The hidden message in letters of recommendation: A natural language processing analysis of letters of recommendation at the United States Military Academy

Benjamin Siegel, Daniel Baller, Joseph Lindquist, James Pennebaker, Diana M. Thomas

Applications for college admission typically require 2-3 letters of recommendation. Colleges consider the letters of recommendation useful for guiding decisions on admission right behind high school grades, curriculum quality and standardized test scores (1). However, little is known on how colleges review letters of recommendation and how the letters of recommendation may add value to existing admissions algorithms (2, 3).

Existing studies on letters of recommendation have relied on human raters to manually assess letters for content (2-4). Letters of recommendation can be challenging to evaluate with this approach. For example, since applicants select their own letter writers, letters are usually positive (2) which makes it difficult to discriminate between letters and correlate letter content to academic outcomes. Moreover, collapsing information across an individual candidate is non-trivial. Two letter writers also may not agree upon qualities for an individual candidate. Finally, manually aggregating information from all submitted letters of recommendation in a form that allows for comparison between candidates is unfeasible. In 2020, Harvard College received 40,248 applications (5). Each applicant is required to submit 3 letters of recommendation (2 from teachers and 1 from a guidance counselor) totaling to 120,744 letters. However, with computational text analysis, we can isolate and differentiate language used in the letters of recommendation and relate these differences to academic performance (6).

Candidates for admission to the United States Military Academy (USMA) are drawn from every state in the nation and application packets require letters of recommendation from a mathematics, English, science, and physical education teacher. The recommendation letters consist of 12 questions with a standard Likert response (strongly agree, agree, neutral, disagree, strongly disagree) and then a free text response to the prompt. Here, we used computational text

analysis from letters of recommendation submitted as part of application packets to the United States Military Academy between the years 2012 to 2018 to address three questions:

- How well can we predict college academic performance from letters of recommendation submitted for admission?
- What type of language within those letters are associated with higher academic achievement?
- What is the added accuracy for predicting college GPA by including the free-form text analysis of candidate letters of recommendation to standard variables already used for making admissions decisions, like standardized test results and high school performance?

RESULTS

In order to pair findings from letters of recommendation to college performance, admissions information and detailed performance metrics recorded after the cadets arrived at USMA was extracted for 8,602 cadets admitted during 2012-2022. The performance variables consist of information such as freshman year GPA, graduating class rank, and USMA-specific outcome variables such as character evaluations, honor violations, and whether the cadet was separated from the Academy. Two important performance variables to note are academic GPA, which is derived solely from a cadet's performance in the classroom, and CQPA, which is a combination of a cadet's academic, military, and physical fitness performances.

Data Preparation

After removing names for privacy reasons, we received three separate data sets from USMA. The first contains information on pre-admission variables, the second contains information on post-admission variables, and the third contains all the recommendation letters. Each data point in the three sets is linked to a unique cadet identification number, so we merged

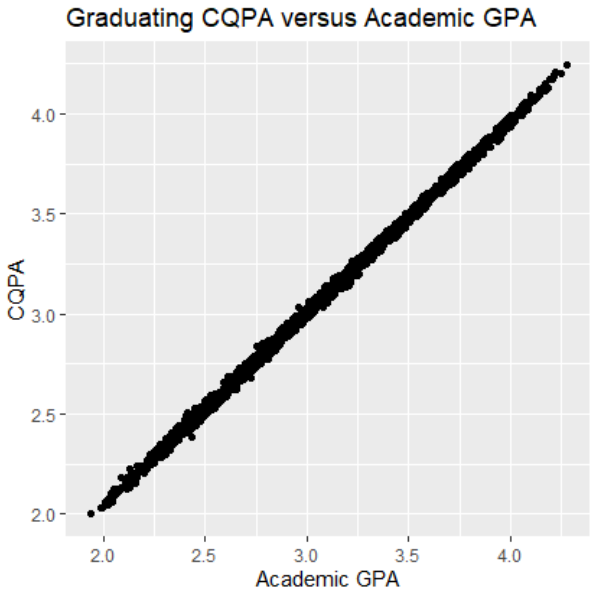
the three data sets into a single database where every row corresponds to a single cadet identification number and the columns correspond to the hundreds of variables of interest.

After merging the data sets, the first data cleaning task related to the recommendation letters. Since many of the goals of this project are not tied to analyzing the differences between letter writers, we combined all three academic recommendations (Math, English, and Science) into a single block of text for each candidate. Many text analysis tools for natural language processing are not compatible with this format, so we converted the raw text into a Corpus object which not only contains a collection of text documents but also metadata about the individual candidates. Using the statistical package, R, we removed all punctuation and whitespace from the Corpus as well as stop words—the words most common in English such as “the”, “at”, and “on.” Finally, we stemmed the remaining words in the Corpus to the root level with the high-performance Hunspell stemmer. For example, the words “loving”, “lovingly”, and “lovely” are all stemmed to the root form “love” with the Hunspell stemmer. After creating, cleaning, and stemming the Corpus, we binded the Corpus content back to the original data set. As a result, we finished the data preparation stage with a single data set containing pre-admission variables, post-admission variables, individual letters of recommendation, and a cleaned version of the combined academic letters of recommendation.

Word Count

With a clean dataset, we began looking for systematic indicators in recommendation letters that suggest an applicant would succeed at USMA. Before finding these systematic indicators, we had to choose a response variable that represents success at USMA. The most obvious metric is graduating academic GPA. However, for reasons that will be discussed momentarily, instead of academic GPA we ultimately decided to use a USMA specific outcome

called CQPA, which is a combination of a cadet's academic, physical, and military performance. Because most of CQPA derives from a cadet's academic performance, Figure 1 demonstrates there is an extremely high correlation between academic GPA and CQPA:



As seen in Figure 1, it seems reasonable to use CQPA instead of academic GPA. The purpose of choosing this alternative metric is because the dataset only contains academic GPA for cadets who graduated. If we used academic GPA as the response, this would automatically exclude cadets from the graduating classes of 2020, 2021, and 2022 as well as cadets who dropped out of West Point after arriving—reducing the data set of 8,602 cadets to a mere 4,220. To avoid dropping half the data set, we decided to use a cadet's CQPA at the end of their freshman year as the response. Although using freshman year CQPA is not perfectly ideal, we have values of this metric for all but 248 cadets. Given the extremely high correlation between graduating CQPA and graduating academic GPA, we infer that freshman year CQPA and freshman year academic GPA are also highly correlated. As a result, it is important to note that

the results of this paper essentially rely on predicting an applicant's success after one year in college.

After settling on a response variable that represents success at USMA, we started analyzing the recommendation letters. One of the first ideas was to simply examine the raw number of words a recommender wrote about an applicant. This idea prompted us to compute four metrics about the letters—the total word count of all three letters, the word count of the shortest letter, the word count of the longest letter, and the average word count across all three letters. We then fit a simple linear regression model to each of the word count metrics using freshman CQPA as the response. Table 1 demonstrates the results of each single-term model:

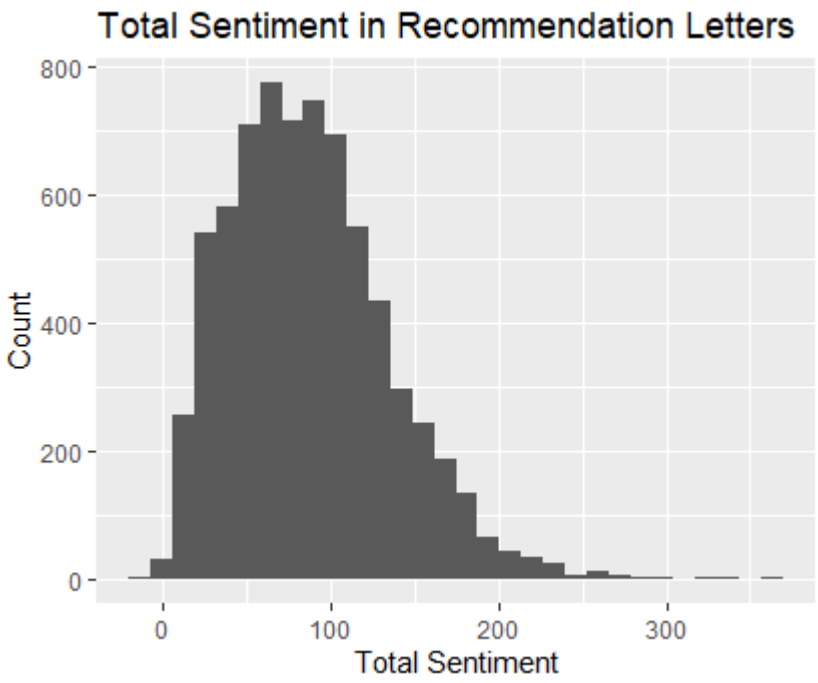
Explanatory Variable	**Estimate**	**P-Value**	**R^2 Value**
Total Word Count	.0001829	<.0001	.01676
Shortest Word Count	.001268	<.0001	.03896
Longest Word Count	.0004241	<.0001	.02155
Average Word Count	.0009391	<.0001	.04259

As seen in Table 1, using any of the word count metrics to predict CQPA is highly significant, and the positive estimates suggest a longer letter is correlated with higher grades. Surprisingly, even an extremely simple model of the average word count across all three letters predicts graduating CQPA with an R^2 value of .04259. The question naturally arises: can we find even more latent indicators in the recommendation letters to increase the predictive power of our model?

Sentiment

Sentiment analysis automates the interpretation and classification of emotions from text. A frequently applied sentiment analysis algorithm involves segmenting text into individual words and then assigning a sentiment score to each word by mapping the word to a pre-designated sentiment dictionary called a lexicon. There are several lexicons available in commonly used programming languages. Here, we describe a few commonly used lexicons in the programming language, R. The "afinn" method assigns each word in the lexicon a score between -5 and 5, with higher scores indicating positive sentiment and lower scores indicating negative sentiment. A coarser method, "bing", classifies words in the lexicon as either positive or negative. The "nrc" method maps each word in the lexicon to one of eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). By summing the individual sentiment scores across a piece of free form writing, one can identify patterns such as largely positive sentiments or even specific emotions such as high frequencies of joy.

Based on our prior research, we decided to use the "afinn" method to describe positive and negative sentiment in the recommendation letters because it maps on a broad -5 to 5 scale rather than a binary positive and negative scale. As a result, we applied the "afinn" lexicon to the cleaned version of the combined academic letters of recommendation Corpus and calculated the total sentiment as the number of positive words minus the number of negative words. Additionally, we created two more sentiment metrics: PPS is the number of words with positive sentiment divided by the total number of words, and PNS is the number of words with negative sentiment divided by the total number of words. Figure 2 demonstrates the distribution of total sentiment in the data set:



As expected, the distribution of sentiment is heavily positive—the average applicant had approximately one hundred more positive words than negative ones. Additionally, while the highest sentiment count was 364, the lowest sentiment count was -12. Furthermore, out of a data set of 8,251 applicants, only seven applicants had recommendations with overall negative sentiment.

After creating the sentiment variables, we fit new simple linear regression models to the new metrics. However, we decided not to use total sentiment because it is highly correlated to word count—a recommender will tend to write more positive words if they are already writing a longer letter. Instead, we used PPS and PNS because the ratios control for the length of a piece of writing. Table 2 demonstrates the results each single-term model:

Explanatory Variable	Estimate	P-Value	R^2 Value
PPS	-.7087	<.0001	.003838
PNS	-1.4147	<.0001	.00197

Although the PPS model is significant, it has a very small R^2 value, suggesting sentiment is not a great predictor. It is interesting to note, however, that the estimate for PPS is negative. This implies a recommender will use fewer positive words for students they feel will perform well. We conjecture this occurs as recommenders use more specific examples of accomplishments instead of non-specific positive language.

LIWC

At this point, we had created a handful of variables—such as word count and sentiment—that did a decent job predicting freshman year GPA. To increase the predictive power of our model, we began searching previous literature to create even more explanatory variables in the recommendation letters. Examining previous research in this area naturally led to the work of social psychologist and language expert James W. Pennebaker, who has spent the past thirty years researching computational linguistics. In his book *The Secret Life of Pronouns*, Dr. Pennebaker describes the development of a computer program called Linguistic Inquiry and Word Count (LIWC). This program contains over eighty different word dictionaries to capture nearly all types of words people use in everyday language. For example, the program contains dictionaries for pronouns, prepositions, verbs, positive emotions, negative emotions, cognitive processes (cognitive words such as “know” and “cause”), perceptual processes (perceptual words such as “look” and “heard”), and comparisons (comparison words such as “bigger” and “best”).

LIWC performs its analysis by comparing every word in a piece of free text to all the dictionaries, and then calculating the percentage of total words that are linked to each dictionary. In the end, after combining each applicant's academic letters of recommendation into a single chunk of text and running this through LIWC, we had created ninety-three additional predictor variables about the recommendation letters. However, some of the LIWC output seemed trivial. For example, punctuation variables such as the percentage of colons or parenthesis did not seem relevant. Additionally, we had concerns about variables with low base rates (less than .5%) because a small number of recommendation letters could unduly impact the model. As a result, we removed the punctuation and low base rate variables in the LIWC output, trimming down to fifty-two LIWC output variables. Now, the biggest question was how to find the most significant terms between the word count, sentiment, and LIWC variables.

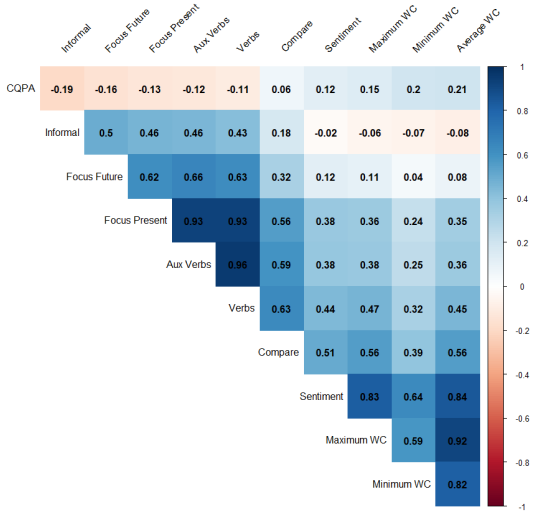
Predictive Model

Out of a pool of fifty-seven (3 word count, 2 sentiment, 52 LIWC) potential explanatory variables, we wanted to pick out a few that best predicted freshman year CQPA. We decided to create a multiple linear regression model and perform best subset selection. To perform best subset selection, the statistical package, R, fits a multiple regression model for each possible combination of predictors in a *k*-term model. In this situation, we let *k* vary from one to eight. Then, it computes performance metrics for each model—usually metrics such as adjusted R^2—and returns the best *k* variables. In other words, R will compute each possible combination of predictors and return the single term model with the highest adjusted R^2, then return the two-term model with the highest adjusted R^2, and so on until reaching the eight-term model. Although this method is exhaustive and computationally difficult, it does pull out the most predictive variables in our data set.

Interpretable Model

The previous model did a good job predicting freshman year CQPA. What it lacks, however, is interpretability—it is not necessarily intuitive why the model picked out certain variables, or even why the coefficients point in a specific direction. As a result, we decided to try creating a second model that made more intuitive sense. Although this new model might not be quite as predictive as the previous one, it would be much more interpretable.

Starting with the same pool of fifty-seven explanatory variables, we began by examining the simple correlations between variables. Although most of the predictors were slightly correlated with freshman year CQPA, only nine were strongly correlated—meaning the absolute value of the correlation is greater than .10. The following graph visualizes the correlations of some of these predictors.



Looking at the top row in the correlation plot, we see several variables are highly correlated with grades. The more important part of the story is that many of the variables are very highly correlated with each other, meaning they are essentially the same concept. For example, the minimum, maximum, and average word counts are extremely highly correlated with one another, as are verbs, auxiliary verbs, and present focus. There is a sense only a few factors are driving all the variation in grades. After some experimentation, we found that just a four-term model of just average word count, proportion of comparative language, verbs, and informal language not only captures most of the factors in the graph but also does a good job predicting freshman CQPA. Furthermore, the coefficients for each of the four terms points in the direction we would expect based off the correlations, meaning this new model is much more interpretable.

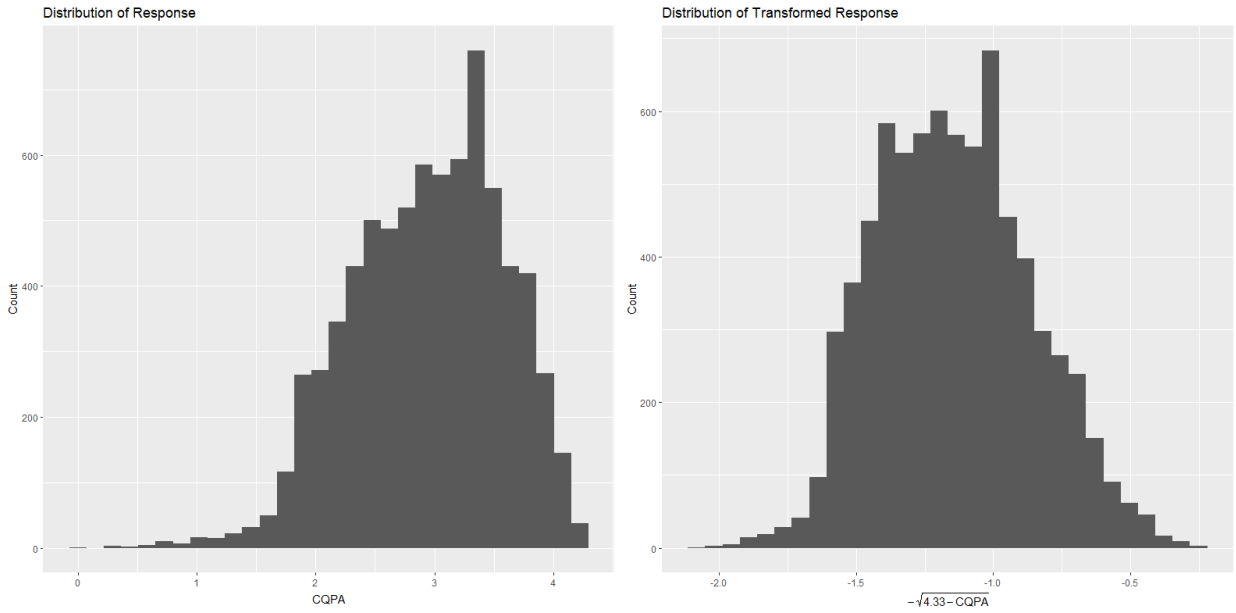
Results

Original Model

To determine how well the automated text analysis model performs, it would be beneficial to first determine the performance of the previous admissions model that uses only survey scores. When we tried fitting this model, however, we noticed several problems occurred such as a non-constant variance of the error terms. Many of these problems happened because the distribution of the response variable (freshman year CQPA) has a strong left skew. To overcome these validation issues, we transformed the response variable according to the following formula:

$$y_{transformed} = -\sqrt{4.33 - y}$$

Performing this transformation removes the left skew and makes the distribution of the response variable mirror the normal distribution. The following plots demonstrate the distribution of the original response variable as well as the transformed response variable.



After transforming the response, we also wanted to determine a metric for evaluating the performance of the model. We decided to randomly split the data set into a 70% training set and a 30% testing set. We trained the model on the training set then evaluated performance by calculating the R^2 value on the testing set. Ultimately, the original model has the form:

$$-\widehat{\sqrt{4.33-y}} = \beta_0 + \beta_1 \times \text{Survey Score} + \epsilon$$

	Estimate	**P-Value**
β_0	-4.3240	<.0001
β_1	.004452	<.0001

Using these coefficients on the testing set, the original admissions model has an R^2 value of .1312.

Predictive Model

The question naturally arises: How much better does automated text analysis predict the performance of freshman year grades? After creating an eight-term multiple linear regression model with best subsets selection on the same training set, we evaluated the performance of this predictive model on the testing set. This new model has the form:

$$-\widehat{\sqrt{4.33 - y}} = \beta_0 + \beta_1 a + \beta_2 b + \beta_3 c + \beta_4 d + \beta_5 e + \beta_6 f + \beta_7 g + \beta_8 h + \beta_9 i + \epsilon$$

Where a = Average word count; b = Analytic writing style; c = Articles; d = Verbs; e = Numbers; f = Insights; g = Certainty; h = Work; i = FAS

	Estimate	**P-Value**
β_0	-2.913	<.0001
β_1	1.191	<.0001
β_2	-.009771	<.0001
β_3	.05069	<.0001
β_4	-.04766	<.0001
β_5	.08217	<.0001
β_6	.06813	<.0001

β_7	.04658	<.0001
β_8	.02357	<.0001
β_9	.008309	<.0001

Using these coefficients on the testing set, the predictive model has an R^2 value of .2188. It is interesting to note that the training set has an R^2 value of .2060. Given the large performance decrease between the training and testing sets, it is likely this model is overfitting the data.

Interpretable Model

As described in the methods section, we found that using average word count, comparative language, verbs, and informal language captures most of the variation in CQPA. As a result, our descriptive model takes the form:

$$-\widehat{\sqrt{4.33 - y}} = \beta_0 + \beta_1 a + \beta_2 b + \beta_3 c + \beta_4 d + \beta_5 e + \epsilon$$

Where a = Average word count; b = Proportion of comparative language; c = Proportion of informal language; d = Proportion of verbs; e = FAS

	Estimate	**P-Value**
β_0	-3.7490	<.0001
β_1	.0003873	<.0001
β_2	.01838	<.0001
β_3	-.02569	<.0001

β_4	-.01137	<.0001
β_5	.003685	<.0001

This model is much simpler than the last one, and it still has an R^2 value of .2067 on the testing set.

We would like to highlight a few important observations about this model. Primarily, this model suggests a student who performs better has longer recommendation letters, has recommendation letters with a higher proportion of comparative language such as "greater" and "best", and has scores higher on the FAS surveys. On the other hand, a student who performs worse has recommendation letters with a larger proportion of verbs and informal writing styles. Second, it is interesting to look at the relative importance of the five predictor variables in this model:

Variable	**Relative Importance**
Average word count	19.4%
Proportion of comparative language	3.0%
Proportion of informal language	9.5%
Proportion of verbs	10.4%
FAS	57.6%

It appears as if nearly 50% of the prediction power comes from automated text analysis, and the other 50% of the prediction power comes from the survey scores. As a result, our recommendation for implementing this type of model in an admissions context is to weight half

of the recommendation score from the surveys and half of the recommendation score from the text analysis.

Finally, we would like to explore two additionally model performance metrics besides the R^2 value. This model has a root mean square error (RMSE) of .5814. In the context of the current problem, this statistic suggests the typical prediction will be off by .5814 CQPA. One additional way we can determine the accuracy of this model is if the prediction is within a "sign" of actual performance. For example, if a student is predicted to achieve a B average, but scores a B- or B+, we consider this to be successful. This model successfully predicts 42.2% of the time, and successfully predicts within two "signs" 74.8% of the time. For comparison, the base rate accuracy of assuming every student will achieve a 3.0 CQPA is 34.8%.

Conclusions

In this paper, we used automated text analysis to create a model that assesses thousands of USMA recommendation letters in a matter of seconds. The previous model used to assess recommendation letters has an R^2 value of .13, while our new and easily interpretable model has an R^2 value of .21. In other words, by utilizing previously latent data in the recommendation letters, we were able to increase the predictive R^2 value by 57.5%. When combined with traditional performance metrics such as high school GPA, class rank, and standardized test scores, colleges will be able to make even more objectively informed admissions decisions.

REFERENCES

1. Clinedinst M & Melissa C (2019 STATE OF COLLEGE ADMISSION REPORT. *The Journal of college admissions* e-pub prior to print(246).
2. Kuncel NR, Kochevar RJ, & Ones DS (2014) A Meta-analysis of Letters of Recommendation in College and Graduate Admissions: Reasons for hope. 22(1):101-107.
3. Baxter JC, Brock B, Hill PC, & Rozelle RM (1981) Letters of recommendation: A question of value. (American Psychological Association), pp 296-301.
4. Aamodt MG, Bryan DA, & Whitcomb AJ (1993) Predicting Performance with Letters of Recommendation. 22(1):81-90.
5. Anonymous (2020) Admissions Statistics.
6. Pennebaker JW (2011) *The secret life of pronouns : what our words say about us* (Bloomsbury Press, New York) 1st U.S. Ed pp xii, 352 p.
7. Anonymous (2020/10/18/) Colleges Will Consider the Full Picture. The New York Times, p L10(L).
8. Vigdor N & Diaz J (May 21, 2020) More Colleges Are Waiving SAT and ACT Requirements. New York Times.
9. Hubler S (2020/05/24/) Why the Tests Instituted To Lift College Diversity Are Falling Out of Favor. The New York Times, p A22(L).
10. Pennebaker JW, Chung CK, Frazee J, Lavergne GM, & Beaver DI (2015) When Small Words Foretell Academic Success: The Case of College Admissions Essays. *PLoS One* 9(12):e115844.