**Data Exploration and Analysis for the**

**Hemingway Measure of Adult Connectedness**

Gildardo Bautista-Maya, Dr. Ping Ye, Dr. Diane Cook

Mathematics Department

University of North Georgia Gainesville

Abstract:

We consider the dataset collected from students participating in the Boy With A Ball (BWAB) program, a faith-based community outreach group, through the Hemingway *Measure of Adult Connectedness*©, a questionnaire measuring the social connectedness of adolescents. First, we approach the data in the conventional method provided by the Hemingway website. We then identify which questions are strong determiners in deciding whether a student has completed the BWAB program or not. With the goal of utilizing the logistic regression, we reduce the set of questions to those only identified as significant in other methods. These methods include linear regression, decision tree, and random forest. The results are explained from a psychological perspective of social adolescent development.

Keywords: connectedness, adolescent, development, Hemingway, linear, logistic, psychological

Introduction:

Boy With A Ball (BWAB) is a faith-based community building focused on relationship building and improving the situation of rising adults. These solutions include mentoring, faith-based camps, scholarships and community development. Utilizing data gathered through the Hemingway *Measure of Adult Connectedness*© questionnaire, administered by BWAB in 2013, we analyze question importance through linear regression, decision tree, random forest, and logistic regression. Furthermore, we use the Hemingway scoring method to compare participants who have completed the program to those who have not to see which aspects of social connectedness separate the two.

The dataset contained the question answers for 220 9th grade students, now called participants, who were referred to BWAB through their school to participate in the BWAB program. 38 of the 220 have participated in BWAB in the past and were participating again, assigned a ‘Program’ value of 1. 182 of the 220 have not completed the BWAB program, assigned a ‘Program’ value of 0. For future reference, the implication of which group a participant is in is relevant in the sense that someone who has not completed the program still needs it, while those who have do not. Furthermore, the group assigned with a value of 0 included those who were new to BWAB as well as those who did not complete the program in its entirety. Those who have been assigned a ‘Program’ value of 1 completed the questionnaire as a post-test, while those assigned a value of 0 completed the questionnaire completed it as either a pre-test.

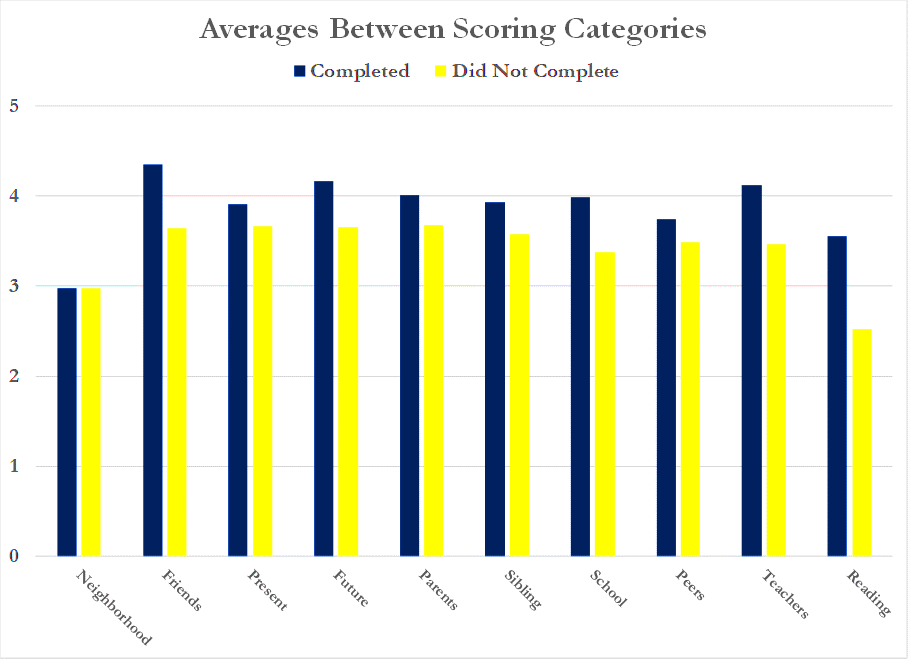
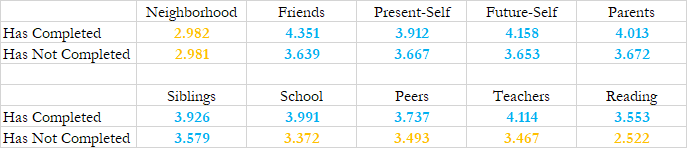
Each question was answered as ‘Not at all true’, ‘Not really true’, ‘Sort of true’, ‘True’, ‘Very true’, ‘Unclear’. Generally, ‘Not at all true’ was assigned a score of 1, ‘Not really true’ was assigned a score of 2, ‘True’ was assigned a score of 4, while ‘Very true’ was assigned a score of 5; however, if the question is worded in such a way as to be reverse scored, ‘Not at all true’ was assigned a score of 5, ‘Not really true’ was assigned a score of 4, ‘True’ was assigned a score of 2, while ‘Very true’ was assigned a score of 1. In both cases, ‘Sort of true’ and ‘Unclear’ were assigned a value of 3, whether it was graded reversed or not.

Analysis of Category Means:

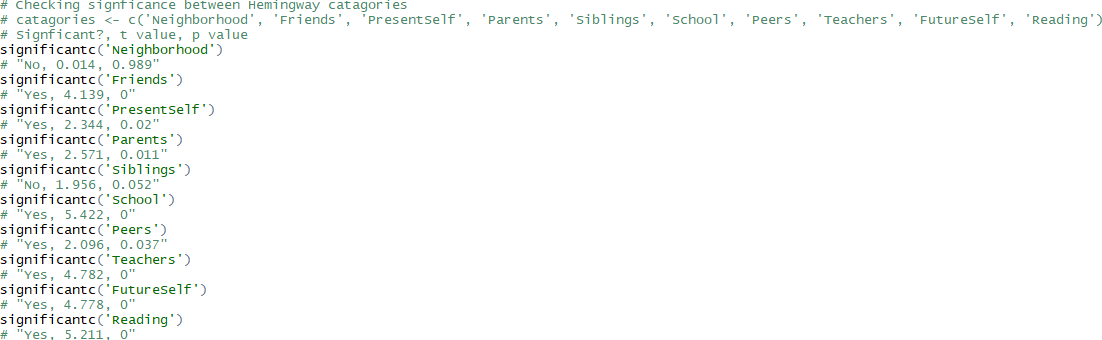
By comparing the resulting means, we can visualize the categories in which the groups coincide or differ. Furthermore, we can observe the categories in which each group scores low or high on connectedness. As far as the Hemingway *Measure of Adult Connectedness*© questionnaire, the scope of the study will supersede past its intended purpose.

We begin by using the Hemingway scoring method, which consists of scoring the participants based on question categories. The 57 questions are broken up into 10 different aspects of social connectedness; Neighborhood, Friends, Present-Self, Parents, Siblings, School, Peers, Teachers, Future-Self, and Reading. Upon finding the average score for each category, we determine whether the difference between those who have completed the program (1) to those who have not (0), and whether the measured difference is significant.

Question values were set to the Hemingway standard from 1 to 5 and the values of reverse-graded questions inverted. Thus, the mean score for each question was taken from each group. Questions were then grouped according to their categories and given a cumulative mean. Without loss of generality, question numbers ending in 1 were from into the category of ‘Neighborhood’, 2 from ‘Friends’, 3 from ‘Present-Self’, 4 from ‘Parents’, 5 from ‘Siblings’, 6 from ‘School’, 7 from ‘Peers’, 8 from ‘Teachers’, 9 from ‘Future-Self’, and 0 from ‘Reading’.



Above, are the measured means between those who have completed the program to those who have not. As stated by the Hemingway manual, scores measuring below a score of 3.5 denote ‘low connectedness’, while scores at or above 3.5 denote ‘high connectedness’. For the group that has completed the program, only one category was marked as denoting low connectedness which is ‘Neighborhood’. Meanwhile, the group that has not completed the program has four categories denoting low connectedness which are ‘Neighborhood’, ‘School’, ‘Peers’, and ‘Teachers’.

After computing the mean of each category, the significance of each was determined. To the left are answers to whether the difference in a category was significant or not, the t-value, and the p-value corresponding to that t-value. Although not every individual question is significant to the 5% confidence interval, we notice that the combination of multiple questions results in most differences in the categories being significant.

Even if a category has a noticeable difference, it does not necessarily imply that the value of the questions was an important in determining which group a participant was in. Although the ‘Reading’ category denotes a large and significant difference in between those who have completed the program and those who have not, no question from that category was found to be ultimately relevant in predicting whether a participant was in the program or not.

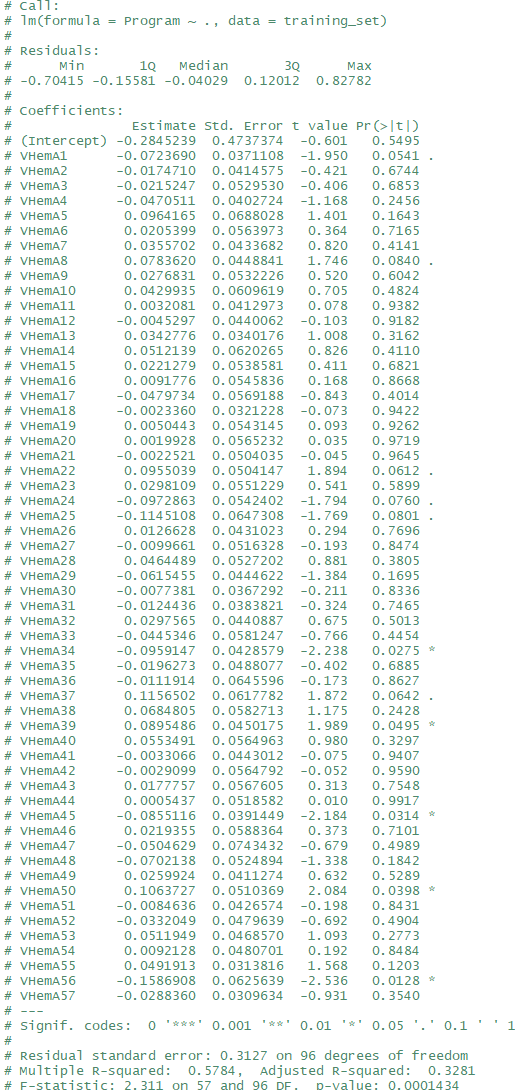
The results are particularly useful in terms of the observing the dataset from a psychological aspect. We observe that there is a clear difference between the social connectedness of those who have completed the program and those who have not. These observations will be interpreted in the later part of this report.

Identifying Significant Predictors:

By developing models to select which questions are the best predictors of which group a participant is in, we can narrow the wide range of data to a few influential questions. Furthermore, we can observe the category in which these questions originate from, thus finding an aspect of social connectedness in which the program can focus its efforts to.

The ideal model when predicting a categorical variable such as trying to predict whether the participant was in the program before or not would be the logistic regression; however, logistic regression does not make sense as the number of initial questions and participants means that any difference could be due to variance. Thus, if we want to use logistic regression, we must reduce the number of independent variables by removing questions which are not impactful in deciding which participant is in which grouping.

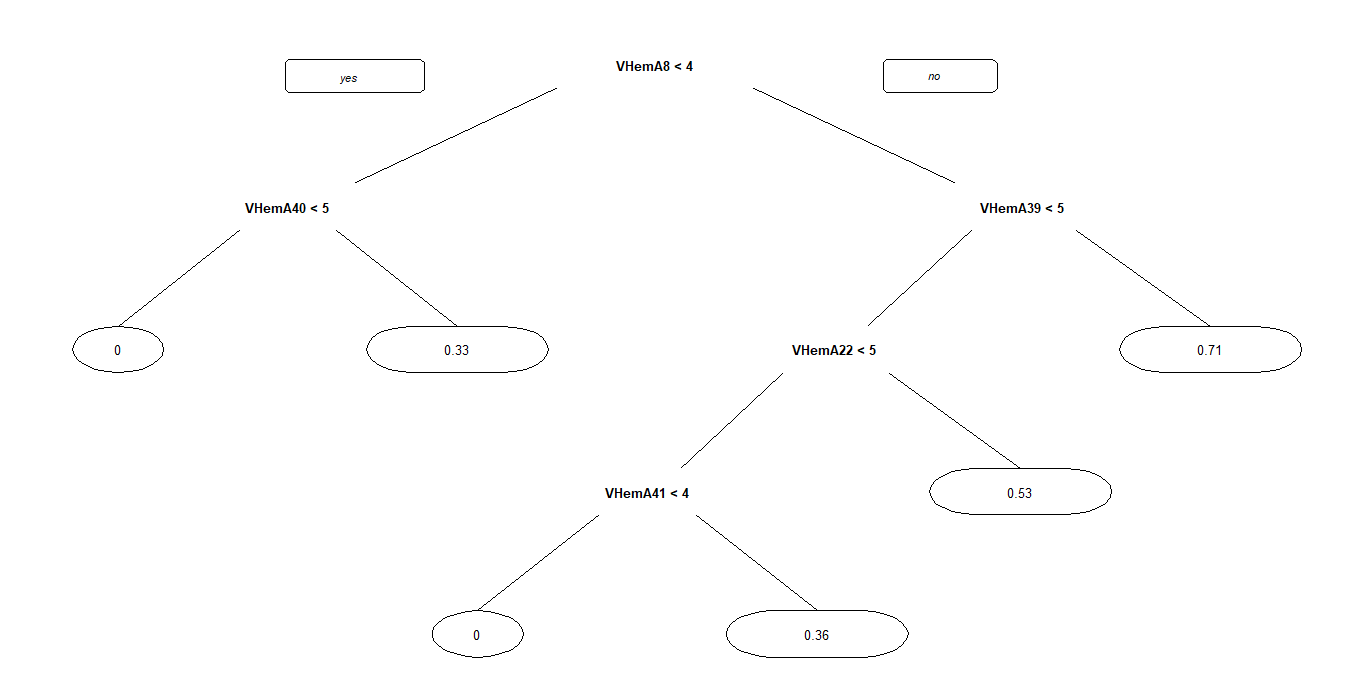
For all models, all used the same training and testing set. The training and testing set was split such that the training set contained 70% of the dataset and had a size of 154 while the testing set contained the remaining 30% and had a size of 66.

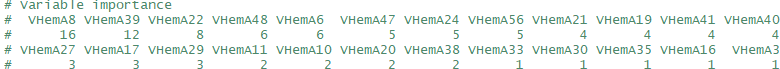
The first model used is the linear regression, which is not as optimal as logistic regression. By estimating the categorical variable, ‘Program’, as a quantitative variable. Thus, instead of either 0 or 1, we attempt to predict ‘Program’ as a value between 0 and 1.

After computing a model on the training set, it was used to predict the value of ‘Program’ on the testing set. The linear regression model predicted with an accuracy of 0.86364 or 86.364%.

As part of the process to reducing the number of independent variables, we will only use variables with a p-value less than 0.05 or 5%. Thus, questions 34, 39, 45, 50, and 56 will be taken for the logistic regression using the linear model.

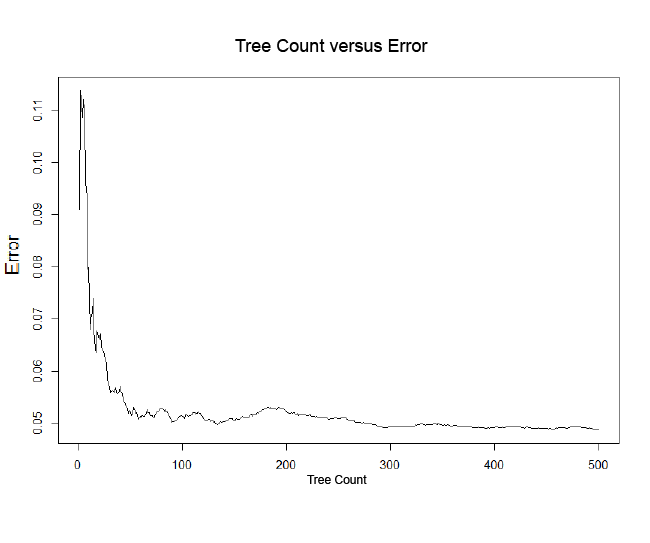
One of the issues of using linear regression is that the model may overshoot and predict values that are above the maximum value or below the minimum value. As seen in our model, we see that over half the entries were assigned a value below 0, even though as a categorical variable it would never be anything less than 0.

The second model used is the decision tree, which operates under the linear model. In a decision tree, the end goal is to assign the entry a probability of whether a participant was in the program previously or not. To do this, the decision tree has branches, the end of each assigns a prediction based on their answer to a few questions. Below is the resulting decision tree.

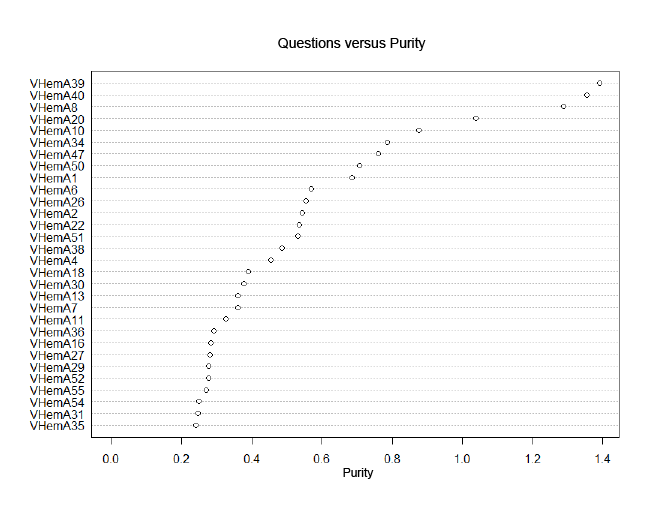
After computing a model on the training set, it was used to predict the value of ‘Program’ on the testing set. The decision tree predicted with an accuracy of 0.81818 or 81.818%.

Through R, we can see which questions affect the model the most, which is shown above. While there is no clear cutoff value, we will only use variables with a variable importance of 5 or larger. Thus, questions 6, 8, 22, 24, 39, 47, 48, and 56 will be taken for the logistic regression using this decision tree.

It is worth noting that while a question may not be present on the decision tree visual, it does not imply that the question was an irrelevant predictor. Furthermore, just because a question is present on the decision tree does not imply that the question was a good enough predictor. For example, take the branches that split for a participant’s response to question 41. It did not matter which response they may or may not have marked, if the participant reached that part of the decision tree, they would have been assigned a value of 0 through utilizing our testing set.

The third model used is the random forest, which is essentially creating multiple decision trees where we will use the aggregate to find the best predictors. For this, we created a random forest of 500 trees.

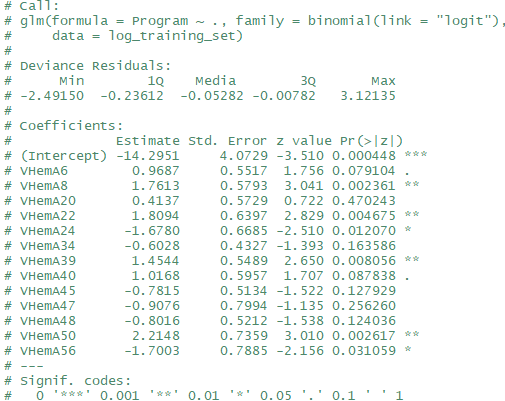
As the number of trees increase, the error shrinks considerably; however, one of the issues with a large tree count is the possibility of over-fitting the data to the training set. By using the random forest to predict the values on the testing set, we find that the model predicted results with an accuracy of 0. 90909 or 90.909%.



Here we see the ‘purity’ of each question, a measure of how influential a variable was to the model. While there is no clear cutoff to say which questions are more telling, we will only use variables whose purity is greater than 1. Thus, questions 8, 20, 39, and 40 will be taken for the logistic regression from the random forest model.

By using the previous methods, we take the best predictors from each model to use for the logistic regression. By doing so, we have reduced the number of independent variables from 57 to just 13. Below is a table with contains the questions marked significant from each model, thus showing that some questions were marked significant in more than one or in all methods.

|  |  |  |  |
| --- | --- | --- | --- |
| Question | Linear Regression | Decision Tree | Random Forest |
| 6 |  | • |  |
| 8 |  | • | • |
| 20 |  |  | • |
| 22 |  | • |  |
| 24 |  | • |  |
| 34 | • |  |  |
| 39 | • | • | • |
| 40 |  |  | • |
| 45 | • |  |  |
| 47 |  | • |  |
| 48 |  | • |  |
| 50 | • |  |  |
| 56 | • | • |  |



The result of the reduction in independent variables is shown. On the left, we have discovered 6 questions out of the 13 candidates which have shown to be useful in determining whether a participant has been in BWAB previously or whether they are new or have left. Furthermore, we see whether the correlation between the question and a participant’s ‘Program’ value.

With the logistic modelling complete, we can see which questions the most influential predictors. Significant questions in order of lowest to highest p-value: Question 8 (0.236%), Question 50 (0.262%), Question 22 (0.468%), Question 39 (0.806%), Question 24 (1.207%), and Question 56 (3.106%). The categories in which these questions came out of are ‘Teachers’, ‘Friends’, ‘Parents’, ‘Future-Self’, ‘School’. Two of those, ‘School’ and ‘Teachers’, were categories in which those who have not finished the program received a mean score considered ‘low connectedness’, which those who have finished the program did not.

Conclusion:

With the results of our logistic regression, we can reduce the size of the questionnaire from 57 questions to a compact 6. By observing the questions that separate the two groups, we can assess the overall impact of the program on student’s social connectedness. Furthermore, we can identify students who may benefit from participating in the BWAB program.

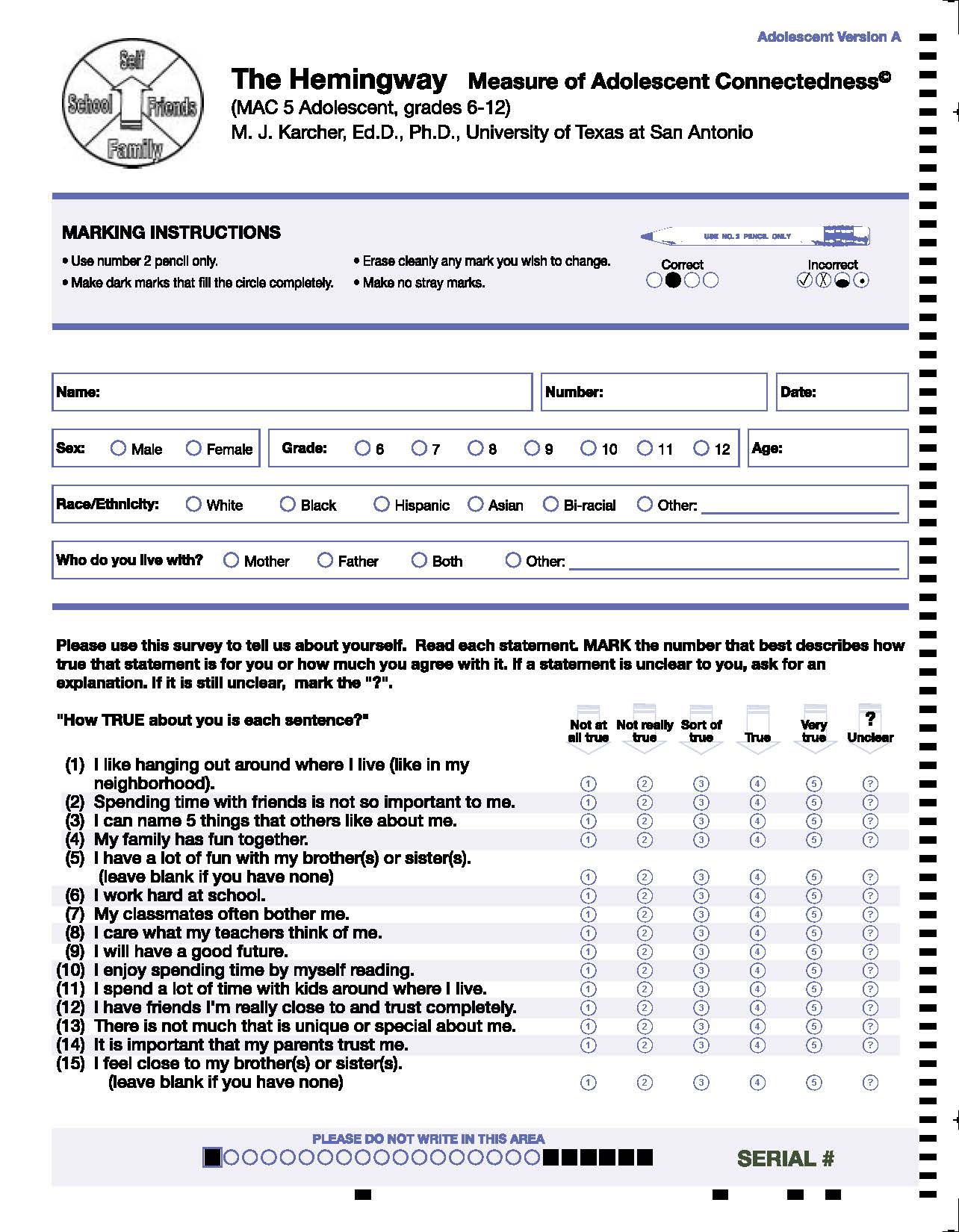
On the other hand, the results from determining the significance of mean differences demonstrates that while the difference in a category may be significant, this does not imply that there exists a question that is a strong determiner within that category. Take the category ‘Reading’, which denoted the largest difference in any scoring category. Despite this, neither of the 6 questions marked as ultimately significant by the logistic regression originated from ‘Reading’.

A notable observation is the accuracy of our decision tree and random forest. From the accuracy of obtained from performing the modelling on the testing set denotes that this accuracy is not a sign of overfitting, but rather, that there exist questions whose responses are strong determiners of which group the participant is in, thus reinforcing our motive of isolating these questions.

From our mean comparison, the group of participants who have not completed the program obtained lower measure of connectedness in all categories when compared to those who have completed the program. Notably, the scores received in the categories ‘School’, ‘Peers’, ‘Teachers’ were marked as ‘low connectedness’ in the group of those who have not completed the program, while those who have completed the program were marked as ‘high connectedness’. These categories clearly have a real-world connection in the sense that these categories are shaped by what the participant experiences at school. The significant questions 8 (“I care what my teachers think of me”) and 50 (“I usually like my teachers”) stress the importance of teacher interactions in a student’s life, with 8 questioning how much the participant values their teacher’s opinion and 50 asking how the participant feels about their teachers. Additionally, the highest scoring category for any mean was the score for ‘Friends’ from the group that finished the program, who had a mean score of 4.351. A friend can be someone who surrounds the participant at school, in the program, in their neighborhood, or all three. Significant question 22 (“Spending time with my friends is a big part of my life”) highlights the importance of spending time with friends and as Dr. Cook states, “Friends may be more important than siblings in the adolescence period, as the focus is on expanding beyond the family”.

While it is easy to focus on what is significant, let us consider a category that was not significant, ‘Neighborhood’. Both groups scored almost identically in the category of ‘Neighborhood’ with those who have completed the program scored a 2.982; Meanwhile those who have not, scored a 2.981. Furthermore, these scores are considered a sign of ‘low connectedness’ in the Hemingway. While the striking similarity between the groups is puzzling, it is worth considering that the data was collected in 2013, a time whether technology does not require participants to be physically associated in terms of the questions asked in the questionnaire.

Without doubt, there is more than meets the eye in any data analysis. Though the group consisting of those who completed the program scored higher than those who have not completed the program, we cannot establish a that participation in the BWAB program definitively caused this change; however, what we can say is that those students with low scores in certain categories could benefit from the program.

Appendix: Hemingway *Measure of Adult Connectedness*© Survey

