $p = 0.046$. Participants in the niacin condition ($M = 0.44$) had a higher average sequential response similarity than those in the psilocybin condition ($M = 0.40$; Figure 3.2). This suggests that participants who received psilocybin treatment were exhibiting more global and exploratory search behaviors.

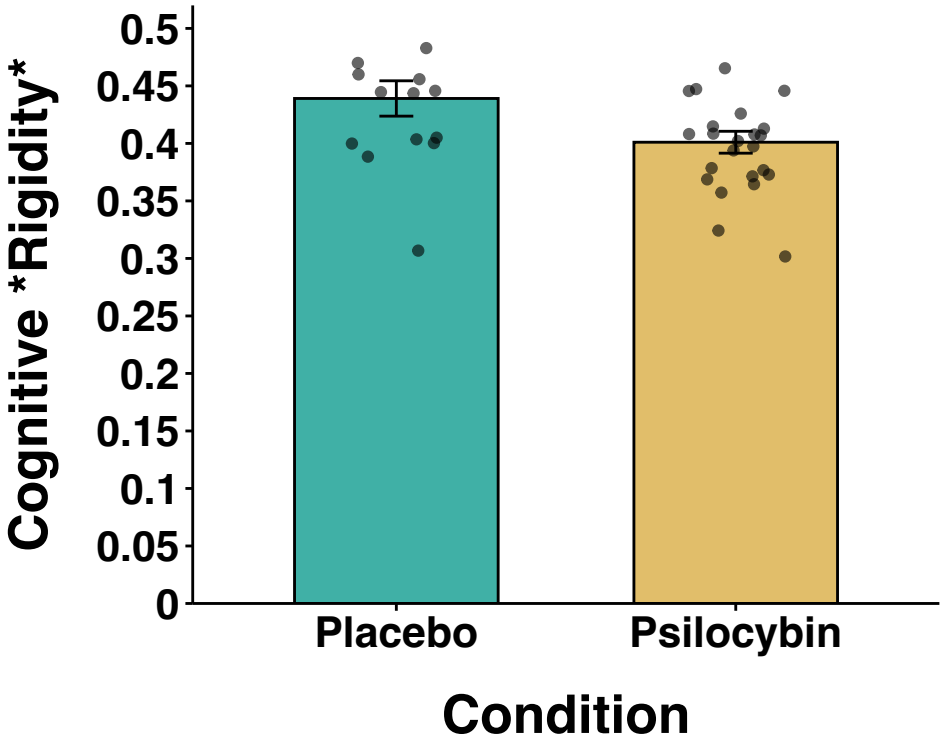


Figure 3.2 Comparison of similarity between sequentially generated responses across conditions.

We next examined whether this increase in exploration by participants in the psilocybin condition led to increased performance. Specifically, whether they guessed the target word correct more often than participants in the niacin condition. No random effects could be fit due to model singularity and convergence issues. Alternatively, to account for the repeated measurements from each subject, we computed the total amount of correctly guessed words (out of 5) for each participant. Because this measure is bounded (between 0 and 5), we fit a binomial logistic regression. Significance of the main effect was assessed with nested model comparisons.

Results showed that while participants in the niacin condition were more likely to correctly guess the target word (M = 3) than those in the psilocybin condition (M = 2), this difference was not significant, $\chi^2(1) = 1.51$, $p = 0.22$. Thus, while participants in the psilocybin condition engaged in broader search, this did result in increased performance.

Discussion. In this experiment a novel task assessing reasoning under dynamic constraints was used to investigate whether psilocybin treatment will impact search and sampling behaviors. It was hypothesized that participants treated with psilocybin will exhibit more global, exploratory, search and sampling behaviors. We found that, compared to the niacin condition, participants in the psilocybin condition engaged in more global sampling, having less average similarity between sequentially generated responses. This did not, however, lead to increased task performance, as there were no differences in how many words participants guessed correctly between the two conditions. An important consideration here is the model singularity and convergence issues that were encountered. While it is a recommended practice to fit the most complete random effect structure possible without encountering such issues (Muradoglu et al., 2023), it is ideal to fit the full model. One possible reason this may have occurred, outside of the smaller sample size, is the limited number of trials that participants completed. If some participants happened to correctly identify the target word after only a few guesses, there may be insufficient variability in their data to correctly model. To account for this in the future, when time constraints are a limiting factor, a lower guess limit but increased number of trials can be implemented. This would provide greater variance in the trial and target word estimations, and likely resolve any singularity or convergence issues. Though increasing the sample size may alone be sufficient. Overall, these results suggest that psilocybin treatment results in changes to people's search and sampling strategies, making them broader and more exploratory.

Experiment 2.3: Spatially Correlated Multi-Armed Bandit

While exploration is a complex behavior, three central components have been identified that may underly the developmental differences outlined above. The first concerns *random exploration*, through which a noisy random sampling process is employed when an agent is learning about a novel environment. This random exploration is most consistent with the analogy of development as a process of simulated annealing, whereby children employ a higher temperature search, sampling broadly and randomly, and this process "cools off" throughout their lifespan (Gopnik, 2020). The second concerns *directed exploration*. In contrast to a random sampling strategy, exploration can also be systematic and directed, where options with greater uncertainty are preferentially sampled to gain more information about the environment. In this way, information itself is viewed as intrinsically rewarding. Directed exploration has been formalized by incorporating an uncertainty bonus in reward estimation, where a choice's reward is weighted by its estimated uncertainty (Auer, 2002), and is thus more sophisticated and nuanced than random exploration. The third and final is distinct from exploration altogether, instead concerning *generalization*. Search strategies require the ability to make inferences about unobserved outcomes and are thus guided by inductive beliefs formed from observed outcomes. Structured knowledge about one's environment has been shown to guide exploration (Acuna & Schrater, 2008; Schulz, et al., 2018), so it is possible that search behavior could be impacted by this knowledge's quality or how it is utilized for generalizing across experiences.

Schulz and colleagues (2019) developed a novel paradigm to test whether and how these three components are related to the observed developmental differences in exploration behaviors. Specifically, they used a modified multi-armed bandit task where rewards were spatially correlated on a grid, and searching behaviors involved selecting tiles along that grid with known

or unknown reward values. The tiles' reward values followed a randomized continuous gradient, where similar values were clustered together on the grid. Having the rewards be spatially correlated provides participants with information they can use to generalize the reward values from observed to unobserved options. Through both behavioral outcomes and computational modeling approaches, this paradigm affords the ability to characterize the relative contributions of each component of exploration into a unified formal model (for full modeling details, see (Giron et al., 2023; Schulz et al., 2019). Overall, this paradigm provides a much more robust, comprehensive, and quantitative investigation of exploration behaviors.

The consensus of their results was that children engage in more directed exploration than adults, generalize less, and there were no differences in random exploration. In considering the behavioral results, adults sampled more locally and less unique options than children. Learning curves were also generated by averaging earned rewards over trials and showed that adults learned faster while children engaged in more exploratory sampling. Regarding the computational modeling, first the fit of predictions from two models were assessed to all participant data: one model that afforded generalization across options and another that did not. The model that did afford generalization fit all participant data better than the alternative model. This better fitting model included parameter estimates for all three components of exploration behaviors. The average generalization parameter value was higher for adults than children, indicating that adults generalized more. The average uncertainty bonus parameter value was higher for children than for adults, indicating that children valued uncertainty more than adults. The extent to which uncertainty is valued informs whether options with higher uncertainty will be sampled, and therefore corresponds to directed exploration (individuals seek out the options with more uncertainty). Finally, and importantly, there were no differences in the average search

temperature parameter value, indicating that there were no developmental differences in the amount of random exploration. Overall, both behavioral and computational modeling approaches converged on the same conclusion: children explore more than adults, and in particular engage in directed exploration where options with higher uncertainty are privileged rather than sampling more randomly.

This same experimental paradigm and analysis plan was adapted to test for the differences in exploration behaviors between adults after receiving psilocybin or niacin. It was hypothesized that participants who receive psilocybin treatment will perform more like children, engaging in more directed exploration and less generalization. Participants who receive niacin will perform similarly to adults in previous work, exhibiting less directed exploration and more generalization. Thus, compared to the niacin condition, participants in the psilocybin condition should earn lower rewards overall, sample more unique options and have increased distance between their sampled options, have higher average values for the best fitting exploration parameter, and lower average values for the best fitting generalization parameter.

Method.

Materials. Participants were presented with a series of 8x8 2D grids (64 tiles total), with one randomly selected tile's reward value shown at the start. One of 40 underlying reward environments, defining a bivariate function on the grid which mapped tiles to expected reward values, was randomly selected without replacement for each round. The reward values of the tiles in all environments were spatially correlated: tiles closer together had more similar reward values (Appendix E). Each round had a maximum reward value randomly sampled from a uniform distribution $U(30, 40)$, and all tiles on the grid for that round had their values scaled to this maximum to obfuscate the value of the round-specific global optima. To prevent reward

values less than 0, each rescaled value was shifted by +5. Further, the tiles on the 8x8 grid also incorporated normally distributed noise $\varepsilon \sim N(0, 1)$.

Procedure. Participants completed 10 rounds total, each with a unique randomly generated 8x8 grid, and were instructed to try to gain as many points as possible. Points were earned by clicking on the tiles, where the tile's reward value would be added to the running total of points accumulated in a given round. Participants had the opportunity to make 25 selections per round and were free to choose between observed tiles (re-selecting a previously revealed tile with known reward value) or unobserved tiles (those which had not been revealed and whose reward value remains unknown). Upon clicking on a previously unobserved tile, that tile's reward value would be revealed and displayed numerically within the tile in addition to a color corresponding to that value (darker colors for higher values). Because of the incorporated noise, tiles that were re-clicked could show some variation in reward value. Only the most recent value was numerically displayed.

The first round served as a tutorial, where participants were introduced to the grids and provided with instructions about how the task worked. Three comprehension questions were presented upon completion which needed to be answered correctly to proceed. Specifically, these questions asked about the goal of the task, how points are earned, and how the point values were distributed. Following the tutorial, participants were presented with eight test rounds. The tenth and final round was a special bonus round. Here, participants first made 15 selections as in previous rounds before the round paused. At this point, participants had to make predictions for the reward value of 5 randomly selected unobserved tiles and certainty ratings for those predictions (rated from 0 to 10). After making the reward predictions and certainty ratings, they were prompted to select one of those 5 tiles to be revealed, and then proceeded to make the rest

of their selections for that round as normal (that is, they were not required to select the other 4 tiles they made predictions about). Full instruction and question texts are listed in Appendix E.

Computational Models.

Model Definitions. Here I provide only a brief overview of the models implemented in this work (For full details, see: Giron et al., 2023; Schulz, et al., 2019; Wu et al., 2018). To assess the relative contributions of each component of exploration, I utilized models previously outlined by Giron et al. (2023) and Schulz et al. (2019). These models were constructed by combining a model of learning, to account for generalization, with a sampling strategy, to account for both random and directed exploration. Together, these models are able to use the history of previous observations to make predictions about participants' subsequent search behaviors. A Gaussian Process (GP) regression is used as the learning model to characterize generalization, or the extent to which the presumed reward value of a novel choice is influenced by previously observed reward values. The level of reward similarity between options exponentially changes with their distance on the grid, such that the closer two options are, the more similar their reward will be. In other words, the reward similarity between two tiles is correlated with their distance. The degree of spatial generalization among tiles is defined by the λ parameter. Specifically, λ determines the degree to which the correlation between distance and reward decays, with slower decay (stronger correlations and more generalization) for larger λ values. In the extreme case of $\lambda = 0$, each reward value is thought to be completely independent – there is no generalization from one observation to another.

To map the beliefs of the GP onto specific reward estimations for each tile, an Upper Confidence Bound (UCB) sampling function is used. UCB has been routinely used in studies of explore-exploit dilemmas (Srinivas et al., 2012) by explicitly incorporating uncertainty in its

reward calculations. The reward value of a given option is defined by the weighted sum of its uncertainty and expected reward. The degree to which resolving uncertainty is valued is determined by the β parameter. β thus corresponds to directed exploration, because when the β value is high, options with greater uncertainty are preferentially sampled.

Finally, while this GP-UCB model can provide value estimates of the different options, for the model to make specific predictions it must convert these values into a probability distribution over the options. A softmax decision policy is implemented here to transform values into choice probabilities, assigning higher probabilities to higher-valued options but still allowing for some probability of selecting lower-valued options. The τ parameter determines the amount of random exploration. Specifically, τ controls amount of stochasticity in sampling behavior. As τ increases, the decision policy becomes increasingly more random.

To summarize, the GP-UCB model combines a GP learning model with a UCB sampling function to model participant choice behaviors. The λ parameter characterizes the extent of generalization between options, where higher values correspond to greater assumed similarity between option distance and reward value. The β parameter characterizes directed exploration, where higher values correspond to greater value of resolving uncertainty. The τ parameter characterizes random exploration. Higher τ values result in higher temperature search, where sampling will be more random. Lower τ values result in lower temperature search, where higher value options are preferentially sampled.

Lesioned Models. In order to assess whether all three components are needed to adequately characterize participant behavior, the performance of the GP-UCB model was compared to several competing models which each lesioned a different parameter. The τ-lesioned model uses an epsilon-greedy policy in place of the softmax policy. With the epsilon-

greedy policy, the probability of sampling a random option is p(ε) and the probability of selecting the option with the highest expected value is p(1 - ε). Thus, the softmax policy allows for graded exploration by providing a smooth probability distribution over all options based on their value, while the epsilon-greedy policy is binary (either the highest value option is selected or a random option is selected). The β-lesioned model sets the value of β to 0, preventing directed exploration by entirely removing the value of resolving uncertainty. The λ-lesioned model replaces the GP learning model with a Bayesian reinforcement learning model. This alternative, the Bayesian Mean Tracker (BMT), models the options' reward values as independent distributions. Generalization is inherently removed as each option is considered entirely independent from the others.

Model Parameters. Model parameter estimates were computed via a leave-one-out cross-validation. Training sets were created for each participant by omitting one of the 8 test rounds and a Maximum Likelihood Estimate (MLE) was computed for each training set. The MLE quantifies the parameter values that maximize the likelihood of the observed data (a participant's choice behavior). The model then uses the MLE-derived best-fitting parameters to make out-of-sample predictions for the left-out round. This process of computing MLE and generating predictions on the left-out round was conducted on all combinations of test and training sets, such that each round serves as the test set once. The parameter estimates were constrained to positive values only, and one set was generated per round per participant.

Results.

Behavioral Results. All behavioral results were analyzed using independent sample t-tests (unless otherwise specified). We first looked at overall participant performance, defined by the average reward earned by each condition (Figure 3.3A). Participants in the niacin condition

(M = 43.0) and psilocybin condition (M = 42.3) earned similar rewards on average, $t(7) = 0.75$, $p = 0.48$, $d = 0.50$, $BF = 0.60$. The average maximum reward was also analyzed as a measure of exploration efficiency (i.e., improved exploration outcomes in terms of increased maximum reward). Kendall rank correlation tests revealed no significant differences between the niacin and psilocybin conditions, $r_\tau = 0.00$, $p > 0.05$, $BF = 0.25$. We also considered participants' averaged reward over trials (learning curves), to assess whether they learned more quickly in one condition vs the other (Figure 3.3B). For each participant, the relationship between trials and earned reward was quantified by Spearman's rho. There were no differences in learning rate over trials between conditions, $t(7) = -0.65$, $p = 0.54$, suggesting that participants in both conditions learned at similar rates.

We next looked at participants' sampling behaviors. When considering the distance between consecutive choices, participants in the niacin condition sampled an average of 1.62 options apart and those in the psilocybin condition sampled an average of 1.55 options apart. This difference was not significant, $t(7) = 0.18$, $p = 0.86$, $d = 0.12$, $BF = 0.51$. Choices were also classified into three distance types: Far (distance > 1), Near (distance = 1), and Repeat (distance = 0; the same choice as the previous trial was repeated). To test whether there were differences in the probability of choice distance selection between conditions, we used a two-way ANOVA (Figure 3.3C). A linear mixed-effects model including the random effect of participant resulted in a singular model fit. However, the inclusion of this term did not result in model improvement ($p > 0.05$). The main effect of condition was omitted because it cannot be sensibly interpreted, thus the full model included the main effect of choice type and its interaction with condition. This revealed no significant interaction of condition and choice type, $F(3, 20) = 0.10$, $p = 0.96$, but the main effect of choice type was significant, $F(2, 20) = 11.52$, $p < 0.001$, $BF > 100$.

Post-hoc pairwise comparisons using the Tukey method for multiple comparisons indicated participants were more likely to select a Repeat option (M = 0.52) than a Near option (M = 0.22), $t(20) = 4.45$, $p < 0.001$, $d = 2.27$, BF = 77.18. They were also more likely to select a Repeat option than a Far option (M = 0.29), $t(20) = 3.59$, $p = 0.005$, $d = 1.73$, BF = 17.12. However, there was no difference in the likelihood of selecting a Near vs Far option, $t(20) = 0.96$, $p = 0.61$. This suggests that participants in both conditions tended to exploit previously selected high reward options.

Considering the number of unique options sampled, another indicator of exploration, participants in the niacin condition sampled an average of 10.41 unique options and those in the psilocybin condition sampled an average of 10.78. This difference was not significant, $t(7) = 0.17$, $p = 0.87$, suggesting that by this measure there were no differences in exploration between conditions.

Finally, a Bayesian hierarchical regression was used to examine how participants' choice distances were affected by the reward value of previous choices (Figure 3.3D). In this model, previous reward value, condition, and their interaction, were used to predict search distance, with participant included as a random effect. Significance of the main effects and interaction were assessed with nested model comparisons. Results revealed a significant main effect of previous reward on distance, $\beta = -8.13$, 95% CI [-11.02, -5.26], BF > 100. This suggests that a higher previous reward resulted in a decrease in distance to the subsequent choice. The main effect of condition was not significant, $\beta = -0.15$, 95% CI [-3.77, 3.57]. There was no difference in search distance between the niacin and psilocybin conditions. Additionally, the interaction between previous reward and condition was not significant, $\beta = 0.03$, 95% CI [-3.71, 3.77]. The effect of previous reward on search distance did not differ between conditions. However, the model that

included the interaction of condition and previous reward was favored over a lesioned model without the interaction (BF = 7.96). This discrepancy is likely a result of the small sample size, as Bayesian analyses explicitly account for uncertainty in the parameter estimates.

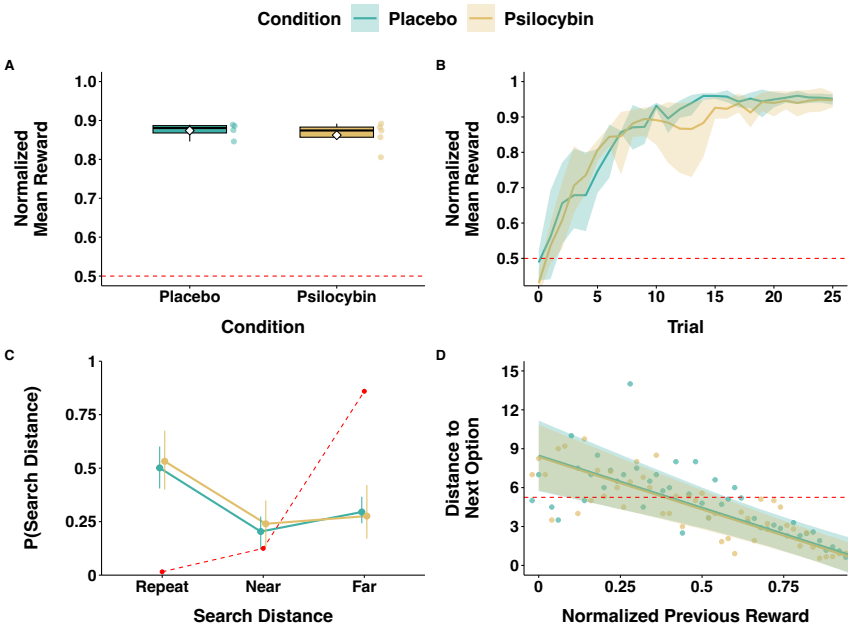


Figure 3.3 Main Behavioral Results. Red dashed line in all graphs represents the expected results from a fully random model. A: Normalized mean reward between conditions. Box plots represent the median and IQR (interquartile range), white diamonds represent group means, and each point is an individual participant's score. B: Learning curves displaying the normalized mean reward across trials for each condition. Lines represent group means and the shaded ribbon represents the 95% CI. C: Proportion of choices based on distance for each condition. Points represent group means with 95% CIs. D: Distance between sequential choices as a function of the reward value of the previous option. Points represent means of the raw data, lines correspond to fixed effects from a hierarchical Bayesian regression with 95% CI shaded regions.

Bonus Round. For predictions made in the bonus round, we first looked at participants' prediction error, defined by the mean absolute difference between actual reward values and participants' predictions (Figure 3.4A). Participants in both the niacin (M = 5.0) and psilocybin (M = 5.3) conditions had similar and relatively low error in their predictions, t(7) = -0.10, p = 0.92. Next, we considered how certain participants were about their reward value predictions (Figure 3.4B). Participants in the niacin condition (M = 5.45) were more confident in their predictions than those in the psilocybin condition (M = 4.88), but this difference was not significant, t(7) = 0.29, p = 0.78. Finally, we investigated how a participant's choice among the

five unknown options was influenced by their reward value predictions and certainty judgements (Figure 3.4C). To do so, judgements of reward value and certainty for chosen options were standardized within-subjects: the chosen option's predicted reward and certainty were divided by them sum of all a participant's predictions and certainty judgements. One participant rated all of their certainty judgements as 0, which resulted in an error in computing their standardized certainty rating (from dividing by 0). Zero (0) was thus used as their standardized certainty rating. Participants in the niacin (M = 0.27) and psilocybin (M = 0.25) conditions did not differ in terms of their choices' predicted values, $t(7) = 0.57$, $p = 0.58$. Participants in the niacin condition (M = 0.22) and psilocybin condition (M = 0.19) chose options for which they were similarly certain of the reward value, $t(7) = 0.47$, $p = 0.65$.

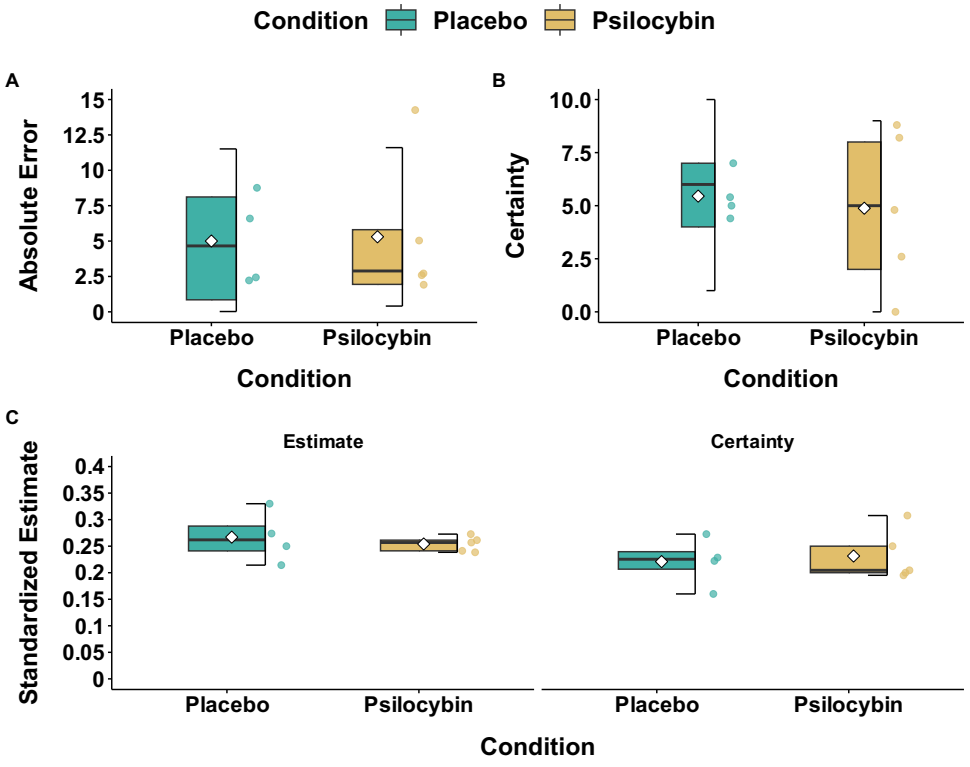


Figure 3.4 Bonus Round Results. In all graphs, box plots represent the median and IQR, white diamonds represent group means, and points represent individual participant responses. A: Absolute error of predictions about the reward value of unobserved choices. B: Participant ratings of certainty about their predicted reward values. C: Standardized reward value predictions (left) and certainty judgements (right) about chosen unobserved options.

Modeling Results.

Model Comparisons. Comparative model fit was assessed in two ways. First, we simulated learning curves for each model by using the median participant parameter estimates. Specifically, each model produced outputs given the same environments that participants saw. Both the full GP-UCB model and β lesioned models produced responses similar to participants', with the full model being slightly more similar (Figure 3.5A). Next, we computed the protected exceedance probability (PXP) using hierarchical Bayesian model selection. In general, PXP measures the probability that a particular model or hypothesis is best among a set of alternatives, accounting for the possibility of differences being due to chance. Here, a higher PXP indicates greater certainty that a given model is superior to the alternatives, and thus characterizes which model is most likely in the population. This analysis revealed that the full GP-UCB model was best across all participants (PXP = 0.92) and both the niacin (PXP = 0.63) and psilocybin (PXP = 0.75) conditions individually (Figure 3.5B).

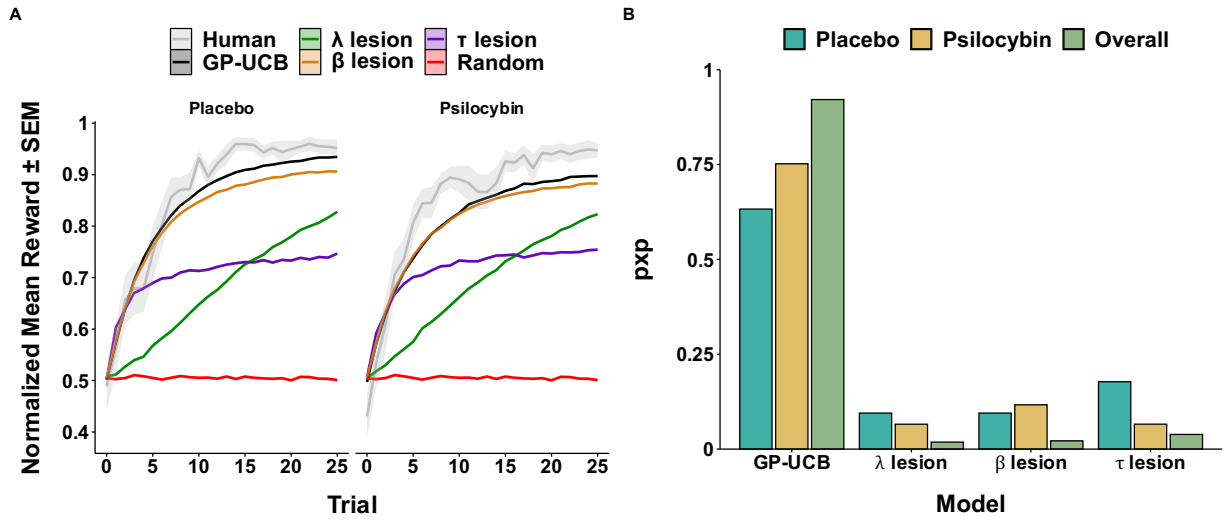


Figure 3.5 Model Fit Results. A: Simulated learning curves for each model compared to participants' responses in both conditions. B: Exceedance probability of different models for each condition and the data overall, reflecting the probability of that model being most common in the population.

Parameter Estimates. Mean parameter estimates of the GP-UCB model for each participant were compared to assess whether participants in each condition exhibited differences in the three components of exploration behavior outlined above. Given the sample size of the current work, Mann-Whitney-U tests were conducted to test for differences between conditions and Kendall's r_τ is used as a measure of effect size. These analyses (Figure 3.6) revealed that participants in the niacin condition (M = 1.13) had a larger λ parameter estimate than those in the psilocybin condition (M = 0.83), indicating that they tended to generalize more. However, this difference was not significant, W = 14, p = 0.41. Participants in the niacin condition (M = 0.18) had a smaller β parameter estimate than those in the psilocybin condition (M = 0.22), indicating that they tended to engage in less directed exploration. Again, this difference was not significant, W = 8, p = 0.73. Finally, participants in the niacin (M = 0.04) and psilocybin (M = 0.03) conditions had similar τ parameter estimates, engaging in similar amounts of random exploration, W = 15, p = 0.29.

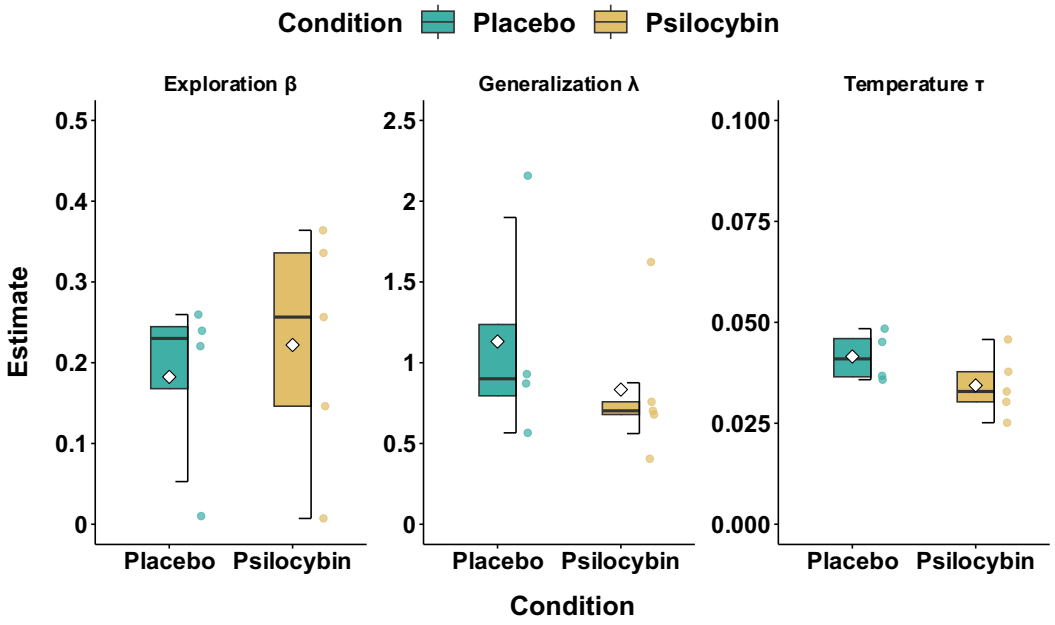


Figure 3.6 Cross-validated parameter estimates for the GP-UCB model. Each point represents an individual participant's median parameter estimate. White diamonds represent condition means, and box plots represent condition medians and IQRs.

Discussion. In this experiment a spatially correlated multi-armed bandit task was used to investigate whether psilocybin treatment will lead to more childlike search and sampling strategies. Specifically, three components of exploration were assessed: random exploration, directed exploration, and generalization. Previous work has shown that, compared to adults, children engage in more directed exploration, sampling more broadly and generalizing less, leading to lower overall rewards earned. However, both adults and children engage in similarly low levels of random exploration. It was hypothesized that participants treated with psilocybin will exhibit similar differences from baseline adults. Participants in both the psilocybin and niacin conditions earned similar amounts of reward overall, exploited known high reward options by sampling locally, generalized from previous options such that higher reward values lead to less distance in subsequent samples, and learned at similar rates across trials. In the bonus round, participants in both groups made similarly accurate reward predictions and moderate certainty judgements. They also tended to pick options with similar reward predictions and certainty judgements. A computational model was also fit with parameters to instantiate each of these three components. Differences in the average parameter estimates between conditions suggest that participants in the niacin condition tend to generalize more and engage in less directed exploration compared to those in the psilocybin condition, though these differences were not significant.

Overall, these results largely suggest that there are no differences in search and sampling strategies between conditions. However, given the sample size limitations, the results from formal statistical analyses must be interpreted with caution. In considering the trends in the raw data and parameter estimates from the computational models, there are some indications that participants treated with psilocybin may be behaving more like children. Beyond the parameter

estimates outlined above, the learning curves show that participants in the psilocybin condition start to earn less reward between trials 10 to 15. One interpretation of this is that after they perform an initial search to locate the high reward value area of the grid, they engage in a brief additional period of exploration before beginning to exploit the discovered high reward options. This would represent a different pattern of directed exploration compared to children, whose learning curves decay across trials, and should be investigated further in future work.

Experiment 2.4: Approach-Avoid Decision Making Task

While much research has provided evidence for the developmental changes in how people resolve the exploration-exploitation dilemma, this has been conducted almost exclusively using multiarmed bandit tasks (like that in Experiment 2.3 above). Critically, in this context, even "bad" options are still rewarding – just less so than other options. Thus, there is no actual cost to exploration, just the opportunity cost of less overall reward. Real-world contexts are more complex, where individuals are tasked not only with differing rewards but also real costs which may be non-trivial (e.g., getting food poisoning from eating spoiled food). Results from these multiarmed bandit tasks are less ecological valid, and the developmental differences in search behaviors in environments with costly options are less clear. However, there exists an alternative paradigm that more closely resembles the complex situations found in the real world where options have varying reward structures, some of which are costly.

In "approach-avoid" decision making tasks, participants are shown a stimulus on each trial that is associated with an initially unknown reward or cost and must decide whether to approach or avoid it. By approaching the stimulus, the participant will receive any reward or incur any cost associated with that stimulus. By avoiding the stimulus, the participant will avoid incurring any costs associated with it, but also forgo any rewards. Thus, the only way

participants can receive any information and learn about the underlying reward structure is by approaching the stimuli and risking any incurred costs of doing so. Previous research has demonstrated that adults tend to overgeneralize from negative outcomes, whereby a single negative experience results in them inferring a general rule about other associated stimuli (Rich & Gureckis, 2018). As a result, they fail to approach other related stimuli, get disconfirming evidence for their initially inferred generalization, and learn the true complex reward structure of their environment, a phenomenon known as a "learning trap" (Rich & Gureckis, 2015).

Approach-avoid decision making task paradigms represent a more ecologically valid way of examining the developmental differences in explore/exploit behaviors. Liquin and Gopnik (2022) adopted a child-friendly version of Rich and Gureckis' (2018) learning trap experiment. They showed that: 1. Even in a child-friendly task, adults still fell into learning traps. 2. Compared to adults, children are both more exploratory and resistant to learning traps. 3. Children and adults do not make different initial inferences, and both expect uncertain actions will be costly if similar to previously costly actions. However, despite predicting an incurred cost of doing so, children are more likely than adults to risk exploring an uncertain action and thus acquire more relevant information and overall learn more. Together, their findings suggest that children are more exploratory than adults in an unknown but ecologically valid environment, this greater exploration allows them to avoid learning traps and learn more effectively than adults, and importantly this difference in exploration is not the result of differences in prior beliefs.

To complement and extend the findings from the spatially correlated multi-armed bandit task above, I adapted this approach-avoid decision making task from Liquin and Gopnik (2022) and tested whether adults treated with psilocybin will likewise be more exploratory and avoid learning traps compared to baseline adults. Specifically, it was hypothesized that participants in

the psilocybin condition would approach costly options more often than those in the niacin condition. This increased exploration should lead them to avoiding learning traps and learning the rule governing the reward structure. However, they should not have different initial expectations than those in the niacin condition.

Method.

Materials. This experiment was conducted electronically using the Qualtrics survey platform. However, several physical stimuli were constructed and implemented in the creation of the images and videos used in the task (Appendix F). A "zaff machine" was constructed using an Apple Macbook Pro laptop with a silver painted shoebox lid on top covering the keyboard. This computer was controlled with a hidden remote to activate the machine with the appropriate response based on which object was placed on top of it. The blocks placed on top of the machine were sixteen yellow painted wooden blocks. These blocks varied along two dimensions: design color (black vs white) and design pattern (striped vs spotted), for a total of 4 different block types. Block types were ascribed to the "zaff" or "non-zaff" category based on a two-dimensional category rule. For example, blocks with a black spotted design are non-zaffs and all other blocks are zaffs. Thus, of the four block types, one was always a non-zaff and the other three were zaffs. The specific rule, i.e., which block type was the non-zaff, was counter-balanced across participants. Four additional blocks were constructed using the same block types outlined above, but were designed on blue blocks rather than yellow.

Procedure. Participants first watched an introductory video where they were introduced to a set of stimuli and a machine. The stimuli set consisted of 4 different types of yellow wooden blocks which varied along two dimensions: design pattern (striped vs spotted) and design color (white vs black). There were 4 of each type of block, for 16 stimuli total, which were all laid out

on a table. On the table was also a "special machine" (laptop with a shoebox covering the keyboard). The experimenter in the video explained that some of the blocks were "zaffs" and some of the blocks were not zaffs, and that the machine was activated only by zaffs. Specifically, when a zaff is put on the machine it would light up with a green smiley face. When a non-zaff is put on the machine, it would light up with a red frowning face. Participants were then told that they would be shown each block one at a time and decide whether or not to put that block on the machine. Participants would have 4 stars when the game begins. Whenever they chose to put a block that was a zaff on the machine they would gain a one star reward. Whenever they chose to put a block that was a non-zaff on the machine, they would incur a two star loss. If they chose not to put a block on the machine, their score would remain unchanged, neither gaining nor losing any stars. Upon the completion of the instructions video, 3 attention check questions were presented which asked what would happen in each of the 3 options participants had per trial (putting a block on the machine when it was a zaff, when it was not a zaff, and not putting it on the machine). All 3 questions needed to be correctly answered before proceeding. If a participant answered any incorrectly, they would be shown the instructions video again.

Following the instructions video, the participants were presented with the approach-avoid phase. There were 16 trials total, one for each of the blocks, and in each trial participants decided whether to approach the block (put it on the machine and risk either the reward or cost associated with it) or avoid the block (put it away and eschew any potential risk or reward). If the participant decided to approach the block, a brief video of the experimenter placing the block on the machine and the resulting machine action (contingent on whether the block was a zaff or non-zaff) would be shown. The experimenter would also narrate the outcome, stating whether or not the block was a zaff and the resulting reward/cost from having approached it. If the

participant decided to avoid the block, the text, “Okay, we’ll put that block away” was displayed on the screen. The 16 trials were organized into 4 trial sets, and within each set one of each block type was shown in a randomized order. In the first trial set, the first trial always contained the zaff that deviated from the non-zaff along both dimensions. For example, if the non-zaff was the block with white stripes, the first trial would contain the black spotted block. The second trial in the first trial set would always contain the non-zaff. The third and fourth trials always contained a zaff that matched the first block shown along only one dimension. This ensured that participants would have the opportunity to see a positive and negative example in the first two trials. This also encouraged them to explore the non-zaff early, be exposed to a negative outcome that they could generalize from, and put them in a position to fall into a learning trap. Additionally, in the first trial set, after making their approach/avoid decision for each block but before being shown the outcome, participants were asked to guess whether or not the block was a zaff. This was included to assess whether participants in each condition made different initial inferences for which objects were zaffs. In other words, to determine whether they may be more exploratory because they do not anticipate a cost for approaching, or if they still approach an object despite inferring a cost for doing so. To summarize, in the first trial set, participants A. saw one of the four block types, B. made an approach/avoid decision, C. made an inference for whether or not that block was a zaff, and D. were shown the outcome of that block if they chose to approach it. For following three trial sets, the four block types were presented in a random order and participants only made approach/avoid decisions. Throughout the approach-avoid phase, participants saw their cumulative total score and number of remaining trials.

After the approach-avoid phase, there were two final phases. First, in the test phase, participants were shown pictures of each type of block one-by-one in a randomized order and

were asked to decide whether each was a zaff or non-zaff. The test phase was used to assess participant learning. Following the test phase, in the generalization phase, participants were shown pictures of novel objects. While the design patterns and colors were identical, the blocks themselves were blue rather than yellow. The pictures were shown one-by-one in a randomized order, and for each the participants were asked whether to indicate whether they thought it was a zaff or non-zaff. All video scripts and question text can be found in Appendix F.

Results.

Exploration. To start, we tested for differences in exploration behaviors between conditions (psilocybin vs niacin), following the analysis plan in the original work (Liquin & Gopnik, 2022). As participants are first ignorant to which objects were zaffs, they should begin by approaching the first objects. After incurring a costly outcome, they should fall into a learning trap, whereby they generalize to a one-dimensional rule based on either the color or pattern match to the costly object (i.e., seeing a white striped non-zaff and inferring that zaffs are black or spotted). If a participant fell into a learning trap, they should be increasingly less likely throughout the task to approach non-zaffs (p(approach | non-zaff) across trial sets). Further, they should persist in avoiding several blocks which were in fact zaffs (i.e., avoiding a white spotted block or black striped block after seeing a white striped non-zaff), never approaching them throughout the entirety of the task (p(approach | zaff) across trial sets).

To test for condition differences in approach behaviors of non-zaff trials across trial sets (Figure 3.7), we fit a mixed-effects logistic regression. Condition (psilocybin vs niacin), trial set (1, 2, 3, and 4), and their interaction were included as fixed effects. Condition was dummy coded and all fixed effects were mean centered. Random intercepts for participant and random slopes for trial set were also included in the model. The resulting full model was as follows: approach ~

condition * Trial_Set + (1+ Trial_Set | participant_id). Significance of the main effects and interaction were assessed with nested model comparison. Results showed that there was no significant interaction, $\chi^2(1) = 0.15$, $p = 0.70$, OR = 0.77, 95% CI [0.21, 2.84]. The difference in probability of approaching non-zaffs across trial sets did not differ between conditions. There was also no main effect of trial set, $\chi^2(1) = 0.008$, $p = 0.93$, OR = 0.94, 95% CI [0.21, 4.08]. Unlike in previous work, across all participants there was not a decrease in the probability of approaching non-zaffs across trial sets. Finally, there was no main effect of condition, $\chi^2(1) = 0.14$, $p = 0.71$, OR = 2.04, 95% CI [0.05, 92.9]. The overall probability of approaching non-zaffs did not differ between conditions.

To test for differences in approach behaviors of zaff trials across trial sets (Figure 3.7), we first calculated the proportion of zaffs approached within trial sets for each participant. We then fit a linear mixed-effects model. Condition (psilocybin vs niacin), trial set (1, 2, 3, and 4), and their interaction were included as fixed effects. Condition was dummy coded and all fixed effects were mean centered. Random intercepts for participant and random slopes for trial set were also included in the model. The resulting full model was as follows: approach ~ condition * Trial_Set + (1+ Trial_Set | participant_id). Significance of the main effects and interaction were assessed with nested model comparisons. Results showed that there was no significant interaction of condition and trial set, $\chi^2(1) = 0.07$, $p = 0.79$. The probability of approaching zaffs across trial sets did not differ between conditions. Consistent with previous work, there was no significant main effect of trial set, $\chi^2(1) = 0.07$, $p = 0.79$. There was no change in the probability of approach zaffs across trial sets. The identical test statistic values and p-values between the main effect of trial set and its interaction with condition are likely due to the fact that the slope for trial set in the niacin condition was zero, but non-zero in the psilocybin condition. Any effect

of trial set only occurred in the psilocybin condition. Thus, the effect of trial set is identical to its interaction with condition. Finally, there was no significant main effect of condition, $\chi^2(1) = 0.06$, $p = 0.80$. The overall probability of approaching zaffs did not differ between conditions.

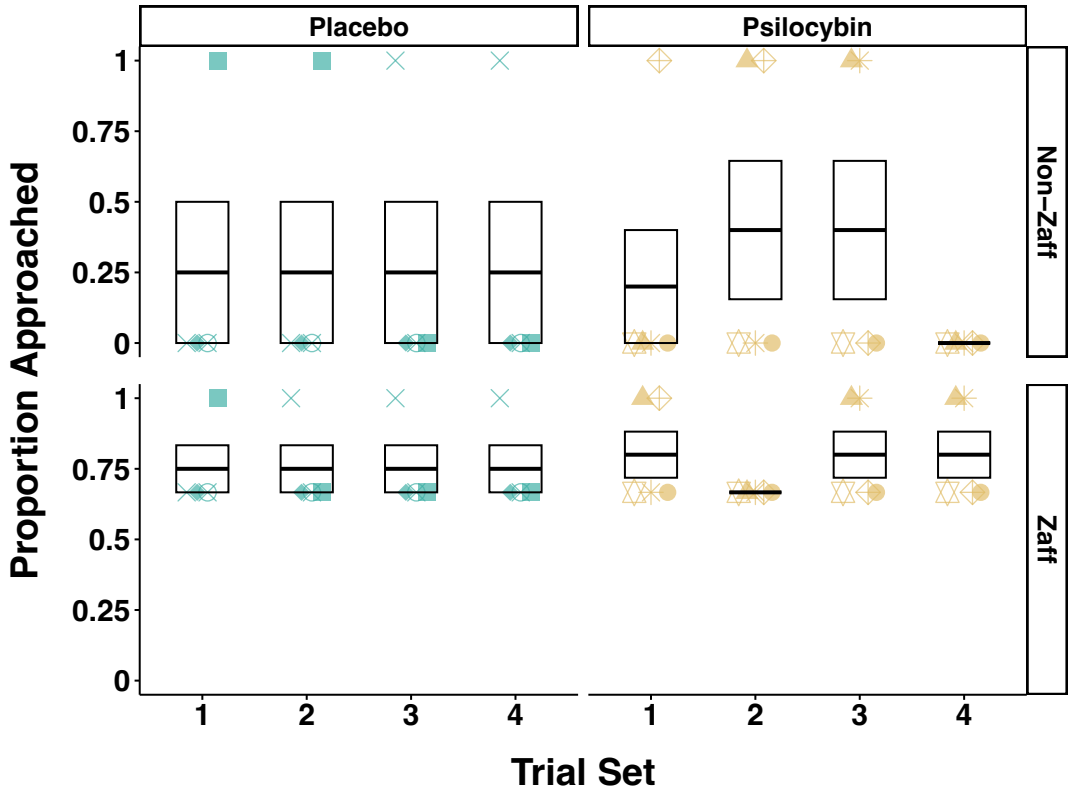


Figure 3.7 Participants' approach-avoid decisions across each of the 4 trial sets. Each trial set contained one zaff and three non-zaffs. Points correspond to individual participants' responses. Box plots represent means and bootstrapped 95% CIs.

Learning. Participants' responses to each object type were used to categorize their learning at test and generalization. Specifically, these responses were coded according to one of four categories: 1. A one-dimension color rule, where zaffs were defined by their color (either white or black), 2. A one-dimensional pattern rule, where zaffs were defined by their pattern (either striped or spotted), 3. A two-dimensional rule, where zaffs were a single object type and all other types were non-zaffs, 4. All other response patterns that do not fit in one of the three previous categories. We compared the difference in proportion of participant responses that

followed each rule category between conditions at both test and generalization (Table 3.2). Interestingly and unlike previous work, their responses at generalization did not follow a similar pattern as at test.

Table 3.2 Number of participants from each condition at Test and Generalization who responded according to each rule.

	One-Dimensional Color Rule		One-Dimensional Pattern Rule		Two-Dimensional Rule		No Discernable Rule	
Condition	Test	Gen.	Test	Gen.	Test	Gen.	Test	Gen.
Placebo	2	0	1	0	0	1	1	3
Psilocybin	1	0	0	1	1	1	3	3

To explicitly investigate participants' susceptibility to learning traps and whether this differed between conditions, we looked at their likelihood of responding according to a one-dimensional rule vs two-dimensional rule. To do so, we fit logistic regression models predicting the learned rule by condition (psilocybin vs niacin) for both generalization and test. In both models, condition was dummy coded and mean centered. Significance of the main was assessed with nested model comparison. At both test, $\chi^2(1) = 2.23$, $p = 0.14$, and generalization, $\chi^2(1) = 1.05$, $p = 0.31$, there were no differences between conditions.

As a final test of learning, we compared the overall reward earned between conditions. If participants had correctly learned the two-dimensional rule, they should approach more zaffs and avoid more non-zaffs, resulting in a higher total earned reward. A one-way analysis of variance compared the average reward earned between the two conditions, and revealed a non-significant main effect of condition, $F(1, 7) = 0.05$, $p = 0.83$, $d = 0.15$, $BF = 0.52$. Participants in the niacin ($M = 11$) and psilocybin ($M = 11.2$) conditions earned similar rewards.

Expectations. In trial one of the first trial set, participants always saw an object that was a zaff. In trial two of the first trial set, participants always saw the non-zaff object. To assess

whether participants in the niacin and psilocybin conditions formed different initial expectations (predictions for whether or not the object was a zaff), we tested for differences in their predictions made on trials three and four of the first trial set (Figure 3.8). At this point, participants had been exposed to both one zaff and the non-zaff, so initial predictions they make on the trials which immediately follow may have influenced their behavior in subsequent trials. We fit a logistic regression to assess the effect of condition (psilocybin vs niacin) on predictions on the third and fourth trial. The mixed-effects logistic regression including a random effect for participant resulted in a singular fit model. However, the inclusion of this random effect did not significantly improve model fit ($p > 0.05$), and the results from both models were identical. Thus, we report the results from the logistic regression. Condition was dummy coded and mean centered, and significance of the main was assessed with nested model comparison. There was no significant main effect of condition, $\chi^2(1) = 0.75$, $p = 0.39$, OR = 1.53, 95% CI [0.59, 4.24]. Participants in both conditions formed similar predictions on these trials.

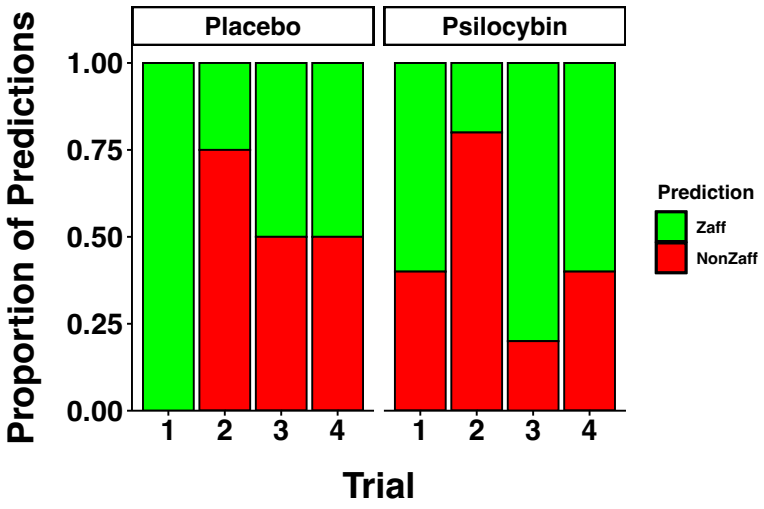


Figure 3.8 Participant predictions on trials three and four of the first trial set. Trials one, three, and four, always contained zaffs, and trial two always contained a non-zaff.

We next looked at whether participants' decisions to approach or avoid a block were influenced by their predictions about that block, and whether this differed between conditions

(Figure 3.9). Mixed-effects logistic regressions resulted in singularity issues with all optimization algorithms attempted. Thus, to account for the repeated measures from each subject, a Bayesian mixed-effect logistic regression was fit with condition (psilocybin vs niacin), prediction (zaff vs non-zaff), and their interaction, as a fixed effects, and a random effect for participant. The full model fit was: approach ~ condition*prediction + (1 | participant_id). As the credible interval included 0, this indicates that there was no significant main effect of condition, $\beta = 4.56$, 95% CI [-33.28, 57.53]. Participants in both conditions always approached objects they expected to be zaffs. However, while participants in the niacin condition (M = 14.3%) approached objects they expected to be non-zaffs less than those in the psilocybin condition (M = 22%), this was not a significant difference.

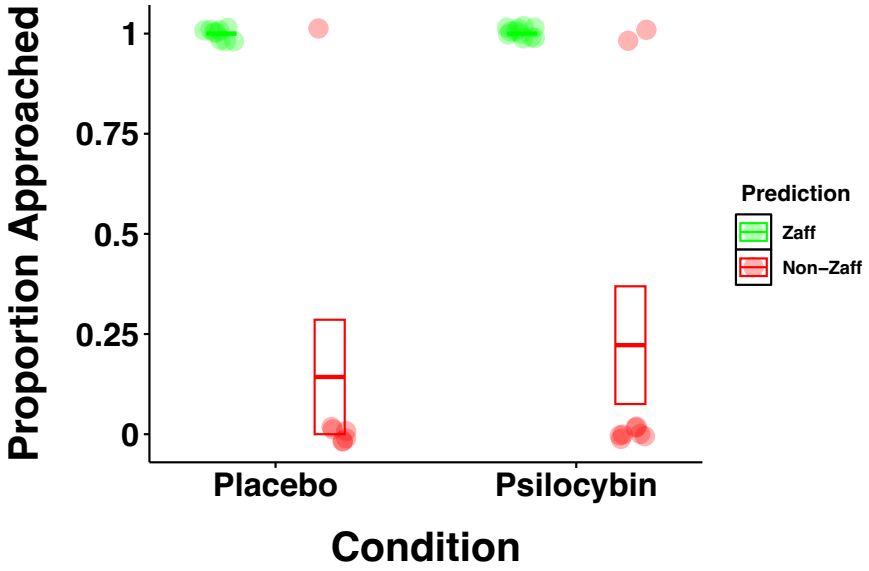


Figure 3.9 Participant approach decisions as a function of their predictions. Individual points represent participants' choices (each represented four times, representing the four trials in which predictions were made), Box plots represent means with boot strapped 95% CIs. Note: all participants approached all objects they predicted were zaffs, thus the box plot is represented by a single horizontal line.

Discussion. In this experiment we investigated whether psilocybin treatment changes exploration behaviors. We used an approach-avoid task where choices can have costs beyond the opportunity cost of forgoing higher reward options. Previous work has shown children are more

exploratory than adults, even when they expect to incur costs for doing so, and this increased exploration leads them to more effective learning. It was hypothesized that participants treated with psilocybin will exhibit similar differences from baseline adults, exploring more, acknowledging costs for doing so, and learning more effectively. However, we found no differences in any measure of exploration or learning between participants treated with niacin or psilocybin. Given that there were no differences between conditions, we would expect that participants as a whole should replicate the findings from adults in prior work (Liquin & Gopnik, 2022), especially as the same stimuli were implemented here. Importantly, this was not the case.

In prior studies, adults were found to only approach 59% of true zaffs and 17.9% of non-zaffs on average, approach non-zaffs less across trial sets, and consistently avoid objects they expected to be zaffs (approaching only 6.3% on average). The majority of adults demonstrating one of the one-dimensional rules at both test (69%) and generalization (62%), and were more likely to learn the correct two-dimensional rule as the number of objects they approached increased. Participants in this study did not exhibit any of these behaviors. They explored much more overall, approaching 75.9% of true zaffs and 25% of non-zaffs on average. This increased amount of exploration and approaching of both zaffs and non-zaffs should indicate that participants would correctly learn the underlying two-dimensional rule, but this was not the case as only one participant at test and two participants at generalization exhibited behavior consistent with this rule. Additionally, they approached non-zaffs at similar rates across trial sets and there was no association between the number of objects approached and exhibiting behavior consistent with the two-dimensional rule, further demonstrating a lack of learning. Finally, participants here approached objects they expected to be non-zaffs 18.3% of the time, more than adults (6.3%) in prior work but less than children (4-5 y/o = 75.7%, 6-7 y/o = 28.5%).

These differences are likely due in part to the limited sample size employed in the pilot data presented here, but may also be due to motivational differences between the samples. In the original work, participants were recruited from Amazon Mechanical Turk (mTurk) and received monetary compensation proportional to their performance. In addition to this, poor performance on an mTurk task could result in a participant getting a bad review. This would lower their "worker rating" and potentially prevent them from participating in other studies, and thus earning income from that participation. Together, this would lead to these participants being highly motivated to perform well, and in this context that corresponds to learning the rules governing reward. Due to IRB and budgetary constraints, in the present work we were unable to implement a similar monetary compensation based on performance. Psilocybin has been shown to elicit increases in gratitude and prosocial behaviors (Griffiths et al., 2018; Noorani et al., 2018), and it was hypothesized that these increases in participants who received psilocybin, or expected to in a compassionate crossover once the study was complete, would provide sufficient motivation to perform well. This may not have been the case and resulted in the behaviors we observed. Future work should incorporate performance-based monetary compensation to provide similar motivation to the original studies. At present, the current work provides no indication that psilocybin changes search and exploration when choices have associated costs.

Experiment 3: Does Psychedelic Treatment Change Exogenous Attention?

Prior work has provided evidence for differences in children and adults' exogenous attention, as outlined in Chapter 1. Adult attention is guided by a top-down process known as selective attention, which narrowly concentrates attentional resources on goal relevant aspects of a task. As a consequence, goal irrelevant aspects of tasks are learned to be ignored, a phenomenon known as learned inattention (Heckler et al., 2006). Learned inattention is thought

to be an indicator of learning, as people optimize the allocation of their attentional resources to what they believe will facilitate learning and success (Yim, 2011). This learned inattention can be costly in the event of goal switching or if one has misinterpreted what are goal-relevant features. Children, on the other hand, have a much more diffuse allocation of attentional resources (Gopnik, 2009), and exhibit less learned inattention. Despite being cued to particular stimuli, children often remember non-cued stimuli equally as well, and better than adults (Deng & Sloutsky, 2016). They also are better at detecting changes to non-cued stimuli (Plebanek & Sloutsky, 2017). This difference in the allocation of attentional resources could be the result of stronger priors or more local perceptual sampling (Gelpi, 2021), whereby this shift between local and global strategies pervades both high-level conceptual inference as well as low-level perceptual processing. Alternatively, it could solely be related to the differences in inhibition and control afforded by prefrontal cortex development (Thompson-Schill et al., 2009).

To formally test whether psilocybin treatment results in changes to one's allocation of attentional resources, I adapted an experimental paradigm used by Plebanek and Sloutsky (2017). In this work, participants were shown a series of overlapping images as target stimuli, one of which was cued, followed by a mask and then a test item. In the test item, either the cued image changed, the un-cued image changed, or neither changed. Participants made familiarity judgements for the cued shapes and change detection judgements for the test images, where they were asked whether or not the test item was identical to the target stimulus. They found an interaction between age and trial type. While adults had greater change detection accuracy for trials where the cued image changed, children had greater change detection accuracy for trials where the un-cued image changed. That children outperformed adults on trials where the un-cued

image changed indicates that they were allocating their attentional resources more broadly than adults.

It was hypothesized that psilocybin will change people's allocation of attentional resources, making them more broadly distributed. Thus, participants in the psilocybin condition should perform like children in previous work and have high change detection accuracy on trials where the un-cued image changes. However, unlike children, they should still have similarly high accuracy on trials where the cued image changes. Participants in the niacin condition should have a narrower allocation of attentional resources and have lower change detection accuracy on trials where the un-cued image changes, performing like adults in previous work. Alternatively, if psilocybin does not change people's allocation of attentional resources, the changed detection accuracy on trials where the un-cued image changes should be similar between the two conditions and in line with results from adults in previous studies.

Method

Materials. The image stimuli utilized throughout the trials were 26 pairs of overlapping shapes, one red and one green (see Appendix G). The mask image was a large black square, and the fixation point was a 60px crosshair ("+" sign).

Procedure. Participants first saw an instructions page which explained the details of the task. They were told that they would see images made up of a green shape and a red shape, and that they should pay very close attention to only the red shape. In this way, the red shape served as the cued shape, and the green shape served as the un-cued shape. They were also told that they would be asked whether or not the red shapes were something they had seen before, and whether the images they saw in each trial changed. After verifying that they read and understood the instructions, they moved to the cuing phase, which was implemented to facilitate their attention

to the cued (red) shape. Here, participants saw five trials which all took a similar form. Each trial began with a fixation point shown in the middle of the screen for 1000ms and then the target shape for an another 1000ms. Following the target shape, a masking image was shown for 500ms before the test shape for a final 1000ms (Appendix G). After all images for a trial had been shown, participants made a two-alternative forced choice familiarity judgement for the cued image (familiar vs new), which served to further direct their attention towards it. They also made a change detection judgement where they were asked whether or not the test item was identical to the target stimulus (yes vs no). As the purpose of the cuing phase was to further direct attention to the cued shapes, in all trials in the cuing phase, the red shape changed from the target to the test stimuli while the green shape remained the same.

Immediately upon completing the cuing phase, participants began the test phase. The test phase contained three different trial types. In Cued Change trials, the cued shape (red) changed from the target to test stimuli. In Un-cued Change trials, the un-cued shape (green) changed from target to test. In the No Change trials, neither shapes change, and the test stimuli was identical to the target (Appendix G). There were five total trials of each trial type, for a total of fifteen trials overall in the test phase. Trials were presented in a fully randomized order.

Results

Change detection accuracy was measured using A', the non-parametric equivalent of the signal detection statistic d'. A' was computed individually for both Cued Change trials and Un-cued Change trials. For both Cued Change and Un-cued Change trials, hits were defined as "change" responses to the change detection question. False alarms were defined as "change" responses on No Change trials. Thus, the false alarm rate was the same for both. The No Change trials had no changes to be detected, and they are thus omitted from the present analyses.

To assess whether change detection accuracy differed across conditions (psilocybin vs niacin) and trial type (cued shape changed vs un-cued shape changed), we fit a linear mixed-effects model. Condition, trial type, and their interaction, were included as fixed effects. Condition and trial type were dummy coded and mean centered. Participant was included in the model as a random effect. The resulting full model was as follows: a_prime ~ condition*trial_type + (1 | participant_id). Significance of the main effects and interaction were assessed using a Type III Analysis of Variance with Satterthwaite's method. Results revealed no significant interaction of condition and trial type, $F(1,7) = 0.23$, $p = 0.65$, and no significant main effect of condition, $F(1, 7) = 0.002$, $p = 0.97$. This indicates that participants in the psilocybin and niacin conditions did not differ in their change detection accuracy. There was, however, a significant main effect of trial type, such that change detection accuracy was higher on trials where the cued shape changed ($A' = 0.94$) than trials where the un-cued shape changed ($A' = 0.70$), $F(1, 7) = 6.73$, $p = 0.036$, $d = 1.23$, $BF = 3.44$ (Figure 3.10). This is consistent with previous work showing that adults have better change detection accuracy for the cued shapes.

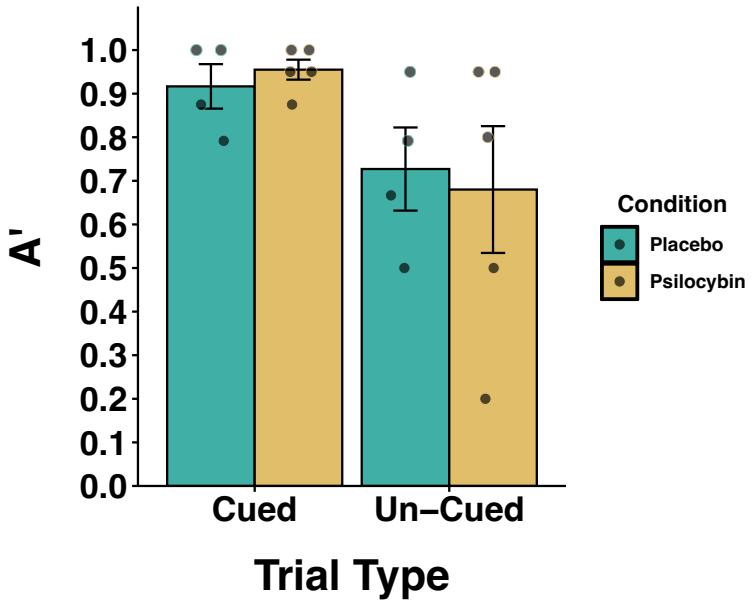


Figure 3.10 Participants' change detection accuracy (A') between conditions and trial types. Bars represent means and error bars represent ±1 Standard Error of the Mean (SEM).

Discussion

In this experiment a change detection task was used to investigate whether psilocybin treatment affects one's allocation of attentional resources. Previous work has shown that, compared to adults, children have higher change detection accuracy for un-cued images. It was hypothesized that participants treated with psilocybin will exhibit similar differences from baseline adults. However, participants in both the psilocybin and niacin conditions had similar change detection accuracy across both cued and un-cued trial types. Interestingly, participants in both conditions had higher change detection accuracy on un-cued trials than adults in previous work. Plebanek and Sloutsky (2017) report an average A' of 0.634 for the adults in their experiment. Here, the psilocybin condition had an average of 0.68 and the niacin condition had an average of 0.73. It is possible that participants here may have been less focused to the cued image. In the original work, participants earned course credit for their participation, which may have led to increased motivation to closely follow the instructions. Future studies can consider offering similar extrinsic motivation to increase fidelity to the instructions.

Given that the perceptual effects of psychedelics occur during the acute drug effects, it is not necessarily surprising that there are no post-hoc attentional shifts. It is therefore also possible that any effects on exogenous attention may occur solely during the acute drug effects (Gopnik, 2018). Further supporting this notion, previous work has demonstrated that impairments in working and episodic memory and visual perception following the administration of psilocybin occur during this time (Barrett et al., 2018). Thus, future studies should consider administering this task while participants are experiencing the acute drug effects. Overall, these results suggest that there are no changes to exogenous attention following the administration of psilocybin.

General Discussion

Across six experiments, the present work sought to unpack the mechanisms underlying the clinical benefits of psilocybin therapy. Following the proposed developmental model in Chapter 1, we investigated whether and how psilocybin affects cognitive flexibility. Three specific features of cognition, symptomatic of being more cognitively flexible, were tested which have previously shown developmental differences. The overall prediction across all experiments was that, after receiving psilocybin treatment, adults will perform more like children on these tasks compared to those who received niacin.

We found initial evidence that psilocybin reduces the strength of people's prior beliefs, making them more sensitive to current evidence. Specifically, participants in the psilocybin condition, like children, were more likely to revise a high prior belief when shown counterevidence. Mixed evidence was found for changes to search and sampling behaviors. In a novel semantic search task with multiple dynamic constraints, participants who received psilocybin were more exploratory, having less similarity between sequentially generated samples. While not statistically significant, model parameter estimates from a spatially correlated multi-armed bandit task indicated that, like children, psilocybin resulted in more exploration and less generalization. However, in a single-constraint serial production task and an approach-avoid task, no differences in exploration and search strategies were observed between conditions. Finally, in a change detection task, psilocybin did not result in increased exogenous attention.

While individual discussion sections highlight limitations specific to the individual task paradigms, several features of this pilot work may have impacted the observed results. Most salient among these is the sample size. While most studies of psilocybin and other psychedelic drugs do not have sample sizes that rival other psychology and drug research, the cognitive tasks

implemented in this work are typically completed with more participants (M = 46.8, SD = 19.5). The results from the formal statistical analyses should thus not be viewed as definitive, but rather a preliminary indication of which avenues may be most promising for future research. Several of the analyses encountered model fitting issues, including singular fits and convergence issues. Though in all cases these were able to be circumvented through the use of different optimization algorithms or Bayesian models, caution should be exercised in interpreting these results until a larger sample from Phase 2 of the broader study is completed.

It is also worth considering the timeframe relative to drug administration that these tasks were completed. Previous work implementing more coarse measures of cognitive flexibility have found different effects depending on when the tasks were administered. For example, some studies have reported increases during the acute effects of ayahuasca (Kuypers et al., 2016) and the day after administering psilocybin, stabilizing one week later (Mason et al., 2019). While others have found decreases during the acute effects (Mason et al., 2021). Others still have found increases that persist outwards of one-month (Doss et al., 2021). Additionally, there are distinct differences in the neural effects of psilocybin, with decreases in within-network DMN connectivity followed by post-acute increases. Psilocybin therapy has thus been likened to a 'reset' mechanism, where acute disintegration affords post-acute re-integration leading to overall improvements in normal functioning (Carhart-Harris et al., 2017). It is thus unclear how the present results may have changed if the tasks were administered during the acute effects of psilocybin, or even one week later. Further, these tasks were only administered a single time. This was in large part due to a lack of previous research to demonstrate whether these tasks were susceptible to practice effects (see Chapter 4). Any potential baseline or demographic differences between participants would be mitigated through effective random condition assignment.

However, collecting data pre- and post-intervention would afford more statistical power, something particularly useful in these contexts.

Finally, there may have been motivational differences between participants in the original studies and those in the present work. In most cases, participants in the original studies were recruited from platforms like mTurk, which provide performance-based monetary compensation for their participation. They also experience pressure to maintain a high user rating so that participants do not get barred from other potentially high paying tasks. While we anticipated that participants here would be more diligent than undergraduate subjects from university research pools often are, there still may have been important differences in extrinsic motivation which may have affected the results.

It is critical to contextualize the findings from the literature on psychedelics overviewed throughout this dissertation thus far in terms of the demographics of their participant samples, and the limitations that may impose on generalizability. Most clinical trials investigating the therapeutic potential of psychedelics are conducted almost exclusively using white, educated, industrialized, rich, and democratic (WEIRD) populations (Michaels et al., 2018), as was the case in the work presented here. However, the therapeutic benefits of psychotropic drugs can be influenced by ethnoracial- and sociocultural-factors (Ninnemann, 2012). For example, culturally-informed interpretations of experiences and epigenetics. There is particular emphasis placed on "set" (the participant's psychology state and motivations) and "setting" (the environment in which the treatment takes place) in psychedelic studies, and it is thought that these components may contribute to the safety and clinical efficacy of psychedelic treatment (Johnson et al., 2008). Survey studies have shown that set and setting are correlated with the outcomes of psychedelic use (Borkel et al., 2023), and most clinical trials implement some variation of the same model

which has been “optimized” to prioritize participant safety and reduce the likelihood of negative or harmful experiences (Johnson et al., 2008). However, both set and setting are particularly influenced by cultural factors (Hartogsohn, 2016). The currently widespread paradigm may not only be *optimized* just for WEIRD populations but could be implementing aspects that reduce safety or increase risk for non-WEIRD participants. Further, the pharmacokinetic and pharmacodynamic properties of psychotropic drugs are influenced by many factors, several of which are ethnoracially- or socioculturally-influenced (Mark von Zastrow, 2018). Thus, when considering the literature on the clinical benefits of psychedelic drug treatment, it is critical to be mindful of the demographics of the participant populations utilized in these studies. For a more comprehensive account, see the review by Fogg and colleagues (2021).

Though there are several legitimate caveats to acknowledge, the present work represents an important step forward for future research by highlighting the most promising targets for investigating changes in cognitive flexibility. If future studies with larger sample sizes are able to replicate the results reported here, or discover differences which were not significant, the increased knowledge about *how* psilocybin therapy imparts its widely observed therapeutic benefits could be used to either predict who may be most responsive to it and/or develop alternative treatments targeting an identical mechanism for individuals who cannot safely or practically receive this treatment.

Acknowledgements

Chapter 3 is currently being prepared for submission for publication of the material. Hurwitz, Ethan; Dean, Jon; Brockbank, Erik; McKinty, Arwynn; Farrell, Briana; Gopnik, Alison; and Walker, Caren. The dissertation author was the primary investigator and author of this material.

Chapter 4 How Flexible is Flexible? Exploring Practice Effects in Cognitive Flexibility Tasks

As highlighted throughout this dissertation, behavioral measures are critical to neuropsychological and cognitive assessment in both clinical and research settings. Rather than relying on an individual's ability to accurately *tell* you about their own characteristics through self-report, behavioral measures require them to *demonstrate* these characteristics. Thus, behavioral measures are thought to provide more objective assessments which are less susceptible to measurement artifacts such as socially desirable responding and demand characteristics (Baumeister et al., 2007; Eisenberg & Fabes, 1990; Schwarz, 1999). However, behavioral measures also have a significant limitation: repeated administrations can result in *practice effects*, changes in scores across task completions in the absence of an intervention (Bartels et al., 2010).

Practice effects result from the development of strategies, memory for test items, and increased comfort with the task; which, critically, can occur independently of change in the measured construct (Calamia et al., 2012). As a result, practice effects seriously undermine the accuracy of conclusions drawn from an individual's performance. Practice may lead to ceiling or floor effects, potentially obscuring the effects of an intervention (Bartels et al., 2010), or result in misattribution of performance increases to an intervention (Calamia et al., 2012). They may also be difficult to identify, as the extent of practice effects has been found to vary both between different domains and even between different tasks within the same domain (Calamia et al., 2012). Perhaps most alarmingly, meta-analyses have shown that practice effects across a wide variety of cognitive domains reach about a quarter of a standard deviation, a common magnitude of effect size for many interventions themselves (Calamia et al., 2012).

Of course, in some cases, practice effects do not pose an issue: if researchers administer their task at a single time point after an intervention, as in the studies outlined in Chapter 3. However, this requires relying on random condition assignment, which is sometimes not possible in clinical contexts. In most cases it is optimal, if not necessary, to collect clinical data within-subjects, either to achieve statistical power when working with populations that are small and difficult to recruit or to account for critical individual differences impacting the measured construct. For example, in neuropsychiatric research and clinical neuropsychology, subtle differences in a patient's condition can lead to different prognoses, and these can be further impacted by other cultural, demographic, or behavioral differences between patients. Thus, both fields have long prioritized accounting for practice effects in their assessments (McCaffrey & Westervelt, 1995).

The current study extends the examination of practice effects to cognitive flexibility tasks. Cognitive flexibility is impaired in many mood disorders (Abend et al., 2018; Everaert et al., 2018; Kraus et al., 2021; Mennen et al., 2019), and neuropsychiatric conditions (Milders et al., 2008; Whiting et al., 2017). Leveraging recent advances in modern computing power, cognitive flexibility tasks applying computational methodology have been able to more quantitatively characterize aspects of human cognition (e.g., Bonawitz et al., 2014; Dasgupta et al., 2017; Lake et al., 2017; Lucas & Griffiths, 2010; Tenenbaum et al., 2011). Applying these tasks to clinical studies, especially those implementing therapeutic interventions, would allow for a more comprehensive appreciation of the condition's underlying mechanisms and the intervention's clinical efficacy. However, no work to date has investigated how performance on these tasks changes with repeated administrations. It is thus unclear whether these tasks are susceptible to practice effects and suitable for such applications. The present work aimed to

address this gap by testing whether three different cognitive flexibility tasks are susceptible to practice effects. Each task assesses a different dimension along which someone can be more cognitively flexible. Specifically, we investigated (1) A change detection task where previous work has shown that greater flexibility corresponds to more diffuse exogenous attention and greater change detection accuracy for un-cued images (Plebanek & Sloutsky, 2017). (2) A spatially correlated multi-armed bandit task where previous work has shown that greater flexibility corresponds to broader search and sampling strategies and less generalization (Giron et al., 2023; Schulz, et al., 2019). (3) An approach-avoid decision making task where previous work has shown that greater flexibility corresponds to better learning and resistance to learning traps (Liquin & Gopnik, 2022). Our results offer the first documentation of the varying degrees of practice effects in these tasks and provide suggestions for the future implementation of measures of cognitive flexibility in the assessment of clinical interventions.

Method

Participants

All participants were recruited from the University of California, San Diego, undergraduate research pool. Participants signed up for one of three different versions of the study, each containing a different counterbalanced order of stimuli sets per time point. Each time point was spaced one-week apart. To account for expected attrition, up to 60 participants from each version were originally recruited with the aim of having 30 completing each. 142 participants originally signed up to participate in this study. 2 participants failed to complete all tasks at the first time point. Of the remaining 140 participants, 42 dropped out after the first time point (failing to complete the second and third), and 2 additional participants failed to complete all the tasks at time point 2. Of the remaining 96, 20 dropped out after the second time point

(failing to complete the third), and an additional participant failed to complete all the tasks at time point 3. Of the 75 individuals who completed all tasks at all time points, 7 individuals were excluded for completing at least one task at a specific time point multiple times, and another 5 were excluded for reporting technical difficulties that impacted their data (i.e., stimuli videos not loading or playing). The final dataset utilized for analysis included 63 individuals: 16 participants in the first order, 20 participants in the second order, and 27 participants in the third order. Participants answered basic demographic questions including age (M = 21.33 years, SD = 2.92, 17.46% declined to answer), gender identity (84.1% female, 4.76% male, 1.59% other gender identity, 9.52% declined to answer), and race (23.8% Hispanic, 22.2% East Asian, 15.9% White, 11.1% Multiracial, 7.94% South Asian, 3.17% Black, 1.59% Native Hawaiian or Pacific Islander, 1.59% other racial identity, 12.7% declined to answer).

Materials and Procedure

Change Detection. In the original task, 26 pairs of overlapping shapes were utilized, one red and one green. Two new sets of stimuli were created to have three total unique sets (the original and two new sets). Each new set of overlapping shape pairs was constructed by creating different combinations of the existing shapes from the original stimuli set. Equivalently different colors were selected for each set of overlapping shapes (purple vs yellow and orange vs blue). See Appendix G for example stimuli. To verify that these new stimuli sets replicated the findings of the original, each were piloted with a unique set of 35 participants and there were no significant differences between any set ($p > 0.05$).

Participants first saw an instructions page which explained the details of the task. They were told that they would see images made up of a green/purple/orange shape and a red/yellow/blue shape (depending on the task version), and that they should pay very close

attention to only the red/yellow/blue shape. In this way, the red/yellow/blue shapes served as the cued shape, and the green/purple/orange shapes served as the un-cued shape. They were also told that they would be asked whether or not the red/yellow/blue shapes were something they had seen before, and whether the images they saw in each trial changed. After verifying that they read and understood the instructions, they moved to the cuing phase, which was implemented to facilitate their attention to the cued shape. Here, participants saw five trials which all took a similar form. Each trial began with a fixation point shown in the middle of the screen for 1000ms and then the target shape for an another 1000ms. Following the target shape, a masking image was shown for 500ms before the test shape for a final 1000ms (Appendix G). After all images for a trial had been shown, participants made a two-alternative forced choice familiarity judgement for the cued image (familiar vs new), which served to further direct their attention towards it. They also made a change detection judgement where they were asked whether or not the test item was identical to the target stimulus (yes vs no). As the purpose of the cuing phase was to further direct attention to the cued shapes, in all trials in the cuing phase, the red shape changed from the target to the test stimuli while the green shape remained the same.

Immediately upon completing the cuing phase, participants began the test phase. The test phase contained three different trial types. In Cued Change trials, the cued shape changed from the target to test stimuli. In Un-cued Change trials, the un-cued shape changed from target to test. In the No Change trials, neither shapes changed, and the test stimuli was identical to the target (Appendix G). There were five total trials of each trial type, for a total of fifteen trials overall in the test phase. Trials were presented in a fully randomized order. Full instructions and question text can be found in Appendix G.

Spatially Correlated Multi-Armed Bandit. Participants were presented with a series of 8x8 2D grids (64 tiles total), with one randomly selected tile's reward value shown at the start (Appendix E). One of 40 underlying reward environments, defining a bivariate function on the grid which mapped tiles to expected reward values, was randomly selected without replacement for each round. However, the reward values of the tiles in all environments were spatially correlated: tiles closer together had more similar reward values. Each round had a maximum reward value randomly sampled from a uniform distribution $U(30, 40)$, and all tiles on the grid for that round had their values scaled to this maximum. This was to obfuscate the value of the round-specific global optima. To prevent reward values less than 0, each rescaled value was shifted by +5. Tiles on the 8x8 grid also incorporated normally distributed noise $\varepsilon \sim N(0, 1)$. Given the variability incorporated when generating grids, the likelihood of repeating the same grid twice was low. As a result, the same materials were used at each time point.

Participants completed 10 rounds total, each with a unique randomly generated 8x8 grid, and were instructed to try to gain as many points as possible. Points were earned by clicking on the tiles, where the tile's reward value would be added to the running total of points accumulated in a given round. Participants had the opportunity to make 25 selections per round and were free to choose between observed tiles (re-selecting a previously revealed tile with known reward value) or unobserved tiles (those which had not been revealed and whose reward value remains unknown). Upon clicking on a previously unobserved tile, that tile's reward value would be revealed and displayed numerically within the tile in addition to a color corresponding to that value (darker colors for higher values). Because of the incorporated noise, tiles that were re-clicked could show some variation in reward value. Only the most recent value was numerically displayed.

The first round served as a tutorial, where participants were introduced to the grids and provided with instructions. Three comprehension questions were presented upon completion which needed to be answered correctly to proceed. Specifically, these questions asked about the goal of the task, how points are earned, and how the point values were distributed. Following the tutorial, participants were presented with eight test rounds. The tenth and final round was a special bonus round. Here, participants first made 15 selections as in previous rounds before the round paused. At this point, participants had to make predictions for the reward value of 5 randomly selected unobserved tiles and certainty ratings for those predictions (rated from 0 to 10). After making the reward predictions and certainty ratings, they were prompted to select one of those 5 tiles to be revealed, and then proceeded to make the rest of their selections for that round as normal (that is, they were not required to select the other 4 tiles they made predictions about). Full instructions and question text can be found in Appendix E.

Approach-Avoid Decision Task. The approach-avoid decision making task was conducted electronically using the Qualtrics survey platform. However, several physical stimuli were constructed and implemented in the creation of the images and videos used in the task. A "zaff machine" was constructed using an Apple Macbook Pro laptop with a silver painted shoebox lid on top covering the keyboard. This computer was controlled with a hidden remote to activate the machine with the appropriate response based on which object was placed on top of it (see below). In the original task, the blocks placed on top of the machine were sixteen yellow painted wooden blocks. These blocks varied along two dimensions: design color (black vs white) and design pattern (striped vs spotted), for a total of 4 different block types. Block types were ascribed to the "zaff" or "non-zaff" category based on a two-dimensional category rule. For example, blocks with a black spotted design are non-zaffs and all other blocks are zaffs. Thus, of

the four block types, one was always a non-zaff and the other three were zaffs. The specific rule, i.e., which block type was the non-zaff, was counter-balanced across participants. Four additional blocks were constructed using the same block types outlined above, but were designed on blue blocks rather than yellow.

While the same zaff machine was used in each stimuli set, two new sets of blocks were created to have three total unique sets (the original and two new sets). Each new set utilized different patterns and colors. In one new set, the blocks were painted teal with yellow vs purple design colors and heart vs checkerboard design patterns. Four additional blocks were constructed using these same block types but were designed on green blocks rather than teal. In the other new set, the blocks were painted navy with orange vs green design colors and triangle vs honeycomb design patterns. Four additional blocks were constructed using these same block types but were designed on maroon blocks rather than navy. See Appendix F for full stimuli sets. To verify that these new stimuli sets replicated the findings of the original, each were piloted with a unique set of 20 participants and there were no significant differences between any set ($p > 0.05$).

Participants first watched an introductory video where they were introduced to a set of stimuli and a machine. The stimuli set consisted of 4 different types of yellow wooden blocks which varied along two dimensions: design pattern (striped vs spotted) and design color (white vs black). There were 4 of each type of block, for 16 stimuli total, which were all laid out on a table. On the table was also a "special machine" (laptop with a shoebox covering the keyboard). The experimenter in the video explained that some of the blocks were "zaffs" and some of the blocks were not zaffs, and that the machine was activated only by zaffs. Specifically, when a zaff is put on the machine it would light up with a green smiley face. When a non-zaff is put on the machine, it would light up with a red frowning face. Participants were then told that they would

be shown each block one at a time and decide whether or not to put that block on the machine. Participants would have 4 stars when the game begins. Whenever they chose to put a block that was a zaff on the machine they would gain a one star reward. Whenever they chose to put a block that was a non-zaff on the machine, they would incur a two star loss. If they chose not to put a block on the machine, their score would remain unchanged, neither gaining nor losing any stars. Upon the completion of the instructions video, 3 attention check questions were presented which asked what would happen in each of the 3 options participants had per trial (putting a block on the machine when it was a zaff, when it was not a zaff, and not putting it on the machine). All 3 questions needed to be correctly answered before proceeding. If a participant answered any incorrectly, they would be shown the instructions video again.

Following the instructions video, the participants were presented with the approach-avoid phase. There were 16 trials total, one for each of the blocks, and in each trial participants decided whether to approach the block (put it on the machine and risk either the reward or cost associated with it) or avoid the block (put it away and eschew any potential risk or reward). If the participant decided to approach the block, a brief video of the experimenter placing the block on the machine and the resulting machine action (contingent on whether the block was a zaff or non-zaff) would be shown. The experimenter would also narrate the outcome, stating whether or not the block was a zaff and the resulting reward/cost from having approached it. If the participant decided to avoid the block, the text, “Okay, we’ll put that block away” was displayed on the screen. The 16 trials were organized into 4 trial sets, and within each set one of each block type was shown in a randomized order. In the first trial set, the first trial always contained the zaff that deviated from the non-zaff along both dimensions. For example, if the non-zaff was the block with white stripes, the first trial would contain the black spotted block. The second trial in

the first trial set would always contain the non-zaff. The third and fourth trials always contained a zaff that matched the first block shown along only one dimension. This ensured that participants would have the opportunity to see a positive and negative example in the first two trials. This also encouraged them to explore the non-zaff early, be exposed to a negative outcome that they could generalize from, and put them in a position to fall into a learning trap. Additionally, in the first trial set, after making their approach/avoid decision for each block but before being shown the outcome, participants were asked to guess whether or not the block was a zaff. This was included to assess whether participants in each condition made different initial inferences for which objects were zaffs. In other words, to determine whether they may be more exploratory because they do not anticipate a cost for approaching, or if they still approach an object despite inferring a cost for doing so. To summarize, in the first trial set, participants A. saw one of the four block types, B. made an approach/avoid decision, C. made an inference for whether or not that block was a zaff, and D. were shown the outcome of that block if they chose to approach it. For following three trial sets, the four block types were presented in a random order and participants only made approach/avoid decisions. Throughout the approach-avoid phase, participants saw their cumulative total score and number of remaining trials.

After the approach-avoid phase, there were two final phases. First, in the test phase, participants were shown pictures of each type of block one-by-one in a randomized order and were asked to decide whether each was a zaff or non-zaff. The test phase was used to assess participant learning. Following the test phase, in the generalization phase, participants were shown pictures of novel objects. While the design patterns and colors were identical, the blocks themselves were blue rather than yellow. The pictures were shown one-by-one in a randomized

order, and for each the participants were asked whether to indicate whether they thought it was a zaff or non-zaff. Full instructions and question text can be found in Appendix F.

Results

Change Detection

Change detection accuracy was measured using A', the non-parametric equivalent of the signal detection statistic d'. A' was computed individually for both Cued Change trials and Un-cued Change trials. For both Cued Change and Un-cued Change trials, hits were defined as "change" responses to the change detection question. False alarms were defined as "change" responses on No Change trials. Thus, the false alarm rate was the same for both. The No Change trials had no changes to be detected, and they are thus omitted from the present analyses.

To assess whether change detection accuracy differed across time points (1, 2, and 3) and conditions (cued shape changed vs un-cued shape changed), we fit a linear mixed-effects model. Time point (1, 2, and 3), trial type (cued shape changed vs un-cued shape changed), and their interaction, were included as fixed effects. Time point and trial type were dummy coded and mean centered. Participant was included in the model as a random effect. The resulting full model was as follows: a_prime ~ time_point*trial_type + (1 | participant_id). Significance of the main effects and interaction were assessed using nested model comparisons. Results revealed no significant interaction of trial type and time point, $\chi^2(1) = 0.44$, $p = 0.51$, or main effect of time point, $\chi^2(1) = 2.09$, $p = 0.15$. Participants' accuracy did not change across time points. As expected, there was a significant main effect of trial type, such that change detection accuracy was higher for trials where the cued shape changed ($A' = 0.89$) than trials where the un-cued shape changed ($A' = 0.62$), $\chi^2(1) = 136.55$, $p < 0.001$.

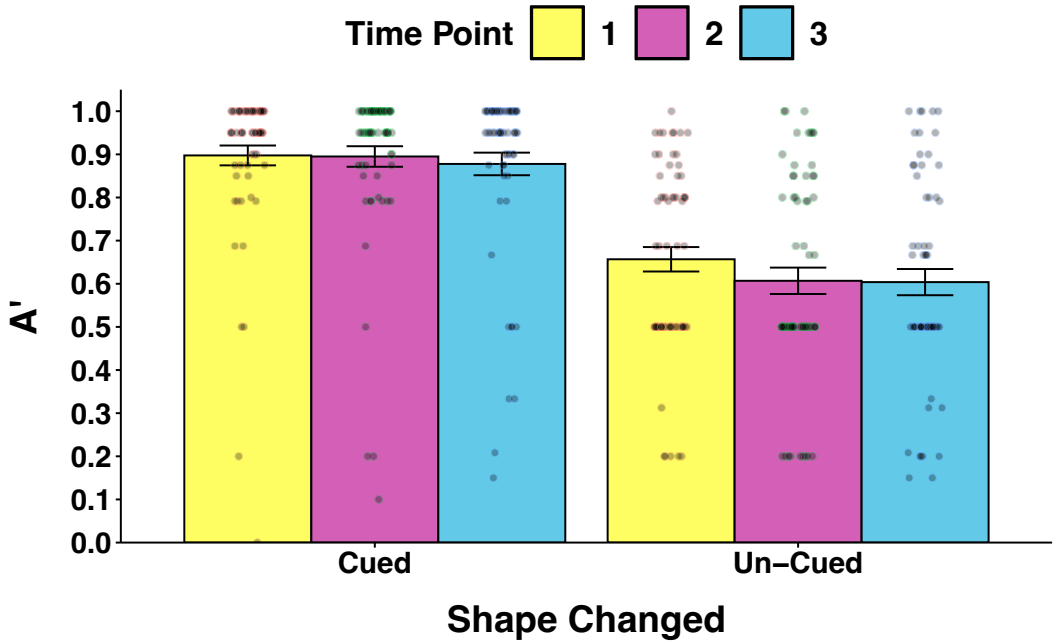


Figure 4.1 Participants' change detection accuracy (A') between time points and trial types. Bars represent means and error bars represent ±1 SEM.

Spatially Correlated Multi-Armed Bandit

Behavioral Results. All behavioral results were analyzed in a similar manner: fitting a linear mixed-effects model including a fixed effect for time point (1, 2, and 3). Time point was dummy coded and mean centered. The models also included random effects for participant and time point (for outcomes that included multiple measurements per subject at each time point) or only participant (for outcomes that included a single measurement per subject at each time point). The significance of the main effects and interactions were assessed with nested model comparisons. Post-hoc pairwise comparisons of the estimated marginal means were performed using the emmeans package with the Tukey method for adjustment. We first looked at overall participant performance, defined as the average reward earned at each time point (Figure 4.2A). There was a significant main effect of time point, $\chi^2(1) = 20.41$, $p < 0.001$. Post-hoc analyses revealed that participants earned 1.63 more points (95% CI=[0.59, 2.67]) at time point 2 than time point 1, $t(62) = 3.78$, $p = 0.001$, $d = 0.36$, BF = 1.73, and 2.22 more points (95% CI=[1.12,

3.31]) at time point 3 than time point 1, $t(62) = 4.87$, $p < 0.001$, $d = 0.54$, $BF = 24.92$, but similarly at time points 2 and 3, $t(62) = 1.47$, $p = 0.26$. We also examined at the averaged reward over trials (learning curves), to assess whether participants learned more quickly at subsequent time points (Figure 4.2B). Importantly, there was no significant main effect of time point, $\chi^2(1)= 3.47$, $p = 0.063$.

We next looked at sampling behaviors (Figure 4.2C). When considering the distance between consecutive choices, there was a significant main effect of time point, $\chi^2(1)= 20.73$, $p < 0.001$. Post-hoc analyses revealed that participants sampled with smaller distances (more locally) at time point 2 ($M = 1.62$) than time point 1 ($M = 1.89$), $t(62) = 4.47$, $p < 0.001$, $d = 0.39$, $BF = 2.27$. They also sampled more locally at time point 3 ($M = 1.53$) than time point 1, $t(62) = 4.92$, $p < 0.001$, $d = 0.53$, $BF = 20.42$, but similarly at time points 2 and 3, $t(62) = 2.14$, $p = 0.09$. When considering the number of unique options sampled, there was a significant main effect of time point, $\chi^2(1)= 17.78$, $p < 0.001$. Post-hoc analyses revealed that participants sampled 2.16 more unique options (95% CI=[0.87, 3.44]) at time point 1 than time point 2, $t(62) = 4.04$, $p = 0.0004$, $d = 0.27$, $BF > 100$, 3.08 more unique options (95% CI=[1.44, 4.73]) at time point 1 than time point 3, $t(62) = 4.50$, $p = 0.0001$, $d = 0.37$, $BF > 100$, and 0.93 more unique options (95% CI [0.04, 1.81] at timepoint 2 than 3, $t(62) = 2.51$, $p = 0.039$, $d = 0.10$, $BF = 0.27$.

Finally, a Bayesian hierarchical regression was used to examine how participants' choice distances were affected by the reward value of previous choices (Figure 4.2D). In this model, previous reward value, time point, and their interaction, were used to predict search distance, with participant included as a random effect. Significance of the main effects and interaction were assessed with nested model comparisons. Results revealed a significant interaction between previous reward and time point was significant, $\beta = -0.45$, 95% CI [-0.54, -0.36], $BF > 100$. As

the previous reward increased, participants were more likely to sample with smaller distances in subsequent time points. However, the effect size is relatively small. The main effect of previous reward on distance was also significant, $\beta = -5.30$, 95% CI [-5.77, -4.81], BF > 100, suggesting that a higher previous reward resulted in a decrease in distance to the subsequent choice. Finally, the main effect of time point was also significant, $\beta = 0.30$, 95% CI [0.20, 0.39], BF > 100, suggesting that in subsequent time points participants tended to sample slightly further away. As with the significant interaction, this is effect size is relatively small.

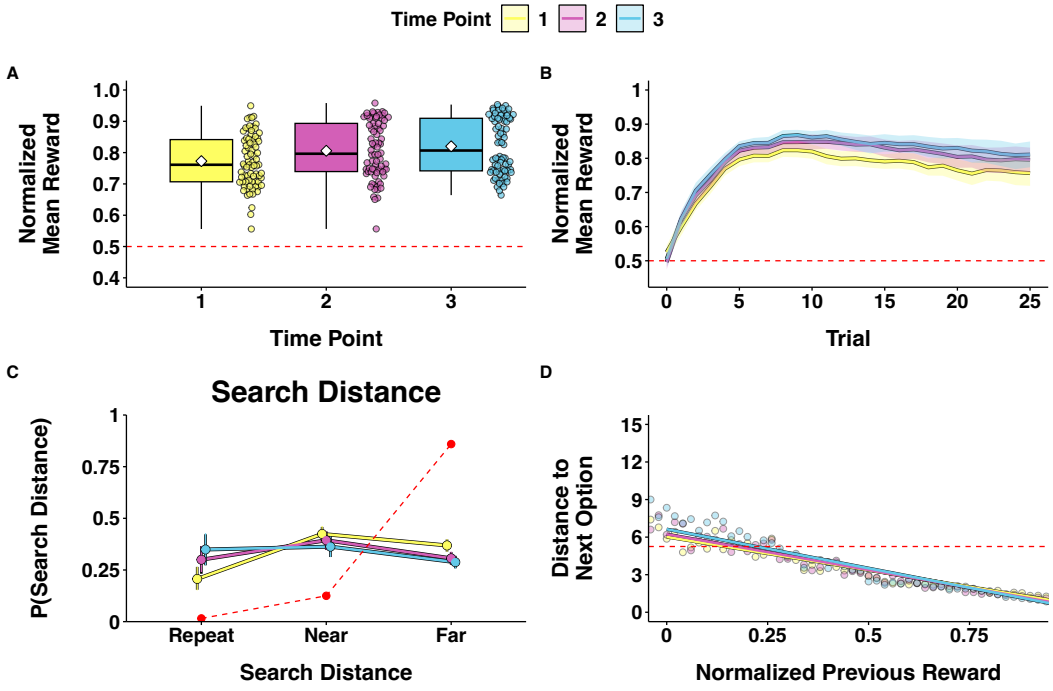


Figure 4.2 Main Behavioral Results. Red dashed line in all graphs represents the expected results from a fully random model. A: Normalized mean reward between time points. Box plots represent the median and IQR (interquartile range), white diamonds represent group means, and each point is an individual participant's score. B: Learning curves displaying the normalized mean reward across trials for each time point. Lines represent group means and the shaded ribbon represents the 95% CI. C: Proportion of choices based on distance for each time point. Points represent group means with 95% CIs. D: Distance between sequential choices as a function of the reward value of the previous option. Points represent means of the raw data, lines correspond to fixed effects from a hierarchical Bayesian regression with 95% CI shaded regions.

Bonus Round. For predictions made in the bonus round, we first looked at participants' prediction error, defined by the mean absolute difference between actual reward values and participants' predictions. There was no significant main effect of time point, $\chi^2(1) = 0.23$, p =

0.63. Participants' predicted reward values of unobserved options had errors of similar magnitude at time points 1 (M = 5.4, SE = 0.38), 2 (M = 5.54, SE = 0.35), and 3 (M = 5.14, SE = 0.36). We next considered how certain participants were about their reward value predictions. There was no significant main effect of time point, $\chi^2(1) = 0.002$, p = 0.97. Participants were similarly certain about their reward value predictions at time points 1 (M = 5.87, SE = 0.18), 2 (M = 5.87, SE = 0.18), and 3 (M = 5.88, SE = 0.19). Finally, we investigated how a participant's choice among the 5 unknown options was influenced by their reward value predictions and certainty judgements. To do so, judgements of reward value and certainty for chosen options were standardized within-subjects at each time point: the chosen option's predicted reward value and certainty were divided by them sum of all a participant's reward value predictions and certainty judgements at that time point. There was no significant main effect of time point for predicted reward value, $\chi^2(1) = 1.50$, p = 0.22. Participants chose options with similar predicted reward values at time points 1 (M = 0.25, SE = 0.001), 2 (M = 0.26, SE = 0.001), and 3 (M = 0.27, SE = 0.001). There was also no main effect of time point for certainty judgements, $\chi^2(1) = 0.61$, p = 0.47. Participants chose options with similar certainty of predicted rewards at time points 1 (M = 0.21, SE = 0.007), 2 (M = 0.22, SE = 0. 007), and 3 (M = 0.22, SE = 0. 007).

Approach Avoid Decision Making Task

Exploration. To start, we tested for differences in exploration behaviors between time points (1, 2, and 3). As participants are first ignorant to which objects were zaffs, they should begin by approaching the first objects. After incurring a costly outcome, they should fall into a learning trap, whereby they generalize to a one-dimensional rule based on either the color or pattern match to the costly object (i.e., seeing a white striped non-zaff and inferring that zaffs are black or spotted). If a participant fell into a learning trap, they should be increasingly less likely

throughout the task to approach non-zaffs (p(approach | non-zaff) across trial sets). Further, they should persist in avoiding several blocks which were in fact zaffs (i.e., avoiding a white spotted block or black striped block after seeing a white striped non-zaff), never approaching them throughout the entirety of the task (p(approach | zaff) across trial sets). If participants learned strategies that they implemented when repeating the task, we should see differences in these approach rates across time points.

To test for differences in approach behaviors of non-zaff trials across trial sets (Figure 4.3), we fit a mixed-effects logistic regression. Time point (1, 2, and 3), trial set (1, 2, 3, and 4), and their interaction were included as fixed effects. All fixed effects were mean centered. Random intercepts for participant and random slopes for trial set and time point were also included in the model. The model specifying the full nesting structure of trial set within time point resulted in a singular fit, so this was dropped. The resulting full model was as follows: approach ~ Time_Point * Trial_Set + (1+ Time_Point + Trial_Set | participant_id). Significance of the main effects and interaction were assessed with nested model comparisons. Results showed that there was no significant interaction, $\chi^2(1) = 0.65$, $p = 0.42$. The difference in probability of approaching non-zaffs across trial sets did not differ across time points. There was no main effect of time point, $\chi^2(1) = 0.42$, $p = 0.52$. The overall probability of approaching non-zaffs did not differ between conditions. However, there was a main effect of trial set, $\beta = -0.33$, $\chi^2(1) = 14.6$, $p < 0.001$, OR = 0.72, 95% CI [0.61, 0.85]. Consistent with previous work, across all time points there was a decrease in the probability of approaching non-zaffs across trial sets.

To test for differences in approach behaviors of zaff trials across trial sets (Figure 4.3), we first calculated the proportion of zaffs approached within trial sets for each participant at each timepoint. We then fit a linear mixed-effects model. Time point (1, 2, and 3), trial set (1, 2, 3,

and 4), and their interaction included as fixed effects, and all fixed effects were mean centered. Random intercepts for participant and random slopes for time point and trial set were also included in the model. The resulting full model was as follows: approach ~ Time_Point * Trial_Set + (1+ Time_Point/Trial_Set | participant_id). Significance of the main effects and interaction were assessed using nested model comparisons. Results showed that there was no significant interaction of time point and trial set, $\chi^2(1) = 0.62$, $p = 0.43$. The probability of approaching zaffs across trial sets did not differ between time points. There was no significant main effect of time point, $\chi^2(1) = 0.12$, $p = 0.73$. The overall probability of approaching zaffs did not differ between time points. Further, unlike in previous work, there was a significant main effect of trial set, $\beta = -0.015$, $\chi^2(1) = 6.78$, $p = 0.009$. Post-hoc pairwise comparisons using the Tukey method for multiple comparisons indicated that participants were more likely to approach zaffs in trial set 1 ($M = 0.70$) than in trial set 2 ($M = 0.65$), $t(438) = 2.72$, $p = 0.034$. Participants were also more likely to approach zaffs in trial set 1 than in trial set 4 ($M = 0.65$), $t(438) = 3.11$, $p = 0.011$. No other pairwise comparisons were significant ($p > 0.05$).

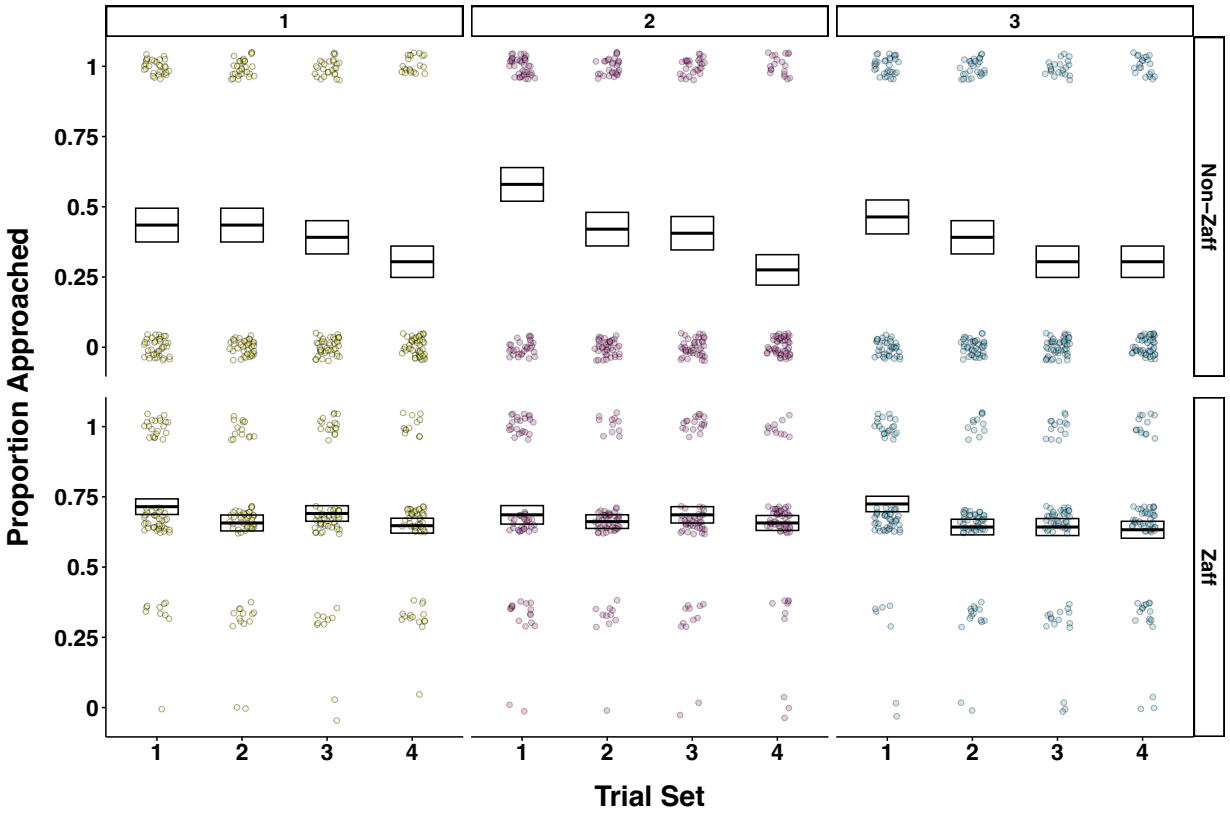


Figure 4.3 Participants' approach-avoid decisions across each of the 4 trial sets. Each trial set contained one zaff and three non-zaffs. Points correspond to individual participants' responses. Box plots represent means and bootstrapped 95% CIs.

Learning. Participants' responses to each object type were used to categorize their learning at test and generalization. Specifically, these responses were coded according to one of four categories: 1. A one-dimension color rule, where zaffs were defined by their color (either white or black), 2. A one-dimensional pattern rule, where zaffs were defined by their pattern (either striped or spotted), 3. A two-dimensional rule, where zaffs were a single object type and all other types were non-zaffs, 4. All other response patterns that do not fit in one of the three previous categories. We compared the difference in proportion of participant responses that followed each rule category between time points at both test and generalization (Table 4.1). Unlike in previous work, at all time points, the majority of participants during generalization and slightly less than half at test responded in ways inconsistent with having learned any rule.

Table 4.1 Proportion of participants from each time point at Test and Generalization who responded according to each rule.

	One-Dimensional Color Rule		One-Dimensional Pattern Rule		Two-Dimensional Rule		No Discernable Rule	
Time Point	Test	Gen.	Test	Gen.	Test	Gen.	Test	Gen.
1	4.7%	7.8%	28.1%	9.4%	23.4%	14.1%	43.7%	68.8%
2	4.7%	9.4%	29.7%	6.3%	17.2%	4.7%	48.4%	79.7%
3	4.7%	9.4%	35.9%	7.8%	17.2%	3.1%	42.2%	79.7%

To explicitly investigate participants' susceptibility to learning traps and whether this differed between time points, we looked at their likelihood of responding according to a one-dimensional rule vs two-dimensional rule. To do so, we fit logistic regression models predicting rule by time point for both generalization and test. Time point was mean centered and included as a fixed effect, and participant was included as a random effect. Significance of the main effects were assessed using nested model comparisons. At both test, $\chi^2(1) = 0.47$, $p = 0.49$, and generalization, $\chi^2(2) = 0.59$, $p = 0.75$, there were no differences between time points.

As a final test of learning, we compared the overall reward earned between time points. If participants had correctly learned the two-dimensional rule, they should approach more zaffs and avoid more non-zaffs, resulting in a higher total earned reward. To do so, we fit a linear mixed-effects model with time point (1, 2, and 3) included as a fixed effect and participant as a random effect. The resulting full model was: reward ~ time_point + (1 | participant_id). Significance of the main effect was assessed using nested model comparisons. This revealed a significant main effect of time point, $\chi^2(1) = 4.70$, $p = 0.030$. Post-hoc pairwise comparisons using the Tukey method for multiple comparisons indicated that participants earned higher rewards at time point 3 ($M = 11.03$) than at time point 2 ($M = 9.84$), $t(124) = 2.96$, $p = 0.01$, $d = 0.41$. Neither the

average reward between time point 1 (M = 10.14) and time point 2, $t(124) = 0.75$, $p = 0.73$, or between time point 1 and time point 3, $t(124) = -2.21$, $p = 0.07$, differed significantly.

Expectations. In trial one of the first trial set, participants always saw an object that was a zaff. In trial two of the first trial set, participants always saw the non-zaff object. To assess whether formed different initial expectations (predictions for whether or not the object was a zaff) at different time points, we tested for differences in their predictions made on trials three and four of the first trial set. At this point, participants had been exposed to both one zaff and the non-zaff, so initial predictions they make on the trials which immediately follow may have influenced their behavior in subsequent trials. We fit a logistic regression to model the relationship between participant's initial predictions by time point (1, 2, and 3). Time point was mean centered and included as a fixed effect. A mixed-effects logistic regression including a random effect for participant resulted in a singular fit model. However, the inclusion of this random effect did not significantly improve model fit ($p > 0.05$), and the results from both models were otherwise identical. Thus, following the approach from the original work, we report the results from the logistic regression. Significance of the main was assessed with nested model comparisons. There was no significant main effect of time point, $\chi^2(1) = 0.02$, $p = 0.90$. Participants formed similar predictions on these trials at all time points (Figure 4.4).

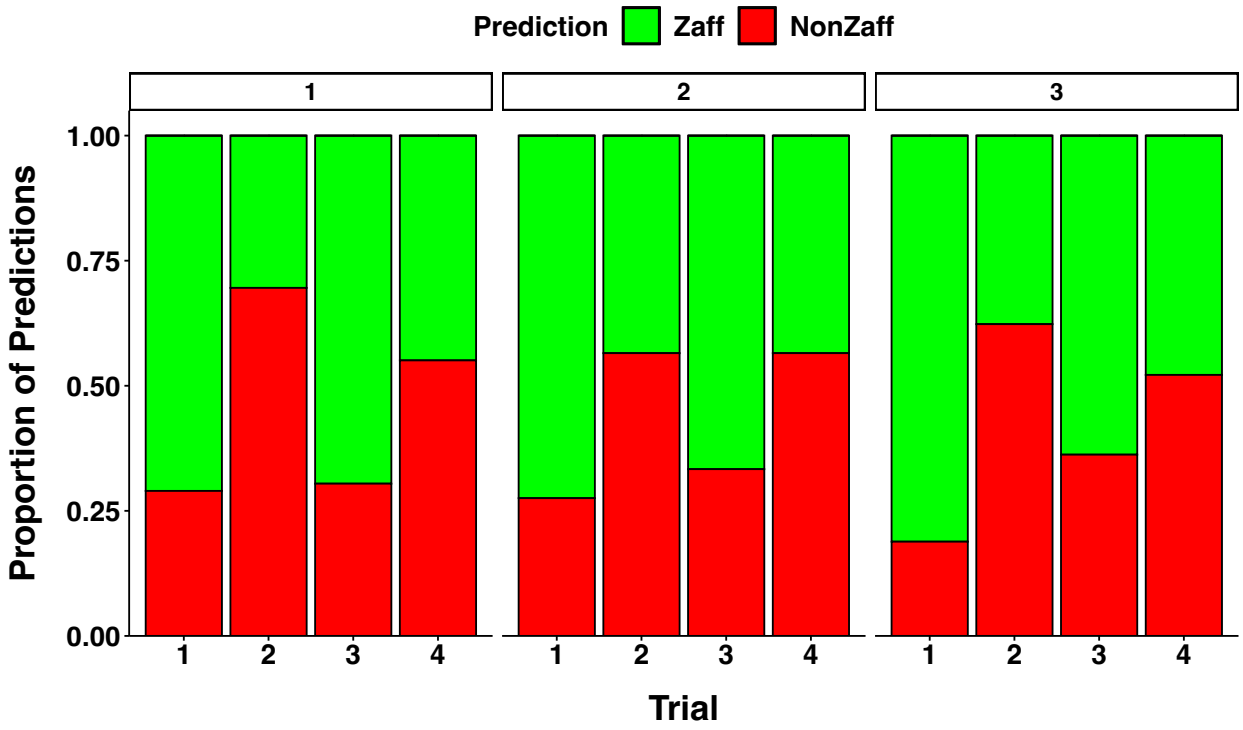


Figure 4.4 Participant predictions on trials three and four of the first trial set. Trials one, three, and four, always contained zaffs, and trial two always contained a non-zaff.

We next looked at whether participants' decisions to approach or avoid a block were influenced by their predictions about that block, and whether this differed across time points (Figure 4.5). A mixed-effects logistic regressions was fit predicting approach by time point (1, 2, and 3), prediction (zaff vs non-zaff), and their interaction as fixed effects. Time point and prediction were dummy coded and mean centered. Participant was included as a random effect. The resulting full model was as follows: Approach ~ Time_Point * Prediction + (1 | participant_id). Significance of the main effects and interaction were assessed using nested model comparisons. There was no significant interaction between time point and prediction, $\chi^2(1) < 0.001$, $p = 0.99$, or significant main effect of time point, $\chi^2(1) < 0.001$, $p = 0.90$. There was, however, a main effect of prediction, $\chi^2(1) = 207.27$, $p < 0.001$, OR = 0.12, 95% CI [0.09, 0.17]. Participants at all time points approached nearly every block they predicted was a zaff (time point 1 = 95.2%, time point 2 = 96.0%, time point 3 = 95.4%), and approached

approximately one quarter of blocks they predicted were non-zaffs (time point 1 = 25.0%, time point 2 = 23.6%, time point 3 = 23.6%). Notably, this extent of approaching objects predicted to be non-zaffs is higher than adults in previous studies.

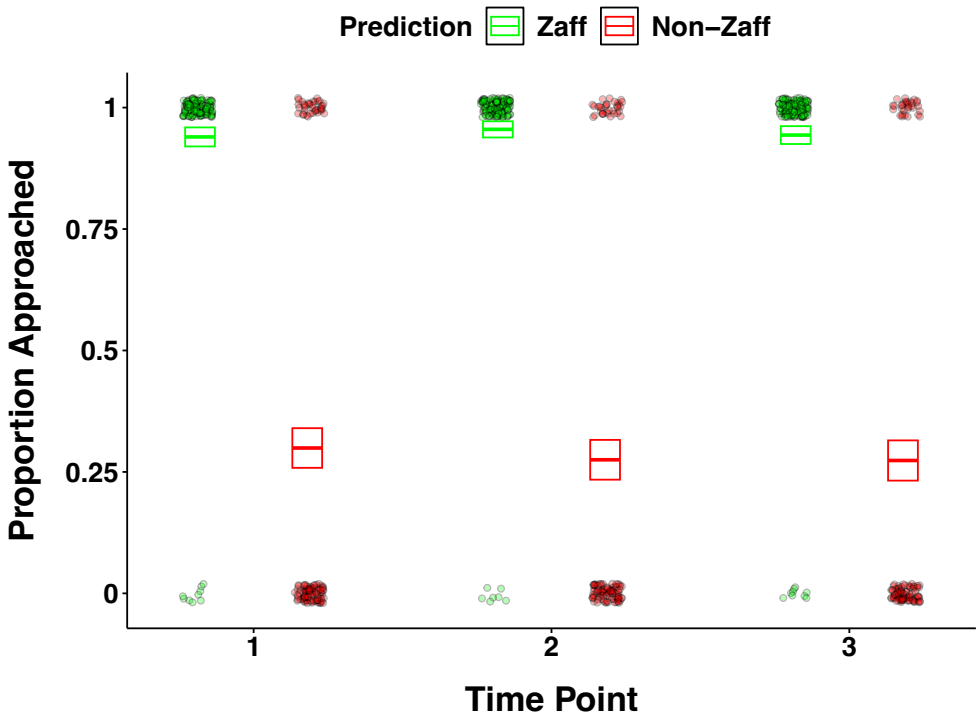


Figure 4.5 Participant approach decisions as a function of their predictions. Individual points represent participants' choices (each represented four times, representing the four trials in which predictions were made), Box plots represent means with boot strapped 95% CIs.

Discussion

Previous research has shown that cognitive flexibility changes across development (Giron et al., 2023; Liquin & Gopnik, 2022; Plebanek & Sloutsky, 2017; Schulz, et al., 2019) and is impaired in many clinical conditions (Abend et al., 2018; Everaert et al., 2018; Kraus et al., 2021; Mennen et al., 2019). As such, there is great interest in using behavioral measures to track individuals' cognitive flexibility over time. However, when an individual completes the same task multiple times, they may develop practice effects in the form of strategies, memory for test items, and increased comfort with the task. Practice effects can occur independent of changes in

the measured construct (Calamia et al., 2012), and as a result, can seriously undermine the accuracy of conclusions drawn from an individual's performance. The present work investigated whether cognitive flexibility tasks are susceptible to practice effects by having participants complete three tasks which each assessed a different dimension of flexibility at three time points spaced one-week apart. Specifically, participants completed a spatially correlated multi-armed bandit task, an approach-avoid decision making task, and a change detection task. Results suggest that these behavioral measures of cognitive flexibility are resistant to all but the most minor practice effects, with similar performance across time points.

Two of the three tasks showed no evidence for practice effects of any kind. In the approach-avoid decision making task, there was no change in either search strategy or learning: participants approached rewarding and costly objects with similar frequency, learned the true rule at similar rates, and formed similar predictions at all time points. Likewise, in the change detection task, participants' accuracy was similar at each time point for both cued and un-cued images. Only in the spatially correlated multi-armed bandit task was there any evidence of minor practice effects. Specifically, participants sampled more locally and earned higher rewards at time points 2 and 3 compared to time point 1. While statistically significant, three key points raise questions about the practical significance and extent of concern warranted of these findings. First, these differences were of low relative magnitude and may not represent meaningful learning or differences in search and sampling strategies. The largest difference in average reward earned between time points was only 2.22 more points. In most rounds of the task, where participants choose 25 tiles to earn their score, the minimum possible reward value for a tile was ~5 and the maximum was ~50. If participants were truly applying different strategies across time points, it is reasonable to expect that the average point differences would be larger than ~2,

especially as this small difference could be explained by the normally distributed noise added to reward values on each trial. This is additionally supported when considering the differences in sampling distance, which characterizes participant search strategies. The maximum possible distance between tiles on the game boards is 14 units. However, the largest difference in average sampling distance was 0.37 units. Given the spatial correlation of tiles' reward values, smaller distances do not equate to large differences in reward values, and average distance differences of less than one unit suggests participants are not exhibiting much learning or change in sampling strategy. Second, despite these changes, participants were no more accurate in predicting the reward values of unobserved choices nor were they more certain in their predictions, suggesting that they were not learning and applying any information about the task between time points. Third, these practice effects were only observed between the first time point to the second and third. That is, after completing the task once, participants' scores stabilized across subsequent time points. Thus, even these minor practice effects which are seemingly not practically significant might be controlled for by having participants complete a practice time point before their baseline scores are recorded.

The current results offer important insights and implications for researchers investigating clinical interventions aimed at changing cognition or cognitive abilities. They offer strong evidence for the test-retest and parallel forms reliability of these three tasks, indicating that they can successfully be administered at multiple time points to assess changes to cognitive flexibility over time and treatment. These more computational approaches may thus be confidently employed in the place of unreliable self-reports or coarser behavioral measures, allowing specific cognitive processes and changes to be more accurately characterized and studied. By administering these tasks at multiple time points, particularly pre- and post-intervention,

researchers can better assess changes in one's cognitive functioning over time and the efficacy of different treatment options.

These results also further our understanding of cognition in general. Repeated exposure to information, or opportunities to get information, does not appear to be helpful in facilitating learning. This has important implications for educational settings by highlighting the need for diverse and adaptive learning strategies. Additionally, that people's strategies did not change over time in these tasks shows they may be used to help better understand strategic adaption. For example, participant performance can be assessed after providing them more elaborative feedback (Van der Kleij et al., 2015), or perhaps by increasing metacognitive awareness by prompting them to list the strategies they employed (Siegler & Jenkins, 2014). Both strategies have been shown to improve learning outcomes in classroom settings (e.g., Guo, 202l; Schute, 2007). Finally, that participants repeatedly succumbed to the same learning trap demonstrates how deeply ingrained learning traps may be. This may be particularly important for treating certain clinical conditions. For example, learned helplessness, a characteristic feature of depression (Klein et al., 1976; Miller & Seligman, 1975), has been framed as a particular type of learning trap (Erev, 2014; Rich, 2018). By better understanding how these maladaptive cognitive patterns are formed, clinicians can develop more effective treatments and therapies to help patients unlearn these behaviors and improve their general functioning.

Overall, this work demonstrates, for the first time, the enhanced psychometric properties of these cognitive flexibility tasks, highlighting their suitability for a variety of research purposes beyond the original work. While much research to date has shown the rigidity of baseline adult cognition in general, this work demonstrates a different dimension of rigidity: the stability of

strategies employed over time. When paired with interventions designed to change people's strategies, these tasks may be used to better understand the mechanisms underlying cognition.

Acknowledgements

Chapter 4 is currently being prepared for submission for publication of the material. Hurwitz, Ethan; Brockbank, Erik; Walker, Caren. The dissertation author was the primary investigator and author of this material.

Chapter 5 General Discussion

In recent years there has been a resurging interest in studying psychedelic drugs and their potential to both shed light on consciousness and be used as a novel therapeutic. While compelling evidence has been amassed for their potential to treat a variety of clinical conditions, there is comparatively little evidence to indicate the mechanisms underlying the observed clinical benefits. Although some theories derived from neuroimaging data have been proposed, such as increased neural plasticity, they can provide little insight into how or in what ways subjective experience is changed and do not afford causal claims. A promising cognitive-behavioral account has also been proposed—that psychedelics increase cognitive flexibility—but has largely used coarse behavioral measures and yielded inconsistent results. Thus, the question of how psychedelics engender their clinical benefits largely remains unclear. The work in this dissertation addresses this gap by applying modern cognitive science and computational methodology to test whether psilocybin results in changes to specific behaviors symptomatic of increased cognitive flexibility. These behaviors are gleaned from an unlikely source: children, a population known to be more cognitively flexible than adults in general.

First, Chapter 2 introduces a novel serial production task which successfully measures cognitive flexibility by characterizing reasoning under dynamic constraints. The extant tools available to quantify and better understand this behavior are limited, failing to explicitly capture the filtering and constraint satisfaction aspects of these problems, incorporating only static constraints (if any), and have been long criticized for psychometric shortcomings. Based on the popular internet game *Contexto*, in this novel task, participants must try to guess a secret target word on each trial. After submitting a guess, they are shown that word's semantic similarity rank to the target word. In this way, participants' own guesses provide feedback which should serve to

constrain their subsequent guesses. We demonstrate that people are indeed sensitive to this feedback, producing sequential guesses that are more similar to each other and to the target word the closer in semantic distance the guess was. With several design parameters that can vary, such as the guess limit per trial, success criteria, or semantic content of target words, this task can be used by researchers to better characterize different dimensions of cognitive flexibility. For example, researchers can investigate whether participants have specific lexical biases by assessing the types of guesses they produce and how this may be affected by the types of target words presented (i.e., would participants produce more adjective guesses after seeing a target word that was an adjective?), or how participant search strategies may be affected by the goal on each trial (i.e., guessing the exact target word vs. guessing a word that is "close enough"). Researchers can also apply computational models with this task to better characterize the specific search and sampling strategies people employ. For example, performance can be compared to creative foraging tasks and the predictions from different foraging algorithms to assess the extent to which people deviate from optimal search. Further, given that this task is in the form of a game, it yields both more and higher quality data than comparative non-game task-based measures (Allen et al., 2023; Hartshorne et al., 2019). Finally, by demonstrating that this task is not susceptible to practice effects, it can be used in clinical contexts to validate interventions intended to affect cognitive flexibility or track changes to cognitive flexibility over time in patients with cognitive deficits.

Chapter 3 presents pilot data from six experiments to assess whether and how psilocybin affects cognitive flexibility. As research in the field of cognitive development has routinely demonstrated, children exhibit greater cognitive flexibility than adults. Here, we test three specific dimensions which have previously shown developmental differences. We first tested

whether psilocybin reduces the strength of people's prior beliefs, making them more sensitive to current evidence, through a causal reasoning task. Leveraging the high prior that adults have on disjunctive causes of action, we demonstrate that 100% of adults who received psilocybin, compared to 25% of those who received niacin, correctly inferred a conjunctive rule when shown evidence consistent with this rule. We next tested whether psilocybin changes people's hypothesis search and sampling strategies, making them more board, exploratory, and resistant to learning traps. Utilizing the Contexto task from Chapter 2, we demonstrate that participants who received psilocybin engaged in broader search compared to those who received niacin. Further, the model parameter estimates from a spatially correlated multi-armed bandit task suggest that participants who received psilocybin engaged in more exploration and less generalization than those who received niacin. However, in a single cue SPT and an approach-avoid decision making task, no differences were found between conditions. Finally, we tested whether psilocybin changes people's exogenous attention, making it more diffuse, through a change detection task. We found that there were no differences in change detection accuracy for cued or un-cued images between participants who received psilocybin or niacin.

Given the small sample sizes of the psilocybin (n=5) and placebo (n=4) groups from this pilot study, the results from the formal statistical analyses should not be viewed as definitive. However, the results across these six experiments provide a preliminary indication of the specific ways in which psilocybin may affect cognitive flexibility. Previous experiments investigating the effects of psilocybin on cognitive flexibility have relied on paradigms employing summary score metrics, which quantify responses rather than the cognitive processes underlying their production (Hass, 2017). Instead, the current work employed paradigms which specifically assessed behaviors symptomatic of different cognitive processes. Consistent with prior proposals of

developmental models developmental models of psychedelics (Gopnik, 2018), the pilot data outlined here suggests that psilocybin decreases the strength of prior beliefs and changes search and sampling strategies to be broader and more exploratory. This could explain the previous research demonstrating that psilocybin elicits mystical-type and insight experiences, which are correlated with their clinical benefit (Griffiths et al., 2008, 2016; Noorani et al., 2018; Roseman et al., 2018). Reducing the strength of people's priors and changing their hypothesis search strategies may result in perspective-shifting breaks from the maladaptive and ruminative thought patterns endemic to conditions like depression and anxiety. It will also be informative to examine whether the observed differences in performance on these assessments are related to functional and structural changes in neural activity. Given that the brain regions showing the greatest changes in activity in response to psilocybin are those that are related to cognitive flexibility, demonstrating a relationship between these neural effects and cognitive performance would help address the mind-brain gap that characterizes the literature on the effects of psilocybin. Further, testing whether scores on these assessments are correlated with clinical outcomes would help elucidate how psychedelics impart their clinical benefit. Together, the pilot results from this battery of cognitive assessments provide a promising direction to guide future research into the mechanisms underlying the clinical benefits of psychedelics.

Finally, Chapter 4 provides evidence for the test-retest and parallel forms reliability of a subset of these experimental paradigms: the change detection task, spatially correlated multi-armed bandit task, and approach-avoid decision making task. New stimuli sets were created for the latter two tasks, and initial pilot studies demonstrated they produce results identical to the original paradigms. Then, a new set of participants completed each task at three time points spaced one-week apart. The results showed that there were no practice effects on the approach-

avoid decision making task or change detection task. While the spatially correlated multi-armed bandit task did demonstrate some minor practice effects, these did not appear to indicate a practical change in search or sampling strategy, and performance stabilized between all administrations after the first time point.

These findings have important implications for future clinical research. In place of unreliable self-reports or coarser behavioral measures, these tasks may be confidently employed in clinical populations, allowing specific cognitive changes to be more accurately characterized over time. By administering these tasks at multiple time points, particularly pre- and post-intervention, researchers can also better assess the efficacy of different treatment options in increasing cognitive flexibility. Further, a major criticism of the extant research on psilocybin and cognitive flexibility is that participants in those studies exhibited practice effects, obscuring the interpretation of the changes in task performance pre- and post-intervention (Mason et al., 2021). Beyond clinical applications, these findings also have important implications for research on cognition in general. Given that people's strategies are consistent over time, the development of these strategies can be more clearly investigated. Further, how these strategies can be changed in the absence of an intervention as powerful as administering a psychedelic drug. For example, perhaps raising an individual's metacognitive awareness by prompting them to list the strategies they employed, and whether that impacts performance during subsequent task administrations.

Overall, the present work provides a novel tool and further validates existing tools to assess changes in cognitive flexibility, and presents preliminary pilot data suggesting that psilocybin increases cognitive flexibility. Collectively, these findings position the field of psychedelic science to identify the specific cognitive mechanisms underlying psychedelic therapy. Future fully powered experiments can utilize the same methods outlined here,

administered at multiple time points, to identify the specific dimensions of cognitive flexibility that psilocybin may increase. With this, it would be possible to better identify who, and what clinical conditions, may be receptive to psychedelic therapy. It would also be possible to develop alternative treatments targeting these same mechanisms for individuals who are not suited for psychedelic treatment. Overall, this work represents an important step forward toward our understanding of psychedelic treatment and cognitions as a whole.

REFERENCES

Abend, R., de Voogd, L., Salemink, E., Wiers, R. W., Pérez-Edgar, K., Fitzgerald, A., White, L. K., Salum, G. A., He, J., Silverman, W. K., Pettit, J. W., Pine, D. S., & Bar-Haim, Y. (2018). Association between attention bias to threat and anxiety symptoms in children and adolescents. *Depression and Anxiety*, *35*(3), 229–238. https://doi.org/10.1002/da.22706

Acuna, D., & Schrater, P. R. (2008). Structure Learning in Human Sequential Decision-Making. *Advances in Neural Information Processing Systems*, *21*.

Akbari Chermahini, S. A., Hickendorff, M., & Hommel, B. (2012). Development and validity of a Dutch version of the Remote Associates Task: An item-response theory approach. *Thinking Skills and Creativity, 7*, 177–186. doi:10.1016/j.tsc.2012.02.003

Allen, K. R., Brändle, F., Botvinick, M. M., Fan, J., Gershman, S. J., Gopnik, A., Griffiths, T. L., Hartshorne, J. K., Hauser, T. U., Ho, M. K., De Leeuw, J. R., Ma, W. J., Murayama, K., Nelson, J. D., Van Opheusden, B., Pouncy, H. T., Rafner, J., Rahwan, I., Rutledge, R., Sherson, J., Özgür, S., Spiers, H., Summerfield, C., Thalmann, M., Vélez, N., Watrous, A., Tenenbaum, J., Schulz, E. (2023). Using Games to Understand the Mind. *Nature Human Behaviour*, 1-9. https://doi.org/10.31234/osf.io/hbsvj

Andersen, K. A. A., Carhart-Harris, R., Nutt, D. J., & Erritzoe, D. (2020). Therapeutic effects of classic serotonergic psychedelics: A systematic review of modern-era clinical studies. *Acta Psychiatrica Scandinavica*. https://doi.org/10.1111/acps.13249

Anderson, B. J., & Meyer, J. G. (2022). *Finding the optimal human strategy for Wordle using maximum correct letter probabilities and reinforcement learning* (arXiv:2202.00557). arXiv. http://arxiv.org/abs/2202.00557

Andersson, M., Persson, M., & Kjellgren, A. (2017). Psychoactive substances as a last resort—A qualitative study of self-treatment of migraine and cluster headaches. *Harm Reduction Journal*, *14*(1), 60. https://doi.org/10.1186/s12954-017-0186-6

Andrews, F. (1975). Social and psychological factors which influence the creative process. In 1. A. Taylor & J. W. Getzels (Eds.), *Perspectives in Creativity*. Chicago: Adline

Arrieta, Ó., Angulo, L. P., Núñez-Valencia, C., Dorantes-Gallareta, Y., Macedo, E. O., Martínez-López, D., Alvarado, S., Corona-Cruz, J.-F., & Oñate-Ocaña, L. F. (2013). Association of Depression and Anxiety on Quality of Life, Treatment Adherence, and Prognosis in Patients with Advanced Non-small Cell Lung Cancer. *Annals of Surgical Oncology*, *20*(6), 1941–1948. https://doi.org/10.1245/s10434-012-2793-5

Auer, P. (2002). Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*, *3*(Nov), 397–422.

Baer, J. (2016). Creativity Doesn't Develop in a Vacuum: Creativity Doesn't Develop in a Vacuum. *New Directions for Child and Adolescent Development*, *2016*(151), 9–20. https://doi.org/10.1002/cad.20151

Barrett, F. S., Carbonaro, T. M., Hurwitz, E., Johnson, M. W., & Griffiths, R. R. (2018). Double-blind comparison of the two hallucinogens psilocybin and dextromethorphan: Effects on cognition. *Psychopharmacology*, *235*(10), 2915–2927. https://doi.org/10.1007/s00213-018-4981-x

Barrett, F. S., Doss, M. K., Sepeda, N. D., Pekar, J. J., & Griffiths, R. R. (2020). Emotions and brain function are altered up to one month after a single high dose of psilocybin. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-59282-y

Barrett, F. S., & Griffiths, R. R. (2018). Classic Hallucinogens and Mystical Experiences: Phenomenology and Neural Correlates. *Current Topics in Behavioral Neurosciences*, *36*, 393–430. https://doi.org/10.1007/7854_2017_474

Barrett, F. S., Johnson, M. W., & Griffiths, R. R. (2015). Validation of the revised Mystical Experience Questionnaire in experimental sessions with psilocybin. *Journal of Psychopharmacology*, *29*(11), 1182–1190. https://doi.org/10.1177/0269881115609019

Barron, F. (1969). Creative person and creative process. New York. Holt, Rinehart & Winston

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., & Ehrenreich, H. (2010). Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. *BMC Neuroscience*, *11*(1), 118. https://doi.org/10.1186/1471-2202-11-118

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the Science of Self-Reports and Finger Movements: Whatever Happened to Actual Behavior? *Perspectives on Psychological Science*, *2*(4), 396–403. https://doi.org/10.1111/j.1745-6916.2007.00051.x

Beaty, R. E., Benedek, M., Silvia, P. J., & Schacter, D. L. (2016). Creative cognition and brain network dynamics. *Trends in Cognitive Sciences*, *20*(2), 87–95.

Beaty, R. E., Chen, Q., Christensen, A. P., Kenett, Y. N., Silvia, P. J., Benedek, M., & Schacter, D. L. (2020). Default network contributions to episodic and semantic processing during divergent creative thinking: A representational similarity analysis. *NeuroImage*, *209*, 116499. https://doi.org/10.1016/j.neuroimage.2019.116499

Becker, A. M., Holze, F., Grandinetti, T., Klaiber, A., Toedtli, V. E., Kolaczynska, K. E., Duthaler, U., Varghese, N., Eckert, A., Grünblatt, E., & Liechti, M. E. (2022). Acute Effects of Psilocybin After Escitalopram or Placebo Pretreatment in a Randomized, Double-Blind, Placebo-Controlled, Crossover Study in Healthy Subjects. *Clinical Pharmacology and Therapeutics*, *111*(4), 886–895. https://doi.org/10.1002/cpt.2487

Beittel, K. R. (1964). Creativity the visual arts In higher education: Criteria, predictors, experimentation and their Interactions. In C. W. Taylor(Ed.), *Widening horizons in creativity*. New York: Wiley

Bennett, P. M., & Harvey, P. H. (1985). Brain size, development and metabolism in birds and mammals. *Journal of Zoology*, *207*(4), 491–509. https://doi.org/10.1111/j.1469-7998.1985.tb04946.x

Benveniste, A., & Frere, J. (2022, February 10). A Collection of the Best Wordle Tips and Tricks. *The New York Times*. https://www.nytimes.com/2022/02/10/crosswords/best-wordle-tips.html

Berg, E. A. (1948). A Simple Objective Technique for Measuring Flexibility in Thinking. *The Journal of General Psychology*, *39*(1), 15–22. https://doi.org/10.1080/00221309.1948.9918159

Bigler, R. S., & Liben, L. S. (1992). Cognitive mechanisms in children's gender stereotyping: Theoretical and educational implications of a cognitive-based intervention. *Child Development*, *63*(6), 1351-1363.

Blanco, N., & Sloutsky, V. (2020). Attentional mechanisms drive systematic exploration in young children. *Cognition*, *202*, 104327. https://doi.org/10.1016/j.cognition.2020.104327

Blanco, N. J., & Sloutsky, V. M. (2021). Systematic exploration and uncertainty dominate young children's choices. *Developmental Science*, *24*(2), e13026.

Bogenschutz, M. P., Forcehimes, A. A., Pommy, J. A., Wilcox, C. E., Barbosa, P. C. R., & Strassman, R. J. (2015). Psilocybin-assisted treatment for alcohol dependence: A proof-of-concept study. *Journal of Psychopharmacology*, *29*(3), 289–299.

Bogenschutz, M. P., Ross, S., Bhatt, S., Baron, T., Forcehimes, A. A., Laska, E., Mennenga, S. E., O'Donnell, K., Owens, L. T., Podrebarac, S., Rotrosen, J., Tonigan, J. S., & 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. https://doi.org/10.1001/jamapsychiatry.2022.2096

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology*, *74*, 35–65. https://doi.org/10.1016/j.cogpsych.2014.06.003

Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: Sampling in cognitive development. *Trends in Cognitive Sciences*, *18*(10), 497–500. https://doi.org/10.1016/j.tics.2014.06.006

Bonnelle, V., Ham, T. E., Leech, R., Kinnunen, K. M., Mehta, M. A., Greenwood, R. J., & Sharp, D. J. (2012). Salience network integrity predicts default mode network function after traumatic brain injury. *Proceedings of the National Academy of Sciences*, *109*(12), 4690–4695. https://doi.org/10.1073/pnas.1113455109

Borkel, L. F., Rojas-Hernández, J., Henríquez-Hernández, L. A., Santana Del Pino, Á., & Quintana-Hernández, D. J. (2023). Set and setting predict psychopathology, wellbeing and meaningfulness of psychedelic experiences: A correlational study. *Expert Review of Clinical Pharmacology*, 1–12. https://doi.org/10.1080/17512433.2023.2295997

Brändle, F., Allen, K. R., Tenenbaum, J., & Schulz, E. (2021). Using games to understand intelligence. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43). https://escholarship.org/uc/item/17z2q92d

Brewer, J. A., Worhunsky, P. D., Gray, J. R., Tang, Y.-Y., Weber, J., & Kober, H. (2011). Meditation experience is associated with differences in default mode network activity and connectivity. *Proceedings of the National Academy of Sciences*, *108*(50), 20254–20259.

Brittain, W. L., & Beittel, K. R. (1961). A study of some tests of creativity in relationship to performances in the visual arts. *Studies in Art Education*, *2*(2), 54-65.

Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Neuroscience*, *18*(4), 590–596. https://doi.org/10.1038/nn.3961

Bryl, K., Wenger, S., Banz, D., Terry, G., Ballester, D., Bailey, C., & Bradt, J. (2021). Power over pain – An interprofessional approach to chronic pain: Program feedback from a medically underserved community. *Journal of Evaluation in Clinical Practice*, *27*(6), 1223-1234. https://doi.org/10.1111/jep.13552

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Annals of the new York Academy of Sciences*, *1124*(1), 1-38.

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring Higher the Second Time Around: Meta-Analyses of Practice Effects in Neuropsychological Assessment. *The Clinical Neuropsychologist*, *26*(4), 543–570. https://doi.org/10.1080/13854046.2012.680913

Carbonaro, T. M., Bradstreet, M. P., Barrett, F. S., MacLean, K. A., Jesse, R., Johnson, M. W., & Griffiths, R. R. (2016). Survey study of challenging experiences after ingesting psilocybin mushrooms: Acute and enduring positive and negative consequences. *Journal of Psychopharmacology*, *30*(12), 1268–1278.

Carbonaro, T. M., Johnson, M. W., & Griffiths, R. R. (2020). Subjective features of the psilocybin experience that may account for its self-administration by humans: A double-

blind comparison of psilocybin and dextromethorphan. *Psychopharmacology*, *237*(8), 2293–2304. https://doi.org/10.1007/s00213-020-05533-9

Carbonaro, T. M., Johnson, M. W., Hurwitz, E., & Griffiths, R. R. (2018). Double-blind comparison of the two hallucinogens psilocybin and dextromethorphan: Similarities and differences in subjective experiences. *Psychopharmacology*, *235*(2), 521–534.

Carhart-Harris, R. L., Fortier, M., & Millière, R. (2017). Psychedelics and consciousness: an interview with Robin Carhart-Harris. *ALIUS Bulletin*, *1*, 1-16.

Carhart-Harris, R., Giribaldi, B., Watts, R., Baker-Jones, M., Murphy-Beiner, A., Murphy, R., Martell, J., Blemings, A., Erritzoe, D., & 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., Bolstridge, M., Rucker, J., Day, C. M. J., Erritzoe, D., Kaelen, M., Bloomfield, M., Rickard, J. A., Forbes, B., Feilding, A., Taylor, D., Pilling, S., Curran, V. H., & Nutt, D. J. (2016). Psilocybin with psychological support for treatment-resistant depression: An open-label feasibility study. *The Lancet Psychiatry*, *3*(7), 619–627. https://doi.org/10.1016/S2215-0366(16)30065-7

Carhart-Harris, R. L., Erritzoe, D., Williams, T., Stone, J. M., Reed, L. J., Colasanti, A., Tyacke, R. J., Leech, R., Malizia, A. L., Murphy, K., Hobden, P., Evans, J., Feilding, A., Wise, R. G., & Nutt, D. J. (2012). Neural correlates of the psychedelic state as determined by fMRI studies with psilocybin. *Proceedings of the National Academy of Sciences*, *109*(6), 2138–2143. https://doi.org/10.1073/pnas.1119598109

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., Kaelen, M., Whalley, M. G., Bolstridge, M., Feilding, A., & Nutt, D. J. (2015). LSD enhances suggestibility in healthy volunteers. *Psychopharmacology*, *232*(4), 785–794. https://doi.org/10.1007/s00213-014-3714-z

Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., Chialvo, D. R., & Nutt, D. (2014). The entropic brain: A theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00020

Carhart-Harris, R. L., Roseman, L., Bolstridge, M., Demetriou, L., Pannekoek, J. N., Wall, M. B., Tanner, M., Kaelen, M., McGonigle, J., Murphy, K., Leech, R., Curran, H. V., & Nutt, D. J. (2017). Psilocybin for treatment-resistant depression: fMRI-measured brain mechanisms. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-13282-7

Castellanos, J. P., Woolley, C., Bruno, K. A., Zeidan, F., Halberstadt, A., & Furnish, T. (2020). Chronic pain and psychedelics: a review and proposed mechanism of action. *Regional Anesthesia & Pain Medicine, 45*(7), 486-494.

Chamberlain, S. R., Fineberg, N. A., Blackwell, A. D., Robbins, T. W., & Sahakian, B. J. (2006). Motor inhibition and cognitive flexibility in obsessive-compulsive disorder and trichotillomania. *American Journal of Psychiatry*, *163*(7), 1282–1284.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367.

Chrysikou, E. G., Hamilton, R. H., Coslett, H. B., Datta, A., Bikson, M., & Thompson-Schill, S. L. (2013). Noninvasive transcranial direct current stimulation over the left prefrontal cortex facilitates cognitive flexibility in tool use. *Cognitive Neuroscience*, *4*(2), 81–89. https://doi.org/10.1080/17588928.2013.768221

Cosimano, M. P. (2021). The role of the guide in psychedelic-assisted treatment. *Handbook of Medical Hallucinogens*, 377–394.

Cropley, A. J. (2000). Defining and measuring creativity: *Are creativity tests worth using? Roeper Review*, *23*(2), 72–79. https://doi.org/10.1080/02783190009554069

Csikszentmihalyi, M., & Getzels, J. W. (1973). The personality of young artists: An empirical and theoretical exploration. *British Journal of Psychology*, *64*(1), 91-104.

d'Aquili, E. G., & Newberg, A. B. (1998). The neuropsychological basis of religions, or why God won't go away. *Zygon®*, *33*(2), 187–201.

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive Psychology*, *96*, 1–25. https://doi.org/10.1016/j.cogpsych.2017.05.001

Davis, A. K., Barrett, F. S., & Griffiths, R. R. (2020). Psychological flexibility mediates the relations between acute psychedelic effects and subjective decreases in depression and anxiety. *Journal of Contextual Behavioral Science*, *15*, 39–45. https://doi.org/10.1016/j.jcbs.2019.11.004

Davis, A. K., Barrett, F. S., May, D. G., Cosimano, M. P., Sepeda, N. D., Johnson, M. W., Finan, P. H., & Griffiths, R. R. (2020). Effects of Psilocybin-Assisted Therapy on Major Depressive Disorder: A Randomized Clinical Trial. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2020.3285

Davis, A. K., Barrett, F. S., So, S., Gukasyan, N., Swift, T. C., & Griffiths, R. R. (2021). Development of the Psychological Insight Questionnaire among a sample of people who have consumed psilocybin or LSD. *Journal of Psychopharmacology*, *35*(4), 437–446. https://doi.org/10.1177/0269881120967878

Davis, A. K., Xin, Y., Sepeda, N. D., Garcia-Romeu, A., & Williams, M. T. (2021). Increases in Psychological Flexibility Mediate Relationship Between Acute Psychedelic Effects and Decreases in Racial Trauma Symptoms Among People of Color. *Chronic Stress*, *5*, 24705470211035607. https://doi.org/10.1177/24705470211035607

Davis, G., & Belcher, T. L. (1971). How shall creativity be measured? Torrance tests, RAT, Alpha Biographical and IQ. *Journal of Creative Behavior, 5*, 153-161

Daws, R. E., Timmermann, C., Giribaldi, B., Sexton, J. D., Wall, M. B., Erritzoe, D., Roseman, L., Nutt, D., & Carhart-Harris, R. (2022). Increased global integration in the brain after psilocybin therapy for depression. *Nature Medicine*, *28*(4), Article 4. https://doi.org/10.1038/s41591-022-01744-z

Deák, G. O. (2003). The development of cognitive flexibility and language abilities. *Advances in child development and behavior*, *31*, 273-328.

Deng, W. (Sophia), & Sloutsky, V. M. (2016). Selective attention, diffused attention, and the development of categorization. *Cognitive Psychology*, *91*, 24–62. https://doi.org/10.1016/j.cogpsych.2016.09.002

Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013). Rational variability in children's causal inferences: The Sampling Hypothesis. *Cognition*, *126*(2), 285–300. https://doi.org/10.1016/j.cognition.2012.10.010

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Depaulis, T. (2020). Board Games Before Ur? *Board Game Studies Journal*, *14*(1), 127–144. https://doi.org/10.2478/bgs-2020-0007

Dillehunt, H. Q. (1973). Creativity in children: A comparison of creativity tests and naturalistic measures of creativity, anxiety, and achievement motivation (Doctoral dissertation, California School of Professional Psychology, 1972). *Dissertation Abstracts International, 33*, 3282B

Doss, M. K., Barrett, F. S., & Corlett, P. R. (2022). Skepticism about Recent Evidence That Psilocybin "Liberates" Depressed Minds. *ACS Chemical Neuroscience*, *13*(17), 2540–2543. https://doi.org/10.1021/acschemneuro.2c00461

Doss, M. K., Považan, M., Rosenberg, M. D., Sepeda, N. D., Davis, A. K., Finan, P. H., Smith, G. S., Pekar, J. J., Barker, P. B., Griffiths, R. R., & Barrett, F. S. (2021). Psilocybin therapy increases cognitive and neural flexibility in patients with major depressive disorder. *Translational Psychiatry*, *11*(1), 574. https://doi.org/10.1038/s41398-021-01706-y

Eisenberg, N., & Fabes, R. A. (1990). Empathy: Conceptualization, measurement, and relation to prosocial behavior. *Motivation and Emotion*, *14*(2), 131–149. https://doi.org/10.1007/BF00991640

Elias, S., Spivak, S., Alverez, A., Olivares, A. G., Ferrol, M., & Keenan, J. P. (2023). Self-perception and self-recognition while looking in the mirror on psilocybin. *Journal of Psychedelic Studies*, *7*(2), 129–134. https://doi.org/10.1556/2054.2023.00251

Ellis, A. (1987). The impossibility of achieving consistently good mental health. *American Psychologist*, *42*(4), 364.

Ellison, B. (1973). Creativity in Black Artists: A Comparison of Selected Creativity Measures Using Judged Creativity as a Criterion. *Journal of Non-White Concerns in Personnel and Guidance*, *1*(3), 150-157.

Erev, I. (2014). Recommender Systems and Learning Traps. In *DMRS* (pp. 38-41).

Everaert, J., Bronstein, M. V., Cannon, T. D., & Joormann, J. (2018). Looking Through Tinted Glasses: Depression and Social Anxiety Are Related to Both Interpretation Biases and Inflexible Negative Interpretations. *Clinical Psychological Science*, *6*(4), 517–528. https://doi.org/10.1177/2167702617747968

Fair, D. A., Dosenbach, N. U. F., Church, J. A., Cohen, A. L., Brahmbhatt, S., Miezin, F. M., Barch, D. M., Raichle, M. E., Petersen, S. E., & Schlaggar, B. L. (2007). Development of distinct control networks through segregation and integration. *Proceedings of the National Academy of Sciences*, *104*(33), 13507–13512. https://doi.org/10.1073/pnas.0705843104

Fanciullacci, M., Bene, E. D., Franchi, G., & Sicuteri, F. (1977). Phantom Limb Pain: Sub-Hallucinogenic Treatment With Lysergic Acid Diethylamide (LSD-25). *Headache: The Journal of Head and Face Pain*, *17*(3), 118–119. https://doi.org/10.1111/j.1526-4610.1977.hed1703118.x

Feldhusen, J. F., & Goh, B. E. (1995). Assessing and accessing creativity: An integrative review of theory, research, and development. *Creativity research journal*, *8*(3), 231-247.

Filipowicz, A., Anderson, B., & Danckert, J. (2016). Adapting to change: The role of the right hemisphere in mental model building and updating. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *70*(3), 201–218. https://doi.org/10.1037/cep0000078

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*, *98*(1), 39–82.

Fogg, C., Michaels, T. I., De La Salle, S., Jahn, Z. W., & Williams, M. T. (2021). Ethnoracial health disparities and the ethnopsychopharmacology of psychedelic-assisted

psychotherapies. *Experimental and Clinical Psychopharmacology*, *29*(5), 539–554. https://doi.org/10.1037/pha0000490

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, *102*(27), 9673–9678.

Garcia-Romeu, A., Barrett, F. S., Carbonaro, T. M., Johnson, M. W., & Griffiths, R. R. (2021). Optimal dosing for psilocybin pharmacotherapy: Considering weight-adjusted and fixed dosing approaches. *Journal of Psychopharmacology*, 0269881121991822.

Garcia-Romeu, A., Davis, A. K., Erowid, E., Erowid, F., Griffiths, R. R., & Johnson, M. W. (2020). Persisting Reductions in Cannabis, Opioid, and Stimulant Misuse After Naturalistic Psychedelic Use: An Online Survey. *Frontiers in Psychiatry*, *10*, 955. https://doi.org/10.3389/fpsyt.2019.00955

Gelpi, R. A. (2021). *Hot, then cold: Predicting hypothesis revision strategies across development with rational process models*. University of Toronto (Canada).

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193.

Gerstenberg, T., & Tenenbaum, J. B. (2017). *Intuitive Theories*. Oxford University Press.

Giron, A. P., Ciranka, S., Schulz, E., Van Den Bos, W., Ruggeri, A., Meder, B., & Wu, C. M. (2023). Developmental changes in exploration resemble stochastic optimization. *Nature Human Behaviour*, *7*(11), 1955–1967. https://doi.org/10.1038/s41562-023-01662-1

Gittins, J. C., & Jones, D. M. (1979). A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, *66*(3), 561–565.

Gough, H. G. (1976). Studying creativity by means of word association tests. *Journal of Applied Psychology*, *61*(3), 348.

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2008). A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science*, *32*(1), 108–154. https://doi.org/10.1080/03640210701802071

Goodwin, G. M., Aaronson, S. T., Alvarez, O., Arden, P. C., Baker, A., Bennett, J. C., Bird, C., Blom, R. E., Brennan, C., Brusch, D., Burke, L., Campbell-Coker, K., Carhart-Harris, R., Cattell, J., Daniel, A., DeBattista, C., Dunlop, B. W., Eisen, K., Feifel, D., Forbes, M., Haumann, H., Hellerstein, D., Hoppe, A., Husain, M., Jelen, L., Kamphuis, J., Kawasaki, J., Kelly, J., Key, R., Kishon, R., Knatz Peck, S., Knight, G., Koolen, M., Lean, M., Licht, R., Maples-Keller, J., Mars, J., Marwood, L., McElhiney, M., Miller, T., Mirow, A., Mistry, S., Mletzko-Crowe, T., Modlin, L., Nielsen, R., Nielson, E., Offerhaus, S., O'Keane, V., Páleníček, T., Printz, D., Rademaker, M., van Reemst, A., Reinholdt, F.,

Repantis, D., Rucker, J., Rudow, S., Ruffell, S., Rush, J., Schoevers, R., Seynaeve, M., Shao, S., Soares, J., Somers, M., Stansfield, S., Sterling, D., Strockies, A., Tsai, J., Visser, L., Wahba, M., Williams, S., Young, A., Ywema, P., Zisook, S., Malievskaia, E. (2022). Single-Dose Psilocybin for a Treatment-Resistant Episode of Major Depression. *New England Journal of Medicine*, *387*(18), 1637–1648. https://doi.org/10.1056/NEJMoa2206443

Goodwin, G. M., Aaronson, S. T., Alvarez, O., Atli, M., Bennett, J. C., Croal, M., DeBattista, C., Dunlop, B. W., Feifel, D., Hellerstein, D. J., Husain, M. I., Kelly, J. R., Lennard-Jones, M. R., Licht, R. W., Marwood, L., Mistry, S., Páleníček, T., Redjep, O., Repantis, D., Schoevers, R., Septimus, B., Simmons, H., Soares, J., Somers, M., Stansfield, S., Stuart, J., Tadley, H., Thiara, N., Tsai, J., Wahba, M., Williams, S., Winzer, R., Young, A., Young, W., Zisook, S., Malievskaia, E. (2023). Single-dose psilocybin for a treatment-resistant episode of major depression: Impact on patient-reported depression severity, anxiety, function, and quality of life. *Journal of Affective Disorders*, *327*, 120–127. https://doi.org/10.1016/j.jad.2023.01.108

Goodwin, G. M., Croal, M., Feifel, D., Kelly, J. R., Marwood, L., Mistry, S., O'Keane, V., Peck, S. K., Simmons, H., Sisa, C., Stansfield, S. C., Tsai, J., Williams, S., & Malievskaia, E. (2023). Psilocybin for treatment resistant depression in patients taking a concomitant SSRI medication. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, *48*(10), 1492–1499. https://doi.org/10.1038/s41386-023-01648-7

Goolsby JR, T. M., & Helwig, L. D. (1975). Concurrent validity of the Torrance tests of creative thinking and the Welsh figural preference test. *Educational and Psychological Measurement*, *35*(2), 507-508.

Gopnik, A. (1996). The scientist as child. *Philosophy of Science*, *63*(4), 485–514.

Gopnik, A. (2009). *The philosophical baby: What children's minds tell us about truth, love & the meaning of life*. Random House.

Gopnik, A. (2012). Scientific Thinking in Young Children: Theoretical Advances, Empirical Research, and Policy Implications. *Science*, *337*(6102), 1623–1627. https://doi.org/10.1126/science.1223416

Gopnik, A. (2018, July 18). For Babies, Life May Be a Trip. *Wall Street Journal*. https://www.wsj.com/articles/for-babies-life-may-be-a-trip-1531932587

Gopnik, A. (2020). Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B*, *375*(1803), 20190502. https://doi.org/10.1098/rstb.2019.0502

Gopnik, A., (2024, July 24-27). *Explore, exploit, empower: Three ages and three intelligences*. The 46th Annual Meeting of the Cognitive Science Society, Rotterdam, The Netherlands.

Gopnik, A., Frankenhuis, W. E., & Tomasello, M. (2020). Introduction to special issue:'Life history and learning: how childhood, caregiving and old age shape cognition and culture in humans and other animals'. *Philosophical Transactions of the Royal Society B, 375*(1803), 20190489.

Gopnik, A., Griffiths, T. L., & Lucas, C. G. (2015). When Younger Learners Can Be Better (or at Least More Open-Minded) Than Older Ones. *Current Directions in Psychological Science*, *24*(2), 87–92. https://doi.org/10.1177/0963721414556653

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., Aboody, R., Fung, H., & Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, *114*(30), 7892–7899. https://doi.org/10.1073/pnas.1700811114

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms and the theory theory. *Psychological Bulletin*, *138*(6), 1085–1108. https://doi.org/10.1037/a0028044

Gouzoulis-Mayfrank, E., Thelen, B., Habermeyer, E., Kunert, H. J., Kovar, K.-A., Lindenblatt, H., Hermle, L., Spitzer, M., & Sass, H. (1999). Psychopathological, neuroendocrine and autonomic effects of 3, 4-methylenedioxyethylamphetamine (MDE), psilocybin and d-methamphetamine in healthy volunteers Results of an experimental double-blind placebo-controlled study. *Psychopharmacology*, *142*(1), 41–50.

Grant, J. A., & Zeidan, F. (2019). Employing pain and mindfulness to understand consciousness: A symbiotic relationship. *Current Opinion in Psychology*, *28*, 192–197. https://doi.org/10.1016/j.copsyc.2018.12.025

Griffiths, R. R., Johnson, M. W., Carducci, M. A., Umbricht, A., Richards, W. A., Richards, B. D., Cosimano, M. P., & Klinedinst, M. A. (2016). Psilocybin produces substantial and sustained decreases in depression and anxiety in patients with life-threatening cancer: A randomized double-blind trial. *Journal of Psychopharmacology*, *30*(12), 1181–1197. https://doi.org/10.1177/0269881116675513

Griffiths, R. R., Johnson, M. W., Richards, W. A., Richards, B. D., Jesse, R., MacLean, K. A., Barrett, F. S., Cosimano, M. P., & Klinedinst, M. A. (2018). Psilocybin-occasioned mystical-type experience in combination with meditation and other spiritual practices produces enduring positive changes in psychological functioning and in trait measures of prosocial attitudes and behaviors. *Journal of Psychopharmacology*, *32*(1), 49–69.

Griffiths, R. R., Johnson, M. W., Richards, W. A., Richards, B. D., McCann, U., & Jesse, R. (2011). Psilocybin occasioned mystical-type experiences: Immediate and persisting dose-related effects. *Psychopharmacology*, *218*(4), 649–665. https://doi.org/10.1007/s00213-011-2358-5

Griffiths, R. R., Richards, W. A., Johnson, M. W., McCann, U. D., & Jesse, R. (2008). Mystical-type experiences occasioned by psilocybin mediate the attribution of personal meaning and spiritual significance 14 months later. *Journal of Psychopharmacology (Oxford, England)*, *22*(6), 621–632. https://doi.org/10.1177/0269881108094300

Griffiths, R. R., Richards, W. A., McCann, U., & Jesse, R. (2006). Psilocybin can occasion mystical-type experiences having substantial and sustained personal meaning and spiritual significance. *Psychopharmacology*, *187*(3), 268–283.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging Levels of Analysis for Probabilistic Models of Cognition. *Current Directions in Psychological Science*, *21*(4), 263–268. https://doi.org/10.1177/0963721412447619

Grob, C. S., Danforth, A. L., Chopra, G. S., Hagerty, M., McKay, C. R., Halberstadt, A. L., & Greer, G. R. (2011). Pilot Study of Psilocybin Treatment for Anxiety in Patients With Advanced-Stage Cancer. *Archives of General Psychiatry*, *68*(1), 71. https://doi.org/10.1001/archgenpsychiatry.2010.116

Grof, S., Goodman, L. E., Richards, W. A., & Kurland, A. A. (1973). LSD-assisted psychotherapy in patients with terminal cancer. *International Pharmacopsychiatry*, *8*, 129–144.

Guilford, J. P. (1950). Creativity. *American Psychologist*, *5*(9), 444.

Guilford, J. P. (1956). The structure of intellect. *Psychological Bulletin*, *53*(4), 267.

Guilford, J. P. (1967). The nature of human intelligence.

Gukasyan, N., Davis, A. K., Barrett, F. S., Cosimano, M. P., Sepeda, N. D., Johnson, M. W., & Griffiths, R. R. (2022). Efficacy and safety of psilocybin-assisted treatment for major depressive disorder: Prospective 12-month follow-up. *Journal of Psychopharmacology (Oxford, England)*, *36*(2), 151–158. https://doi.org/10.1177/02698811211073759

Haddon, F. A., & Lytton, H. (1971). Primary education and divergent thinking abilities—four years on. *British Journal of Educational Psychology*, *41*(2), 136-147.

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Wedeen, V. J., & Sporns, O. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biology*, *6*(7), e159.

Harman, W. W., McKim, R. H., Mogar, R. E., Fadiman, J., & Stolaroff, M. J. (1966). Psychedelic agents in creative problem-solving: A pilot study. *Psychological reports*, *19*(1), 211-227.

Harrington, D. M. (1975). Effects of explicit instructions to "be creative" on the psychological meaning of divergent thinking test scores. *Journal of Personality*, *43*(3), 434–454. https://doi.org/10.1111/j.1467-6494.1975.tb00715.x

Hart, Y., Mayo, A. E., Mayo, R., Rozenkrantz, L., Tendler, A., Alon, U., & Noy, L. (2017). Creative foraging: An experimental paradigm for studying exploration and discovery. *PloS One*, *12*(8), e0182133.

Hartogsohn, I. (2016). Set and setting, psychedelics and the placebo response: An extra-pharmacological perspective on psychopharmacology. *Journal of Psychopharmacology*, *30*(12), 1259–1267. https://doi.org/10.1177/0269881116677852

Hartshorne, J. K., De Leeuw, J. R., Goodman, N. D., Jennings, M., & O'Donnell, T. J. (2019). A thousand studies for the price of one: Accelerating psychological science with Pushkin. *Behavior Research Methods*, *51*(4), 1782–1803. https://doi.org/10.3758/s13428-018-1155-z

Hass, R. W. (2017). Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications. *Memory & Cognition*, *45*(2), 233–244. https://doi.org/10.3758/s13421-016-0659-y

Heapy, A. A., Stroud, M. W., Higgins, D. M., & Sellinger, J. J. (2006). Tailoring cognitive-behavioral therapy for chronic pain: A case example. *Journal of Clinical Psychology*, *62*(11), 1345–1354. https://doi.org/10.1002/jclp.20314

Heckler, A. F., Kaminski, J. A., & Sloutsky, V. M. (2006). Differential cue salience, blocking and learned inattention. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 28, No. 28).

Hennessey, B. A., & Amabile, T. M. (2010). Creativity. *Annual Review of Psychology*, *61*, 569–598. https://doi.org/10.1146/annurev.psych.093008.100416

Herbst, E., Lucas, C., & Buchsbaum, D. (2017). Investigating the explore/exploit trade-off in adult causal inferences. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 39).

Hermle, L., Fünfgeld, M., Oepen, G., Botsch, H., Borchardt, D., Gouzoulis, E., Fehrenbach, R. A., & Spitzer, M. (1992). Mescaline-induced psychopathological, neuropsychological, and neurometabolic effects in normal subjects: Experimental psychosis as a tool for psychiatric research. *Biological Psychiatry*, *32*(11), 976–991.

Hocevar, D. (1981). Measurement of Creativity: Review and Critique. *Journal of Personality Assessment*, *45*(5), 450–464. https://doi.org/10.1207/s15327752jpa4505_1

Hood Jr, R. W. (1975). The construction and preliminary validation of a measure of reported mystical experience. *Journal for the Scientific Study of Religion*, 29–41.

Hood, R. W. (2005). Mystical, spiritual, and religious experiences. *Handbook of the Psychology of Religion and Spirituality*, 348–364.

Huang, P. S., Chen, H. C., and Liu, C. H. (2012). The development of Chinese word remote associates test for college students. *Psychological Testing, 59*, 581–607.

Hung, S. P., Huang, P. S., & Chen, H. C. (2016). Cognitive Complexity in the Remote Association Test - Chinese Version. *Creativity Research Journal*, *28*(4), 442–449. https://doi.org/10.1080/10400419.2016.1229988

Huttenlocher, P. R. (1990). Morphometric study of human cerebral cortex development. *Neuropsychologia*, *28*(6), 517–527.

Huttenlocher, P. R., & Dabholkar, A. S. (1997). Regional differences in synaptogenesis in human cerebral cortex. *The Journal of Comparative Neurology*, *387*(2), 167–178.

Isbell, H. (1959). Comparison of the reactions induced by psilocybin and LSD-25 in man. *Psychopharmacologia*, *1*(1), 29–38. https://doi.org/10.1007/BF00408109

Jaquish, G. A., & Ripple, R. E. (1985). A Life-Span Developmental Cross-Cultural Study of Divergent Thinking Abilities. *The International Journal of Aging and Human Development*, *20*(1), 1–11. https://doi.org/10.2190/RNJJ-NBD0-4A3K-0XPA

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Johns Hopkins University. (2023). *Effects of Psilocybin in Post-Treatment Lyme Disease* (Clinical Trial Registration NCT05305105). clinicaltrials.gov. https://clinicaltrials.gov/study/NCT05305105

Johnson, M. W., Garcia-Romeu, A., & Griffiths, R. R. (2017). Long-term Follow-up of Psilocybin-facilitated Smoking Cessation. *The American Journal of Drug and Alcohol Abuse*, *43*(1), 55–60. https://doi.org/10.3109/00952990.2016.1170135

Johnson, M. W., & Griffiths, R. R. (2017). Potential Therapeutic Effects of Psilocybin. *Neurotherapeutics*, *14*(3), 734–740. https://doi.org/10.1007/s13311-017-0542-y

Johnson, M. W., Richards, W. A., & Griffiths, R. R. (2008). Human hallucinogen research: Guidelines for safety. *Journal of Psychopharmacology*, *22*(6), 603–620.

Jones, G., Ricard, J. A., Lipson, J., & Nock, M. K. (2022). Associations between classic psychedelics and opioid use disorder in a nationally-representative U.S. adult sample. *Scientific Reports*, *12*(1), 4099. https://doi.org/10.1038/s41598-022-08085-4

Jordan, L. A. (1975). Use of canonical analysis in Cropley's "A five-year longitudinal study of the validity of creativity tests." *Developmental Psychology, 11*, 1-3

Kanter, S. (1984). *An investigation of the concurrent validity of selected subtests of Torrance Tests of Creative Thinking* [Unpublished doctoral dissertation]. University of Pittsburgh.

Karlins, M., Schuerhoff, C., & Kaplan, M. (1969). Some factors related to architectural creativity in graduating architecture students. *The Journal of General Psychology*, *81*(2), 203-215.

Kast, E. (1966). LSD and the dying patient. *The Chicago Medical School Quarterly*, *26*(2), 80–87.

Kast, E. (1967). Attenuation of anticipation: A therapeutic use of lysergic acid diethylamide. *The Psychiatric Quarterly*, *41*(4), 646–657. https://doi.org/10.1007/BF01575629

Kast, E. C., & Collins, V. J. (1964). Study of lysergic acid diethylamide as an analgesic agent. *Anesthesia & Analgesia*, *43*(3), 285–291.

Katz, A. N., & Poag, J. R. (1979). Sex differences in instructions to "be creative" on divergent and nondivergent test scores. *Journal of Personality*, *47*(3), 518–530. https://doi.org/10.1111/j.1467-6494.1979.tb00630.x

Kerns, J. G., Cohen, J. D., MacDonald, A. W., Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior Cingulate Conflict Monitoring and Adjustments in Control. *Science*, *303*(5660), 1023–1026. https://doi.org/10.1126/science.1089910

Kervadec, E., Fauvel, B., Strika-Bruneau, L., Amirouche, A., Verroust, V., Piolino, P., Romeo, B., & Benyamina, A. (2023). Reduction of alcohol use and increase in psychological flexibility after a naturalistic psychedelic experience: A retrospective survey. *Alcohol and Alcoholism*, *59*(2), agad078. https://doi.org/10.1093/alcalc/agad078

Kim, C., Johnson, N. F., & Gold, B. T. (2012). Common and Distinct Neural Mechanisms of Attentional Switching and Response Conflict. *Brain Research*, *1469*, 92–102. https://doi.org/10.1016/j.brainres.2012.06.013

Kinderlehrer, D. A. (2023). The Effectiveness of Microdosed Psilocybin in the Treatment of Neuropsychiatric Lyme Disease: A Case Study. *International Medical Case Reports Journal*, *Volume 16*, 109–115. https://doi.org/10.2147/IMCRJ.S395342

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science*, *220*(4598), 671–680.

Klein, D. C., Fencil-Morse, E., & Seligman, M. E. (1976). Learned helplessness, depression, and the attribution of failure. *Journal of personality and social psychology, 33*(5), 508.

Kogan, N., & Pankove, E. (1974). Long-term predictive validity of divergent-thinking tests: some negative evidence. *Journal of educational psychology*, *66*(6), 802.

Kometer, M., Pokorny, T., Seifritz, E., & Volleinweider, F. X. (2015). Psilocybin-induced spiritual experiences and insightfulness are associated with synchronization of neuronal oscillations. *Psychopharmacology*, *232*(19), 3663–3676. https://doi.org/10.1007/s00213-015-4026-7

Kraus, N., Niedeggen, M., & Hesselmann, G. (2021). Trait anxiety is linked to increased usage of priors in a perceptual decision making task. *Cognition*, *206*, 104474. https://doi.org/10.1016/j.cognition.2020.104474

Kuromaru, S., Okada, S., Hanada, M., Kasahara, Y., & Sakamoto, K. (1967). The effect of LSD on the phantom limb phenomenon. *The Journal-Lancet*, *87*(1), 22–27.

Kuypers, K. P. C. (2018). Out of the box: A psychedelic model to study the creative mind. *Medical Hypotheses*, *115*, 13–16. https://doi.org/10.1016/j.mehy.2018.03.010

Kuypers, K. P. C., Riba, J., de la Fuente Revenga, M., Barker, S., Theunissen, E. L., & Ramaekers, J. G. (2016). Ayahuasca enhances creative divergent thinking while decreasing conventional convergent thinking. *Psychopharmacology*, *233*(18), 3395–3403.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

Lapidow, E., & Walker, C. M. (2020). Informative experimentation in intuitive science: Children select and learn from their own causal interventions. *Cognition*, *201*, 104315. https://doi.org/10.1016/j.cognition.2020.104315

Laughlin, P. R., Doherty, M. A., & Dunn, R. P. (1968). Intentional and incidental concept formation as a function of motivation, creativity, intelligence, and sex. *Journal of Personality and Social Psychology, 8*, 401– 409. doi:10.1037/h0025522

Leary, T., Litwin, G. H., & Metzner, R. (1963). Reactions to psilocybjn administered in a supportive environment. *The Journal of Nervous and Mental Disease*, *137*(6), 561-573.

Lebedev, A. V., Acar, K., Horntvedt, O., Cabrera, A. E., Simonsson, O., Osika, W., Ingvar, M., & Petrovic, P. (2023). Alternative beliefs in psychedelic drug users. *Scientific Reports*, *13*(1), 16432. https://doi.org/10.1038/s41598-023-42444-z

Lebedev, A. V., Lövdén, M., Rosenthal, G., Feilding, A., Nutt, D. J., & Carhart-Harris, R. L. (2015). Finding the self by losing the self: Neural correlates of ego-dissolution under psilocybin: Finding the Self by Losing the Self. *Human Brain Mapping*, *36*(8), 3137–3153. https://doi.org/10.1002/hbm.22833

Lee, C. S., & Therriault, D. J. (2013). The cognitive underpinnings of creative thought: A latent variable analysis exploring the role of intelligence and working memory in three creative thinking processes. *Intelligence, 41*, 306 –320. doi:10.1016/j.intell.2013.04.008

Lee Bae, C., Huggins-Manley, A., & Therriault, D. (2014). A Measure of Creativity or Intelligence? Examining Internal and External Structure Validity Evidence of the Remote Associates Test. *Psychology of Aesthetics Creativity and the Arts*, *8*. https://doi.org/10.1037/a0036773

Lee, J. K., & Orsillo, S. M. (2014). Investigating cognitive flexibility as a potential mechanism of mindfulness in generalized anxiety disorder. *Journal of Behavior Therapy and Experimental Psychiatry*, *45*(1), 208–216.

Leibenluft, E., & Pine, D. S. (2013). Resting State Functional Connectivity and Depression: In Search of a Bottom Line. *Biological Psychiatry*, *74*(12), 868–869. https://doi.org/10.1016/j.biopsych.2013.10.001

Lemons, G. (2011). Diverse Perspectives of Creativity Testing: Controversial Issues When Used for Inclusion Into Gifted Programs. *Journal for the Education of the Gifted*, *34*(5), 742–772. https://doi.org/10.1177/0162353211417221

Lewis, C. R., Preller, K. H., Kraehenmann, R., Michels, L., Staempfli, P., & Vollenweider, F. X. (2017). Two dose investigation of the 5-HT-agonist psilocybin on relative and global cerebral blood flow. *NeuroImage*, *159*, 70–78. https://doi.org/10.1016/j.neuroimage.2017.07.020

Linville, T. M. (2016). Project MKULTRA and the search for mind control: Clandestine use of LSD within the CIA. https://digitalcommons.cedarville.edu/history_capstones/6/

Liquin, E., & Gopnik, A. (2022). Children are more exploratory and learn more than adults in an approach-avoid task. *Cognition*, *218*, 104940. https://doi.org/10.1016/j.cognition.2021.104940

Long, H. (2014). An Empirical Review of Research Methodologies and Methods in Creativity Studies (2003–2012). *Creativity Research Journal*, *26*(4), 427–438. https://doi.org/10.1080/10400419.2014.961781

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299. https://doi.org/10.1016/j.cognition.2013.12.010

Lucas, C. G., & Griffiths, T. L. (2010). Learning the Form of Causal Relationships Using Hierarchical Bayesian Models. *Cognitive Science*, *34*(1), 113–147. https://doi.org/10.1111/j.1551-6709.2009.01058.x

Ly, C., Greb, A. C., Cameron, L. P., Wong, J. M., Barragan, E. V., Wilson, P. C., Burbach, K. F., Zarandi, S. S., Sood, A., & Paddy, M. R. (2018). Psychedelics promote structural and functional neural plasticity. *Cell Reports*, *23*(11), 3170–3182.

MacLean, K. A., Johnson, M. W., & Griffiths, R. R. (2011). Mystical experiences occasioned by the hallucinogen psilocybin lead to increases in the personality domain of openness. *Journal of Psychopharmacology*, *25*(11), 1453–1461.

Malitz, S., Esecover, H., Wilkens, B., & Hoch, P. H. (1960). Some observations on psilocybin, a new hallucinogen, in volunteer subjects. *Comprehensive Psychiatry*.

Mark von Zastrow, M. D. (2018). Drug receptors & pharmacodynamics. *Basic & Clinical Pharmacology*, 15.

Marschall, J., Fejer, G., Lempe, P., Prochazkova, L., Kuchar, M., Hajkova, K., & van Elk, M. (2022). Psilocybin microdosing does not affect emotion-related symptoms and processing: A preregistered field and lab-based study. *Journal of Psychopharmacology*, *36*(1), 97–113. https://doi.org/10.1177/02698811211050556

Martin, J. D., Blair, G. E., & Herrmann, W. J. (1981). Correlations between Scores on Torrance Tests of Creative Thinking and Ingenuity Subtest of the Flanagan Aptitude Classification Tests. *Psychological Reports, 48*(1), 195-198.

Martin, M. M., & Rubin, R. B. (1995). A New Measure of Cognitive Flexibility. *Psychological Reports*, *76*(2), 623–626. https://doi.org/10.2466/pr0.1995.76.2.623

Mason, N. L., Kuypers, K. P. C., Reckweg, J. T., Müller, F., Tse, D. H. Y., Da Rios, B., Toennes, S. W., Stiers, P., Feilding, A., & Ramaekers, J. G. (2021). Spontaneous and deliberate creative cognition during and after psilocybin exposure. *Translational Psychiatry*, *11*(1). https://doi.org/10.1038/s41398-021-01335-5

Mason, N. L., Mischler, E., Uthaug, M. V., & Kuypers, K. P. C. (2019). Sub-Acute Effects of Psilocybin on Empathy, Creative Thinking, and Subjective Well-Being. *Journal of Psychoactive Drugs*, *51*(2), 123–134. https://doi.org/10.1080/02791072.2019.1580804

Mason, N. L., Szabo, A., Kuypers, K. P. C., Mallaroni, P. A., De La Torre Fornell, R., Reckweg, J. T., Tse, D. H. Y., Hutten, N. R. P. W., Feilding, A., & Ramaekers, J. G. (2023). Psilocybin induces acute and persisting alterations in immune status in healthy volunteers: An experimental, placebo-controlled study. *Brain, Behavior, and Immunity*, *114*, 299–310. https://doi.org/10.1016/j.bbi.2023.09.004

McCaffrey, R. J., & Westervelt, H. J. (1995). Issues associated with repeated neuropsychological assessments. *Neuropsychology Review*, *5*(3), 203–221. https://doi.org/10.1007/BF02214762

McCrae, R. R. (1987). Creativity, Divergent Thinking, and Openness to Experience. *Journal of Personality and Social Psychology*, *52*(6), 1258–1265.

McGovern, H. T., Grimmer, H. J., Doss, M., Hutchinson, B., Timmermann, C., Lyon, A., Corlett, P., Laukkonen, R. (2023). The power of insight: How psychedelics solicit false beliefs. *PsyArXiv preprint: https://psyarxiv. com/97gjw*.

McKenna, T. (1999). *Food of the gods: The search for the original tree of knowledge: a radical history of plants, drugs and human evolution*. Random House.

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., & Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3), 191–215. https://doi.org/10.1037/dec0000033

Mennen, A. C., Norman, K. A., & Turk-Browne, N. B. (2019). Attentional bias in depression: Understanding mechanisms to improve training and treatment. *Current Opinion in Psychology*, *29*, 266–273. https://doi.org/10.1016/j.copsyc.2019.07.036

Michaels, T. I., Purdon, J., Collins, A., & Williams, M. T. (2018). Inclusion of people of color in psychedelic-assisted psychotherapy: A review of the literature. *BMC Psychiatry*, *18*(1), 245. https://doi.org/10.1186/s12888-018-1824-6

Milders, M., Ietswaart, M., Crawford, J. R., & Currie, D. (2008). Social behavior following traumatic brain injury and its association with emotion recognition, understanding of intentions, and cognitive flexibility. *Journal of the International Neuropsychological Society, 14*(2), 318-326.

Miller, W. R., & Seligman, M. E. (1975). Depression and learned helplessness in man. *Journal of Abnormal Psychology, 84*(3), 228.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Moore, A., & Malinowski, P. (2009). Meditation, mindfulness and cognitive flexibility. *Consciousness and Cognition*, *18*(1), 176–186. https://doi.org/10.1016/j.concog.2008.12.008

Muradoglu, M., Cimpian, J. R., & Cimpian, A. (2023). Mixed-Effects Models for Cognitive Development Researchers. *Journal of Cognition and Development*, *24*(3), 307–340. https://doi.org/10.1080/15248372.2023.2176856

Newberg, A. B., & d'Aquili, E. G. (2000). The neuropsychology of religious and spiritual experience. *Journal of Consciousness Studies*, *7*(11–12), 251–266.

Nichols, D. E. (2004). Hallucinogens. *Pharmacology & Therapeutics*, *101*(2), 131–181. https://doi.org/10.1016/j.pharmthera.2003.11.002

NIH launches HEAL Initiative, doubles funding to accelerate scientific solutions to stem national opioid epidemic. (2018, April 3). National Institutes of Health (NIH). https://www.nih.gov/news-events/news-releases/nih-launches-heal-initiative-doubles-funding-accelerate-scientific-solutions-stem-national-opioid-epidemic

Ninnemann, K. M. (2012). Variability in the Efficacy of Psychopharmaceuticals: Contributions from Pharmacogenomics, Ethnopsychopharmacology, and Psychological and Psychiatric Anthropologies. *Culture, Medicine, and Psychiatry*, *36*(1), 10–25. https://doi.org/10.1007/s11013-011-9242-y

Noorani, T., Garcia-Romeu, A., Swift, T. C., Griffiths, R. R., & Johnson, M. W. (2018). Psychedelic therapy for smoking cessation: Qualitative analysis of participant accounts. *Journal of Psychopharmacology*, *32*(7), 756–769. https://doi.org/10.1177/0269881118780612

Nour, M. M., & Carhart-Harris, R. L. (2017). Psychedelics and the science of self-experience. *British Journal of Psychiatry*, *210*(03), 177–179. https://doi.org/10.1192/bjp.bp.116.194738

Nour, M. M., Evans, L., Nutt, D., & Carhart-Harris, R. L. (2016). Ego-Dissolution and Psychedelics: Validation of the Ego-Dissolution Inventory (EDI). *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00269

O'Donnell, K. C., Mennenga, S. E., Owens, L. T., Podrebarac, S. K., Baron, T., Rotrosen, J., Ross, S., Forcehimes, A. A., & Bogenschutz, M. P. (2022). Psilocybin for alcohol use disorder: Rationale and design considerations for a randomized controlled trial. *Contemporary Clinical Trials*, *123*, 106976. https://doi.org/10.1016/j.cct.2022.106976

Ott, J. (1996). Entheogens II: On entheology and entheobotany. *Journal of Psychoactive Drugs*, *28*(2), 205–209.

Pahnke, W. N. (1969). Psychedelic drugs and mystical experience. *International Psychiatry Clinics*, *5*(4), 149–162.

Palm, K. M., & Follette, V. M. (2011). The roles of cognitive flexibility and experiential avoidance in explaining psychological distress in survivors of interpersonal victimization. *Journal of Psychopathology and Behavioral Assessment*, *33*(1), 79–86.

Pang, E. W., Dunkley, B. T., Doesburg, S. M., Da Costa, L., & Taylor, M. J. (2016). Reduced brain connectivity and mental flexibility in mild traumatic brain injury. *Annals of Clinical and Translational Neurology*, *3*(2), 124–131. https://doi.org/10.1002/acn3.280

Parker, J. (1980). *The predictive validity of creativity and intelligence tests administered at age five* [Unpublished doctoral dissertation]. University of Georgia.

Peck, S. K., Shao, S., Gruen, T., Yang, K., Babakanian, A., Trim, J., Finn, D. M., & Kaye, W. H. (2023). Psilocybin therapy for females with anorexia nervosa: A phase 1, open-label feasibility study. *Nature Medicine*, *29*(8), 1947–1953. https://doi.org/10.1038/s41591-023-02455-9

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://aclanthology.org/D14-1162.pdf

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*(3), 302–321.

Pisano, V. D., Putnam, N. P., Kramer, H. M., Franciotti, K. J., Halpern, J. H., & Holden, S. C. (2017). The association of psychedelic use and opioid use disorders among illicit users in the United States. *Journal of Psychopharmacology*, *31*(5), 606–613. https://doi.org/10.1177/0269881117691453

Plebanek, D. J., & Sloutsky, V. M. (2017). Costs of Selective Attention: When Children Notice What Adults Miss. *Psychological Science*, *28*(6), 723–732. https://doi.org/10.1177/0956797617693005

Popperova, M. (1971). Some methodological problems relating to tests of creative thinking. *Czskoslovenská Psychologie, 15*(4), 39 1-397

Preller, K. H., Duerler, P., Burt, J. B., Ji, J. L., Adkinson, B., Stämpfli, P., Seifritz, E., Repovš, G., Krystal, J. H., Murray, J. D., Anticevic, A., & Vollenweider, F. X. (2020). Psilocybin Induces Time-Dependent Changes in Global Functional Connectivity. *Biological Psychiatry*, *88*(2), 197–207. https://doi.org/10.1016/j.biopsych.2019.12.027

Price, D. D., Harkins, S. W., & Baker, C. (1987). Sensory-affective relationships among different types of clinical and experimental pain. *Pain*, *28*(3), 297–307.

Raison, C. L., Sanacora, G., Woolley, J., Heinzerling, K., Dunlop, B. W., Brown, R. T., Kakar, R., Hassman, M., Trivedi, R. P., Robison, R., Gukasyan, N., Nayak, S. M., Hu, X., O'Donnell, K. C., Kelmendi, B., Sloshower, J., Penn, A. D., Bradley, E., Kelly, D. F., Kelmendi, B., Sloshower, J., Penn, A., Bradley, E., Kelly, D., Mletzko, T., Nicholas, C., Hutson, P., Tarpley, G., Utzinger, M., Lenoch, K., Warchol, K., Gapasin, T., Davis, M., Nelson-Douthit, C., Wilson, S., Brown, C., Linton, W., Johnson, M., Ross, S., Griffiths,

R. R. (2023). Single-Dose Psilocybin Treatment for Major Depressive Disorder: A Randomized Clinical Trial. *JAMA*, *330*(9), 843. https://doi.org/10.1001/jama.2023.14530

Ramachandran, V., Chunharas, C., Marcus, Z., Furnish, T., & Lin, A. (2018). Relief from intractable phantom pain by combining psilocybin and mirror visual-feedback (MVF). *Neurocase*, *24*(2), 105–110.

Ramaekers, J. G., Hutten, N., Mason, N. L., Dolder, P., Theunissen, E. L., Holze, F., Liechti, M. E., Feilding, A., & Kuypers, K. P. (2021). A low dose of lysergic acid diethylamide decreases pain perception in healthy volunteers. *Journal of Psychopharmacology*, *35*(4), 398–405.

Reverberi, C., Toraldo, A., D'Agostini, S., & Skrap, M. (2005). Better without (lateral) frontal cortex? Insight problems solved by frontal patients. *Brain*, *128*(12), 2882–2890. https://doi.org/10.1093/brain/awh577

Rhodes, M. (1961). An analysis of creativity. *The Phi Delta Kappan*, *42*(7), 305–310.

Rich, A. (2018). *Causes and Consequences of Exploratory Choice*. New York University.

Rich, A. S., & Gureckis, T. M. (2015). The Attentional Learning Trap and How to Avoid It. *Proceedings of the 37th Annual Conference of the Cognitive Science Society, 1973-1978*. Pasadena, CA: Cognitive Science Society.

Rich, A. S., & Gureckis, T. M. (2018). The limits of learning: Exploration, generalization, and the development of learning traps. *Journal of Experimental Psychology: General*, *147*(11), 1553.

Richards, W. A., Rhead, J. C., DiLeo, F. B., Yensen, R., & Kurland, A. A. (1977). The peak experience variable in DPT-assisted psychotherapy with cancer patients. *Journal of Psychedelic Drugs*, *9*(1), 1–10.

Roseman, L., Leech, R., Feilding, A., Nutt, D. J., & Carhart-Harris, R. L. (2014). The effects of psilocybin and MDMA on between-network resting state functional connectivity in healthy volunteers. *Frontiers in Human Neuroscience*, *8*, 204.

Roseman, L., Nutt, D. J., & Carhart-Harris, R. L. (2018). Quality of acute psychedelic experience predicts therapeutic efficacy of psilocybin for treatment-resistant depression. *Frontiers in Pharmacology*, *8*, 974.

Ross, S., Bossis, A., Guss, J., Agin-Liebes, G., Malone, T., Cohen, B., Mennenga, S. E., Belser, A., Kalliontzi, K., & Babb, J. (2016). Rapid and sustained symptom reduction following psilocybin treatment for anxiety and depression in patients with life-threatening cancer: A randomized controlled trial. *Journal of Psychopharmacology*, *30*(12), 1165–1180.

Roweton, W. E., Farless, J. E., Donham, R., Wleklinski, D. J., & Spencer, H. L. (1975). Indices of classroom creativity. *Child Study Journal, 5*, 151-162.

Said-Metwaly, S., Noortgate, W. V. D., & Kyndt, E. (2017). Approaches to Measuring Creativity: A Systematic Literature Review. *Creativity. Theories – Research - Applications*, *4*(2), 238–275. https://doi.org/10.1515/ctra-2017-0013

Sampedro, F., de la Fuente Revenga, M., Valle, M., Roberto, N., Domínguez-Clavé, E., Elices, M., Luna, L. E., Crippa, J. A. S., Hallak, J. E. C., de Araujo, D. B., Friedlander, P., Barker, S. A., Álvarez, E., Soler, J., Pascual, J. C., Feilding, A., & Riba, J. (2017). Assessing the Psychedelic "After-Glow" in Ayahuasca Users: Post-Acute Neurometabolic and Functional Connectivity Changes Are Associated with Enhanced Mindfulness Capacities. *International Journal of Neuropsychopharmacology*, *20*(9), 698–711. https://doi.org/10.1093/ijnp/pyx036

Sanborn, A. N. (2017). Types of approximation for probabilistic cognition: Sampling and variational. *Brain and Cognition*, *112*, 98–101.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167. https://doi.org/10.1037/a0020511

Schindler, E. A. D., Sewell, R. A., Gottschalk, C. H., Luddy, C., Flynn, L. T., Lindsey, H., Pittman, B. P., Cozzi, N. V., & D'Souza, D. C. (2021). Exploratory Controlled Study of the Migraine-Suppressing Effects of Psilocybin. *Neurotherapeutics: The Journal of the American Society for Experimental NeuroTherapeutics*, *18*(1), 534–543. https://doi.org/10.1007/s13311-020-00962-y

Schindler, E. A. D., Sewell, R. A., Gottschalk, C. H., Luddy, C., Flynn, L. T., Zhu, Y., Lindsey, H., Pittman, B. P., Cozzi, N. V., & D'Souza, D. C. (2022). Exploratory investigation of a patient-informed low-dose psilocybin pulse regimen in the suppression of cluster headache: Results from a randomized, double-blind, placebo-controlled trial. *Headache*, *62*(10), 1383–1394. https://doi.org/10.1111/head.14420

Schindler, E. A., Gottschalk, C. H., Weil, M. J., Shapiro, R. E., Wright, D. A., & Sewell, R. A. (2015). Indoleamine hallucinogens in cluster headache: Results of the Clusterbusters medication use survey. *Journal of Psychoactive Drugs*, *47*(5), 372–381.

Schneier, F. R., Feusner, J., Wheaton, M. G., Gomez, G. J., Cornejo, G., Naraindas, A. M., & Hellerstein, D. J. (2023). Pilot study of single-dose psilocybin for serotonin reuptake inhibitor-resistant body dysmorphic disorder. *Journal of Psychiatric Research*, *161*, 364–370. https://doi.org/10.1016/j.jpsychires.2023.03.031

Schultes, R. E. (1940). Teonanacatl: The Narcotic Mushroom of the Aztecs. *American Anthropologist*, *42*(3), 429–443.

Schultes, R. E., Hofmann, A., & Rätsch, C. (2001). *Plants of the gods: Their sacred, healing, and hallucinogenic powers*. Healing Arts Press Rochester, VT.

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2018). Putting bandits into context: How function learning supports decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(6), 927–943. https://doi.org/10.1037/xlm0000463

Schulz, E., Wu, C. M., Ruggeri, A., & Meder, B. (2019). Searching for rewards like a child means less generalization and more directed exploration. *Psychological science*, *30*(11), 1561-1572.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*(2), 93.

Scott, W. A. (1962). Cognitive Complexity and Cognitive Flexibility. *Sociometry*, *25*(4), 405–414. https://doi.org/10.2307/2785779

Seiver, E., Gopnik, A., & Goodman, N. D. (2013). Did She Jump Because She Was the Big Sister or Because the Trampoline Was Safe? Causal Inference and the Development of Social Attribution. *Child Development*, *84*(2), 443–454. https://doi.org/10.1111/j.1467-8624.2012.01865.x

Seminowicz, D. A., Mayberg, H. S., McIntosh, A. R., Goldapple, K., Kennedy, S., Segal, Z., & Rafi-Tari, S. (2004). Limbic–frontal circuitry in major depression: A path modeling metanalysis. *NeuroImage*, *22*(1), 409–418. https://doi.org/10.1016/j.neuroimage.2004.01.015

Sessa, B. (2020). *The psychedelic renaissance: Reassessing the role of psychedelic drugs in 21st century psychiatry and society*. Aeon Books.

Sewell, R. A., Halpern, J. H., & Pope, H. G. (2006). Response of cluster headache to psilocybin and LSD. *Neurology*, *66*(12), 1920–1922.

Sheline, Y. I., Barch, D. M., Price, J. L., Rundle, M. M., Vaishnavi, S. N., Snyder, A. Z., Mintun, M. A., Wang, S., Coalson, R. S., & Raichle, M. E. (2009). The default mode network and self-referential processes in depression. *Proceedings of the National Academy of Sciences*, *106*(6), 1942–1947. https://doi.org/10.1073/pnas.0812686106

Sherman, L. E., Rudie, J. D., Pfeifer, J. H., Masten, C. L., McNealy, K., & Dapretto, M. (2014). Development of the Default Mode and Central Executive Networks across early adolescence: A longitudinal study. *Developmental Cognitive Neuroscience*, *10*, 148–159. https://doi.org/10.1016/j.dcn.2014.08.002

Shute, V. J. (2007). Focus on formative feedback. *ETS Research Report Series*, (1), i-47.

Siegel, J. S., Subramanian, S., Perry, D., Kay, B. P., Gordon, E. M., Laumann, T. O., Reneau, T., Metcalf, N., Chacko, R., Gratton, C., Horan, C., Krimmel, S., Shimony, J., Schweiger, J., Wong, D., Bender, D., Scheidter, K., Whiting, F., Padawer-Curry, J., Shinohara, R., Chen, Y., Moser, J., Yacoub, E., Nelson, S., Vizioli, L., Fair, D., Lenze, E., Carhart-Harris, R., Raison, C., Raichle, M., Snyder, A., & Dosenbach, N. U. (2024). Psilocybin desynchronizes the human brain. *Nature*, 1-8.

Siegler, R. & Jenkins, E. (2014). *How children discover new strategies*. Psychology Press.

Silvia, P. J., Nusbaum, E. C., Berg, C., Martin, C., & O'Connor, A. (2009). Openness to experience, plasticity, and creativity: Exploring lower-order, high-order, and interactive effects. *Journal of Research in Personality*, *43*(6), 1087–1090. https://doi.org/10.1016/j.jrp.2009.04.015

Skager, R. W., Klein, S. P., & Schultz, C. B. (1967). The prediction of academic and artistic achievement at a school of design. *Journal of Educational Measurement*, *4*(2), 105-117.

Sloshower, J., Skosnik, P. D., Safi-Aghdam, H., Pathania, S., Syed, S., Pittman, B., & D'Souza, D. C. (2023). Psilocybin-assisted therapy for major depressive disorder: An exploratory placebo-controlled, fixed-order trial. *Journal of Psychopharmacology*, *37*(7), 698–706. https://doi.org/10.1177/02698811231154852

Smigielski, L., Scheidegger, M., Kometer, M., & Vollenweider, F. X. (2019). Psilocybin-assisted mindfulness training modulates self-consciousness and brain default mode network connectivity with lasting effects. *NeuroImage*, *196*, 207–215. https://doi.org/10.1016/j.neuroimage.2019.04.009

Smith, H. (1964). Do drugs have religious import? *The Journal of Philosophy*, *61*(18), 517–530.

Smith, K. A., Huber, D. E., & Vul, E. (2013). Multiply-constrained semantic search in the Remote Associates Test. *Cognition*, *128*(1), 64–75. https://doi.org/10.1016/j.cognition.2013.03.001

Snell-Rood, E. C. (2013). An overview of the evolutionary causes and consequences of behavioural plasticity. *Animal Behaviour*, *85*(5), 1004–1011.

Spitzer, M., Thimm, M., Hermle, L., Holzmann, P., Kovar, K.-A., Heimann, H., Gouzoulis-Mayfrank, E., Kischka, U., & Schneider, F. (1996). Increased activation of indirect semantic associations under psilocybin. *Biological Psychiatry*, *39*(12), 1055–1057. https://doi.org/10.1016/0006-3223(95)00418-1

Srinivas, N., Krause, A., Kakade, S. M., & Seeger, M. (2012). Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. *IEEE Transactions on Information Theory*, *58*(5), 3250–3265. https://doi.org/10.1109/TIT.2011.2182033

Stamets, P. (1996). *Psilocybin mushrooms of the world: An identification guide*. Ten Speed Press Berkeley, CA.

Strassman, R. (2014). *DMT and the soul of prophecy: A new science of spiritual revelation in the Hebrew Bible*. Simon and Schuster.

Strickland, J. C., Garcia-Romeu, A., & Johnson, M. W. (2021). Set and Setting: A Randomized Study of Different Musical Genres in Supporting Psychedelic Therapy. *ACS Pharmacology & Translational Science*, *4*(2), 472–478. https://doi.org/10.1021/acsptsci.0c00187

Suchow, J. W., Griffiths, T., & Hartshorne, J. K. (2020). Workshop on Scaling Cognitive Science. *CogSci*. https://cognitivesciencesociety.org/cogsci20/papers/0006/0006.pdf

Sumner, E., Li, A. X., Perfors, A., Hayes, B., Navarro, D., & Sarnecka, B. W. (2019). The Exploration Advantage: Children's instinct to explore allows them to find information that adults miss. https://doi.org/10.31234/osf.io/h437v

Sumner, E., Steyvers, M., & Sarnecka, B. W. (2019). It's not the treasure, it's the hunt: Children are more explorative on an explore/exploit task than adults. *Proceedings of the 41st Annual Conference of the Cognitive Science Society,* 2891–2897. Montreal, CA: Cognitive Science Society.

Supekar, K., Uddin, L. Q., Prater, K., Amin, H., Greicius, M. D., & Menon, V. (2010). Development of functional and structural connectivity within the default mode network in young children. *NeuroImage*, *52*(1), 290–301. https://doi.org/10.1016/j.neuroimage.2010.04.009

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Bradford Book.

Tagliazucchi, E., Carhart-Harris, R. L., Leech, R., Nutt, D., & Chialvo, D. R. (2014). Enhanced repertoire of brain dynamical states during the psychedelic experience. *Human Brain Mapping*, *35*(11), 5442–5456.

Tagliazucchi, E., Roseman, L., Kaelen, M., Orban, C., Muthukumaraswamy, S. D., Murphy, K., Laufs, H., Leech, R., McGonigle, J., Crossley, N., Bullmore, E., Williams, T., Bolstridge, M., Feilding, A., Nutt, D. J., & Carhart-Harris, R. (2016). Increased Global Functional Connectivity Correlates with LSD-Induced Ego Dissolution. *Current Biology*, *26*(8), 1043–1050. https://doi.org/10.1016/j.cub.2016.02.010

Taft, R., & Rossiter, J. (1966). The Remote Associates Test: Divergent or convergent thinking? *Psychological Reports, 19*, 1313–1314. https://doi.org/10.2466/pr0.1966.19.3f.1313

Taylor, C. W., Smith, W. R., & Ghiselin, B. (1963). The creative and other contributions of one sample of research scientists. *Scientific Creativity: Its recognition and development*. New York: Wiley.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, *115*(1), 155.

Thompson-Schill, S. L., Ramscar, M., & Chrysikou, E. G. (2009). Cognition Without Control: When a Little Frontal Lobe Goes a Long Way. *Current Directions in Psychological Science*, *18*(5), 259–263. https://doi.org/10.1111/j.1467-8721.2009.01648.x

Todd, A. R., Forstmann, M., Burgmer, P., Brooks, A. W., & Galinsky, A. D. (2015). Anxious and egocentric: How specific emotions influence perspective taking. *Journal of Experimental Psychology: General*, *144*(2), 374.

Torrado Pacheco, A., Olson, R. J., Garza, G., & Moghaddam, B. (2023). Acute psilocybin enhances cognitive flexibility in rats. *Neuropsychopharmacology*, *48*(7). https://doi.org/10.1038/s41386-023-01545-z

Torrance, E. P. (1966). Torrance tests of creative thinking. *Educational and Psychological Measurement*. https://psycnet.apa.org/doiLanding?doi=10.1037/t05532-000

Uddin, L. Q., Supekar, K. S., Ryali, S., & Menon, V. (2011). Dynamic Reconfiguration of Structural and Functional Connectivity Across Core Neurocognitive Brain Networks with Development. *Journal of Neuroscience*, *31*(50), 18578–18589. https://doi.org/10.1523/JNEUROSCI.4465-11.2011

University of Wisconsin, Madison. (2024). *Phase I Study of the Safety and Adjunctive Effects of Psilocybin in Adults With Opioid Use Disorder Maintained on a Buprenorphine/Naloxone Formulation* (Clinical Trial Registration NCT04161066). clinicaltrials.gov. https://clinicaltrials.gov/study/NCT04161066

Vadivelu, N., Kai, A. M., Kodumudi, V., Sramcik, J., & Kaye, A. D. (2018). The Opioid Crisis: A Comprehensive Overview. *Current Pain and Headache Reports*, *22*(3), 16. https://doi.org/10.1007/s11916-018-0670-z

Vaidya, V. A., Marek, G. J., Aghajanian, G. K., & Duman, R. S. (1997). 5-HT2A receptor-mediated regulation of brain-derived neurotrophic factor mRNA in the hippocampus and the neocortex. *Journal of Neuroscience*, *17*(8), 2785–2795.

Van Amsterdam, J., & Van Den Brink, W. (2022). The therapeutic potential of psilocybin: A systematic review. *Expert Opinion on Drug Safety*, *21*(6), 833–840. https://doi.org/10.1080/14740338.2022.2047929

Vollenweider, F. X., Leenders, K. L., Scharfetter, C., Maguire, P., Stadelmann, O., & Angst, J. (1997). Positron emission tomography and fluorodeoxyglucose studies of metabolic hyperfrontality and psychopathology in the psilocybin model of psychosis. *Neuropsychopharmacology*, *16*(5), 357–372.

Vollenweider, F. X., & Preller, K. H. (2020). Psychedelic drugs: Neurobiology and potential for treatment of psychiatric disorders. *Nature Reviews Neuroscience*, *21*(11), 611–624. https://doi.org/10.1038/s41583-020-0367-2

Vollenweider, F. X., Vollenweider-Scherpenhuyzen, M. F. I., Bäbler, A., Vogel, H., & Hell, D. (1998). Psilocybin induces schizophrenia-like psychosis in humans via a serotonin-2 agonist action: *NeuroReport*, *9*(17), 3897–3902. https://doi.org/10.1097/00001756-199812010-00024

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and Done? Optimal Decisions From Very Few Samples. *Cognitive Science*, *38*(4), 599–637. https://doi.org/10.1111/cogs.12101

Vul, E., & Pashler, H. (2008). Measuring the Crowd Within: Probabilistic Representations Within Individuals. *Psychological Science*, *19*(7), 645–647. https://doi.org/10.1111/j.1467-9280.2008.02136.x

Walker, C., Rett, A., & Bonawitz, E. (2020). Design drives discovery in causal learning. *Psychological Science, 31*(2), 129-138.

Wallach, M. A., & Kogan, N. (1965). A new look at the creativity-intelligence distinction. *Journal of Personality*, *33*(3), 348–369. https://doi.org/10.1111/j.1467-6494.1965.tb01391.x

Wasson, R. G. (1980). *The wondrous mushroom: Mycolatry in Mesoamerica*. McGraw-Hill New York.

Weisbecker, V., & Goswami, A. (2010). Brain size, life history, and metabolism at the marsupial/placental dichotomy. *Proceedings of the National Academy of Sciences*, *107*(37), 16216–16221. https://doi.org/10.1073/pnas.0906486107

Whiting, D. L., Deane, F. P., Simpson, G. K., McLeod, H. J., & Ciarrochi, J. (2017). Cognitive and psychological flexibility after a traumatic brain injury and the implications for treatment in acceptance-based therapies: A conceptual review. *Neuropsychological rehabilitation, 27*(2), 263-299.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915–924.

Wu, C.-L., Huang, S.-Y., Chen, P.-Z., & Chen, H.-C. (2020). A Systematic Review of Creativity-Related Studies Applying the Remote Associates Test From 2000 to 2019. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.573432

Yim, H., Best, C., & Sloutsky, V. (2011). Cost of attention as an indicator of category learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature Protocols*, *1*(1), 297–301.

Appendix A

Contexto instruction:

"How to play:
On each trial, there is a randomly selected secret word.
Your goal is to try and figure out that secret word.

The secret word will always be a noun. You have a maximum of 30 guesses per trial.

The words were sorted by an artificial intelligence algorithm according to how similar they were to the secret word.

After submitting a word, you will see its position. The secret word is number 1.

You will earn points based on the closest word you can get on each trial.
If you get the secret word, you will earn a bonus for that trial.

*At the start of each trial you will be given one word to get you started. This word is randomly selected, so it may be close to the target or far away."

* Only included in the version that provided a start word.

Contexto Word Lists

Experiment 1:

Target	Close	Medium	Far
cookie	dessert	yogurt	penny
flower	orchid	girl	spider
barbecue	propane	pasta	boat
tree	vine	natural	hunt
moose	reindeer	frog	sculpture
pancake	oatmeal	tortilla	queen
camera	video	microscope	machine
car	taxi	toy	summer
pencil	ink	clay	latte

Experiment 2:
Set 1 -- 'flower', 'moose', 'cookie', 'pencil', 'car'
Set 2 -- 'barbecue', 'pancake', 'camera', 'banana', 'tuna'
Set 3 -- "pumpkin", "mushroom", "aquarium", "frog", 'clock'

Appendix B

Phantom Limb Pain Study Measures:

Pain ratings – Visual Analogue Scale and Numerical Rating Scale for pain intensity, unpleasantness, and overall pain rating, for phantom limb pain, residual limb pain, and other pain, in last week, 24 hours, and currently. Frequency of pain occurrences and average duration for all 3 in last week and 24 hours.

Brief Pain Inventory (BPI)
Short-Form McGill Pain Questionnaire (SF-MPQ)
Trinity Amputation and Prosthesis Scale (TAPES-R)
Beck Depression Inventory (BDI)
Post-Traumatic Stress Disorder Check List (PCL-8)
Interpersonal Reactivity Index (IRI)
Compassion Scale (CS)
Self-Compassion Scale (SCS)

Mystical Experience Questionnaire (MEQ)
Challenging Experience Questionnaire (CEQ)
5-Dimensional Altered States of Consciousness Rating Scale (5D-ASC)

Appendix C

Experiment 2.1 Stimuli

Blicket Machine:

Blicket Containers (instructions video):

Training Trial Blickets:

Test Trial Blickets:

Experiment 2.1 Video Scripts and Question Text

Intro video:

“In this study, we are going to figure out what things are blickets. You can’t tell what a blicket is just by looking at it. But, they do have something called blicketness inside of them. *Pulls out machine* -- here I have a machine. Blicketness will make this machine light up. I want to know which of my things are blickets. Only a few of them are blickets and most of them are not

blickets. It is important to know which are which. Now, *Pulls out containers of blickets and non-blickets*, look at these baskets. People have put things that are blickets into one and things that are not blickets into the other. There are many more things that are not blickets than things that are blickets. Let's count them. 'One, two, three, four.' *pulls objects out and places in front of basket while counting* There are four things that are not blickets *puts non-blickets back into container*. 'One'. There is one thing that is a blicket. *Puts blicket back into container. Almost all the things we counted are not blickets, and only one thing was a blicket. Okay, now let's see the ones that are not sorted."

Training Trials:

Video starts with the machine and 3 objects on the table in front of the experimenter. Let's try these 3 things. Here are 3 things. It is very important to know which are blickets. Let's call this one *picks up first object from the left* trapezoid. Let's call this one *picks up second object from the left* oval. And let's call this one *picks up third object from the left* diamond. Let's see what happens when we put trapezoid onto the machine. *Puts trapezoid on the machine* It did not turn on. *removes trapezoid and places back down.* Let's see what happens when we put oval onto the machine. *puts oval on the machine*. It did not turn on *removes oval and places back down*. Let's see what happens when we put diamond onto the machine *puts diamond on the machine*. It did not turn on *removes diamond and places back down*. Let's see what happens when we put trapezoid and oval onto the machine *places both trapezoid and oval on the machine*. It did not turn on *removes both objects and places back down*. Let's see what happens when we put trapezoid and diamond onto the machine *places both trapezoid and diamond on the machine*. It turned on! *removes both objects and places back down.* Let's see what happens when we put oval and diamond onto the machine *places both oval and diamond on the machine*. It did not turn on *puts both objects back down*.

Test Trials:

Video starts with the machine and 3 objects on the table in front of the experimenter. Here are 3 things. It is very important to know which are blickets. Let's call this one *picks up first object from the left* pentagon. Let's call this one *picks up second object from the left* rectangle. And let's call this one *picks up third object from the left* triangle. *picks up pentagon* Let's see what happens when we put pentagon onto the machine *places on the machine*. It did not turn on *removes object and places it back*. Let's see what happens when *picks up rectangle* we put rectangle onto the machine. *places rectangle on the machine* it did not turn on *removes object and places back down*. Let's see what happens when *picks up pentagon* we put pentagon *picks up triangle* and triangle onto the machine *places both objects on the machine*. It turned on! *removes objects and places them back*. Let's see what happens when we put pentagon *points to pentagon*, rectangle *points to rectangle*, and triangle *points to triangle* onto the machine. *picks up all 3 and places on the machine*. It turned on! *removes all 3 objects and places back down*. Let's see what happens when *picks up pentagon* we put pentagon *picks up triangle* and triangle onto the machine *places both objects on the machine*. It turned on! *removes objects and places them back*.

Question Text:

Judgement Questions

"Now that you've seen these objects on the machine, do you think that (Object Shape) is a blicket?"

- Yes, it is a blicket, No, it is not a blicket

"How confident are you about whether (Object Shape) is a blicket?"

- 0-100 slider scale with the following labels at 25 point increments:
    - Not confident at all, Not very confident, Kind of confident, Confident, Extremely confident

Intervention Questions

"Now it's your turn. Which of these would you use to turn on the machine yourself? Use only what you need to activate the machine."

"How confident are you about whether your choice will activate the machine?"

- 0-100 slider scale with the following labels at 25 point increments:
    - Not confident at all, Not very confident, Kind of confident, Confident, Extremely confident

Appendix D

Single Cue Serial Production Task Predetermined Anchor Words

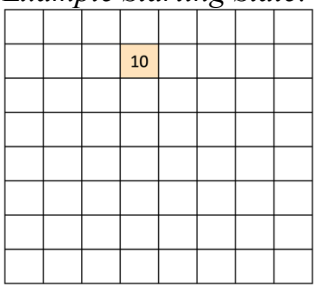
Bunny, cat, dog, bird, dolphin, turtle, fish, cow, elephant, zebra, giraffe

Appendix E

Spatially Correlated Multi-Armed Bandit Task Stimuli, Instructions Text, and Question Text

Grid Environments:

Example Starting State:



Example Full Environment:

14	13	11	6	6	6	7	11
16	16	13	10	6	6	8	10
17	18	15	12	10	10	13	16
19	21	17	16	18	16	18	21
19	22	25	25	23	22	25	26
20	22	27	29	31	31	32	32
21	24	28	31	34	37	38	40
23	25	30	33	36	38	39	43

Instructions Text:

"In each round of this game you will see a game board like the one you see here. At the beginning of each round, one box on the game board is already revealed. Clicking on a box will reveal its score. In each round you have 25 clicks and your goal is to collect as many points as possible!"

"Now you can practice on this game board. Before each click you have to decide whether to reveal a new box or click again on a box that already has its score revealed. When you choose to reveal a new box, you don't know how many points you'll get, but it could be worth many points! If you click a box that is already revealed, the score may change slightly, but will stay about the same. **Heres an additional tip:** On the game boards, high scores often appear next to other high scores, and low scores often appear next to other low scores! So a box's score depends on where it is located on the board!"

"Great job! In the next rounds you can collect up to 5 stars on each game board. The more points you collect in a round, the more stars you get for that round. The goal is to earn as many stars as possible!"

Round Completion Message:

"You have finished this round and have collected X stars. You have X rounds remaining!"

Bonus Round Instructions Text:

“This last round is a special bonus round. Here, you can initially only reveal 15 boxes. After clicking 15 times, you’ll see five boxes that haven’t been revealed yet. For each of these boxes you will estimate how many points you would get if you clicked on it. See if you can get them all correct! After that, you’ll answer a few questions and then you can continue playing as usual until all 25 clicks are used up.”

After making their initial 15 choices:
"I'm going to ask you questions about five different boxes that you haven't revealed yet. For each box, you will estimate how many points you would get if you clicked on it. After that, you will pick one of the five boxes to reveal and then the game will continue normally. You will continue playing as usual until all 25 clicks are used up.”

Question Text:

Attention Check:
“What is your task?”
- Learn Colors, Collect as many points as possible, Find the boxes with the most points, I have no idea

“How can you earn points?”
- By clicking only on hidden boxes, By clicking only on revealed boxes, By clicking on hidden or revealed boxes

“How are points distributed?”
- Randomly, High scores are **never** next to each other, High scores are **always** in the same area, High scores are **often** in the same area

Bonus Round:
“How many points will you get if you click here? What do you think?”

“How sure are you?”

Appendix F

Approach-Avoid Decision Making Task Stimuli, Scripts, and Question Text

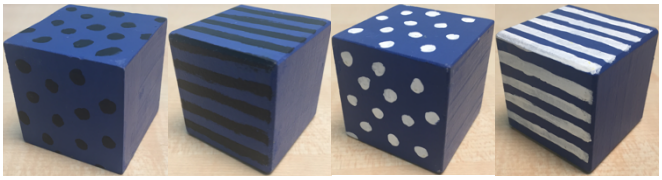
Stimuli

Zaff Machine:

Set 1

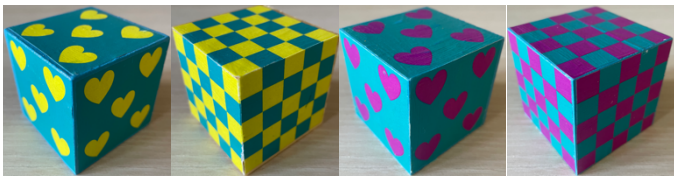
Training and Test Blocks:

Generalization Blocks:

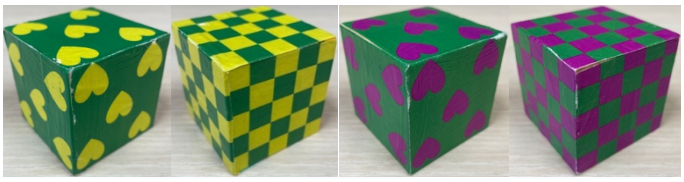


Set 2 (Used in Chapter 4)

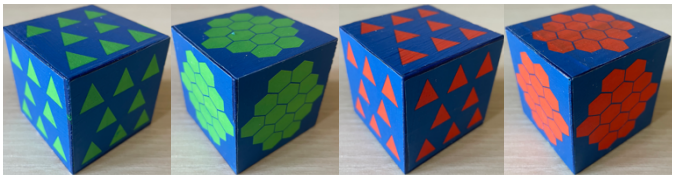
Training and Test Blocks:



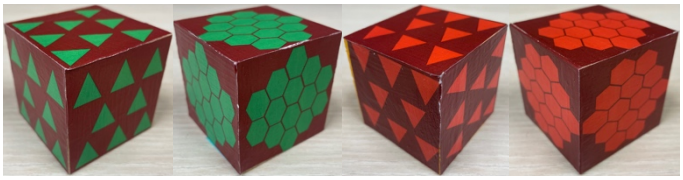
Generalization Blocks:



Set 3 (Used in Chapter 4)
Training and Test Blocks:



Generalization Blocks:



Scripts:

Intro Video:

"See all these blocks? Some of these blocks are zaffs but some of these blocks are not zaffs. And this is a zaff machine. How my zaff machine works is that if we put a block that is a zaff on the machine, the screen lights up green and a happy face comes up. But if we put a block that's not a zaff on the machine, the screen lights up red and a sad face comes up. So that's how my zaff machine works and we're going to play a game with my zaff machine. In this game, you get to start with 4 stars. There are ways to get more stars, but there are also ways to lose stars. I'm going to take one block at a time, and you get to decide whether or not to put the block on the machine. If you put the block on the machine and it is a zaff, you get one more star. But if you put the block on the machine and it's not a zaff, we'll take two stars away. But you can also decide not to put a block on the machine and then we'll just put the block away and nothing happens. You don't get any stars and you don't lose any stars. So that's how this game works."

Zaff Video:

"It's a zaff, so you get one more star!"

Non-Zaff Video:

"It's not a zaff, so we'll take two stars away."

Questions

Training Phase:
"Would you like to put this block on the machine?"

Prediction Question (first trial set only):
"Do you think this block is a zaff or not a zaff?"

Test and Generalization:
"Is this block a zaff or not a zaff?"

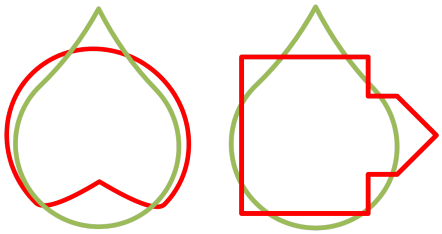
Appendix G

Change Detection Task Stimuli, Instructions Text, and Question Text

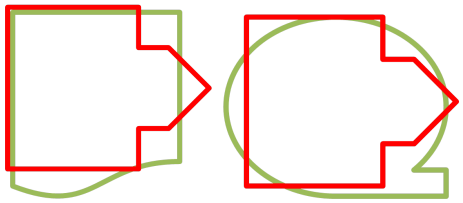
Example Stimuli:

Set 1:

Cue Change Trial

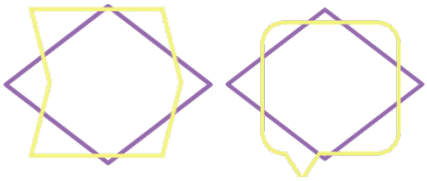


Un-Cue Change Trial

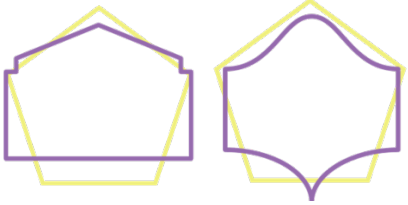


Set 2 (Used in Chapter 4):

Cue Change Trial

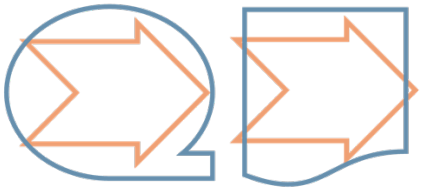


Un-Cue Change Trial

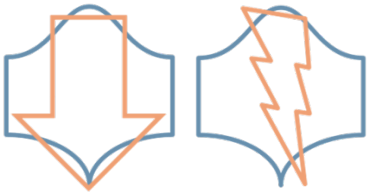


Set 3 (Used in Chapter 4):

Cue Change Trial



Un-Cue Change Trial



Instructions Text

"In this task you will be shown a series of trials that contain images made up of a (un-cued color) part and a (cued color) part. **Pay very close attention to only the (CUED COLOR) part.** You will be asked whether the **(CUED COLOR)** part is new or has been shown on a previous trial. You will also be asked if the picture within each trial changes."

Question Text

Familiarity Judgement
"Press S if the **(CUED COLOR)** image is familiar and K if the **(CUED COLOR)** image is new."

Change Judgement
"Did the image change? Press S if yes and K if no."