OPEN-LABEL PLACEBO COGNITIVE TRAINING MAY BE INEFFECTIVE FOR COGNITIVE IMPROVEMENT

by

Michael Joseph Waltrip
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Psychology

Committee:

____ Director

____

____

____ Department Chairperson

____ Program Director

____ Dean, College of Humanities and Social Sciences

Date: ___________________________________ Fall Semester 2020
George Mason University
Fairfax, VA

Open-Label Placebo Cognitive Training May Be Ineffective for Cognitive Improvement

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Michael Joseph Waltrip
Master of Arts
George Mason University, 2018
Bachelor of Arts
California State University San Marcos, 2016

Director: Patrick E. McKnight, Professor
Department of Psychology

Fall Semester 2020
George Mason University
Fairfax, VA

Acknowledgements

Thank you to my advisor, Dr. McKnight, and to my other awesome committee members, Dr. Helton and Dr. Shaw, for your friendship and your mentorship. Thank you to the administrative staff in the Psychology Department for answering my many questions. Thank you to everyone that listened to my ideas and gave their feedback. Thank you to my family and other loved ones (Michelle ☺) for your support during this process. Thank you to my roommates: Scott, DaSean, Gentian, Jingchao, and Victoria for making our house a place of wonderful work/fun balance. Finally, thank you to everyone not mentioned here who made my time in graduate school so amazing.

Table of Contents

List of Figures

Abstract

OPEN-LABEL PLACEBO COGNITIVE TRAINING MAY BE INEFFECTIVE FOR COGNITIVE IMPROVEMENT

Michael Joseph Waltrip, MA

George Mason University, 2020

Dissertation Director: Dr. Patrick E. McKnight

Recent research suggests that deception is not required to elicit placebo effects. So-called "open-label" placebos are honestly given placebo treatments with beneficial effects, primarily on self-report outcomes. In a randomized control trial, I used open-label placebo cognitive training to try and improve cognitive performance. The strengths of the experiment come from a large sample, matched protocols across conditions, and the inclusion of relevant constructs to test mechanistic explanations. However, no effects were observed on either objective or subjective outcomes as a result of the open-label placebo treatment. There were some unexpected performance differences between participants recruited from a flyer advertising the brain-training component of the study, compared to participants recruited from a control flyer. The results of the primary hypotheses (open-label placebo effectiveness) may reflect a true null result, an ineffective manipulation, or noise from an overly complex experiment. The main conclusion is that

some skepticism towards the generalizability of open-label placebo effects is warranted. I discuss if that conclusion stems from the current failure of the effect to replicate or from acknowledging the design components that may have caused its failure.

Chapter One: Introduction

Placebo treatments given without deception - called “open-label placebos" - offer weaker effects on only a subset of outcomes compared to the traditional blinded placebo (i.e., with deception). Despite a growing body of research, evidence remains limited regarding these effects. I begin with a short overview of open-label placebo research that includes a comparison with traditional placebos. Following, I outline the rationale of the current study.

Overview of Open-Label Placebo Research

Shifting from Deceptive Placebos to Open-Label Placebos

Traditional placebos cause ethical concerns. A search for “placebo ethics” on Google Scholar returns more than one-million results. Patients expect honest communication from medical professionals and taking a placebo treatment under false pretenses can degrade trust. In addition, accidental unblinding of a placebo treatment was thought to negate its effects which made long term placebo use difficult (Beecher, 1955). As a result, researchers looked towards open-label placebos as an ethical and practical alternative. In a reflection on his research, Kaptchuk, a prominent scientist in the field, describes the path that led him towards open-label placebos (2018). He cites inspiration from early work by Park and Covi (1965) where they unintentionally explored the open-label placebo effect. Kaptchuk’s main motivation, however, came from qualitative reports during a placebo acupuncture study for irritable bowel syndrome. Against common beliefs at the time, participants did not report positive expectations from placebos. Instead, they were worried, confused, and unfooled by concealment. Because placebos still worked despite the negative emotions, he thought the mechanisms underlying placebo effects might be "more non-conscious, sensory-motor, and embodied than everyone assumed" (313). He wondered if honest discussion and prescription of placebos would overcome those issues.

Differential Evidence Supporting Open-Label Placebo and Placebo Effectiveness

The effectiveness of open-label placebos is demonstrated across a new but diverse body of literature. One recent meta-analysis claims an average standardized mean difference of 0.88, 95% CI 0.62 to 1.14, $I^2 = 1\%$ (Charlesworth et al., 2017). Across five experiments, open-label placebos decreased symptoms in patients with allergic rhinitis, major depressive disorder, irritable bowel syndrome, and other ailments. In experiments with healthy patients, open-label placebos improved sleep quality, increased positive emotions, lowered experimentally induced heat-pain, and reduced emotional and physical distress (Locher et al., 2017; El Brihi, Horne, & Faasse, 2018). The background provided here is far from exhaustive, but the selected articles demonstrate the existence of the open-label placebo effect. For a more detailed review and commentary on the literature, see Blease, Bernstein, and Locher (2019).

Despite the empirical support for open-label placebos with patient and healthy samples, the effects observed are limited to subjective outcomes. From the patient samples, the positive open-label placebo effects came from self-reported symptoms (e.g., allergies, depression, pain; Schaefer, Harke, & Denke, 2016; Schaefer, Sahin, & Berstecher, 2018; Kelley, Kaptchuk, Cusin, Lipkin, & Fava, 2012; Locher et al., 2017); none involve objective outcomes. Similarly, the studies focused on healthy adults produced positive open-label placebo effects only with self-reported measures (e.g., sleep quality, emotional distress, and physical distress; El Brihi et al., 2018). Open-label placebo effects may be uniquely constrained to measures susceptible to bias from conscious or unconscious expectations

In contrast, traditional placebos have characteristics that open-label placebos do not, or have not yet been observed. Placebos can affect objective measures like the opioid response, athletic performance, cognitive performance, and many others (Benedetti, & Amanzio, 1997; Beedie & Foad, 2009; Foroughi, Monfort, Paczynski, McKnight, & Greenwood, 2016). Placebo effects may also be stronger than their open-label counterparts: one meta-analysis for placebos as antidepressants found an effect size of Cohen's d = 1.85 (Rief et al., 2009). Another unique characteristic of placebos are tiered dose-responses: 4 open-label placebo pills had the same effect as one pill, whereas 2 blinded placebo pills resulted in a stronger effect than just 1 pill (El Brihi et al., 2018; de Craen et al., 1999). The exact cause and nature of the differences between concealed placebos and open-label placebos are not yet understood.

Individual Differences in Open-Label Placebo and Placebo Effectiveness

There is much more to be discovered for open-label placebos. Kaptchuk argues that 4 tenets facilitate the effect: 1) removing the stigma of placebo effects, 2) describing the automatic nature of placebo responses, 3) showing that there is no requirement to believe, and, 4) instructing that following the prescribed protocol is critical (2018). Though sensible, Kaptchuk's tenets have little systematic manipulation (Blease et al., 2019).

And, studies that do manipulate them report conflicting results: experimentally manipulated expectations moderate the effect in Locher et al., (2017) and Schafer et al. (2018), but not in Leibowitz, Hardebeck, Goyer, & Crum (2019). In two other studies, expectations predicted improvement, but only for participants in the open-label placebo group, and not the control groups (El Brihi et al., 2018; Meeuwis et al., 2018). The amount and quality of patient to caregiver (or researcher) contact is inconsistent or insufficiently detailed in reports too. Additional points of uncertainty include how adequately matched control groups fare against open-label placebo groups, how the effect differs across self-report and non self-report measures, and how psychology and physiology influence one another.

Chapter Two: Methods

The goal for the experiment was to 1) test open-label placebos on objective and subjective outcomes, 2) compare theoretical explanations of the effect , 3) match protocols between experimental treatment and control arms, and 4) include a larger sample than used in most open-label placebo research.

Overview

I used a randomized controlled trial of open-label placebo cognitive training on cognitive performance. Critically, recent evidence suggests that placebo effects underlie any cognitive improvement after cognitive training (Foroughi et al., 2016; Clark, Lawlor-Savage, & Goghari, 2017). In an online experiment, participants took a baseline cognitive assessment (pretest), were randomized into one of two cognitive training tasks, and then completed a parallel form cognitive assessment (posttest). Participants also reported cognitive workload and effort throughout. Lastly, the experiment finished with measures related to beliefs about intelligence, need for cognition, and demographic information.

Procedure

Undergraduate psychology students from George Mason University's recruitment pool, SONA, completed the study online for course credit. Participants were recruited from one of two SONA flyers. The first flyer was generic and advertised for a cognitive psychology experiment: "*Complete and earn 3 credits*". The second flyer advertised a study on "Brain Training & Cognitive Enhancement: *Numerous studies have shown working memory training can increase fluid intelligence (e.g., Au et al., 2014; Jaeggi et* *al., 2008*; *Rudebeck et al., 2012*)". The manipulation intended to incentivize self-selection of people with strong brain-training beliefs, or to influence participants' beliefs about brain training's effectiveness.

In the experiment, participants first provided consent. Next, they responded to a manipulation check gauging their beliefs on brain training's effectiveness. They rated their level of agreement to the statement: "You can train your brain to increase your general intelligence" on a 5 point scale ranging from "Strongly Disagree" to "Strongly Agree". The manipulation check was presented amongst four throwaway items, masking its intentions.

Following, all participants were taught about open-label placebo effects via a 2-minute video describing Kaptchuk's 4 tenets edited to fit the protocol of my study (2018). The script is below.

"Placebo effects demonstrate the power of our own minds. Consider the fact that for many health problems, sugar pills produce effects comparable to real medicine. In fact, even knowing that you are taking a sugar pill can produce a strong effect; five recent studies demonstrated these "open-label" placebo effects. Those who respond to placebos are not necessarily easily fooled nor are they faking the outcomes. Many are skeptical of the effects of the sugar pill and yet experience these benefits. The effects of these placebos are real and felt by people with legitimate problems. Moreover, the effects happen automatically, and everyone may benefit from them. The purpose of the current study is to demonstrate an open-label placebo effect with cognitive training.

Prior work on the placebo effect imposed a deception procedure. Participants would unknowingly receive a fake treatment and respond. Those studies mentioned previously clearly demonstrated that deception was no longer a required element to produce the effect; hence, the open-label placebo effect was borne.

Researchers do not know how or why these placebo effects persist without deception. What we do know is that participants must participate knowingly in a study, complete the study protocol as prescribed, and engage in all aspects with equal vigor. Our study here was designed to test these aspects to better understand the effect. We welcome your participation and hope with your assistance, our results shed great light on these challenging problems. Thank you for your attention and thank you for agreeing to participate."

The remainder of the experiment was broken up into four blocks (Figure 1).

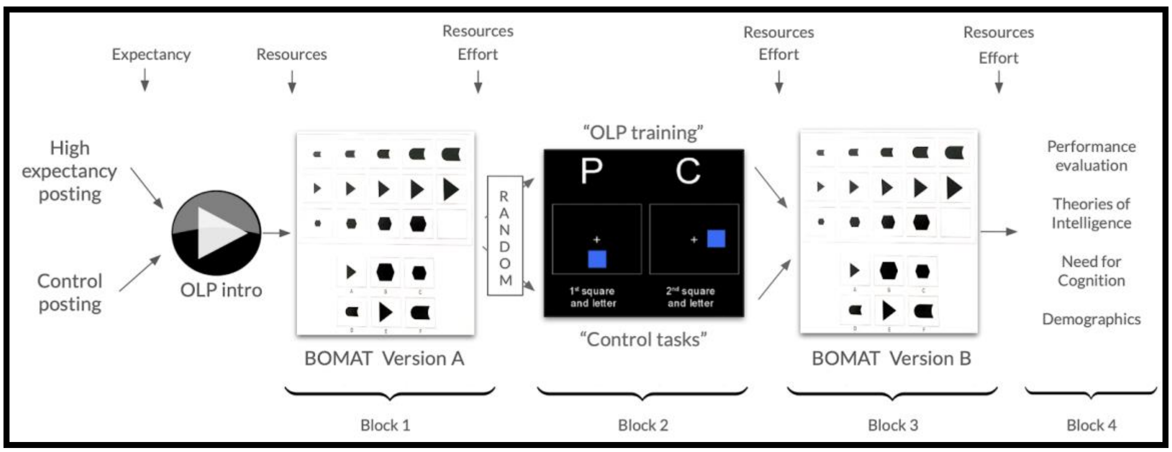

Figure 1. Procedural flow chart for the current experiment.

Block 1 - Pretest

Participants completed a one-item cognitive resource measure called the Gas Tank Questionnaire (Monfort, Graybeal, Harwood, McKnight, & Shaw, 2018). The Gas Tank Questionnaire states: "Think about your brain as an engine… indicate how much gas you have left right now". Participants responded on a 0-100 slider with scale anchors labelled "Empty" and "Full". Following, participants completed the pretest: the Bochum Matrices Test Advanced - Short Version (BOMAT) (Hossiep, Turck, & Hasella, 1999). The BOMAT consists of 27 problems (26 are scored). In each problem, there are 15 patterns arranged in a 3 × 5 matrix with one pattern missing. The goal of the task is to choose the correct pattern out of six possible options. After taking the pretest, participants responded to a second gas tank questionnaire, and a one-item effort measure: "How much effort did you use during the last task?" on a 10-point scale ranging from "No effort" to "Maximal effort".

Block 2 - Manipulation

Participants were randomly assigned with replacement to the control or open-label placebo group and instructed accordingly. In the open-label placebo group, participants were told that they would complete the open-label placebo cognitive training. In the control group, participants were told that they would complete some control processing speed tasks. After the manipulation, however, the experiment was identical for both groups. All participants completed the dual adaptive n-back task[3]- a commonly used task in brain-training literature (Kirchner, 1958; Jaeggi et al., 2010; Millisecond Software, 2015). In the dual adaptive n-back task, participants responded to visual and auditory stimuli “N” number of items back. For example, in an audio stream 1-back, if the letter “r” was presented, the correct response was “r”. Whereas, in a 3-back, if the letters “d”, “r”, and “z” were presented, the correct response was “d”, as that was the letter 3 items back. At the same time, participants completed a visual n-back. In addition to letters, they also remembered the locations of blocks in one of 8 locations on the screen. The adaptive component changes the “N” in n-back dependent on participants’ performance. If they perform well, they recall more items back in the next trial. If they perform poorly, they recall less items back in the next trial. To reiterate, participants in the open-label placebo group and the control group both completed the dual adaptive n-back task - all that differed was the instructions that preceded the task. After completing the n-back, they responded to a second effort measure.

Block 3 - Posttest

Participants completed a third gas tank questionnaire and then the posttest BOMAT parallel form. The protocol here followed past research, using parallel versions of the BOMAT, both consisting of 27 problems (26 scored) (Jaeggi, Buschkuehl, Shah, & Jonides, 2014; Foroughi et al., 2016). After the posttest, participants completed a fourth gas tank questionnaire, and a third effort measure.

Block 4 - Subjective Measures, Relevant Scales, and Demographics

In the final block, participants reported an evaluation of their performance on the posttest: "Compared to the first time you completed this Matrix task, how do you think you performed this time?" on a 5-point scale ranging from "A lot worse" to "A lot better". Next, they completed Dweck’s Theories of Intelligence scale: eight items that measure beliefs regarding the malleability of intelligence on a 6-point scale ranging from "Strongly Disagree" to "Strongly Agree" (Dweck, 2000). Following, they completed Cacioppo's Need for Cognition scale: 18 items that measure "the tendency for an individual to engage in and enjoy thinking" on a 5-point scale ranging from "extremely uncharacteristic" to "extremely characteristic" (Cacioppo & Petty, 1982; Cacioppo, Petty, & Feng Kao, 1984). To finish the experiment, participants reported their gender and age.

Participants

443 participants initially consented for the experiment. After consenting, recall that participants responded to a manipulation check and then watched a video. Many participants dropped out in those steps (159; 35.9%). For the remaining three blocks, there was an average drop off of 18 participants per block (6.7%). Following, 15 participants (6.6%) were removed after logs showed they either skipped a block, repeated a block, or had missing data for the dependent variable. 214 participants finished the experiment and a had a usable dataset.

In the final cleaning step, I dropped participants who failed any one of three manipulation checks. The attention checks were written to lower face validity. Specifically, they asked for participant recall of certain task related features to appear as a memory assessment, rather than an attention assessment. Dropping those who failed the attention checks removed 46 participants (21%).

The final analyzed sample consisted of 168 participants (M_{age} = 20.4, SD_{age} = 2.4, 118 female, 45 male, & 5 prefer not to say). 88 of the participants came from the "brain-training" flyer and 80 came from the control flyer. Across both recruitment methods, 94 of the participants were randomized to the open-label placebo group and 74 were in the control group (for consort diagram, Figure 2). Participants were randomized with replacement because the experiment took place longitudinally and I wanted to avoid any treatment by time interactions. Anecdotal wisdom touts that participants who complete experiments at the beginning of school semesters differ from those who wait until the end.

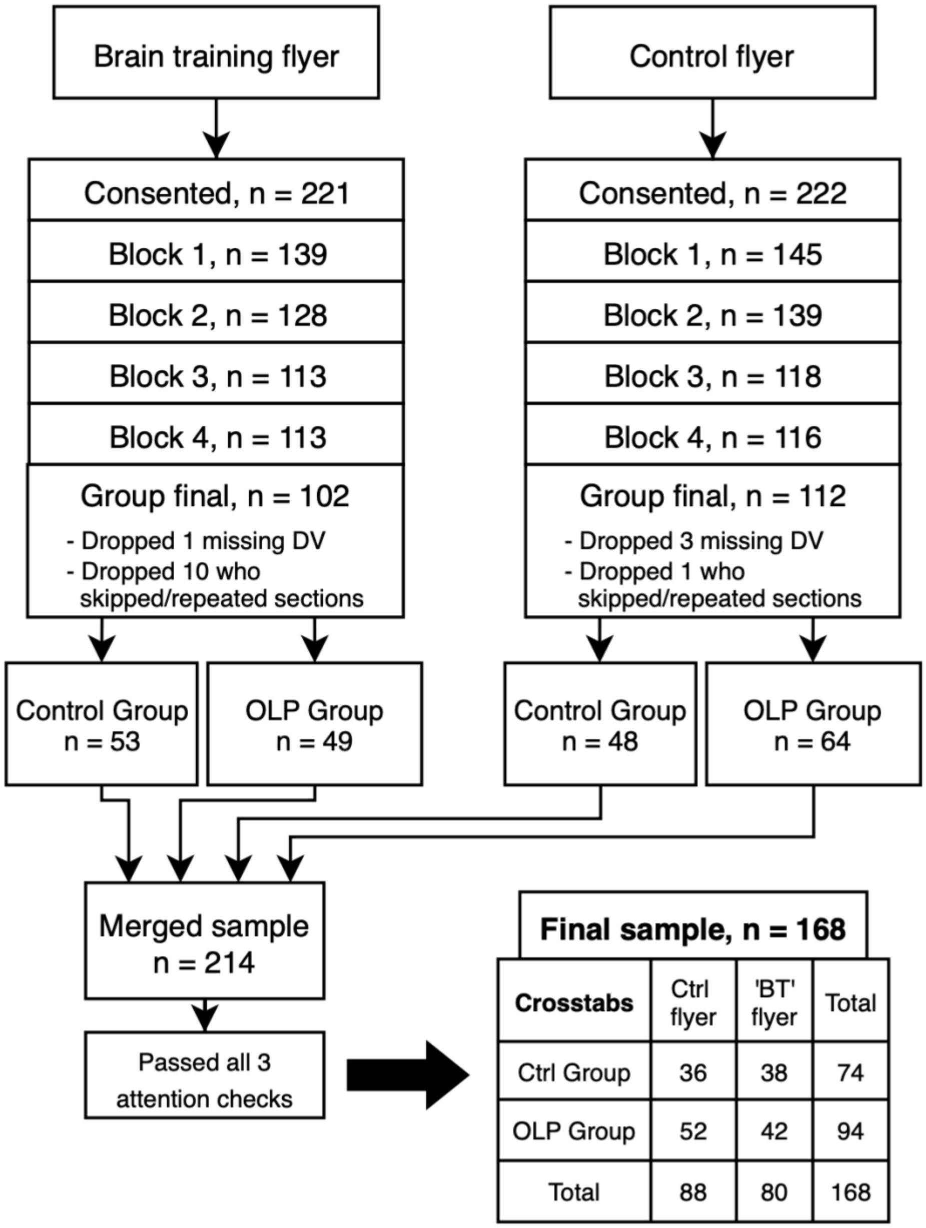

Figure 2. A consort diagram for the current experiment.
Participants were recruited from one of two flyers and further randomized into an open-label placebo or control group. "Ctrl" = Control, and "BT" = brain-training, "OLP" = Open-Label Placebo.

Sample Size and *A Priori* Statistical Power

Sample requirements were informed from previous research and based on the first two hypotheses. The placebo effect resulting from cognitive training in Foroughi et al. (2016) had an effect size of Cohen's d = 0.39 for the BOMAT and Cohen's d = 0.51 for Raven's Advanced Progressive Matrices (another cognitive test; Raven, Court, Raven, & Kratzmeier, 1994). Further, the open-label placebo meta-analysis by Charlesworth et al. (2017) reported an average standardized mean difference equal to 0.88. Because the study was developed online to recruit a larger sample, I had the flexibility to choose a more conservative effect size in the power calculations. With power set to 0.80, alpha set to .05 for a two-tailed test, and an expected effect size of 0.39: the required sample size was determined to be 104 participants per group (Brant, n.d.). The other two effect sizes of .51 and .88 result in required samples of 61 and 21 participants per group, respectively. Here, 214 participants initially completed the experiment. After screening those who failed attention checks, the analyzed sample size of 168 participants was sufficiently powered based on criteria from two of the three referenced studies.

Hypotheses

Participants in the open-label placebo group, relative to the control group, will:

1. Perform better on the posttest (objective open-label placebo test).
2. Evaluate their performance differential as higher between posttest and pretest (subjective open-label placebo test).
3. Report using more effort during the n-back and the posttest (mechanistic test).

4. Have less remaining cognitive resources after the n-back and the posttest (mechanistic test).
5. Score equivalently on the Theories of Intelligence and Need for Cognition scales (confound rule out).

Participants recruited via the “brain training” flyer, relative to the control flyer, will

6. Perform similarly for the remainder of the experiment.

Chapter Three: Results

Overview

The open-label placebo manipulation was not successful in improving objective or subjective performance, compared to the control group. There were no differences between the groups pertinent to hypotheses one through five. Possible explanations are offered in the discussion section.

There were differences between participants recruited from the control flyer, compared to those recruited from the brain training flyer. Notably, the brain training flyer resulted in decreased cognitive performance, negative evaluation of performance, and stronger beliefs about the malleability of intelligence. Self-reported effort and resource use were related to the differences between recruitment groups.

Data Cleanliness and Analysis

14 data points with extreme outliers were replaced with missing values. Specifically, response times for the BOMAT, the n-back training, the Theories of Intelligence scale, and the Need for Cognition scale had outliers as far as 12 standard deviations above the mean (e.g. 9 hours to respond). I assumed such behavior was due to technical issues or distraction from the experiment. I further assumed participants were affected completely at random. Thus, deletion of the data would not produce biased estimates (Bell, Kromrey, & Ferron, 2009; Kang, 2013). Accordingly, for these response time measures only, participants with scores greater than 4 standard deviations were dropped (14 data cells). Behavioral and self-report data were not altered. Lastly, two

participants did not complete block 4 of the experiment; though, they were not deleted as their data were relevant to a number of hypotheses. I ran some analyses before doing the data alteration describe above. I also ran the analyses with more resilient modeling frameworks (e.g. robust regression, Welch's t-test). Regardless of the outliers' inclusion or the modeling framework, the results presented below were consistent.

Analyses were run using R, the statistical computing software (version 3.6.1, 2019-07-05; R Core Team, 2019), R Studio (Version 1.2.5019, 2019-10-24; RStudio Team, 2020), and supplemental downloadable packages (Auguie, 2017; Sjoberg, 2019; Grolemund & Wickham, 2011; Warnes et al., 2016; Wickham, 2016; Wickham, 2007; Wickham, 2019; Wickham & Bryan, 2019; Wickham, François, François, Henry, & Müller, 2020; Willse, 2018; Dowle & Srinivasan, 2019; Revelle, 2018; Behrendt, 2014).

A few of my hypotheses expect no differences between groups. Under a null hypothesis significance testing framework, lack of a significant effect is not the same as evidence for equivalence. Accordingly, a large portion of randomized control trials are made up of "tests of equivalence" and "non-inferiority trials". There are a number of ways to assess equivalency, including conducting two one side tests, determining regions of insignificance, and bayesian inference. Here, for simplicity, I report traditional frequentist tests, but also include effect sizes (cohen's D) and the percentage of the total scale range that the difference between the groups reflects (similar to the reporting from Amorim, Machado, & Maher, 2020). In all of the proposed hypotheses of no difference, the respective effect sizes and the percentage differences between groups are extremely small and fall into a region of insignificance.

Manipulation Checks

Beliefs in Brain Training's Efficacy by Flyer Recruitment Group

Not Supported. A t-test revealed that there was no difference in beliefs about the efficacy of brain training across participants recruited from the brain training flyer and the control flyer, $t(165.96) = -0.35$, $p = .72$. Standardized mean difference between the groups was Cohen's $d = 0.06$ (95% CI: -0.25, 0.36).

Beliefs in the Flexibility of Intelligence by Flyer Recruitment Group (Replicating Foroughi et al. 2016)

Supported. There was a difference between the groups on the Theories of Intelligence scale, $t(162.13) = -2.44$, $p = .02$, Cohen's $d = 0.38$ (95% CI: 0.07, 0.69). Specifically, participants recruited from the brain training flyer believed intelligence was more malleable ($M = 36.25$, $SD = 7.04$) than participants recruited from the control group ($M = 33.62$, $SD = 6.80$).

Hypothesis 1: Improved Objective Cognitive Performance for Open-Label Placebo Group

Not supported. There was no objective open-label placebo effect in cognitive performance. A t-test revealed no differences between the open-label placebo and control groups on the posttest, $t(156.7) = 0.33$, $p = .74$, Cohen's $d = -0.05$ (95% CI: -0.36, 0.25). Following, I ran a series of regression models predicting the difference score between the pretest and posttest while controlling for relevant variables. In short, across 21 models, the BOMAT difference score was related to scores on the pretest, response times for both tests, and self-reported effort after both tests *but not open-label placebo status* (whether someone was in the control group or the open-label placebo group) (Figure 3). The most successful model, determined by adjusted R^2, predicted 41.7% of the variance in the BOMAT difference score, $F(4, 162) = 30.73$, $p < .001$. In the model, score on the pretest ($\beta = -.68$, $t = -9.34$, $p < .001$), effort on the posttest ($\beta = .13$, $t = 2.10$, $p = .04$), and average item response time on the posttest ($\beta = .66$, $t = 8.86$, $p < .001$) were significantly related to the difference score, but open-label placebo status was not ($\beta = .03$, $t = .46$, $p = .65$).

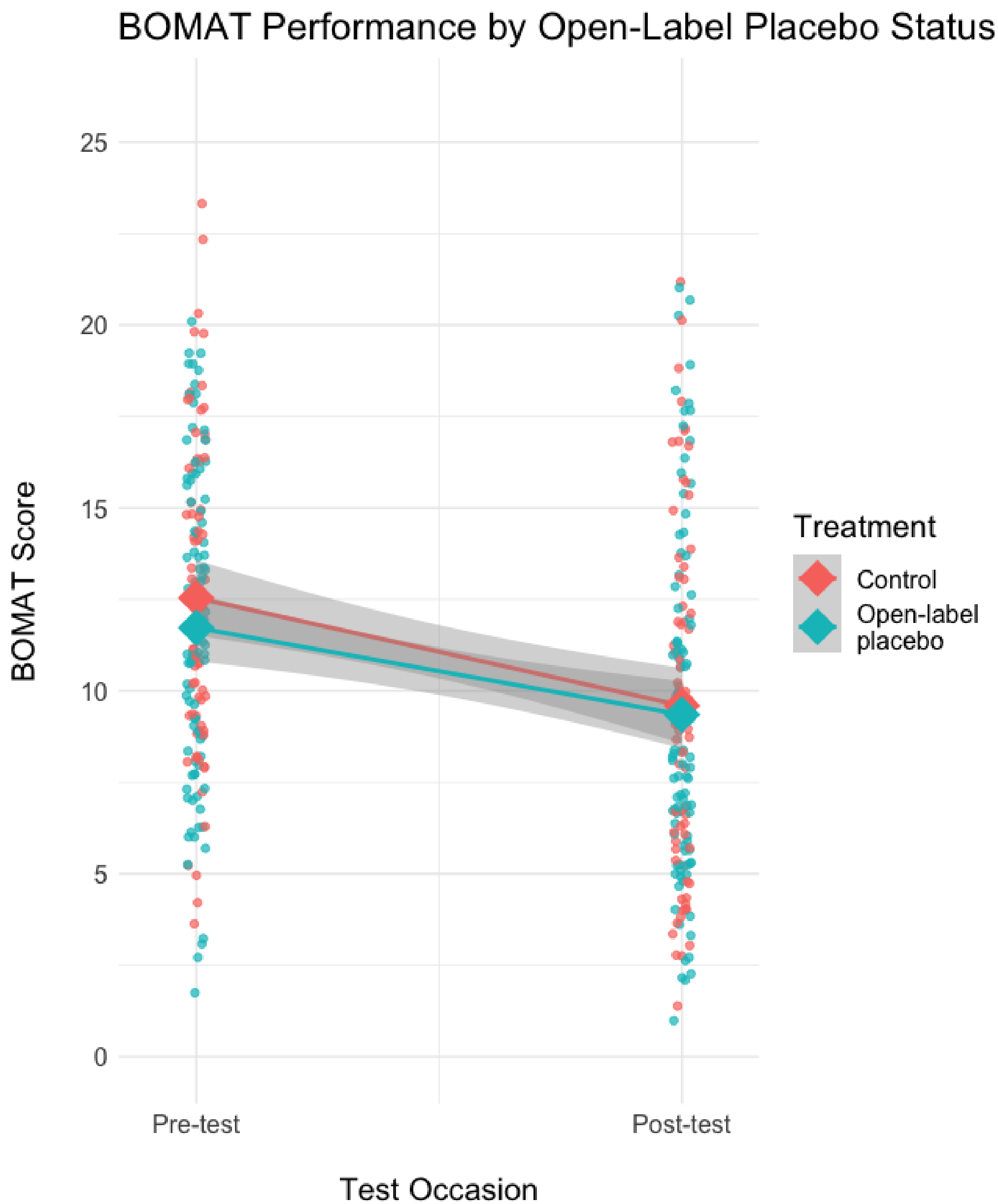

Figure 3. Objective performance and open-label placebos
BOMAT performance on the pretest and the posttest, split by participants in the open-label placebo group and the control group. Diamonds show the mean of each group. Error bands show 95% confidence intervals.

I tested the primary hypothesis another way by looking at the percentage of participants who had positive BOMAT difference scores, split by open-label placebo status. In other words, I looked at how many participants improved their score from pretest to posttest. In the control group, 25.7% of participants improved their score and in the open-label placebo group, 30.9% of participants improved their score. Statistically, there was no difference between groups in BOMAT improvement rates, $\chi^2(1) = 0.32$, p = .57.

Hypothesis 2: Improved Subjective Cognitive Performance for Open-Label Placebo Group

Not supported. There was no subjective open-label placebo effect in cognitive performance. A t-test revealed no difference between the open-label placebo and control groups on their evaluation of how they performed on the posttest, relative to the pretest, $t(155.78) = -0.52$, $p = .60$, Cohen's $d = 0.08$ (95% CI: -0.22, 0.39) (Figure 4). As before, regression models were run to control for relevant variables. Across the models, open-label placebo status never reached significance in predicting self-reported evaluation of performance. Of brief note, significant predictors included objective performance differences between the tests, self-reported effort for both tests, self-reported effort difference scores across the tests, and the percentage of time spent on the posttest, relative to the pretest. However, model statistics for these predictors are not pertinent to the hypothesis and are thus withheld.

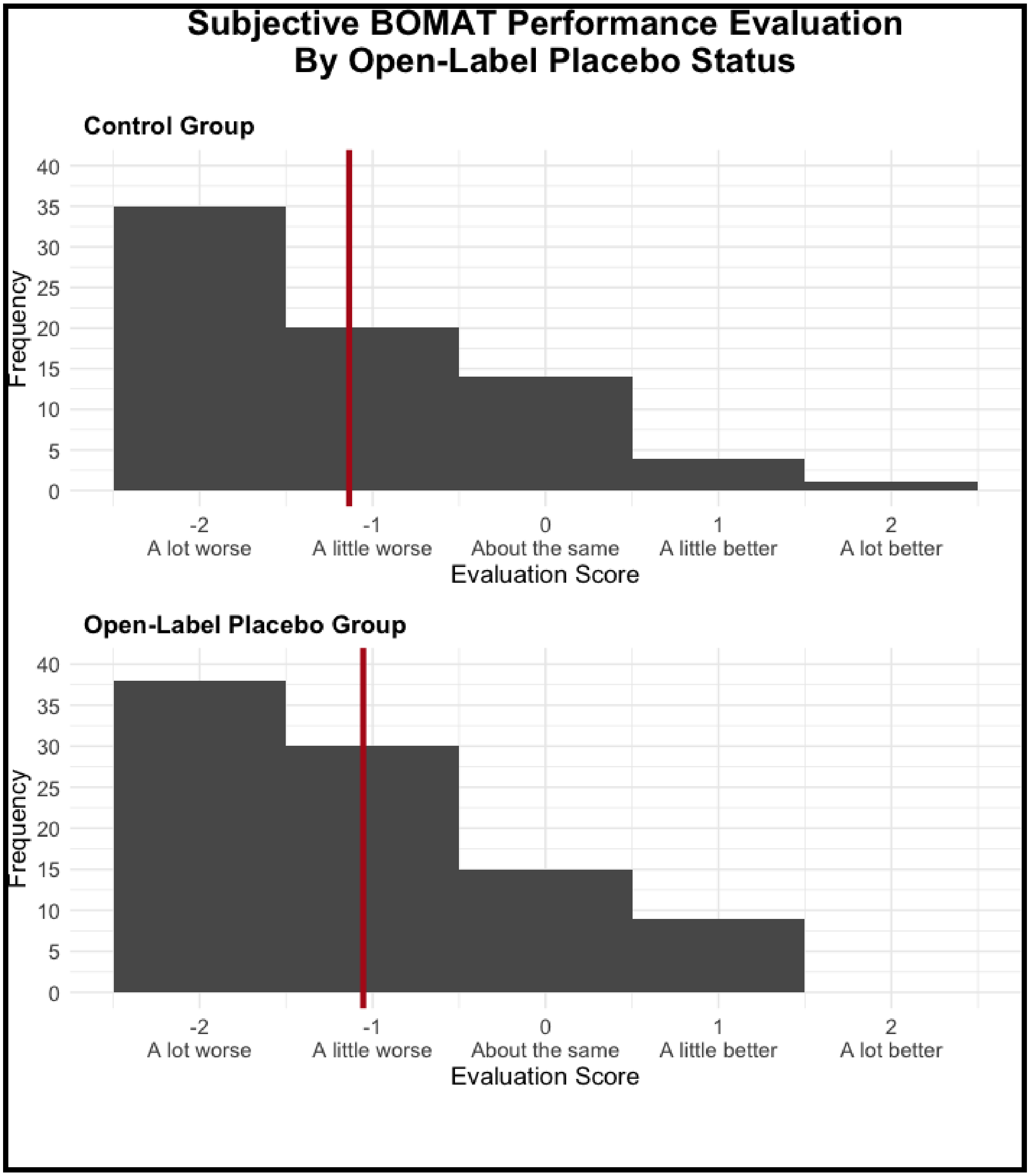

Figure 4. Subjective performance and open-label placebos.
Histograms of subjective performance evaluation for the control group (top) and the open-label placebo group (bottom). Vertical lines show the mean of each group. Recall that the prompt was: "*Compared to the first time you completed this Matrix task, how do you think you performed this time?*"

Hypothesis 3: Increased Effort on N-Back and Posttest for Open-Label Placebo Group

Not supported. There were no significant differences between the open-label placebo and the control groups on self-reported effort during the n-back training, $t(160.72) = -0.03$, $p = .98$, Cohen's $d = 0.005$ (95% CI: -0.30, 0.31), nor on the amount of time spent during the n-back training, $t(144.6) = 1.45$, $p = .15$, Cohen's $d = -0.23$ (95% CI: -0.54, 0.08).

Similarly, there were no significant differences between the open-label placebo and the control groups on self-reported effort during the posttest, $t(150.45) = -0.12$, $p = .90$, Cohen's $d = 0.02$ (95% CI: -0.29, 0.32), nor on the amount of time spent during the posttest, $t(154.41) = 0.25$, $p = .80$, Cohen's $d = -0.04$ (95% CI: -0.35, 0.27) (Figure 5).

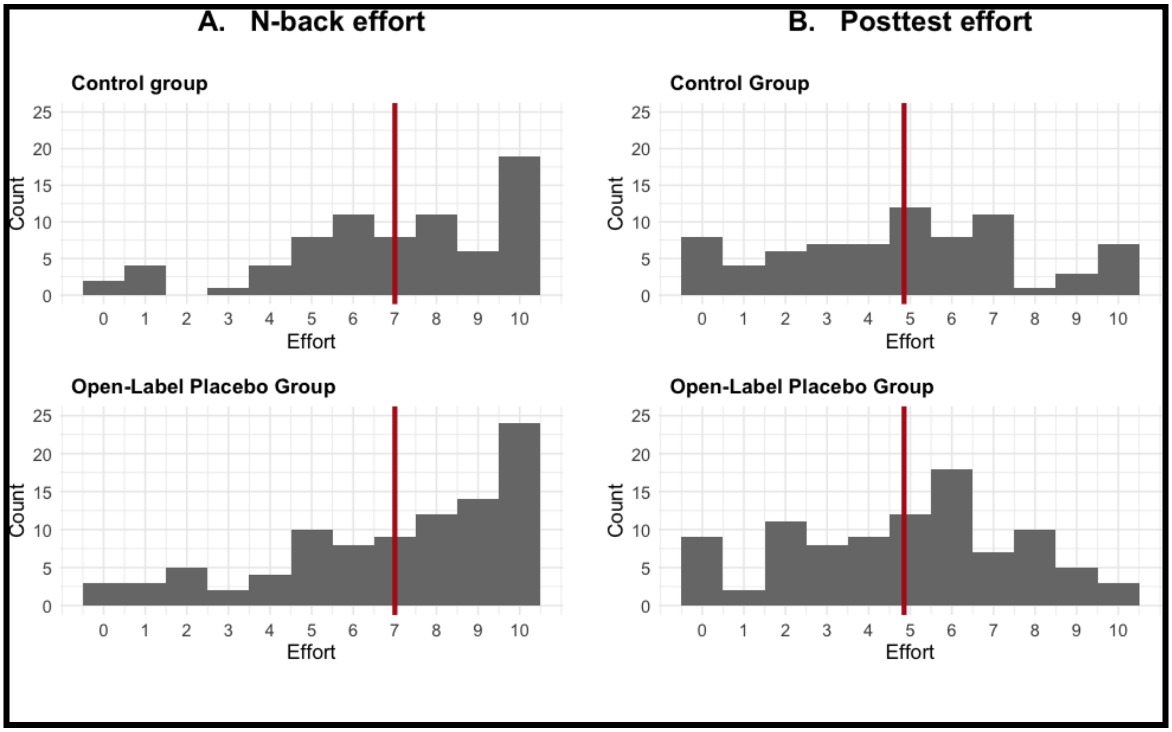

Figure 5. Self-reported effort during n-back and posttest.
Vertical lines show the mean of each group. A: Self-reported effort during the n-back for the control group (top left), and the open-label placebo group (bottom left). B: Self-reported effort during the posttest for the control group (top right), and the open-label placebo group (bottom right).

Hypothesis 4: Increased Resource Use During N-back and Posttest for Open-Label Placebo Group

Not supported. There were no differences in resource use between the open-label placebo and the control groups at any time points in a series of t-tests: *before n-back*: $t(163.37) = -0.59$, $p = .55$, Cohen's $d = 0.09$ (95% CI: -0.21, 0.40) *after n-back & before posttest*, $t(162.44) = -1.47$, $p = .14$, Cohen's $d = 0.23$ (95% CI: -0.08, 0.53), and *after posttest*: $t(153.69) = -0.53$, $p = .59$, Cohen's $d = 0.08$ (95% CI: -0.22, 0.39).

In another approach, two regression models were used to predict remaining resources. In the first regression, resources remaining after the n-back was predicted by resources remaining before the n-back, open-label placebo status, and their interaction, $F(3, 164) = 24.61$, $p < .001$. Resources remaining before the n-back significantly predicted resources remaining after the n-back ($\beta = .61$, $t = 5.83$, $p < .001$), but not open-label placebo status ($t = 1.33$, $p = .19$), nor the interaction term ($t = -0.81$, $p = .42$).

In the second regression, resources remaining after the posttest was predicted by resources remaining before the posttest, open-label placebo status, and their interaction, $F(3, 164) = 65.03$, $p < .001$. Resources remaining before the posttest significantly predicted resources remaining after the posttest ($\beta = .85$, $t = 10.02$, $p < .001$), but not open-label placebo status ($t = -0.83$, $p = .41$), nor the interaction term ($t = -1.71$, $p = .09$) (Figure 6).

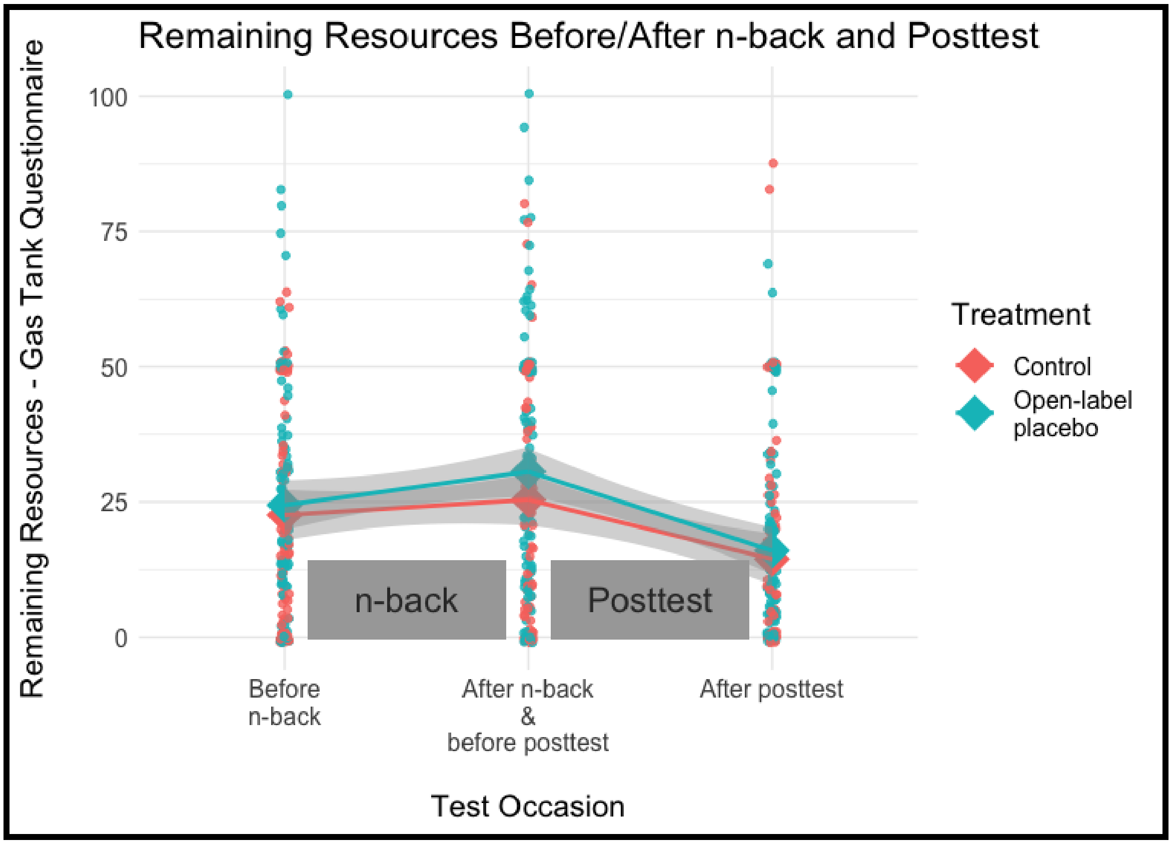

Figure 6. Self-reported resource use during n-back and posttest.
Self-reported remaining resources at different timepoints split across open-label placebo status. Diamonds show the mean of each group. Error bands show 95% confidence intervals.

Hypothesis 5: No Differences on Theories of Intelligence and Need for Cognition Scales by Open-Label Placebo Status.

Supported. A t-test supported that there was no difference between the open-label placebo group and the control group on Theories of Intelligence, $t(152.28) = 0.19$, $p = .85$. Standardized mean difference between the groups was Cohen's $d = -0.03$ (95% CI: -0.34, 0.28). The difference between groups reflects an insignificant 0.49% of the total scale range. A second t-test supported that there was no difference between the open-label placebo group and the control group on Need for Cognition, $t(159) = 0.64$, $p = .52$.

Standardized mean difference between the groups was Cohen's *d* = -0.10 (95% CI: -0.41, 0.21). The difference between the groups reflects an insignificant 1.27% of the total scale range (Figure 7).

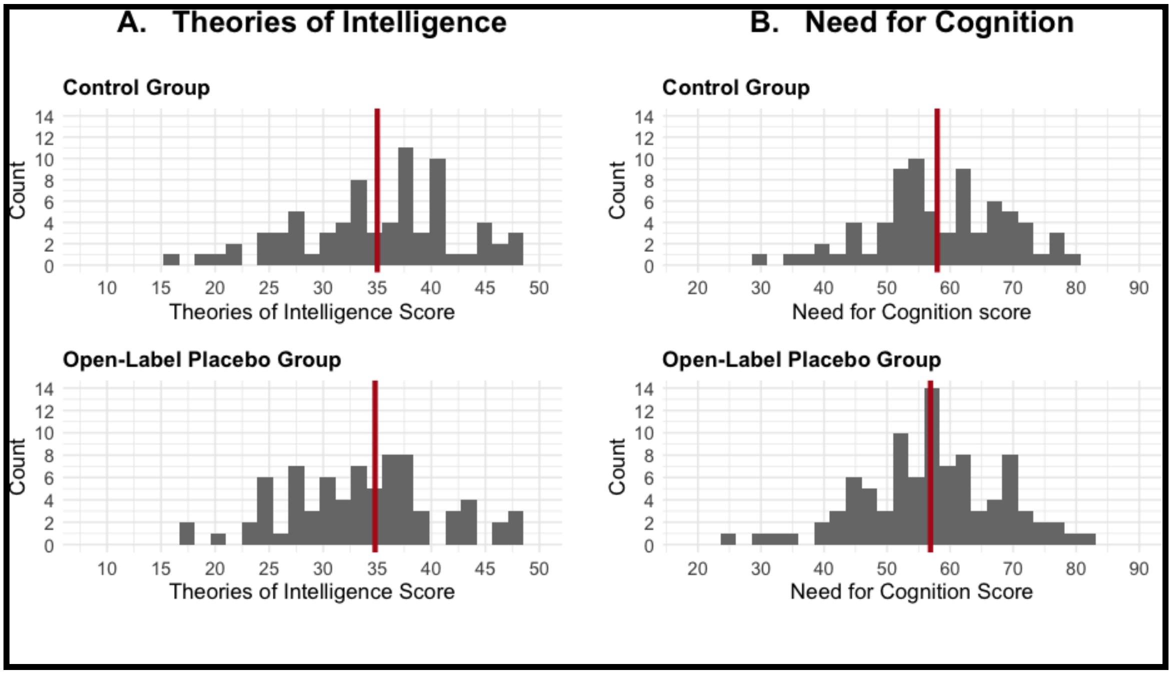

Figure 7. Theories of Intelligence and Need for Cognition scales.
Vertical lines show the mean of each group. A: Theories of Intelligence scores for the control group (top left), and the open-label placebo group (bottom left). B: Need for Cognition scores for the control group (top right), and the open-label placebo group (bottom right).

Hypothesis 6: Experiment Performance by Recruitment Flyer Status

Hypothesis six predicted that participants recruited from the control flyer would perform similarly throughout the experiment as those recruited from the "brain training" flyer (with the exception of the theories of intelligence scale reported above in 'manipulation checks').

Primary Outcomes: Objective and Subjective Performance By Recruitment Flyer Status

There were a few unexpected differences between participants recruited from the brain training flyer and participants recruited from the control flyer. Participants recruited from the brain training flyer had lower BOMAT difference scores (M = -3.31, SD = 3.72) than the control group (M = -2.00, SD = 3.04), t(159.5) = 2.40, p = .02, Cohen's d = -0.38 (95% CI: -0.68, -0.07). Participants recruited from the brain training flyer also evaluated their performance differential as worse (M = -1.30, SD = 0.95) than the control group (M = -0.90, SD = 0.99), t(163.91) = 2.69, p = .01, Cohen's d = -0.42 (95% CI: -0.73, -0.11). Of note, there were no differences between the groups at pretest (p = .30) or at posttest (p = .39). More clearly, BOMAT performance was not different at either test occasion, but, the change between timepoints was worse for those recruited from the brain-training flyer, compared to the control flyer. (Figure 8). Further, those recruited from the brain-training flyer also reported that they felt they performed worse on the posttest relative to the pretest, compared to the control group. (Figure 9).

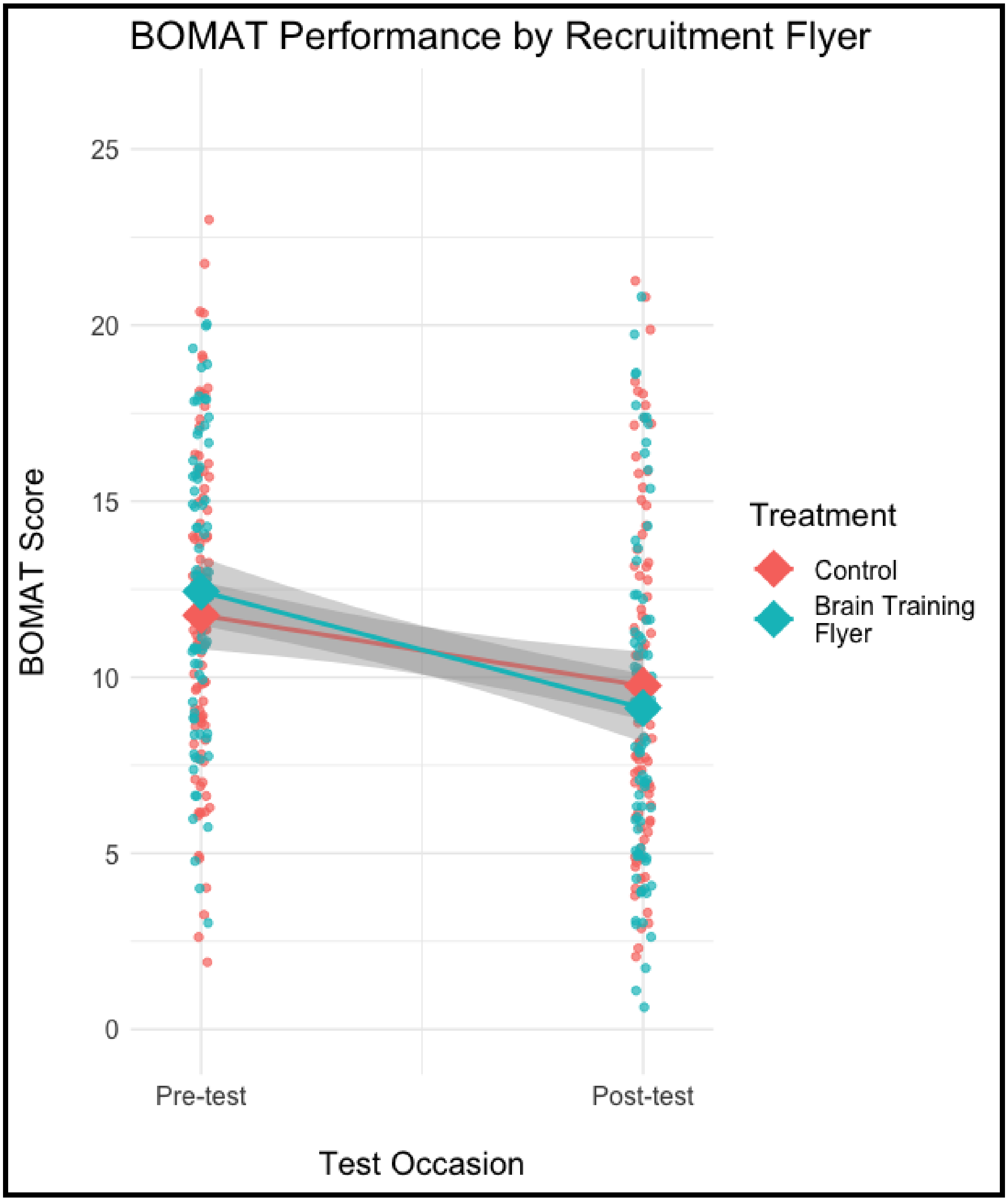

Figure 8. Objective performance and recruitment flyer.
BOMAT performance on the pretest and the posttest, split by participants recruited from the brain-training flyer and the control flyer. Diamonds show the mean of each group. Error bands show 95% confidence intervals.

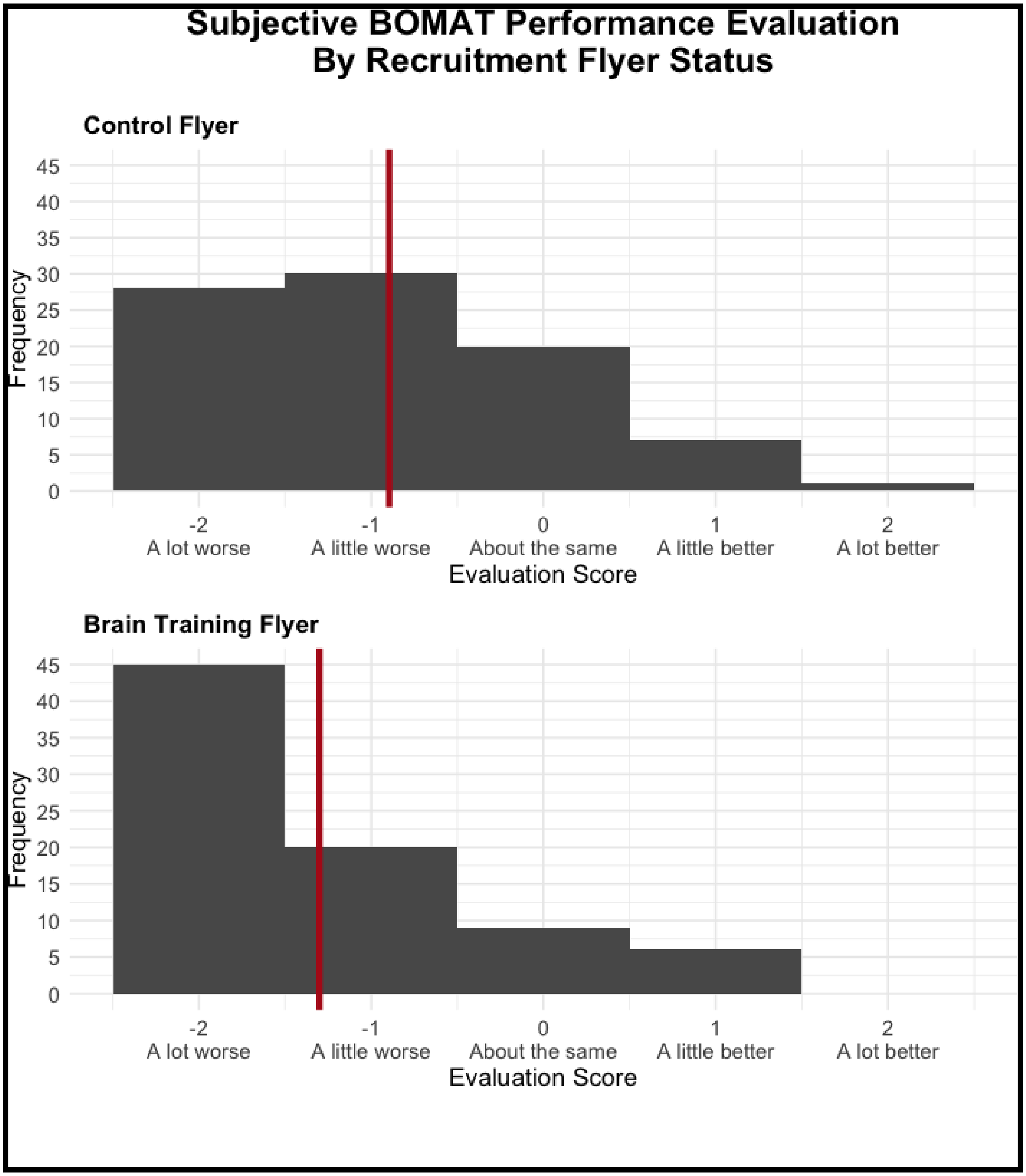

Figure 9. Subjective performance and recruitment flyer.
Histograms of subjective BOMAT performance evaluation for the control flyer (top) and the brain-training flyer (bottom). Vertical lines show the mean of each group. Recall that the prompt was: "*Compared to the first time you completed this Matrix task, how do you think you performed this time?*"

Interaction between the Flyer Manipulation and Open-Label Placebo Manipulation on the Primary Outcomes

I also explored the relationship that recruitment flyer status and open-label placebo status jointly had on either the objective or the subjective outcomes. Two sets of models tested the main effects of each variable, their interactions, and included the same covariates reported in hypothesis 1: reported effort, resource use, response time, etc. The first model had objective difference scores as the dependent variable, and the second model had subjective performance evaluation as the dependent variable. Aside from the effect that recruitment flyer status had on the two outcomes reported above, open-label placebo status never had a significant main effect nor an interaction with flyer recruitment status. As detailed in hypothesis 1, the covariates included in the models remained related to performance but are withheld for non-pertinence.

Explanatory Models: Effort Measures, Resource Measures, and Individual Differences

I followed up on the differences between the recruitment flyer groups on the primary outcomes by looking at self-reported effort and resource use. The group recruited from the brain training flyer reported using more resources during the pretest (M = -33.60, SD = 23.93) than the control group (M = -25.52, SD = 24.64), $t(165.26) = 2.15$, p = .03, Cohen's d = -0.33 (95% CI: -0.64, -0.03). The mean value reflects reported resources remaining before the test, subtracted from resources remaining after the test; negative values are the amount of resources 'lost' during the test. The group recruited from the brain training flyer also reported using less effort during the posttest (M = 4.00, SD =

2.81) than the control group (*M* = 5.58, *SD* = 2.61), $t(161.31) = 3.77$, $p < .001$, Cohen's *d* = -0.59 (95% CI: -0.90, -0.27), compared to the control group. Interestingly, there were no differences between the groups on the inverse measures: reported effort during the pretest ($p = .26$). and resource use during the posttest ($p = .76$).

Two mediation models explored the relationships between BOMAT performance, recruitment flyer status, and resource use/reported effort. Critically, participants reported their effort and resource use during the posttest *after* taking the posttest. Therefore, true mediation is impossible to test. The temporal timing of the events forbid it. For example, rather than effort predicting performance, it is possible that participants who performed worse subsequently reported using less effort as a way to save face. Keep that in mind when reading the following paragraphs. The mediation models were used to explore statistical relationships between the measures - not causal mechanisms.

Mediation models were run with bootstrapped standard errors (1000 iterations) and unstandardized variables. In the first model, reported resource use on the pretest was tested as a mediator between recruitment flyer status and the BOMAT difference score. However, evidence for mediation was non-significant: average indirect effect = -.08 (95% CI: -0.34, 0.08), $p = .41$. Only 5.71% of the total effect went through the mediator and there was still a significant direct effect between the recruitment flyer status and the BOMAT difference score ($p = .03$). In the second model, reported effort on the posttest was tested as a mediator - again, between recruitment flyer status and the BOMAT difference score. Evidence for 'full' mediation was strong: average indirect effect: = -0.39 (95% CI: -0.87, -0.08), $p = .004$. The effort on the posttest was responsible for

30.0% of the total effect and there was no longer a significant direct effect between recruitment flyer status and BOMAT difference score (p = .13).

Two additional mediation models explored why participants recruited from the brain training flyer evaluated their performance as worse. The results on this subjective outcome mirrored the results from the mediation models on the objective outcomes described in the previous paragraph. Specifically, there was weak evidence for resource use on the pretest meditating the relationship between recruitment flyer status and perception of performance: average indirect effect = -0.06 (95% CI: -0.17, .00), p = .052. Resource use on the pretest was responsible for only 16.4% of the total effect and there was still a significant direct effect between recruitment flyer status and perceived performance (p = .03). Finally, in the fourth model, there was strong evidence for effort on the posttest fully meditating the relationship between recruitment flyer status and perception of performance: average indirect effect = -0.18 (95% CI: -0.32, -0.07), p < .001. Effort on the posttest was responsible for 44.9% of the total effect and there was no longer a significant direct effect between recruitment flyer status and perceived performance (p = .14).

Lastly, I looked at the Need for Cognition scale and the Theories of Intelligence scales. First, I tested whether participants recruited from the brain training flyer had differences on the Need for Cognition, compared to those recruited from the control group. There was no difference between groups on Need for Cognition, t(162.66) = .20, p = .84, Cohen's d = -0.03 (95% CI: -0.34, 0.27). Recall from the manipulation check that participants recruited from the brain-training flyer had higher beliefs in the flexibility of intelligence, compared to those recruited from the control flyer. I followed up on the flyer's main effect with a test for interactions. Indeed, recruitment flyer status and beliefs about Theories of Intelligence interacted significantly to predict BOMAT difference scores, $F(3, 162) = 7.34$, $p < .001$, Adj $R^2 = .10$. There was a positive relationship between Theories of Intelligence and BOMAT difference scores for those recruited from the Brain Training flyer, but not the control flyer (Figure 10). Additionally, recruitment flyer status and Theories of Intelligence interacted significantly to predict effort on the posttest, $F(3, 162) = 6.05$, $p < .001$, Adj $R^2 = .08$. The relationship between theories of intelligence and posttest effort was negative for those recruited from the control flyer, but positive for those recruited from the brain training flyer (Figure 11). Finally, in two regression models predicting BOMAT difference scores from Theories of Intelligence and posttest effort, posttest effort was the only significant predictor for those recruited by the control flyer, while Theories of Intelligence was the only significant predictor for those recruited from the brain training flyer.

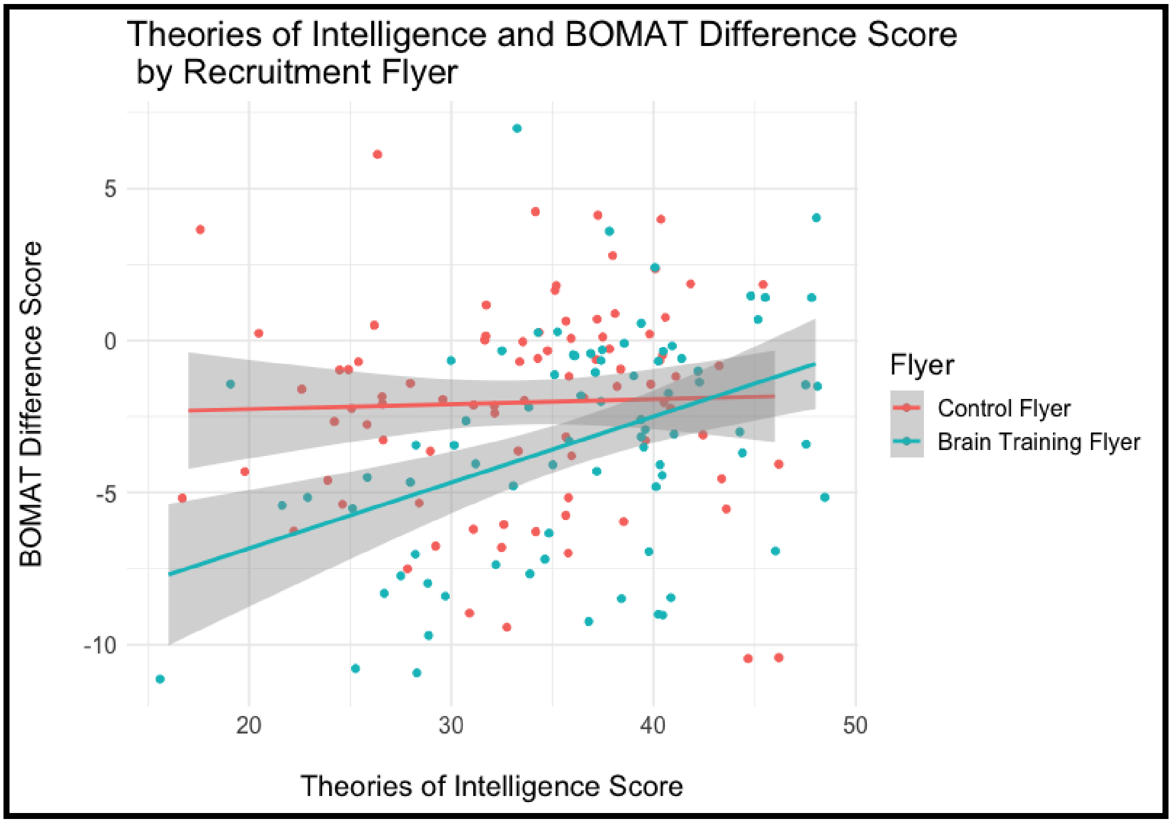

Figure 10. Objective performance, Theories of Intelligence, and recruitment flyer.
The relationship between the Theories of Intelligence scale and BOMAT difference scores, split by recruitment flyer status. Error bands show 95% confidence intervals.

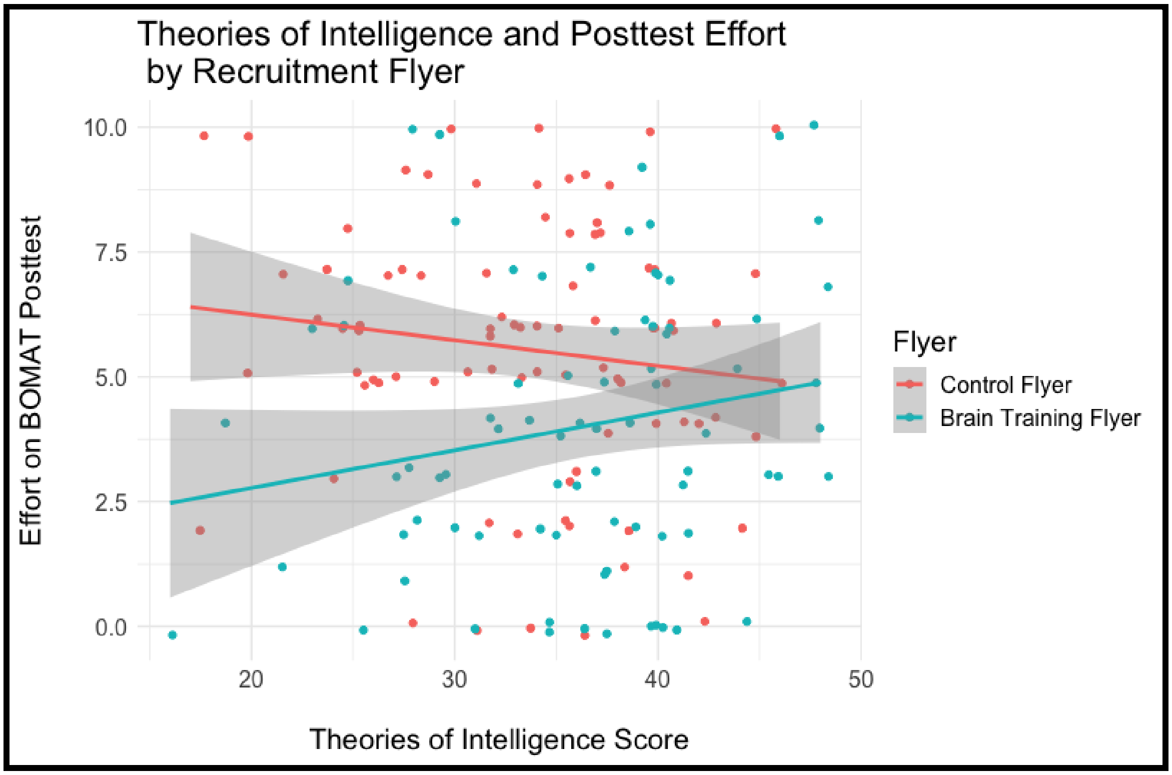

Figure 11. Posttest effort, Theories of Intelligence, and recruitment flyer.
The relationship between the Theories of Intelligence scale and self-reported effort on the BOMAT posttest, split by recruitment flyer status. Error bands show 95% confidence intervals.

Summary for Hypothesis 6

To summarize, participants recruited from the brain training flyer, compared to the control flyer, had worse BOMAT difference scores, and subjectively evaluated their BOMAT difference score as worse. They reported using more resources on the pretest, and less effort on the posttest. Further, reported effort on the posttest explained the relationship between recruitment flyer status and both outcomes (objective and subjective BOMAT performance). There were no interactions between recruitment flyer status and open-label placebo status on objective or subjective outcomes. Lastly, beliefs about the malleability of intelligence, manipulated by (or differential self-selection from) the recruitment flyer, was a protective factor for BOMAT difference scores. The protective factor is partially explained by the positive relationship between Theories of Intelligence and posttest effort that only exists for those recruited from the brain training flyer.

Chapter Four: Discussion

I did not replicate the open-label placebo effect. There were no differences between the open-label placebo group and the control group in the current experiment. The lack of differences between groups on the objective outcomes is consistent with past open-label placebo research. However, the lack of differences between groups for the subjective outcomes is inconsistent with past research. Prior meta-analytic work suggests that open-label placebos can elicit improvement, specifically on subjective symptoms (Charlesworth et al. 2017). In the following paragraphs, I attempt to reconcile my discrepant results. Possible explanations why the results were the way that they were include 1) a true null result, 2) an ineffective manipulation and conditioning protocol, and 3) noisy data from a lengthy self-directed online experiment. The true explanation, of course, is unknown. Nonetheless, this dissertation offers methodological contribution to a burgeoning area of work.

Explanation 1: A True Null Result

With any null findings, the first takeaway ought to be that the results are valid. It is all too easy in hindsight to attack our studies with methodological criticism, explaining away the lack of effect. In the following paragraphs, I defend the null result. In order, I discuss design strengths of sample size and methodological differences, the file-drawer problem, and power.

Sample Size

A common critique of open-label placebo work is that the sample sizes often used in experiments are small (Kaptchuk & Miller, 2018; Amorim et al., 2020). Small samples are subject to sampling error, and thus, the implicit concern is that the open-label placebo effect might fail to replicate in a larger sample. Accordingly, my study had 168 participants and did not observe an effect. My sample was more than double and up to eight times as large as samples in prior work (Sandler & Bodfish, 2008; Kaptchuk et al., 2010, Kelley et al., 2012; Carvalho et al., 2016; Schaefer et al., 2016). Therefore, it is possible that the results here accurately reflect a nonsignificant population level effect. Future large sample open-label placebo research will help determine the validity of the explanation that my large sample was the cause of a null result. To look forward, there is an open-label placebo randomized control trial on premenstrual syndrome in progress that will have up to 150 participants (Nascimento et al., 2020).

Methodological Differences

In addition to sample size, another *a priori* methodological strength of the current study was that the protocols matched between groups. Often in open-label placebo trials, the protocol does not match between groups. Typically, the placebo group does something (i.e. takes pills), while the control group does nothing. In my study, the open-label placebo group and the control group completed nearly identical tasks. Unfortunately, it is difficult to make definitive claims for how methodological rigor relates to the effect (Blease et al., 2019). It appears there are more questions than answers. Contradictory evidence can be found when comparing matched and unmatched protocols with observation of an open-label placebo effect. For example, both Locher et al. (2017) and Kaptchuk et al. (2010) found an effect but the former had matched protocols while the latter did not. On the other hand, the current study and Leibowitz et al. (2019) did not observe a main effect from an open-label placebo treatment, but I used matched protocols and Leibowitz et al. (2019) did not. Unfortunately, the uncertainty remains when considering other explanations too. For example, similarly to the presence of experimentally manipulated expectations, a rationale about open-label placebos provided with the treatment strengthened the effect in Locher et al. (2017), but not in Leibowitz et al. (2019).

Briefly, most traditional open-label placebo interventions help ailed samples get back to baseline. However, here, I tried to *enhance* a healthy sample. The inherent differences between the two populations might be a cause of the discrepant findings. Perhaps sick or otherwise affected participants are more suggestible and therefore more susceptible to benefit from open-label placebo treatments. Intuitively, sick participants benefit directly from improvement. On the other hand, the participants in my study do not - they got no feedback on their BOMAT test scores and have little motivation to improve, regardless of which group they were in. Perhaps, the benefit that a participant perceives a treatment to give them is critical for the effect. And, as was the case here, when there is little incentive to improve, there is also low likelihood of observing the open-label placebo effect.

The 'File-Drawer' Problem

Further, although many studies support the existence of open-label placebo effects, there is no way of knowing how many open-label placebo trials have failed. Referred to as the '*file drawer problem'*, it is unclear if the current study was one of only a few non-significant experiments, or if it was one of many. A metric called fail-safe N calculates how many hypothetical non-significant studies would need to exist in the 'file-drawer' (i.e. not published) to render the conclusions of a meta-analysis non-significant. Here, I calculate three fail-safe N statistics for Charlesworth et al. (2017). Rosenthal's (1979) fail-safe N = 74 null studies would be needed to bring the effect size of 0.88 from Charlesworth et al. (2017) down to a non-significant level. Orwin's (1983) statistic allows determining a clinically *in*significant value to bring the meta-analytic effect down to (e.g. d = 0.20, rather than 0.0). Here, Orwin's statistic = 17 null studies would be needed to bring the meta-analytic effect size down to .2. Finally, Rosenberg's (2005) statistic has the benefit of combining the new hypothetical evidence with a *weighted* meta-analytic average effect size. Accordingly, Rosenerg's statistic = 53 null studies would be needed to nullify the meta-analytic effect.

Importantly, fail-safe N does not account for nor identify publication bias; rather, it is simply an estimate of if publication bias might exist. Here, the meta-analytic effect comes from 5 studies, and the fail-safe N statistics are comparatively small (fail-safe N values in the thousands is common in meta-analytic literature). Evidence for the existence and generalizability of the open-label placebo effect is weakened by the null results from the current study, the comparatively low fail-safe N statistics, and the recency of empirical investigation into the open-label placebo effect.

A Comment on Power

The primary goal of the current study was to determine whether or not an open-label placebo treatment could improve objective or subjective cognitive performance. As I wrote in the *a priori* required sample size estimations, power was set at .8, alpha at .05 for a two-tailed test, and the expected effect size in a range between .39 and .88. A limitation in the required sample calculations was that the effect sizes I used came from two different phenomena. Two of the three referenced effect sizes were reported from traditional placebo work (in the same study). The third referenced effect size was from a meta-analysis on open-label placebo effects. Importantly, my study planned to add an open-label component to a placebo cognitive training effect. Hence, the justification for referencing effect sizes from studies investigating different effects.

With 168 participants, two of the three referenced effect sizes supported that the study was sufficiently powered to find an effect, should one exist. Of course, I did not find an effect. *A posteriori* power was calculated for each hypothesis. For the BOMAT difference score, *a posteriori* power was 17.7% (Kane, 2018). For the subjective performance evaluation, *a posteriori* power 7.5%. Such low power levels imply one of two things: my sample was not large enough, or my expected effect size estimates were inaccurate. Most likely, it was the latter: recall that Cohen's *D* was -0.05 for the BOMAT difference score between open-label placebo and control groups, and 0.08 for the subjective performance evaluation between groups. Both statistics were not significantly different from zero and far from anything practically meaningful. Perhaps, I was faulty in my assumption that the placebo cognitive training and its subsequent open-label placebo component could be combined. Perhaps combining the two effects, as I did, results in a null effect.

Explanation 1 Conclusion

To conclude, the null results from the current study might be explained by a larger sample more accurately describing a population level effect. The results also might reflect the presence of certain design elements like matched controls or the population that the participants were sampled from. My study might be one of many unpublished similar studies and could reflect the correct inferential decision (failure to reject a true null). Put simply, the new empirical excitement about open-label placebo effects might be overstated. However, in the paragraphs below, I present alternative explanations of why my study might have missed an effect that does really exist (a type 2 error).

Explanation 2: Ineffective Manipulation and Conditioning

It is possible that the concept of cognitive training as a tool for cognitive improvement is too distant from the thoughts of a general population. The disconnect between the stimulus and response increases when considering the placebo and open-label placebo nature of my experiment. Brunswick, a famous philosopher of science, refers to the term *ecological validity* as the potential utility of cues in an organism's environment (1956). In other words, it is the conceptual distance between a stimulus and its response. For example, a manipulation using a pill to treat pain has high ecological validity -- it is commonly understood that pills have utility for pain relief.

However, in my study, I introduced participants to open-label placebo effects, then to the idea of cognitive training, and then to the idea that cognitive training is effective, but likely only via a placebo effect. Finally, I tried to elicit an open-label placebo response via open-label placebo cognitive training. In short, there were a lot of cognitive steps to internalizing the design. Perhaps, in the minds of my participants, the stimulus was not close to the intended response: the concept that open-label placebo cognitive training can be used to improve cognitive performance was not sufficiently conditioned. Brunswick might conclude the manipulation suffered from low ecological validity.

The ecological validity argument is strengthened from a conclusion in an essay by Gendler (2008). She argues that disagreements between subconscious and conscious beliefs are typically won by the subconscious. In early trials, participants expressed surprise at the idea of taking open-label placebo pills and that they weren't expecting much (Kaptchuk, 2018). Still, those participants' symptoms decreased. Outwardly, they knew that taking inert pills was silly, but their subconscious had such a strong conditioned response between 'pill' and 'symptom relief' that it overcame the incredulity, resulting in improvement. In the case of my study, the subconscious might not have had a firm grasp on the concept that open-label placebo cognitive training can result in cognitive improvement. Therefore, there was no way for the subconscious to overcome the conscious disbelief.

A final comment on conditioning comes from Leibowitz et al. (2019) in their discussion of their failed open-label placebo manipulation. In their study, participants received a brief description of open-label placebo effects - much shorter than the rationale provided in most other work. The same was true for my study. Here, participants watched a 2-minute video introducing open-label placebo effects, followed by one screen of less than 100 words describing their condition assignment. Leibowitz et al. (2019) make the argument that there might be a relationship between rationale length and presence of an effect (perhaps, the effect also requires a person-to-person interaction, rather than a video and text on a screen). If their conclusion is true, it is unlikely that my open-label placebo manipulation reached the critical threshold. At a deeper level, ignoring placebo and open-label placebo effects, there were no differences in the beliefs about the efficacy of brain training between participants recruited from the brain training flyer compared to the control flyer. The similarity between groups suggests there was insufficient unconscious conditioning, or a lack of difference in participant self-selection. If a manipulation specifically designed to change beliefs about brain training's effectiveness or differentially recruit participants failed to do those things, then the second brief video manipulation describing open-label placebo effects had no framework upon which to build.

Of course, I chose my study's design by weighing certain methodological pros and cons. Specifically, using an online paradigm allowed for faster recruitment of a large and diverse sample, the measurement of objective and subjective outcomes, and a proof of concept paradigm for easily scalable open-label placebo research. All decisions come with trade-offs. Nevertheless, the prevailing weakness of my experiment may be that it lacked ecological validity and a sufficiently detailed explanation of open-label placebo treatments.

Explanation 3: Noisy Data from an Unmotivated Sample

The final explanation for the current results is simple; the experiment required too much work. Participants watched a video, completed three difficult cognitive tasks (the pretest, the n-back, and the posttest), two surveys, and a number of repeated single item measures. Importantly, they completed all tasks unsupervised and on their own time.

To combat inattention, I added non-intrusive attention checks and screened participants accordingly. Specifically, the checks assessed recall of task components (e.g. "*which shapes were the blocks in the n-back task?*"), seemingly assessing memory during a cognitive experiment. Further, I ran many models to try and control for other confounds. As mentioned in the results, I included raw response times, a composite ratio of time spent on both tests, and other metrics. Although these confounds had direct relationships with cognitive improvement and general performance throughout the experiment, none resulted in a significant open-label placebo effect; nor did any of the confounds have significant interactions with open-label placebo status. Despite the inability to control for confounds, it is critical to understand that modeling procedures can only overcome a certain amount of noise. Statistical controls in regression models can help explain some non-random error, but they cannot fix a severely damaged dataset. Below, I provide comparisons that show how severely my data were affected.

I based the majority of my experiment off of the paradigm used in Foroughi et al. (2016). In their experiment, they recruited participants via one of two flyers (just like mine) and saw that participants recruited via the brain training flyer had improved cognitive performance, relative to those recruited from the control flyer (using the same BOMAT and n-back measures). Their effect was not replicated in my experiment. In fact, my results were the opposite. Participants recruited from the brain training flyer performed worse than those recruited from the control flyer. Critically, their experiment was done in person with an experimenter present and over the course of two days. Supervision can facilitate conscientious responding and effort (e.g. social facilitation/Hawthorne/Rosenthal effects). In addition, breaking up the experiment across two days gives participants a chance to rest. Because my experiment did not include rest or supervision, perhaps the participants responded carelessly or without motivation. The data support the carelessness hypothesis. Regardless of any interventions, typical average BOMAT scores are between 17 and 19 (Jaeggi et al., 2014; Foroughi et al., 2016). The average score of my participants was 10.77. Further, participants generally improve from BOMAT pretest to posttest, but mine got worse (Figure 12).

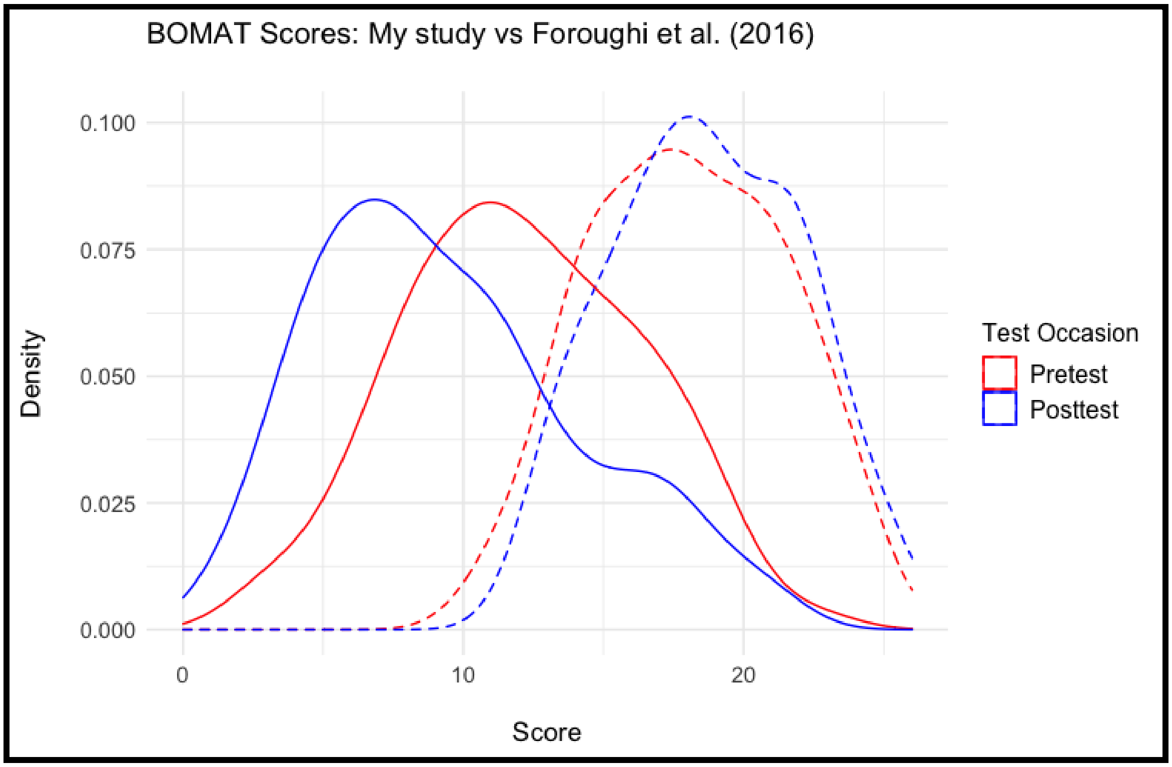

Figure 12. Benchmarking the current results.
Density plot of BOMAT scores for participants in the current study (solid lines), compared to the scores of participants in Foroughi et al. (2016) (dashed lines).

Additionally, regardless of condition, participants from Foroughi et al. (2016) had correlations of r = .95 between BOMAT pretest and posttest scores; the correlation for my participants was r = .69. The correlations in Jaeggi et al. (2014) had a large range between r = .42 and r = .72. Further, among many cognitive tests, they noticed a positive correlation between reliability estimates and effect size estimates: the higher the reliability was for a measure, the stronger the cognitive training effect size was. All of the correlations reported here are confounded by interventions between pretest and post-test; therefore, they might reflect lower bound estimates of reliability. (Jaeggi et al., 2014). As far as I could find, there are no pure test-retest estimates for the BOMAT or psychometric validation studies. Therefore, the reliability and validity of the measure itself are not without question. Lastly, the BOMAT is intended to measure cognitive ability, not effort or fatigue. If my measurements were so heavily impacted by those confounds, then perhaps the BOMAT scores don't reflect cognitive reasoning ability at all. More clearly, the paragraphs above detail how the data from my experiment were messy - this paragraph detailed how the measure itself was messy.

Finally, when looking at average BOMAT response times for each item there is worrisome trend. The expectation was that, for each test, the response time (and variance) would increase with item difficulty and then plateau off. Indeed, the trend did exist in the pretest for the first 20 or so items. However, rather sharply, response time per item drops (and variance decreases) and then stays there for the remainder of the experiment - even for the easy items in the posttest (Figure 13). The drop in response time and variability indicates that most of my participants stopped trying midway through the pretest.

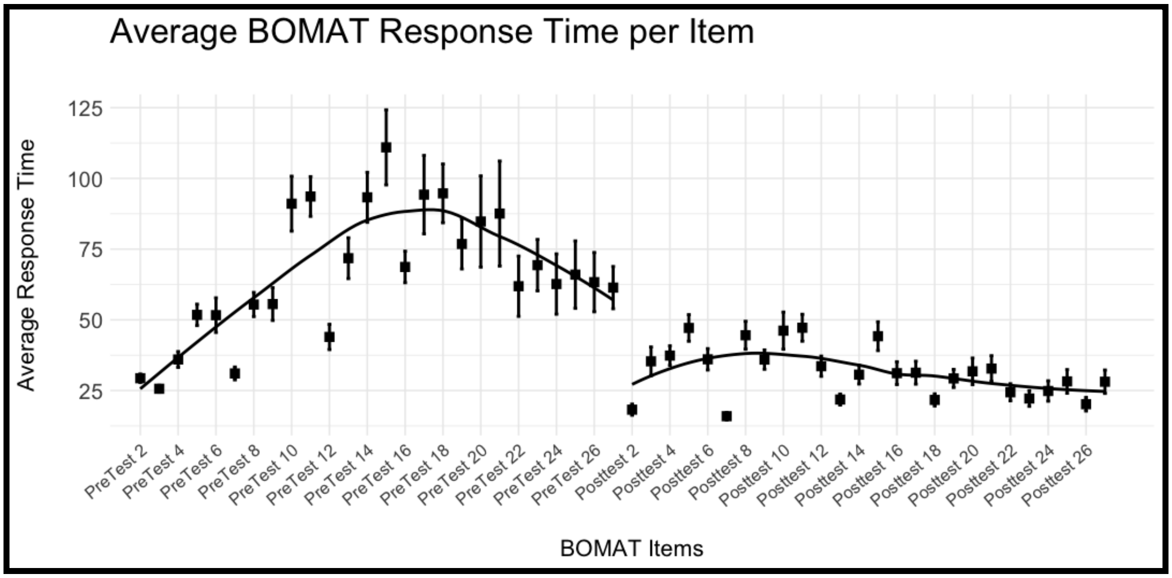

Figure 13. Average BOMAT item response time.
Error bars reflect +/- 1 standard error. As in past research with the BOMAT, item one for each test was dropped prior to analyses (only items 2-26 were scored).

Although it might support open-label placebo theory to explain away the null-results by throwing up my hands and saying the data were simply too messy, there are counterarguments to make. As stated previously, I attempted to model out confounds such as response times and screen via attention checks. Further, certain relationships between measures existed that suggest at least some level of conscientious responding occured. For example, although the correlations between BOMAT scores observed in Foroughi et al. (2016) were higher, I still observed a strong positive relationship. I also observed correlations between the reported effort and resource use participants gave on both tests, BOMAT performance and n-back performance, participants evaluation of their performance and their objective performance, and others. Further, although the open-label placebo manipulation did nothing, the recruitment flyer manipulation resulted in significant differences between groups. Therefore, to claim that the open-label placebo manipulation did nothing because the data were too messy or the tasks too hard, ignores the fact that there was some reliable variance for an effect to partition, should the effect actually exist. Of course, it's possible that all explanations have some truth. Perhaps, open-label placebo effects are not as strong as expected *and*, because of certain design elements, my study was unable to find any effect that might actually exist.

Methodological Offering of the Current Study

Despite the null outcome of the current study, many takeaways and recommendations can be derived from its conclusions. Primarily, open-label placebo effects might have limited generalizability. Perhaps only populations with intrinsic motivation for improvement can experience benefits (e.g. sick patients). Further, they might only experience those benefits on subjective outcomes. Researchers must continue including both objective and subjective outcomes in future research to determine where the effect can be observed. Similarly, it is still unclear which antecedents of the effect are necessary and/or sufficient. The field needs more high-powered studies that investigate the effect of protocol matching, manipulating expectations, intervention length and style, and how well an intervention may internalized (i.e. ecological validity). As an aside on power, take care in determining *a priori* power. Perhaps open-label placebo effect sizes are not as large as previously thought; they may be less than the 0.88 average standardized mean difference from Charlesworth et al. (2017).

Chapter Five: Conclusions

The current paper describes a study that used open-label placebo cognitive training to try and improve cognitive performance. The manipulation was ineffective; there were no differences in objective or subjective cognitive performance between the open-label placebo and the control groups. I provided three explanations for why my study turned out this way. The first was that the null result is valid, the second was that my manipulation lacked ecological validity, and the third was that the experiment was too difficult. In light of past open-label placebo work and the methodological shortcomings of the study, each explanation likely has some truth. The first explanation (that the result is valid) ought to temper the expectations of future research. Open-label placebo effects might be limited in strength and generalizability. However, I am most persuaded by the second and third explanations. Even if open-label placebo effects are limited, the current study was likely not suited to observe one. Researchers in the field ought to learn from the shortcomings of the current study when designing their work. The data and arguments herein exemplify the outcomes that result from certain methodological considerations and describe how those considerations influence open-label placebo treatments.

References

Amorim, A. B., Machado, G. C., & Maher, C. G. (2020). Enthusiastic claims for open-label placebo pills ignore the evidence. Pain, 161(5), 1124. https://doi: 10.1097/j.pain.0000000000001803

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: a meta-analysis. Psychonomic bulletin & review, 22(2), 366-377. https://doi.org/10.3758/s13423-014-0699-x

Auguie, B. (2017). gridExtra: Miscellaneous functions for "grid" graphics. (Version 2.3) [Computer software]. https://CRAN.R-project.org/package=gridExtra

Beecher, H. K. (1955). The powerful placebo. Journal of the American Medical Association, 159(17), 1602–1606. https://doi.org/jama.1955.02960340022006

Beedie, C. J., & Foad, A. J. (2009). The placebo effect in sports performance. Sports Medicine, 39(4), 313-329. https://doi.org/10.2165/00007256-200939040-00004

Behrendt, S. (2014). lm.beta: Add standardized regression coefficients to lm-objects. (Version 1.5-1) [Computer software]. https://CRAN.R-project.org/package=lm.beta

Bell, B. A., Kromrey, J. D., & Ferron, J. M. (2009). Missing data and complex samples: The impact of listwise deletion vs. subpopulation analysis on statistical bias and hypothesis test results when data are MCAR and MAR. Proceedings of the Joint Statistical Meetings, Survey Research Methods Section, 26, 4759-4770.

Benedetti, F., & Amanzio, M. (1997). The neurobiology of placebo analgesia: From endogenous opioids to cholecystokinin. Progress in Neurobiology, 52(2), 109–125. https://doi.org/10.1016/S0301-0082(97)00006-3

Blease, C. R., Bernstein, M. H., & Locher, C. (2019). Open-label placebo clinical trials: Is it the rationale, the interaction or the pill? BMJ Evidence-Based Medicine. http://dx.doi.org/10.1136/bmjebm-2019-111209

Brant, R. (n.d.) Inference for means: Comparing two independent samples. https://www.stat.ubc.ca/~rollin/stats/ssize/n2.html

Brunswik, E. (1956). Perception and the representative design of psychological experiments. (2nd ed.). University of California Press.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. Journal of Personality and Social Psychology, 42(1), 116-131. https://doi.org/10.1037/0022-3514.42.1.116

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. Journal of Personality Assessment, 48(3), 306-307. https://doi.org/10.1207/s15327752jpa4803_13

Carvalho, C., Caetano, J. M., Cunha, L., Rebouta, P., Kaptchuk, T. J., & Kirsch, I. (2016). Open-label placebo treatment in chronic low back pain: A randomized controlled trial. Pain, 157(12), 2766-2772. https://doi.org/10.1097/j.pain.0000000000000700

Charlesworth, J. E., Petkovic, G., Kelley, J. M., Hunter, M., Onakpoya, I., Roberts, N., Miller, F. G., & Howick, J. (2017). Effects of placebos without deception compared with no treatment: A systematic review and meta-analysis. Journal of Evidence-Based Medicine, 10(2), 97-107. https://doi.org/10.1111/jebm.12251

Clark, C. M., Lawlor-Savage, L., Goghari, V. M., (2017). Working memory training in healthy young adults: Support for the null from a randomized comparison to active and passive control groups. PLoS ONE, 12(5), e0177707. https://doi.org/10.1371/journal.pone.0177707

de Craen, A. J. M., Moerman, D. E., Heisterkamp, S. H., Tytgat, G. N. J., Tijssen, J. P. G., & Kleijnen, J. (1999). Placebo effect in the treatment of duodenal ulcer. British Journal of Clinical Pharmacology, 48(6), 853– 860. https://doi.org/10.1046/j.1365-2125.1999.00094.x

Dowle, M. & Srinivasan, A. (2019). data.table: Extension of `data.frame`. (Version 1.12.6) [Computer software]. https://CRAN.R-project.org/package=data.table

Dweck, C. S. (2000). Self-theories: Their roles in motivation, personality, and development. Psychology Press.

El Brihi, J., Horne, R., & Faasse, K. (2018). Prescribing placebos: An experimental examination of the role of dose, expectancies, and adherence in open-label placebo effects. Annals of Behavioral Medicine, 53(1), 16-28. https://doi.org/10.1093/abm/kay011

Foroughi, C. K., Monfort, S. S., Paczynski, M., McKnight, P. E., & Greenwood, P. M. (2016). Placebo effects in cognitive training. Proceedings of the National Academy of Sciences, 113(27), 7470-7474. https://doi.org/10.1073/pnas.1601243113

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. (Version 1.7.4) [Computer Software]. Journal of Statistical Software, 40(3), 1-25. http://www.jstatsoft.org/v40/i03/

Gendler, T. S. (2008). Alief and belief. The Journal of Philosophy, 105(10), 634-663.

Hossiep, R., Turck, D., & Hasella, M. (1999). Bochumer Matrizentest (BOMAT) Advanced.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. Proceedings of the National Academy of Sciences, 105(19), 6829-6833. https://doi.org/10.1073/pnas.0801268105

Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y. F., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning: Implications for training and transfer. Intelligence, 38, 625-635. https://doi.org/10.1016/j.intell.2010.09.001

Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. Memory and cognition, 42(3), 464-480. https://doi.org/10.3758/s13421-013-0364-z

Kane, SP. (2018, November 10). Post-hoc power calculator. ClinCalc. https://clincalc.com/stats/power.aspx

Kang, H. (2013). The prevention and handling of the missing data. Korean Journal of Anesthesiology, 64(5), 402-406. https://doi.org/10.4097/kjae.2013.64.5.402

Kaptchuk, T. J. (2018). Open-label placebo: Reflections on a research agenda. Perspectives in Biology and Medicine, 61(3), 311-334. https://doi.org/10.1353/pbm.2018.0045

Kaptchuk, T. J., & Miller, F. G. (2018). Open label placebo: Can honestly prescribed placebos evoke meaningful therapeutic benefits? BMJ, 363, k3889. https://doi.org/10.1136/bmj.k3889

Kaptchuk, T. J., Friedlander, E., Kelley, J. M., Sanchez, M. N., Kokkotou, E., Singer, J. P., Kowalczykowski, M., Miller, F. G., Kirsch, I., & Lembo, A. J. (2010). Placebos without deception: a randomized controlled trial in irritable bowel syndrome. PloS ONE, 5(12), e15591. https://doi.org/10.1371/journal.pone.0015591

Kelley, J. M., Kaptchuk, T. J., Cusin, C., Lipkin, S., & Fava, M. (2012). Open-label placebo for major depressive disorder: A pilot randomized controlled trial. Psychotherapy and Psychosomatics, 81(5), 312-314. https://doi.org/10.1159/000337053

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. Journal of Experimental Psychology, 55(4), 352–358. https://doi.org/10.1037/h0043688.

Leibowitz, K. A., Hardebeck, E. J., Goyer, J. P., & Crum, A. J. (2019). The role of patient beliefs in open-label placebo effects. Health Psychology, 38(7), 613-622. https://doi.org/10.1037/hea0000751

Locher, C., Nascimento, A. F., Kirsch, I., Kossowsky, J., Meyer, A., & Gaab, J. (2017). Is the rationale more important than deception? A randomized controlled trial of open-label placebo analgesia. Pain, 158(12), 2320-2328. https://doi.org/10.1097/j.pain.0000000000001012

Millisecond Software. (2015). Inquisit 5 dual-n-back task—adaptive [Computer software]. Retrieved from https://www.millisecond.com.

Meeuwis, S. H., Van Middendorp, H., Veldhuijzen, D. S., Van Laarhoven, A. I., De Houwer, J., Lavrijsen, A. P., & Evers, A. W. (2018). Placebo effects of open-label verbal suggestions on itch. Acta Dermato-Venereologica, 98(1-2), 268-274. https://doi.org/10.2340/00015555-2823

Monfort, S. S., Graybeal, J. J., Harwood, A. E., McKnight, P. E., & Shaw, T. H. (2018). A single-item assessment for remaining mental resources: Development and validation of the Gas Tank Questionnaire (GTQ). Theoretical Issues in Ergonomics Science, 19(5), 530-552. https://doi.org/10.1080/1463922X.2017.1397228

Nascimento, A. F., Gaab, J., Kirsch, I., Kossowsky, J., Meyer, A., & Locher, C. (2020). Open-label placebo treatment of women with premenstrual syndrome: Study protocol of a randomised controlled trial. BMJ Open, 10(2), e032868. http://dx.doi.org/10.1136/bmjopen-2019-032868

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. Journal of Educational Statistics, 8(2), 157-159.

Park, L., & Covi, L. (1965). Nonblind placebo trial. Archives of General Psychiatry 12(4), 336–345. https://doi.org/10.1001/archpsyc.1965.01720340008002

R Core Team. (2019). R: A language and environment for statistical computing. (Version 3.6.1) [Computer software]. https://www.R-project.org/

Raven, J. C., Court, J. H., Raven, J., & Kratzmeier, H. (1994). Advanced Progressive Matrices. Oxford Psychologists Press.

Revelle, W. (2020) psych: Procedures for personality and psychological research. (Version 2.0.7) [Computer Software]. https://CRAN.R-project.org/web/packages/psych/index.html

Rief, W., Nestoriuc, Y., Weiss, S., Welzel, E., Barsky, A. J., & Hofmann, S. G. (2009). Meta-analysis of the placebo response in antidepressant trials. Journal of Affective Disorders, 118(1-3), 1-8. https://doi.org/10.1016/j.jad.2009.01.029

RStudio Team. (2020). RStudio: Integrated Development for R. (Version 1.3.959) [Computer software]. RStudio, PBC. http://www.rstudio.com/

Rudebeck, S. R., Bor, D., Ormond, A., O'Reilly, J. X., & Lee, A. C. (2012). A potential spatial working memory training task to improve both episodic memory and fluid intelligence. PloS ONE, 7(11), e50431. https://doi.org/10.1371/journal.pone.0050431

Sandler, A. D., & Bodfish, J. W. (2008). Open-label use of placebos in the treatment of ADHD: A pilot study. Child: Care, Health and Development, 34(1), 104-110. https://doi.org/10.1111/j.1365-2214.2007.00797.x

Schaefer, M., Harke, R., & Denke, C. (2016). Open-label placebos improve symptoms in allergic rhinitis: A randomized controlled trial. Psychotherapy and Psychosomatics, 85(6), 373–374. https://doi.org/10.1159/000447242

Schaefer, M., Sahin, T., & Berstecher, B. (2018). Why do open-label placebos work? A randomized controlled trial of an open-label placebo induction with and without extended information about the placebo effect in allergic rhinitis. PloS ONE, 13(3), e0192758. https://doi.org/10.1371/journal.pone.0192758

Sjoberg, D. (2019). hablar: Non-astonishing results in R. (Version 0.3.0) [Computer software]. https://CRAN.R-project.org/package=hablar

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. Psychological Bulletin, 86(3), 638-641. https://doi.org/10.1037/0033-2909.86.3.638

Rosenberg, M. S. (2005). The file-drawer problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. Evolution, 59(2), 464-468. https://doi.org/10.1111/j.0014-3820.2005.tb01004.x

Warnes, G. R., Bolker, B., Gorjanc, G., Grothendieck, G., Korosec, A., Lumley, T., MacQueen, D., Magnusson, A., & Rogers, J. (2017). Gdata: Various R programming tools for data manipulation. (Version 2.18.0) [Computer Software]. https://CRAN.R-project.org/package=gdata

Wickham, H. (2007). Reshaping data with the reshape package. Journal of Statistical Software, 21(12), 1-20.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis.(Version 3.3.1) [Computer software]. https://CRAN.R-project.org/web/packages/ggplot2/index.html)

Wickham, H. (2019). stringr: Simple, consistent wrappers for common string operations. (Version 1.4.0) [Computer software]. https://CRAN.R-project.org/package=stringr

Wickham, H., & Bryan, J. (2019). readxl: Read excel files. (Version 1.3.1) [Computer software]. https://CRAN.R-project.org/package=readxl

Wickham, H., François, R., Henry, L., & Müller, K. (2020). dplyr: A grammar of data manipulation. (Version 1.0.0) [Computer software]. https://CRAN.R-project.org/package=dplyr

Willse, J. T. (2018). CTT: Classical test theory functions. (Version 2.3.3) [Computer software]. https://CRAN.R-project.org/package=CTT