Section 1

1. Yes, there is evidence of imbalance. In the logit output, there are several variables with p-values below 0.05, including kwh\_2009\_09, question 420, and question 4701.2\_1. Looking at the quick means test output, it appears that there are differences in the means between the treatment and control for some variables (kwh\_2009\_12 and D\_410\_3 and D\_405\_2) that could be statistically significant. To determine whether this is the case, we did a “by-hand” t-test. Our suspicions were confirmed: the variables that we had identified had high t-values in the t-test output. This indicates that the treatment group was not as good as randomly assigned, since certain pre-trial consumption values and survey responses are predictors of assignment to treatment.
2. Some of the variables that were significant in the logit model were also significant in the quick means comparison (kwh\_2009\_12), but others were only significant in the quick means test (kwh\_2009\_10, D\_410\_3). We would expect there to be more significant variables in the quick means test because of multicollinearity. If there is high or imperfect multicollinearity, it does not violate the OLS assumptions but it inflates the standard error, which can cause a variable to appear insignificant when it actually is significant. In other words, the estimates could be precise but not unbiased.
3. The logit check is good because it gives you standard errors and p-values, but you may not trust the estimates in situations with large data sets whose overlapping variables run a higher risk of collinearity. In general, we expect standard errors to decrease with the number of observations. However, we are concerned that some of the survey questions may be similar to one other, which would lead to higher multicollinearity and blow up the standard errors. The quick means test is easy to perform and it is not misled by significance, but it is not as rigorous as the two-sample t-test.
4. Questions 410, 420, 43111 are redundant - they all determine the number of people in the household that are over/under 15. Questions 43521 and 5414 are also somewhat redundant - they both determine expected reduction in energy usage. Variables that should have been included are income level, type of home, type of appliances, and attitude toward conservation.

Section 3

1. The coefficient estimate for the treatment-trial interaction variable was not significant without the weights but became significant with inclusion of the weights.
2. Without weights, logged energy consumption for people in the treatment group and in the trial period (i.e. those who actually received treatment) decreased by 0.008 kwh relative to the control group. Since the results are not statistically significant, however, we cannot conclude that the C4 treatment was effective.
3. With weights, logged energy consumption for people in the treatment group and in the trial period decreased by 0.025 kwh relative to the control group. Assuming that the weights are correct, this means that the regression results were biased downwards in the initial results (without weights); when the weights were applied, the treatment effect increased four-fold. The coefficient estimate for the treatment effect is statistically significant in this case; however, even after the weighting, the coefficient estimates for several of survey response dummy variables were as large or larger than the treatment effect. Therefore, we would conclude that the C4 treatment was not very effective. The results may be statistically significant, but they are not practically significant since the treatment effect appears rather weak.
4. The weights we created theoretically rebalanced the dataset based on observable characteristics, but this dataset of observed characteristics also omitted many variables that likely explain energy consumption. In addition, some of the variables that it did include were collinear. Furthermore, we did not attempt to restore balance based on any unobservables that may be confounding our estimates. In future models, we could use an instrumental variable approach to attempt to limit this potential endogeneity. This suggests that the weights derived from propensity scores may not be entirely reliable, and that our coefficient estimates should be interpreted cautiously.