**Question 1**

Briefly explain the results of each different logit regression with the 2 variables you selected (which converge). Highlight any significant differences between the models and if there is any evidence of imbalance. Please be concise.

In each of the runs A1, A3, B1, B3, using questions 200, 5414, and 405, we did not see any statistically significant P values (with the exception of “kwh\_2009\_07“ in model run A1, however the coefficient of the variable is so small that is likely has not impact of the probability of being selected for the treatment group). This stands to reason because adding more, independent variables enhances that balance of the data. In a runs of A1, A3, B1, B3, using questions 5414, and 405, but not 200, we observed several more statistically significant P values.

**Question 2**

Briefly explain the benefit(s) and potential problem(s) of using ALL the available survey variables in a logit regression. Please be concise.

The biggest potential problem is that, by including all of the variables, you run the risk of over-categorizing the results such that some cells might end up having too small a number or zero observations (for example, there may not be any 1) females 2) who do not have internet and 3) indicated that they did not expect to save any money from the trial,) and the model would therefore not converge. A secondary risk is that if you try to use your model to predict the behavior of individuals or groups that don't precisely match your original group, your predictions will not be accurate.

Further, using all available survey variables risks overfitting the boundary between the two classes. The benefit of this is that your model will explain more of the variance in outcomes. This could be confirmed through cross-validation -- only using half of our data to create the model, and seeing how well our model predicts the outcome variable for the data set that wasn't used.

**Question 3**

Briefly explain when and why (if at all) it would be sufficient to use only a subset of the available survey questions. Please be concise.

Including a large subsets (or all) of the variables increases the risk of having two or more variables that are covariants—two variables that essentially describe the same underlying condition. Adding these variables to to model won’t necessarily improve accuracy of the model but will increase the complexity of the model.

On the other hand, using a subset of survey questions is essentially identical to having never asked the missing questions in the first place (assuming questions don't interact with each other), and so doesn't cause any real problem with the model but will still slightly reduce the model accuracy.