Cognitive Recovery With Cannabis Abstinence Among
High School-Aged Adolescents
NCT03276221
Document Date: 04/28/2023

Cognitive Recovery with Cannabis Abstinence Among High-school-aged Adolescents

Statistical Analysis Plan

The ARCHES Team

NCT: 03276221

First created 2/12/2022

Last revised 04/28/2023

Unblinded analysis 02/03/2023

Version 2.2

Table of Contents

Study Design

The study consisted of a one-month (with a two-month follow-up) longitudinal trial of cannabis abstinence (with data collection from 9/2017 to 12/2022) designed to determine:

1. Whether cognitive performance improves for students abstinent from cannabis use for one month above and beyond any improvement observed for students engaging in continued use.
2. When cognitive performance improves (if improvement during abstinence does occur).
3. Whether cognitive performance in abstinent individuals "returns" to performance levels observed in non-using peers.

Participants

The study planned to recruit 210 adolescents across the following conditions:

1. A control group of non-cannabis users (Non-users; n = 70).
2. Regular users of cannabis (n = 140) randomized to either…
    a. Contingency management intervention that incentivized 4 weeks of cannabis abstinence (CB-Abst; n = 70).
    b. Non-contingent monitoring with no abstinence requirement (CB-Mon; n = 70).

All participants completed cognitive assessments, toxicology testing, self-report questionnaires and semi-structured mood and substance use interviews during the 4-week study as well as one 30-day follow-up visit. Abstinence was verified using the statistical model of Schwilke and colleagues (2011). The approach differentiates residual cannabinoid excretion from new cannabis exposure by comparing observed urine metabolite ratios against expected ratios. Observed ratios that exceed a 95% confidence interval indicate new cannabis use.

Order of data collection and creation of the analysis plan

We acknowledge that under ideal circumstances outcomes should be clearly defined along with a corresponding analysis plan prior to any data collection. We note that for this study, specific primary outcomes were defined and the analytic plan written near the end of the 5-year data collection, and for full disclosure, that 4 papers had already been published on data collected from the study.

- Schuster and colleagues (2020) created a pharmacokinetic model of THC elimination using data on urine metabolites.
- Schuster and colleagues (2021) examined alcohol use over the month of abstinence and the follow-up session.
- Cooke and colleagues (2021) examined anxiety and depression symptoms over the month of abstinence and the follow-up session.
- Savulich and colleagues (2021) examined sex differences over a variety of baseline characteristics, including cognitive outcomes.

Despite this order of events, the outcomes and analysis plan proposed here remain confirmatory in nature, because, in the one case that cognitive outcomes were examined, they were only examined at baseline.

Analyst-blind

To better ensure confirmatory conclusions, and to avoid unconscious bias towards adjustments that favor statistical significance, initial implementation of analyses was 'analyst-blind' (Dutilh, Sarafoglou, & Wagenmakers, 2019). In other words, analyses were first implemented on data where intervention status (whether a participant was in the CB-Abst or CB-Mon group) had been randomly shuffled. Data was shuffled by an individual outside the core ARCHES team. Only after data exclusion criteria had been determined, transformations of

predictor and outcome variables had been finalized, models had been properly specified, and any other unforeseen circumstances addressed, was the un-shuffled data provided to the analyst for the final analysis implementation.

Notable changes made during the analyst-blind period were:

- Removal of number of years spent using cannabis as a covariate. This measure was dropped due to its high correlation with a participant's overall age.
- The decision to only examine a linear trend for change over time. This decision was made due to the small number of time points (a quadratic time trend requires 3 parameters, but there were only 4 time points to estimate this trend).
- The addition of a secondary analysis of observed data only (imputed observations were excluded). This analysis was added to assess how robust our results were to our treatment of missing data.
- The decision to model bounded count data using binomial regression rather than linear regression. This decision was implemented to better deal with non-normal and non-continuous data.
- The decision to compute standard errors and confidence intervals using the robust estimates from the GEE models rather than a combination of GEE and a bootstrap approach. This decision was implemented for computational efficiency, as the bootstrap added little beyond the robustness already provided by the GEE approach.

One additional change was made after the initial unblinded analyses were run. We incorporated a new outcome, total correct for immediate recall from the verbal recognition memory task. While this outcome is not listed as part of the recommended CANTAB measures, it was important to include based on theoretical motivations, as its inclusion allowed us to assess the replicability of the results of Schuster and colleagues (2018).

Primary Outcomes

The 16 primary outcomes for the study were measures taken from 7 modules provided by the Cambridge Neuropsychological Test Automated Battery (CANTAB; Cambridge Cognition, 2019). The set of modules used for assessing each domain of interest (memory and executive functioning, respectively) were determined via CANTAB's recommendations listed on their website (see https://www.cambridgecognition.com/cantab/cognitive-tests/memory/ for modules to assess memory and https://www.cambridgecognition.com/cantab/cognitive-tests/memory/ for modules to assess executive functioning). Note subjects completed one additional module (the Motor Screening Task; MOT) at the start of each session, but since this module was intended primarily for calibration and training, we exclude it from analyses. We also note that while often researchers will collapse multiple outcomes into a smaller set of composite scores for interpretability and to avoid issues with multiple comparisons, it is difficult to do so with the CANTAB modules due a complex hierarchical correlation structure across tasks and domains of interest (e.g., Lenehan, Summers, Saunders, Summers, & Vickers, 2015). However, as an exploratory analysis we do include a replication of the method used for a pilot study of the current project (Schuster et al., 2018) in which a smaller set of CANTAB modules (which partially overlap with those assessed for the current study) was analyzed.

CANTAB tasks

The 3 modules for assessment of memory were…

1. The Paired Associates Learning (PAL) module.

    Description: Subjects were shown a set of boxes. The boxes were opened in random order to display a symbol hidden in the box. Symbols were then shown in the center of the screen. The goal of the task was to select the box that the onscreen symbol was originally hidden in. There were 20 trials in total, 4 trials each for conditions with 2, 4, 6, 8, and 12 boxes onscreen.

2. The Spatial Span (SSP) module.

    Description: Subjects were shown a set of white squares, each of which briefly change color in a set order. The goal of the task was then to select the boxes that changed color either (a) in the order that they changed color (the forward variant), or (b) in the reverse of the order that they changed color (the backward variant). Each variant consisted of 8 trials, starting with 2 boxes and ending at 9 boxes on screen.

3. The Verbal Recognition Memory (VRM) module.

    Description: The subject saw a list of 18 words, shown one at a time on the screen (the study phase). The goal of the task was then to (a) recall as many of the studied words as possible immediately after the study phase (the immediate free recall phase), (b) view a new sequence of 36 words and indicate which of these words were previously studied and which were novel (the immediate recognition phase), and (c) following a 20 minute delay, recall as many of the studied words as possible (the delayed free recall phase), and finally (c) view another sequence of 36 words and again indicate which of these words were previously studied and which were novel (the delayed recognition phase).

The 4 modules for assessment of executive functioning were…

1) The Multitasking Test (MTT) module.

    a) Subjects completed a series of trials in which they viewed an arrow that (a) could point to the left or right, and (b) could appear on the left or right-hand side of the screen. The goal of the task was to either (1) indicate the direction of the arrow irrespective of the side it appeared on, or (2) indicate the side the arrow appeared on, irrespective of the direction it was pointing. Trials could be (a) congruent (the arrow pointed in the same direction of the side it appeared in) or (b) incongruent (the arrow pointed in the opposite

direction it appeared in). Blocks of trials could be either (1) single-task (subjects only had to respond to either the direction or side for all trials) or (2) multi-task (trials could alternate between having to respond to the direction or side). Subjects completed 160 trials in total. Trials consisted of 80 congruent versus 80 incongruent, and 80 that were single-task versus 80 that were multi-task.

2) The One-Touch Stockings of Cambridge (OTS) module.
    a) Subjects saw a top and bottom set of three colored balls distributed across three different stacks. The goal of the task was to determine how to move the balls in the bottom step to match the top set. More specifically, subjects had to indicate the number of the moves it would take to move the balls in the bottom set to match the top set. Subjects completed 15 trials, with a set of 2 trials each for solutions that took 1, 2, or 3 moves, and a set of 3 trials each for solutions that took 4, 5, or 6 moves.
3) The Stop Signal Task (SST) module.
    a) Subjects saw an arrow pointing to the left or right. The goal of the task was to (a) indicate the direction of the arrow as quickly as possible, or (b) if a toned sounded following presentation of the arrow, avoid making a response. Subjects completed 424 trials, 318 in which they indicated the direction of the arrow (Go trials), and 106 in which the tone sounded and they were suppose to inhibit their response (Stop trials).
4) The Spatial Working Memory (SWM) module.
    a) Subjects saw a set of boxes. The goal of the task was to identify via a process of elimination the box containing a hidden token. Once a token once found, a new token was hidden in a different box. Tokens were never hidden in a box that had previously had a token. Subjects completed 30 trials total. There were conditions for 4, 6, 8, and 12 boxes onscreen, with the number of trials per condition equal to the number of boxes.

Selection of outcomes

The modules listed above produce a large number of outcomes (from 10 for the SSP to 58 for the OTS). However, CANTAB provides a set of recommended outcomes for researchers, so for simplicity and consistency with established guidelines, for each module we analyzed the subset of recommended outcomes per CANTAB's manual (Cambridge Cognition, 2021). We also included one measure, total correct for immediate recall from the VRM module, which was not part of the recommended outcomes but was important to include to assess replicability of a prior study (Schuster et al., 2018). Across the 7 modules, there were a total of 17 outcomes, listed below with a brief description and details on the units of measurement, range, and interpretation.

1. Outcomes to assess change in memory functioning:
    1.1. The total number of errors (adjusted for number of uncompleted problems) from the **PAL** module.
        - The number of times the subject chose the incorrect box for a stimulus on assessment problems plus an adjustment for the estimated number of errors they would have made on any problems/attempts/recalls they did not reach.
        - Count variable, ranging from 0 to 70, where higher scores indicate more errors and worse performance.
        - Data were modeled using a binomial distribution using the log-odds link function.
    1.2. The first attempt memory score from the **PAL** module.
        - The number of times a subject chose the correct box on their first attempt when recalling pattern locations – calculated across all trials.
        - Count variable, ranging from 0 to 20, where higher scores indicate better memory performance.
        - Data were modeled using a binomial distribution using the log-odds link function.

1.3. The forward span length from the **SSP** module.

- The longest length of sequences (spans) a subject successfully remembered in the original order.
- Count variable, ranging from 2 to 9, with higher scores indicating better memory performance.
- Data were shifted to range from 0 to 7 and were modeled using a binomial distribution using the log-odds link function.

1.4. The reverse span length from the **SSP** module.

- The longest length of sequences (spans) a subject successfully remembered in the reverse order.
- Count variable, ranging from 2 to 9, with higher scores indicating better memory performance.
- Data were shifted to range from 0 to 7 and were modeled using a binomial distribution using the log-odds link function.

1.5. The total correct – immediate recall measure from the **VRM** module.

- The total number of distinct words a subject correctly recalled from the studied list of 18 words during the immediate free recall phase.
- Count variable, ranging from 0 to 18, with higher scores indicating better immediate recall performance.
- Data were modeled using a binomial distribution using the log-odds link function.

1.6. The total correct – delayed recall measure from the **VRM** module.

- The total number of distinct words a subject correctly recalled from the studied list of 18 words during the delayed free recall phase.
- Count variable, ranging from 0 to 18, with higher scores indicating better immediate recall performance.

- Data were modeled using a binomial distribution using the log-odds link function.

1.7. The total correct in the immediate recognition condition from the **VRM** module.

- The total number of words a subject correctly identified as either previously studied or novel during the immediate recognition phase.
- Count variable, ranging from 0 to 36, with higher scores indicating better immediate recognition memory performance.
- Data were modeled using a binomial distribution using the log-odds link function.

1.8. The total correct in the delayed recognition condition from the **VRM** module.

- The total number of words a subject correctly identified as either previously studied or novel during the delayed recognition phase.
- Count variable, ranging from 0 to 36, with higher scores indicating better recognition memory performance.
- Data were modeled using a binomial distribution using the log-odds link function.

2. Outcomes to assess change in executive functioning:

2.1. The total incorrect from the **MTT** module.

- The total number of trials where a subject made an incorrect response irrespective of condition.
- Count variable, ranging from 0 to 160, with higher scores indicating more errors and worse performance.
- Data were modeled using a binomial distribution using the log-odds link function.

2.2. The median response latency from the **MTT** module.

- The median over the set of response times for all correct responses irrespective of condition.
- Response time variable in milliseconds (ms), ranging from 100 to 2000 ms, with higher values indicating slower performance across all trials.

- Data were modeled using a normal distribution (i.e., the standard linear model).

2.3. The median incongruency cost from the **MTT** module.

- The median of response times for incongruent trials minus the median of response times for congruent trials.
- Difference score in milliseconds (ms), ranging from -1900 to 1900 ms, with higher values indicating slower responses on incongruent trials, suggesting that it takes longer to process conflicting information.
- Data were modeled using a normal distribution (i.e., the standard linear model).

2.4. The median multitasking cost from the **MTT** module.

- The median of response times for the multi-task blocks (trials alternated between having to respond to the direction or position of the arrow) minus the median of response times for the single-task blocks (all trials consisted of either responding to the direction or position of the arrow).
- Difference score in milliseconds (ms), ranging from -1900 to 1900 ms, with higher values indicating slower responses on multi-task blocks, suggesting that it takes longer to process multiple sources of information.
- Data were modeled using a normal distribution (i.e., the standard linear model).

2.5. The number of problems correctly solved on the first choice from the **OTS** module.

- The total number of trials where the subject chose the correct answer on the first attempt.
- Count variable, ranging from 0 to 15, with higher scores indicating better performance.
- Data were modeled using a binomial distribution using the log-odds link function.

2.6. The median latency to first correct choice from the **OTS** module.

- The median of the response times for all trials in which the subject picked the correct choice on the first attempt.
- Response time variable in milliseconds (ms), a positive unbounded variable, with higher values indicating slower performance across correct trials.
- Data were modeled using a normal distribution (i.e., the standard linear model).

2.7. The stop signal reaction time from the Stop Signal Task (**SST**) module.

- The estimate of the duration at which a person can successfully inhibit a response 50% of the time, represents the time before which all actions become ballistic and a subject is unable to cancel a response selection.
- Response time variable in milliseconds (ms), ranging from 0 to 500 ms, with higher values indicating that it takes longer for a person to inhibit a response.
- Data were modeled using a normal distribution (i.e., the standard linear model).

2.8. The total number of between errors from the Spatial Working Memory (**SWM**) module.

- The number of times a subject incorrectly revisited a box that had contained a token in a previous trial across the 4, 6, and 8 box conditions.
- Count variable, ranging from 0 to 63, with higher scores indicating worse performance.
- Data were modeled using a binomial distribution using the log-odds link function.

2.9. The participant's strategy for the 6-8 box conditions from the **SWM** module.

- The number of times subjects began a new search from the same box they started on in previous trials. It is assumed that a person who begins from the same box each time is using a planned strategy to find the tokens.
- Count variable, ranging from 2 to 14, with higher scores indicating low strategy, meaning a subject began searches from many different boxes.

- Data were shifted to range from 0 to 12 and were modeled using a binomial distribution using the log-odds link function.

Statistical model

We analyzed all outcomes via regression models fit using generalized estimating equations (GEE; Liang & Zeger, 1986). Note the GEE approach provides robust estimates of coefficients even when distributional assumptions are violated and when heteroscedasticity is present. We assumed data were clustered over subjects, and that the observations for a subject per each study visit were uncorrelated (The GEE method is also robust to misspecification of the correlation structure for a subject's observations).

Continuous outcomes (e.g., latencies and reaction times) were fit using a standard linear regression model. Bounded count outcomes (e.g., number of errors or correct responses) were fit using binomial regression. Specifically, let $y_{(p,t)}$ refer to the outcome value for participant *p* at time *t*, where $y_{(p,t)} = \{0, 1, \ldots, H\}$ with *H* referring to the highest possible value. We assumed that $y_{(p,t)}$ was distributed according to a binomial process governed by a latent probability $\theta_{(p,t)}$:

$$y_{(p,t)} \sim \text{Binomial}\left(\theta_{(p,t)}\right).$$

The log-odds of this latent probability can then be decomposed into the weighted sum of the predictors of interest.

In contrast to linear regression, which includes a parameter for the residual standard deviation, binomial regression has no additional parameter to handle overdispersion. While failure to account for overdispersion can lead to underestimation of standard errors, as noted before the GEE approach is robust to misspecification of distributional assumptions and corrects for this underestimation, and as such we did not need to implement any additional steps to address overdispersion issues.

Key confirmatory effect

The key confirmatory effect of interest for each outcome will be a dummy-coded contrast between CB-Mon (the referent, coded as 0) and CB-Abst (coded as 1), testing whether a constant effect of abstinence exists, averaged over all time points.

Multiple comparisons

Each module/task from the CANTAB examines a separate cognitive process of interest. We therefore address multiple comparisons issues within each module. As the number of outcomes per module range from 1 to 4, we used a simple Bonferroni correction, dividing the desired error rate of 5% by the number of outcomes per module. Therefore, effects of interest were flagged as statistically significant for $\alpha = 0.05$ (for the SST module), $\alpha = 0.025$ (for the PAL, SSP, OTS, and SWM modules), $\alpha = 0.0167$ (for the VRM module), and $\alpha = 0.0125$ (for the MTT module).

Covariates

The covariates we included for the primary analysis were:

- A coefficient for a linear trend to capture change over time, coded as week 1 = -3, week 2 = -1, week 3 = 1, and week 4 = 3.
- A subject's baseline value for a given outcome, converted to a z-score.

In other words, we assume a conservative additive model, adjusting for baseline levels and with main effects for a) the impact of abstinence and b) change over time, but no treatment by time interaction.

Additionally, to assess the robustness of our conclusions we will incorporate, at a minimum, the following covariates as part of a secondary analysis:

- Age of participant at baseline (in years).
- Biological sex (male versus female).

Baseline scores from the Cannabis Use Disorder Identification Test - Revised (CUDIT-R; Adamson et al., 2010). The CUDIT-R is an 8-item scale with scores ranging from 0 to 32, where higher scores indicate more problematic cannabis use (a score of 8 or higher indicates hazardous use, while a score at or above 12 indicates a possible use disorder).

Missing data

Missing outcome data (e.g., subjects who did not complete all four visits or did not complete all 7 modules during a visit) will be handled using multiple imputation via chained equations (MICE; see for example Azur, Stuart, Frangakis, & Leaf, 2011). Missing outcome values will be imputed 48 times using linear regression and predictive mean matching. We will then apply the statistical model to each combination of observed and imputed data and will pool the resulting set of 48 model results using Rubin's rule. Subjects must have at least one post-baseline data point to be included in an analysis. Note because we are interested in the effect of abstinence, only data with biochemically-verified abstinence will be included – all data for the visit of and subsequent visits in which abstinence was no longer verified will be excluded for CB-Abst subjects. Covariates for imputing missing values will be:

- Non-missing outcome values at other time points;
- Outcome values at the baseline visit;
- The covariates listed above (age in years, biological sex at birth, baseline scores for CUDIT-R, and duration in years spent using cannabis).

Clinical Significance

In addition to determining whether there is a statistically significant different between groups, we assessed clinical significance by computing standardized effect sizes for all contrasts of interest per each outcome. Effect sizes were estimated by dividing the estimate (transformed when necessary to be in the original scale of the measure) by the standard

deviation of the outcome pooling over groups during the baseline visit. In other words, we normalized the contrast of interest by the baseline variation in the outcome.

Secondary analyses

We conducted the following secondary analyses:

1. We tested our assumption of an additive model by fitting a model that includes a group by time interaction (i.e., the product of the contrast between CB-Abst and CB-Mon and the linear time trend variable). We conducted an omnibus test comparing the simpler additive model to the more complex interaction model – if the associated Wald test is significant at $p < \alpha_M$ (i.e., the specified level of significance α for a given module *M*) this will indicate the presence of a group by time interaction. Post-hoc comparisons of the regression coefficients of the interaction model will also indicate, if abstinence was associated with improvements in cognitive performance, when such improvement occurred. This secondary analysis addresses the second research question of the study.
2. We examined how cognitive performance of the two groups of cannabis users (the contingency management and monitoring groups) also compares against the control sample of non-users. This secondary analysis addresses the third research question of the study. We therefore…
    a. Analyzed data from all three groups.
    b. Created two dummy-coded variables to represent the 3 levels (contingency management, monitoring, and non-users) for group status. We used the non-user group as the referent, and had separate variables for the CB-Abst and CB-Mon groups.
    c. Assessed (1) an additive model (main effects for group and time), and (2) an interaction model (also including at the group x time interaction in addition to the main effects).

d. Conducted an omnibus test comparing the simpler additive model to the more complex interaction model – if the associated Wald test is significant at $p < \alpha_M$ (i.e., the specified level of significance α for a given module *M*), this will indicate the presence of a group by time interaction.
e. If a group by time interaction or a main effect of group was found, we ran post-hoc comparisons to examine (a) if the CB-Mon group had worse performance compared to the non-user group, (b) if the CB-Abst group did not differ compared to the non-user group or (c) if performance for the CB-Abst group improved to levels matching the non-user group.

3. We checked whether our conclusions were robust to the inclusion of covariates, examining the inclusion of age, CUDIT-R scores, and biological sex, all assessed at baseline. All covariates were converted into z-scores.
4. We checked whether our conclusions for our primary analysis were robust to our treatment of missing data (i.e., imputation of missing values). We reran the primary analysis using only observed values (we excluded values that were imputed).

Secondary Outcomes

The 2 secondary outcomes for the study were measures taken from the module in the set included in the study that CANTAB designated for assessing the domain of attention (see https://www.cambridgecognition.com/cantab/cognitive-tests/attention). While the MOT module is included in this list, as noted before since it is intended primarily for calibration and training, we excluded it from analyses.

CANTAB tasks

The module for assessment of attention was…

1. The Rapid Visual Information Processing (RVP) module.

   Description: Subjects saw a sequence of digits appear one at a time in the center of the screen, at the rate of 100 digits per minute. The goal of the task was to indicate as quickly as possible when a target sequence of digits (e.g., 2-4-6) appeared onscreen. There were 54 3-digit target sequences, and 546 single-digit distractors.

Selection of outcomes

Again, for the RVP module we only analyzed the subset of recommended outcomes per CANTAB's manual (Cambridge Cognition, 2021). However, the CANTAB recommends both a measure of discriminability, A', using signal detection theory (SDT), along with the proportion of false alarms. Because A' is calculated using the proportion of hits and false alarms, we excluded the proportion of false alarms due to redundancy issues. Therefore, there were a total of 2 recommended outcomes, listed below with a brief description and details on the units of measurement, range, and interpretation.

1. Outcomes to assess change in attention functioning:

   1.1. A measure of discriminability, A' from the **RVP** module.

   - A transformation of the proportion of hits and false alarms, providing an estimate of a subject's ability to detect the target sequences of digits after controlling for

response bias (i.e., biases to respond or not respond irrespective of the stimuli on screen).

- Bounded continuous variable, ranging from 0 to 1, where higher scores indicate a greater ability to discriminate between targets and distractors.
- Data were modeled using a normal distribution (i.e., the standard linear model).

1.2. The median of the response times for hits from the **RVP** module.

- The median over the set of response times for all correct responses (identification of target 3-digit sequences).
- Response time variable in milliseconds (ms), ranging from 100 to 1900 ms, with higher values indicating slower performance in detecting targets.
- Data were modeled using a normal distribution (i.e., the standard linear model).

Statistical model

We will use the same statistical models and approach for primary and secondary analyses as described for the primary outcomes. We will deem effects significant for $\alpha = 0.025$.

Power

Power was determined via an R script conducting a power analysis for a linear model with treatment and time effects, including a treatment by time interaction, and assuming subject-varying intercepts and slopes for change over time (Schoenfeld, 2009). The script can be found at the following link:

- http://hedwig.mgh.harvard.edu/biostatistics/sites/default/files/public/powslopes.r

We assumed each group would have an initial set of 70 subjects. Though we had a retention rate of 93% in our pilot trial with 100% of retained participants with complete data, power calculations assumed a conservative 4% drop-out per week (i.e., just under 85% retention at Week 4). Power was computed using outcomes from a prior study (Schuster et al., 2018):

- The mean response latency (in ms) for correct responses from the Attention Switching Task (AST) module.
- The mean switching cost (in ms; equivalent to the multitasking cost in the MTT module) from the AST module.
- The mean response latency (in ms) to correct choice on a first attempt from the Delayed Match to Sample (DMS) module.

Though based on different modules from the CANTAB than those currently used, these outcomes were chosen because of (a) the availability of prior data, (b) compatibility with the assumptions of the power analysis software, and (c) because of their overlap with the domains of interest, as the AST (which is very similar to the MTT) examines executive functioning and the DMS examines memory.

Power analyses indicated we would have an 80% chance to detect a difference at 2-sided alpha of .05 if the true difference between CB-Abst and CB-Mon (assuming learning in both groups) over 4 weeks of abstinence was:

1. 11 milliseconds per study week (ms/wk) for the mean response latency from the AST, 52% of the 21 ms/wk effect observed in Schuster et al. (2018).
2. 12 ms/wk for the switching cost from the AST, 50% of the 25 ms/wk effect observed in Schuster et al. (2018).
3. 67 ms/wk for the mean response latency to correct from the DMS, 42% of the 160 ms/wk effect observed in Schuster et al. (2018).

Software

All analyses will be conducted using the statistical software R (version 4.1.1; R Core Team, 2021) and integrated development environment RStudio (version 2020.9.0.351; RStudio Team, 2021). Data will be prepared using the R packages ‘dplyr’ (version 1.0.7; Wickham, François, Henry, & Müller, 2021) and ‘tidyr’ (version 1.1.4; Wickham, 2021). Missing data will be imputed using the R package ‘mice’ (version 3.14.0; Van Buuren & Groothuis-Oudshoorn, 2011). Models will be fit using the R package ‘geepack’ (version 1.3.4; Hojsgaard, Halekoh, & Yan, 2006). Reproducible code and de-identified data will be organized using the R package ‘targets’ (version 0.8.1; Landau, 2021).

References

Adamson, S. J., Kay-Lambkin, F. J., Baker, A. L., Lewin, T. J., Thornton, L., Kelly, B. J., & Sellman, J. D. (2010). An improved brief measure of cannabis misuse: The Cannabis Use Disorders Identification Test - Revised (CUDIT-R). *Drug and Alcohol Dependence*, *110*, 137-143. https://doi.org/10.1016/j.drugalcdep.2010.02.017

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20* (1), 40 - 49. https://doi.org/10.1002/mpr.329.

Cambridge Cognition (2019). CANTAB (version V1.9) [Computer software]. www.cantab.com

Cambridge Cognition (2021). *CANTAB Connect research: Admin application user guide*. Cambridge Cognition.

Dutilh, G., Sarafoglou, A., & Wagenmakers, E.-J. (2021). Flexible yet fair: Blinding analyses in experimental psychology. *Synthese*, *198*, 5745 – 5772. https://doi.org/10.1007/s11229-019-02456-7

Hojsgaard, S., Halekoh, U., & Yan, J. (2006)/ The R package geepack for generalized estimating equations. *Journal of Statistical Software*, 15 (2), 1 – 11. https://doi.org/10.18637/jss.v015.i02.

Landau, W. M., (2021). The targets R package: A dynamic Make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, *6* (57), 2959. https://doi.org/10.21105/joss.02959.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73* (1), 13-22. https://doi.org/10.1093/biomet/73.1.13.

Lenehan, M. E., Summers, M. J., Saunders, N. L., Summers, J. J., Vickers, J. C. (2015). Does the Cambridge Automated Neuropsychological Test Battery (CANTAB) distinguish

between cognitive domains in healthy older adults? *Assessment*, 23 (2), 163 – 172. https://doi.org/10.1177/1073191115581474.

Cooke, M. E., Gilman, J. M., Lamberth, E., Rychik, N., Tervo-Clemmens, B., Evins, A. E., & Schuster, R. M. (2021). Assessing changes in symptoms of depression and anxiety during four weeks of cannabis abstinence among adolescents. *Frontiers in psychiatry*, *12*, 1 - 9. https://doi.org/10.3389/fpsyt.2021.689957

R Core Team. (2021). R: A language and environment for statistical computing (version 4.1.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org

RStudio Team. (2021). RStudio: Integrated development environment for R (version 2020.9.0.351) [Computer software]. RStudio, PBC. http://www.rstudio.com

Savulich, G., Rychik, N., Lamberth, E., Hareli, M., Evins, A. E., Sahakian, B. J., & Schuster, R. M. (2021). Sex differences in neuropsychological functioning are domain-specific in adolescent and young adult regular cannabis users. *Journal of the International Neuropsychological Society*, *27* (6), 592-606. https://doi.org/10.1017/S1355617720001435

Schoenfeld, D. (2009). *A program to calculate the power or sample size for a random slopes model*. MGH Biostatistics Center. Retrieved from http://hedwig.mgh.harvard.edu/biostatistics/node/42

Schuster, R. M., Gilman, J., Schoenfeld, D., Evenden, J., Hareli, M., Ulysse, C., Nip, E., Hanly, A., Zhang, H., & Evins, A. E. (2018). One month of cannabis abstinence in adolescents and young adults is associated with improved memory. *The Journal of Clinical Psychiatry*, *79* (6), 1 - 18. https://doi.org/10.4088/JCP.17m11977

Schuster, R. M., Potter, K., Lamberth, E., Rychik, N., Hareli, M., Allen, S., Broos, H. C., Mustoe, A., Gilman, J. M., Pachas, G., & Evins, A. E. (2021). Alcohol substitution during one month of cannabis abstinence among non-treatment seeking youth. *Progress in Neuro-*

Psychopharmacology and Biological Psychiatry, *107*, 110205. https://doi.org/10.1016/j.pnpbp.2020.110205

Schuster, R. M., Potter, K., Vandrey, R. Hareli, M., Gilman, J., Schoenfeld, D. & Evins, A. E. (2020). Urinary 11-nor-9-carboxy-tetrahydrocannabinol elimination in adolescent and young adult cannabis users during one month of sustained and biochemically-verified abstinence. *Journal of Psychopharmacology*, *34* (2). 197 – 210. https://doi.org/10.1177/0269881119872206

Schwilke, E. W., Gullberg, R. G., Darwin, W. D., Chiang, C. N., Cadet, J. L., Gorelick, D. A., Pope, H. G., & Huestis, M. A. (2011). Differentiating new cannabis use from residual urinary cannabinoid excretion in chronic, daily cannabis users. *Addiction*, *106* (3), 499 - 506. https://doi.org/10.1111/j.1360-0443.2010.03228.x

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1-67. https://doi.org/10.18637/jss.v045.i03.

Wickham, H., François, R., Henry, L. & Müller, K. (2021). dplyr: A grammar of data manipulation (version 1.0.7). [Computer software] Retrieved from https://cran.r-project.org/web/packages/dplyr/index.html

Wickham, H. (2021). tidyr: Tidy messy data (version 1.1.4). [Computer software] Retrieved from https://CRAN.R-project.org/package=tidyr