**Kirsten Follow up options:   
  
Key (in no particular order):  
1. T tests between overall low and high drug users**

**2. moderate**Option 1: T test to measure differences in cognitive scores between low and high usersStep 1:  **– you would run a “Latent Profile Analysis (LPA)** (model-based clustering)”  
  
you would feed it which ever “substance abuse” variables you want to be clustered into low and high, and it would automatically assign each row either a low use or high use score  
  
(here I fed it nicotince, marijuana, and alcohol – but up to you)

**-------- Sample Code ------------**

**library(mclust)**

**# Select and scale your variables**

**lpa\_vars <- df %>%**

**select(marijuana\_pca\_score, alcohol\_pca\_score, nicotine\_pca\_score) %>%**

**scale()**

**# Run 2-class latent profile analysis**

**lpa\_model <- Mclust(lpa\_vars, G = 2)**

**# Attach classification to your df**

**df$substance\_use\_class <- factor(lpa\_model$classification)**

**# Optional: make it "Low Use" vs. "High Use"**

**# Check means per class first:**

**aggregate(lpa\_vars, by = list(class = lpa\_model$classification), mean)**

**# Assume class 1 = lower means → label accordingly**

**df$substance\_use\_class <- factor(**

**lpa\_model$classification,**

**labels = c("Low Use", "High Use")**

**)**

**------------------------------------**

Step 2: *run t tests to measure differences in cognition across the 2 groups (you don’t only have to use the measures that were previously related to* *substances in the shapley/correlations – this is a new test! Do whatever!* ***That said – because this technically would be a bunch of individual tests on top of the ones we already did – you may want to only choose like 3 things (maybe just like texi, Zuckerman, RPI)***because you run\*\*\* multiple t-tests you would want to use **significance correcting for multiple comparisons (called** False Discovery Rate (FDR / Benjamini-Hochberg — preferred)))

**-------- Sample Code ------------**

**# Run t-tests across cognitive variables**

**cog\_vars <- c("texi\_total.z", "Flanker\_Z", "delay\_discounting\_indiff\_z", "rpi.z")**

**p\_vals <- sapply(cog\_vars, function(var) {**

**t.test(df[[var]] ~ df$substance\_use\_class)$p.value**

**})**

**# Apply FDR correction**

**p\_adj <- p.adjust(p\_vals, method = "fdr")**

**# View results**

**data.frame(**

**Cognitive\_Measure = cog\_vars,**

**Raw\_p = round(p\_vals, 4),**

**FDR\_Adjusted\_p = round(p\_adj, 4)**

**))**

**------------------------------------**

Option 2: Moderation (interaction) to test whether RPI *may* relate to substance abuse – but may be moderated by impulsivity *(such that if you mathematically were to split impulsivity into 2 levels, might the people at one level show a significance with substance abuse while the people at another level don’t?*  
  
*(we would need to substantiate that we are NOT just searching*  for an rpi significance – p hacking - but that this is an actual *a priori* interesting question **(in fact if we didn’t find a significant moderation….** we would be showing that even in the HIGH IMPULSIVITY group, the RPI *still* didn’t predict substance , it would be just as cool as if it was significant.

**-------- Sample Code ------------**

**lm(substance\_use ~ RPI \* Impulsivity + controls, data = df)**

**------------------------------------**

Option 3: For each cohort, take the most impactful shapley cognition variables, and run them in a model with (PNRT and Wellbeing), to see how cognitive predictors compare with affective behavioral predictors.   
  
(this is just a general linear model – we wouldn’t want to use shapely)

**-------- Sample Code ------------**

**# For one cohort (e.g., cohort\_df)**

**model <- lm(substance\_use ~ top\_cog + PNRT + wellbeing, data = cohort\_df)**

**summary(model)**

**------------------------------------**