United Way Community Impact Grant Evaluation Process Project

United Way of Northeast Georgia
Department of Community Impact

Emma Brown, Mira Patel, and Jonah Pryor
University of Georgia, Department of Statistics

STAT 5020W: Statistics Capstone II
May 10th, 2021

Contents

List of Figures

List of Tables

1 Introduction

1.1 The Client

United Way is a worldwide organization engaged in nearly 1,800 communities across more than 40 countries and territories. The main focus of the organization centers around creating community-based and community-led solutions that strengthen the three cornerstones for a good quality of life: education, financial stability, and health.

United Way of Northeast Georgia is the branch of United Way that serves twelve counties in Northeast Georgia including Banks, Barrow, Clarke, Elbert, Franklin, Greene, Hart, Jackson, Madison, Morgan, Oconee, and Oglethorpe. Their goal is creating a Northeast Georgia region where every man, woman, and child has access to quality education, financial stability, and a healthy lifestyle. They are working to motivate and mobilize resources to meet the highest priority needs of the individuals and families living in the region. United Way of Northeast Georgia focuses on three pillars: basic needs, early childhood success, and workforce development. They follow the model of "investing in impact" through granting funds to programs of local nonprofits, providing resources, and offering training and educational series to nonprofit leaders and organizations [9].

Mark Madison is the Director of Community Impact for United Way of Northeast Georgia. He oversees the Community Impact Grant Application Evaluation Process outlined in Section 1.2. His goal is to improve the Community Impact Grant Application Evaluation Process and ensure that it is designed in a way that gives every applicant the ability to present the best application possible, and every evaluator the ability to score each question as accurately as possible.

1.2 Community Impact Grant Evaluation Process

One of the ways that United Way follows the model of "investing in impact" is granting funds to programs of local nonprofits, which they do by awarding Community Impact Grants. Each year United Way has a certain amount of funding that they award to programs of nonprofit organizations that apply for their Community Impact Grants. Volunteers for United Way score these applications. The evaluation of these applications determines the recipients of the Community Impact Grants.

Programs can apply for a Community Impact Grant in one of three categories, which correspond to the three impact areas United Way of Northeast Georgia focuses on: Basic Needs, Early Childhood Success, and Workforce Development. As part of the application process, volunteer evaluators and applicants are given training on how to grade or fill out the application. This is done to ensure consistency, fairness, and reduce ambiguity in how the questions are interpreted.

The application itself has the following sections that address essential information about the applicant's program(s): Program Description and Need for

Service, Collaboration, Program Performance Measures, Program Budget Request, and Agency Financial Information. Within each of these sections there are several evaluation prompts that the volunteer evaluators look at in order to grade the applications. For each prompt, evaluators select a pre-written response that they feel reflects how the program meets the standards of that specific prompt. Based on the number of choices available, these responses are then translated to a score on either a scale of one-to-five or zero-to-two points. These scores are then weighted corresponding to how important that part of the application is and totaled into a final application score.

1.3 The Data

The available data comes from two sources: the applications submitted by the local programs requesting funding from United Way and the evaluations submitted by the volunteer evaluators. The applications are available as PDF copies of each organization's application. The evaluation data is formatted in an Excel file containing individual data tables for each applicant program. There are nineteen total evaluation prompts among the five application sections specified above and thirty-one applicant programs among the three categories specified above. Fourteen evaluators were assigned to the applications in the Basic Needs application category, fifteen to the Early Childhood Success category, and fourteen to the Workforce Development category. Every evaluator did not evaluate every application in their respective category, and every application did not receive the same number of evaluations. This poses a challenge because since every evaluator does not evaluate every application in their category, we cannot consider the data paired, but since the evaluators for each application are not completely unique, the data will not meet the independence assumptions for most statistical tests.

As mentioned in Section 1.2, within each application section there are several evaluation prompts that the volunteer evaluators look at in order to grade the applications. For each prompt, evaluators select a pre-written response that they feel reflects how the program meets the standards of that specific prompt. This type of data is known as Likert Scale Data. A Likert scale is a psychometric interval survey rating scale, which measures a response to a close-ended question on a rating scale of how much the evaluator agrees or disagrees. Based on the number of choices available, these responses are then translated to a score on either a scale of one-to-five or zero-to-two points. These scores are considered ordinal data. Traditionally, there are certain limitations to the way ordinal data can be presented and analyzed, since it cannot be assumed that the difference between a score of one and two is equivalent to the difference between a score of two and three. However, since the scores given by the set of evaluators on each application are averaged into a final score, and the Community Impact Department compares applications based on differences in these final average scores, the value of a point is given a level of interpretability in this context.

There is no missing data in terms of missing values in the dataset we have. Every submitted score from every evaluator is available for each organization and every organization included in the dataset is paired with a copy of the respective application. However, the data is “incomplete” in the sense that we do not have scoring information or applications for programs that were eliminated in the first round of evaluations. Additionally, there is no record of volunteers who were assigned to evaluate a certain application and failed to submit their evaluation.

1.4 Research Objective

Mark Madison, the Director of Community Impact, is interested in improving the Community Impact Grant Application Evaluation Process. The primary goal is to ensure that the application process is designed in a way that gives every applicant the ability to present the best application possible, and every evaluator the ability to score each question as accurately as possible. Limited funding sets constraints on the amount of applicants that can be chosen for a grant, so it is imperative that the organizations with the greatest potential for high community impact get chosen. The research will aim to inform the Director of Community Impact in future training of the applicants and evaluators and further development of the application and evaluator prompts. The client has introduced the following questions to guide the analysis:

1. **Were applicant responses consistent enough to imply a common understanding of application prompts?**
   If an application question is written in a way that there is not a universal understanding amongst the applicants of what is being asked, this could potentially put some or all applicants at a disadvantage. In Section 2, the consistency of applicant responses is investigated using sentiment analysis to determine if common understanding of application prompts can be assumed.

2. **Were any evaluation prompts consistently scored very poorly or very well?**
   Consistently high-scored prompts could be an indication of a well-worded prompt or successful training of applicants and evaluators in that area of the application, while consistently low-scored prompts could indicate the opposite. In Section 3, score distributions for each evaluation prompt are compared to an average score distribution, to determine if any prompts consistently receive exceptionally low or high scores.

3. **Did any evaluators give consistently low or high scores?**
   A certain amount of variability in how evaluators score is inevitable, but if a particular evaluator is consistently scoring much lower or higher than the others, this could have a significant impact on which programs are chosen to receive the Community Impact Grant funding. In Section 4, score distributions for each evaluator are compared to an average score distribution, to determine if any evaluators consistently give exceptionally low or high scores.

4. **Was the number of evaluators enough for each application that it normalized the possible consistently low or high scoring evaluators?**
   If enough volunteer evaluators are assigned to each application, the average score received will not be as influenced by an extreme-scoring evaluator. In Section 5, the influence of each evaluator on an application's average score is compared based on the number of evaluators assigned to an application, to determine a sufficient number of evaluators to mitigate the effects of extreme scorers.

5. **Were the evaluation prompts adequate for evaluators to make consistent judgements?**
   Evaluation prompts should be written in a manner in which the evaluators can confidently determine whether the question was adequately addressed. In Section 6, instances of inconsistency in judgement were evaluated based on correlation and variance measures to identify any prompts that may be inadequately written or explained.

The analysis conducted in this report will center around these questions which will inform the decisions of the Community Impact Program in their efforts to improve the Community Impact Grant Application and Evaluation process.

1.5 Challenges

There are several challenges posed given the research questions and available data. In order to answer the research questions regarding the consistency of application responses and evaluator scores, methods to gauge or quantify "consistency" must be determined. The first research question centers on the consistency of the applications themselves. To evaluate the consistency of applicant's responses, sentiment analysis should be conducted on the applications. Since the applications are in PDF format, this presents an additional challenge. To facilitate this analysis, the available PDF copies of the applications were converted into a dataset containing identifying information about each organization and program along with their written responses to each application section.

The remaining research questions focus on the numerical score data contained in the original dataset. As previously mentioned, the scores contained in this dataset were given on two different grading scales, one-to-five and zero-to-two. Additionally, some scores are weighted corresponding to their significance. These weighted scores are proportional to the original scores given, but are not the exact values assigned by evaluators in their grading process. While it is important to address the impact the different scales have on the grading process of evaluators and analyze the difference in impact different sections of the application have due to the difference in weights of evaluation prompt scores, the research questions of interest for this project primarily focus on the decisions made by evaluators while scoring the applications. To directly compare the way prompts were scored by evaluators, the consistency of evaluator scoring, and the spread of scores given by each evaluator, it is best to consider these values on the original scale with which they were assigned. To this end, the dataset used for analysis in this report contains the raw, unweighted scores originally given for each evaluation prompt.

As mentioned in Section 1.3, while there are no missing values in the data set, the sample is incomplete in the sense that there exists only records for applications who made it past the first round of evaluations. As a result, applications that scored very poorly and the evaluations for those applications are not included in the score data. This will affect the distribution of scores for each application prompt as well as the distribution of each evaluator's assigned scores. This presents challenges in the interpretation of all the quantitative measures based on score such as the spread of scores assigned by each evaluator or the spread of scores attained on each prompt. Because information is missing regarding how evaluators scored applications that did not make it to this round of evaluations or how those applications would have scored on each of the application prompts, caution must be used in interpretations and decisions of which prompts were consistently scored higher or lower than others and which evaluators gave consistently higher or lower scores. To facilitate these comparative interpretations using only the higher-leaning scores available given the evaluations present in this data set, scores will be standardized based on the average score given on each prompt and the variability between all scores given on each prompt where appropriate.

2 Were applicant responses consistent enough to imply a common understanding of application prompts?

2.1 Exploratory Data Analysis

The first research question proposed by the client addresses whether applicant responses were consistent enough to imply a common understanding of the application prompts. Determining consistency in applicant responses depends solely on the contents of applications, which are contained in text form. In an effort to quantify the contents of these applications in a meaningful way, sentiment analysis can be performed on the data. Consistent sentiment can be considered an indication of general response consistency, implying a common understanding of application prompts.

Sentiment scores for each application can be obtained using the *sentimentr* package in R, which evaluates sentiments on a sentence level, and gives each sentence a score ranging from negative to positive values, corresponding to negative and positive sentiments, respectively [6]. The average sentiment analysis score for each application is calculated as the average of the sentiment scores obtained from every sentence contained in the application.

The main indication of response consistency is an approximately normal distribution of the average sentiment score by application. A Normal QQ plot can be used to evaluate how well a distribution matches a standard normal distribution, by plotting the distribution of sentiment scores against sample quantiles drawn from a normal distribution with the same mean and standard deviation as the distribution of sentiment scores. Figure 1 contains the Normal QQ plot of the distribution of average sentiment analysis scores for each application.

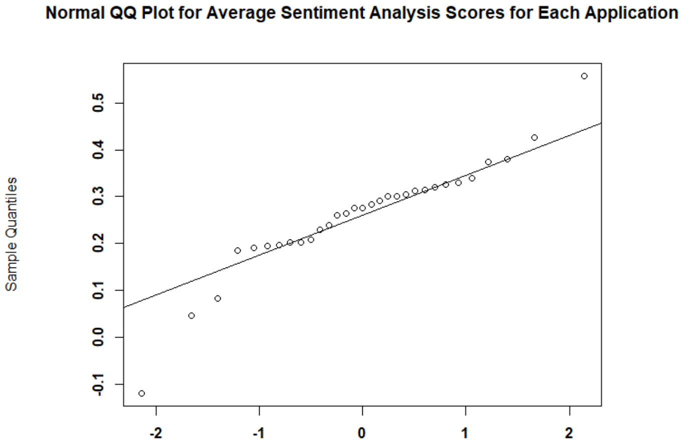


Figure 1: Normal QQ Plot for Average Sentiment Analysis Scores for Each Application

As seen in Figure 1, the sentiment scores generally fall along the QQ line, with the exception of a few potential outliers. This indicates that the average sentiment analysis scores for each application could be distributed approximately normally. Exceptions to this normality could be indications of applications with inconsistent application responses. A Shapiro-Wilk Test for Normality can be applied to the distribution of average sentiment analysis scores to determine if the scores are approximately normally distributed [8]. Additionally, a Rosner's Generalized Extreme Studentized Deviate Test can be applied to the distribution, to identify any outliers which are exceptions to the general normality of the sentiment scores [5].

2.2 Methods

In order to definitively prove whether or not the data follows a normal distribution, a Shapiro-Wilk Test for Normality can be conducted on the data. Many researchers recommend the Shapiro-Wilk Test as the best choice for testing the normality of data [8]. A Shapiro-Wilk Test can easily be conducted on the data using the *shapiro.test* function in R. This test operates under a null and alternative hypothesis. The null hypothesis is that the data comes from a normal distribution, whereas the alternative hypothesis is that the data comes from a non-normal distribution. The *shapiro.test* function in R returns a p-value, and at significance level 0.05, if the p-value is less than 0.05, we reject the null hypothesis in favor of the alternative.

To examine whether the points that vary from the QQ line in Figure 1 are definitively outliers, a Rosner's Generalized Extreme Studentized Deviate (ESD) Test can be performed on the distribution of average sentiment scores. Rosner's Generalized ESD Test is used to detect one or more outliers in an approximately normal univariate dataset [5]. The Generalized ESD test is more flexible than other methods for determining outliers because it only requires an upper bound for the suspected number of outliers rather than an exact amount. For this specified upper bound k, the Generalized ESD Test essentially performs k separate tests: the first for a single outlier, the second for two outliers, up to a kth test for k outliers. The test operates under the null hypothesis that there are no outliers in the dataset, with the alternative hypothesis that there are up to k outliers in the dataset. Rosner indicates in his own simulation studies that the test is very accurate for samples that have 25 or more observations and reasonably accurate for samples that have 15 or more observations [5]. For this particular research question, our sample has 31 observations, which correspond to the average sentiment score for the 31 applicant programs.

The R package *EnvStats* contains the function *rosnerTest* which performs the Rosner's Generalized ESD Test [4]. To identify potential inconsistencies in applicant responses, the *rosnerTest* function in R can be used to perform the Rosner's Generalized Extreme Studentized Deviate Test on the average sentiment scores for each application.

2.3 Analysis

To determine whether the average sentiment analysis scores obtained from each of the 31 applications follow a normal distribution, a Shapiro-Wilk Test for Normality is performed on the distribution of average sentiment scores. Figure A.1 in the Appendix contains the R output results of this testing procedure. The test returned a p-value of approximately 0.02 for the average sentiment score distribution. This indicates that the data does not come from a normal distribution, based on the test hypotheses outlined in Section 2.2. However, as seen in Figure 1, the Normal QQ Plot implies that the data could come from a normal distribution if not for a couple potential outliers.

Therefore, a Rosner's Generalized ESD Test should be performed to see if any of the average sentiment scores are indeed outliers. The sole assumption for Rosner's Generalized ESD Test is approximate normality of the variable being measured. Since the majority of points fall on the QQ line in Figure 1, the Rosner's Generalized ESD Test is safe to perform for this data. In an effort to make sure all outliers are identified, Rosner's Generalized ESD test will be performed *k=5* times, testing for up to five outliers in the distribution of average sentiment scores for the 31 applications. Figure A.2 in the Appendix contains the R output results of this testing procedure. The Rosner's Generalized ESD Test identified one outlier, which corresponds to the observation in the far bottom left-hand corner of Figure 1. The program identified as an outlier by this process will be considered further to determine potential causes of the inconsistency.

Because Rosner's Generalized ESD Test identified only a single outlier, another Shapiro-Wilk Test for Normality is performed on the distribution of average sentiment scores, excluding the one confirmed outlier. Figure A.3 in the Appendix contains the R output results of this testing procedure. For this distribution, the Shapiro-Wilk Test for normality returns a p-value of approximately 0.19. Therefore, we fail to reject the null hypothesis, and conclude that the distribution of average sentiment scores follow a normal distribution at significance level 0.05, with the exception of the single confirmed outlier.

2.4 Results

In Section 2.3, it was found using a Shapiro-Wilk Test for Normality that the average sentiment analysis scores for each application generally followed a normal distribution, indicating reasonable consistency in application responses by the majority of applicant programs. There was, however, one deviation from this normality. One program was identified as an outlier in the distribution of average sentiment score by application using Rosner's Generalized Extreme Studentized Deviate Test. This program had the lowest average sentiment score of all programs, meaning that on average, they had the most "negative" tone in their responses among all applicants. This implies a lack of consistency in this program's responses on their application, but not necessarily a lack of understanding of the prompts. After identifying this program as an outlier, it was discovered that this program had the sixth-highest average evaluator grand total score among all programs. Therefore, even though this program lacked consistent sentiment in their responses, they did not appear to lack understanding of the prompts due to their high average total score. It should be noted, however, that this particular program is a very well-known program, so evaluator bias is a potential cause of the program's exceptionally high score.

3 Were any evaluation prompts consistently scored very poorly or very well?

3.1 Exploratory Data Analysis

To facilitate the identification of any evaluation prompts that were consistently scored very poorly or very well, the distributions of the raw scores given by all evaluators on all applications for each prompt were isolated. The prompts fall into one of two categories: those graded on a zero-to-two point scale and those graded on a one-to-five point scale. An average distribution for the two point scale questions was created using the average score given on every two point prompt for each application by every evaluator and the same was done for the five point prompts. The full prompt names corresponding to the variable names used in the following figures and discussion can be found in Table A.1 in the Appendix.

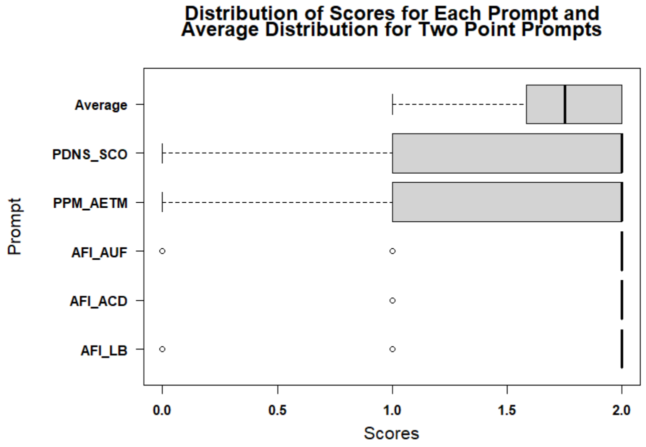


Figure 2: Parallel Boxplots of the Distribution of Scores for each Zero-to-Two Point Scale Prompt

Figure 2 contains parallel boxplots of the distribution of scores for each of the zero-to-two point scale prompts along with the constructed average distribution containing the average score given on every two point scale prompt for each application by every evaluator. The variables PDNS_SCO and PPM_AETM correspond to the fifth prompts in the Program Description and Need for Service section of the application and the Program Performance Measures section of the application, respectively. These two prompts have the exact same distribution where 50% of applications received a score of two, 25% received a score of one, and 25% received a score of zero. The variables AFI_AUF, AFI_ACD, and AFI_LB correspond to the first three prompts in the Agency Financial Information section of the application. These three prompts also have almost identical distributions, as nearly every application received a score of two on these prompts, and any instances where a score of zero or one was assigned are considered outliers.

Figure 3 contains parallel boxplots of the distribution of scores for each of the one-to-five point scale prompts along with the constructed average distribution containing the average score given on every five point scale prompt for each application by every evaluator. The distribution of scores for the variable C_CPKR, which corresponds to the only evaluation prompt in the Collaboration section of the application, is the only prompt which has a right-skewed distribution. The Collaboration section is an optional part of the application.

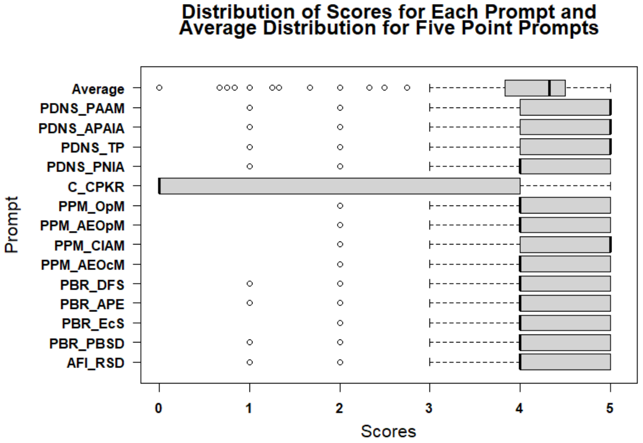


Figure 3: Parallel Boxplots of the Distribution of Scores for each One-to-Five Point Scale Prompt

When this section does not apply to an applicant, they receive a score of zero. The majority of applicants do not include a collaboration element, therefore the right-skewed distribution seen in Figure 3 is not unusual. Every other one-to-five point scale prompt has almost the same distribution, with differences only in whether the median score is a four or five and whether any applications received a score of one. For all the remaining prompts, scores of one and two are considered outliers when they occur, and 75% of the scores assigned are scores of either four or five.

Further analysis must be conducted to distinctly identify which of these prompts were consistently scored higher and lower compared to the other evaluation prompts across all five of these application sections. Comparing the distribution of scores given on each prompt to the average distribution of the corresponding point scale will identify the prompts which were consistently scored lower or higher than average. Two-Sample T-Tests will be used to compare each prompt's distribution to the average distribution for questions of the respective point scale. Those with significantly higher mean scores than the average will be considered consistently high-scoring prompts, and those with significantly lower mean scores than the average will be considered consistently low-scoring prompts.

3.2 Methods

The second research question outlined by the client regards whether any prompts were graded significantly higher or lower than others. This is a question of group comparison. Different methods of group comparison are appropriate for different kinds of data. As can be seen from Figures 2 and 3, the distributions of scores given on each prompt are not normally distributed which is an assumption of a parametric Two-Sample T-Test. Additionally, Two-Sample T-Tests are conducted under the assumption that the data are continuous. This is not the case for the ordinal score data resulting from each evaluation. However, the score data is ultimately averaged for each application, and funding decisions are made based on the total average score for each application. This means that this data has interpretations for between-integer point values that other ordinal data may not. Another assumption of a Two-Sample T-Test is that the data values are independent, meaning that there is no relationship between the observations in each group or between the groups themselves. This assumption is not met for the data because of the nature of the evaluation process, in which every evaluator looks at several applications, but not all of them, and every application is looked at by several evaluators, but not all of them. Additionally, since the distributions being compared with each prompt distribution are created by averaging, the scores given on each prompt are inherently involved in the average distribution they are being compared to. Fortunately, studies have been conducted that indicate that parametric Two-Sample T-Tests are very robust against violations to the fundamental assumptions of data continuity and normality. The study *Five-Point Likert Items: T Test versus Mann-Whitney-Wilcoxon* conducted by de Winter and Doduo, indicates that parametric T-tests yield valid results, even under assumption violations, on Likert scale ordinal response data [2]. Additionally, one of the primary concerns of using a Two-Sample T-Test on ordinal data is the interpretability of the results, since, for example, a value of 3.5 does not have specific meaning in an ordinal data distribution. However, since an application's final score is calculated as the mean of scores given by all evaluators on that application, average scores are interpretable in this context. In conclusion, parametric Two-Sample T-tests can be applied, with caution, on the available score data.

As mentioned in Section 3.1, an average distribution for the two point scale questions was created using the average score given on every two point prompt for each application by every evaluator and the same was done for the five point prompts. Comparing the distribution of scores given on each individual prompt to the respective average distribution using Two-Sample T-Tests will identify the prompts which were consistently scored higher or lower than average. When performing multiple group comparisons using the same data, the error rate can grow exponentially. Therefore, each set of tests will be performed with the Bonferroni Error Rate Correction, which controls the group error rate at a set level.

3.3 Analysis

To identify any prompts that were consistently scored very poorly or very well, Two-Sample T-Tests for differences in means can be conducted between each prompt and a theoretical average prompt of the same scoring scale. As explained in Section 1.3, there are two scoring scales used in this dataset, a one-to-five point scale and a zero-to-two point scale depending on the number of options presented to evaluators for each prompt. Distributions were created to represent average prompts of each size by calculating the average score given on every five point prompt for each application by every evaluator, and likewise for the two point prompts.

Two-Sample T-Tests were conducted between the distribution of scores of each zero-to-two point scale prompt and the theoretical average two point prompt score distribution. Table 1 shows the results of these five T-tests. The p-value represents the probability that the observed difference in sample means is at least as large as observed under the assumption that the true means of the score distributions are equal. The smaller the p-value, the more likely it is that the true means are not equal. Additionally, the upper and lower confidence limits for each 99% confidence interval are displayed in Table 1. The family-wise error rate for this set of t-tests is 0.05, so each individual T-test is conducted with a Bonferroni corrected error rate of 0.01, thus creating the 99% confidence intervals. For each of these confidence intervals, there is a 99% chance that the confidence interval contains the true mean difference between the score of each zero-to-two point scale prompt and the corresponding theoretical average prompt. Since all five confidence intervals contain zero, it is not concluded that any of these five prompts received scores significantly different from average.

Table 1: Two-Sample T-Test for Zero-to-Two Point Prompts

Prompt	P-Value	Lower Bound	Upper Bound
PDNS_SCO	0.36	-0.17	0.08
PPM_AETM	0.67	-0.15	0.11
AFI_AUF	0.10	-0.04	0.19
AFI_ACD	0.05	-0.03	0.19
AFI_LB	0.93	-0.14	0.13

Additionally, Two-Sample T-Tests were conducted between the distribution of scores of each one-to-five point scale prompt and the theoretical average five point prompt score distribution. Table 2 shows the results of these fourteen T-tests. The p-value represents the probability that the observed difference in sample means is at least as large as observed under the assumption that the true means of the score distributions are equal. The smaller the p-value, the more likely it is that the true means are not equal. Additionally, the upper and lower confidence limits for each 99.64% confidence interval are displayed in Table 2. The family-wise error rate for this set of T-tests is 0.05, so each individual T-test is conducted with a Bonferroni corrected error rate of approximately 0.0036, thus creating the 99.64% confidence intervals. For each of these confidence intervals, there is a 99.64% chance that the confidence interval contains the true mean difference between the score of each one-to-five point scale prompt and the corresponding theoretical average prompt. As can be seen in Table 2, there are five prompt distributions that have p-values less than the error rate of approximately 0.0036, indicating that they are significantly different from the theoretical average prompt.

Table 2: Two-Sample T-Test for One-to-Five Point Prompts

Prompt	P-Value	Lower Bound	Upper Bound
PDNS_PAAM	**0.00**	**0.31**	**0.79**
PDNS_APAIA	**0.00**	**0.14**	**0.66**
PDNS_TP	**0.00**	**0.17**	**0.64**
PDNS_PNIA	0.10	-0.12	0.42
C_CPKR	**0.00**	**-3.11**	**-1.96**
PPM_OpM	0.11	-0.12	0.39
PPM_AEOpM	0.01	-0.03	0.43
PPM_CIAM	**0.00**	**0.01**	**0.51**
PPM_AEOcM	0.28	-0.15	0.32
PBR_DFS	0.40	-0.19	0.35
PBR_APE	0.33	-0.16	0.32
PBR_EcS	0.18	-0.13	0.35
PBR_PBSD	0.54	-0.21	0.32
AFI_RSD	0.46	-0.19	0.32

The five variables that were identified as being significantly different than the average five point prompt distribution are PDNS_PAAM, PDNS_APAIA, PDNS_TP, C_CPKR, and PPM_CIAM. The confidence intervals can be used to determine the direction in which each of these variables differs from the average. The confidence intervals for the difference in means between PDNS_PAAM, PDNS_APAIA, PDNS_TP, PPM_CIAM and the theoretical average prompt do not contain zero, and both confidence limits are positive. This indicates that these prompts are scored consistently higher than average. The confidence interval for the difference in means between C_CPKR and the theoretical average prompt does not contain zero, and both confidence limits are negative. This indicates that this prompt is scored consistently lower than average.

3.4 Results

In Section 3.3, five variables corresponding to scores on evaluation prompts were identified as scoring significantly different from average by Two-Sample T-Tests. One variable C_CPKR was identified as being scored significantly lower than average. The variable C_CPKR corresponds to the scores given on the only prompt in the Collaboration section of the application. The collaboration prompt is presented to evaluators as the following question: "Collaboration (if applicable): How confident are you that each collaborator involved plays a key role in the program that leads to participant success?" Evaluators may select from the answers "Very Confident," "Confident," "Moderately Confident," "Slightly Confident," "Not at All Confident," or leave the answer blank, to indicate that the element of collaboration is not relevant to the application being scored. A selection of one of the answers corresponds to a score of one-to-five, while leaving the answer blank corresponds to a score of zero. This prompt is the only one scored on the one-to-five point scale that can receive a score of zero. Many applications did not contain an element of collaboration, and accordingly were correctly assigned a score of zero for this prompt. Therefore, it is expected that this prompt would consistently be scored lower than the others. The four variables PDNS_PAAM, PDNS_APAIA, PDNS_TP, and PPM_CIAM were identified as receiving scores significantly higher than average. The first three of these variables correspond to the first three prompts in the Program Description and Need for Service section of the application, which address Program Alignment with Agency Mission, Activities Purpose Align with Impact Area, and Target Population. The fourth variable corresponds to the third prompt in the Program Performance Measures section of the application, which addresses whether performance measures Contribute to Impact Area Measures. These higher scores could be indications of well written prompts and successful training of both applicants and evaluators, or could simply be correlated with the strengths of the applications that made it to this round of evaluation.

4 Did any evaluators give consistently low or high scores?

4.1 Exploratory Data Analysis

To facilitate the identification of any evaluators that gave consistently low or high scores, the distributions of the standardized scores given by each evaluator on all applications for all prompts were isolated. Figures 4-6 contain boxplots of the distribution of scores given by each evaluator separated by the three focus area of the applications: Basic Needs, Early Childhood Success, and Workforce Development, respectively. To generate these boxplots, the scores are standardized, meaning that each score is centered on the mean score given by all evaluators on a given question and scaled based on the spread of scores given by all evaluators on a given question. Every score given by a specific evaluator on any prompt makes up one of the boxplots. Therefore, the values in the distributions seen are an indication of how each evaluator's scoring compares to average scoring. Additionally, the standardized average score given on each prompt for every application in every category makes up the distribution labeled "Average". This allows us to visually compare what an average evaluator would look like to each of the individual evaluators' score distributions.

Figure 4 contains the distribution of scores given by each evaluator on applications from organizations in the Basic Needs category. In this plot, Evaluator 6 especially stands out as a harsh grader, and Evaluators 4 and 5 are also worth a closer look.

Figure 5 shows the same information for applications in the Early Childhood Success Category. Evaluators 17 and 19 stand out as potential harsh graders, and Evaluators 22 and 29 stand out as potentially generous graders in this parallel boxplot.

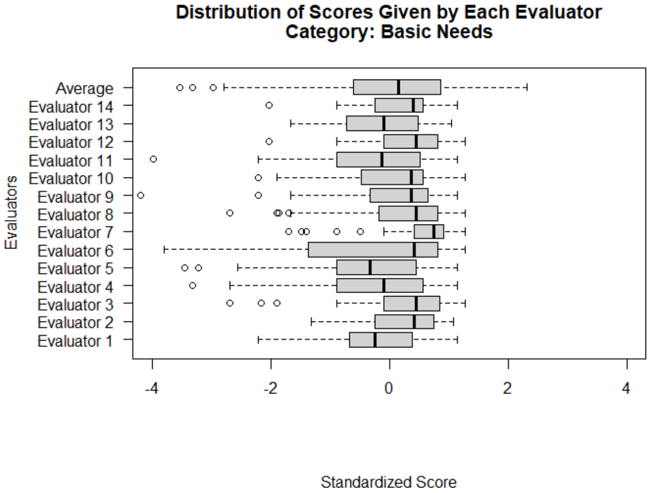


Figure 4: Parallel Boxplots of the Distribution of Scores Given by Each Evaluator in the Basic Needs Application Category

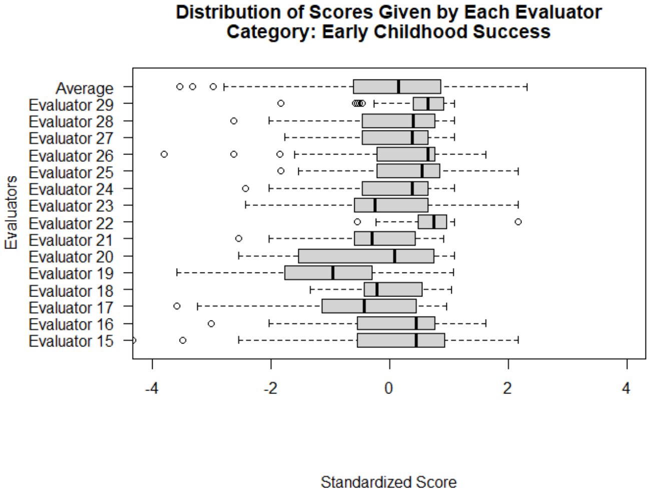


Figure 5: Parallel Boxplots of the Distribution of Scores Given by Each Evaluator in the Early Childhood Success Application Category

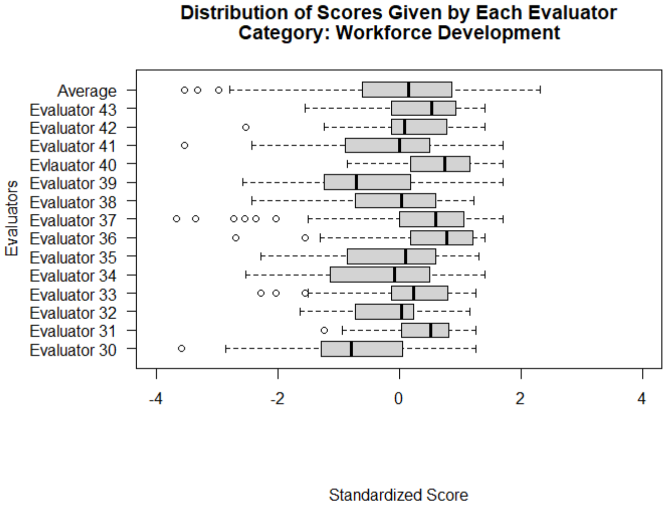


Figure 6: Parallel Boxplots of the Distribution of Scores Given by Each Evaluator in the Workforce Development Application Category

Finally, the distribution of scores given by each evaluator on applications from organizations in the Workforce Development category are shown in Figure 6. In this figure, Evaluators 36 and 40 stand out as potential generous graders, and Evaluators 30 and 39 stand out as particularly harsh graders.

To determine which evaluators' score distributions are statistically significantly different from the average distribution, Two-Sample T-Tests will be used to compare each evaluator's distribution to the average distribution. Those with significantly higher mean scores than the average will be considered generous graders, and those with significantly lower mean scores than the average will be considered harsh graders.

4.2 Methods

The third research question outlined by the client regards whether any evaluators tended to give significantly higher or lower scores than others. This is another question of group comparison. To make this kind of inference a two-sided test, which yields confidence intervals that indicate the direction in which a distribution significantly differs from the average is needed. The parametric Two-Sample T-Test provides these necessary components.

As explained in Section 3.2, the assumptions for parametric Two-Sample T-Tests are not sufficiently met for this data. An assumption of a Two-Sample T-Test is that the data values are independent, meaning that there is no relationship between the observations in each group or between the groups themselves.

This assumption is not met for the data because of the nature of the evaluation process, in which every evaluator looks at several applications, but not all of them, and every application is looked at by several evaluators, but not all of them. Additionally, since the distribution being compared with each evaluator distribution is created by averaging, the scores given by each evaluator are inherently involved in the average distribution they are being compared to. Two-Sample T-tests are also conducted under the assumption that the data in each group are normally distributed. Figures 6-8 show that this assumption is reasonable for the evaluator tests, because the standardization of scores used to generate these distributions yields unimodal and roughly symmetric boxplots for most of the evaluator distributions. Additionally, Two-Sample T-Tests are conducted under the assumption that the data are continuous. This is not the case for the ordinal score data resulting from each evaluation. However, the standardization of each score based on the mean and standard deviation of each individual prompt, results in evaluator distributions that closely mimic truly continuous data. Furthermore, the score data is ultimately averaged for each application, and funding decisions are made based on the total average score for each application. This means that this data has interpretations for between-integer point values that other ordinal data may not. Finally, the study *Five-Point Likert Items: T Test versus Mann-Whitney-Wilcoxon* conducted by de Winter and Doduo, indicates that parametric T-tests yield valid results, even under assumption violations, on Likert scale ordinal response data [2]. Therefore, the T-Tests will be conducted, but results for these tests should be interpreted with caution.

As mentioned in Section 4.1, an average distribution was created using the standardized average score given on each prompt for every application in every category, to represent a theoretical average evaluator. Comparing the distribution of scores given by each individual evaluator to the average distribution using Two-Sample T-Tests will identify the evaluators which consistently give higher or lower scores. When performing multiple group comparisons using the same data, the error rate can grow exponentially. Therefore, this set of tests will be performed with the Bonferroni Error Rate Correction, which controls the group error rate at a set level.

4.3 Analysis

To identify any evaluators that consistently gave higher or lower scores than others, Two-Sample T-Tests for differences in means can be conducted between each evaluator distribution and a distribution representing a theoretical average evaluator. The average distribution was created using the standardized average score given on each prompt for every application in every category, to represent a theoretical average evaluator.

Two-Sample T-Tests were conducted between the distribution of scores given by each individual evaluator and the theoretical average evaluator score distribution. Table 3 shows the results of these 43 T-tests. The p-value represents the probability that the observed difference in sample means is at least as large

Table 3: Two-Sample T-Test for Evaluators

Evaluator	P-Value	Lower Bound	Upper Bound
Evaluator 1	0.01	-0.78	0.07
Evaluator 2	0.08	-0.19	0.59
Evaluator 3	0.02	-0.11	0.67
Evaluator 4	0.12	-0.75	0.28
Evaluator 5	0.01	-0.96	0.11
Evaluator 6	0.01	-0.96	0.11
Evaluator 7	**0.00**	**0.21**	**0.86**
Evaluator 8	0.03	-0.13	0.58
Evaluator 9	0.57	-0.30	0.42
Evaluator 10	0.78	-0.36	0.42
Evaluator 11	0.03	-0.71	0.14
Evaluator 12	**0.00**	**0.02**	**0.67**
Evaluator 13	0.59	-0.85	0.63
Evaluator 14	0.13	-0.23	0.59
Evaluator 15	0.72	-0.51	0.64
Evaluator 16	0.81	-0.44	0.51
Evaluator 17	**0.00**	**-0.97**	**-0.07**
Evaluator 18	0.32	-0.20	0.37
Evaluator 19	**0.00**	**-1.48**	**-0.46**
Evaluator 20	0.22	-0.89	0.42
Evaluator 21	0.12	-1.27	0.51
Evaluator 22	**0.00**	**0.46**	**0.99**
Evaluator 23	0.25	-0.80	0.40
Evaluator 24	0.55	-0.28	0.40
Evaluator 25	**0.00**	**0.03**	**0.72**
Evaluator 26	0.05	-0.19	0.72
Evaluator 27	0.39	-0.25	0.42
Evaluator 28	0.41	-0.31	0.51
Evaluator 29	**0.00**	**0.23**	**0.80**
Evaluator 30	**0.00**	**-1.06**	**-0.18**
Evaluator 31	0.00	-0.01	0.78
Evaluator 32	0.13	-0.50	0.19
Evaluator 33	0.18	-0.24	0.56
Evaluator 34	0.06	-0.69	0.19
Evaluator 35	0.57	-0.61	0.44
Evaluator 36	**0.00**	**0.15**	**0.86**
Evaluator 37	0.25	-0.33	0.67
Evaluator 38	0.43	-0.52	0.32
Evaluator 39	0.00	-1.00	0.01
Evaluator 40	**0.00**	**0.30**	**1.00**
Evaluator 41	0.05	-0.75	0.19
Evaluator 42	0.05	-0.14	0.54
Evaluator 43	0.02	-0.15	0.79

as observed under the assumption that the true means of the score distributions are equal. The smaller the p-value, the more likely it is that the true means are not equal. Additionally, the upper and lower confidence limits for each 99.88% confidence interval are displayed in Table 3. The family-wise error rate for this set of T-tests is 0.05, so each individual T-test is conducted with a Bonferroni corrected error rate of approximately 0.0012, thus creating the 99.88% confidence intervals. For each of these confidence intervals, there is a 99.88% chance that the confidence interval contains the true mean difference between the scores given by each evaluator and the theoretical average evaluator. As can be seen from Table 3, there are ten evaluator score distributions that have p-values less than the error rate of approximately 0.0012, indicating that they are significantly different from the theoretical average evaluator.

The ten evaluators that were identified as being significantly different than the average evaluator distribution are Evaluator 7, Evaluator 12, Evaluator 17, Evaluator 19, Evaluator 22, Evaluator 25, Evaluator 29, Evaluator 30, Evaluator 36, and Evaluator 40. The confidence intervals can be used to determine the direction in which each of these evaluators differs from the average. The confidence intervals for the difference in means between evaluators 7, 12, 22, 25, 29, 36, 40 and the theoretical average evaluator do not contain zero, and both confidence limits are positive. This indicates that these evaluators consistently gave scores higher than average. The confidence intervals for the difference in means between evaluators 17, 19, 30 and the theoretical average evaluator do not contain zero, and both confidence limits are negative. This indicates that these evaluators consistently gave scores lower than average.

4.4 Results

In Section 4.3, the distributions of individual evaluator scores were compared to an average distribution to address which evaluators tended to score differently from average. The results of the Two-Sample T-Tests conducted identified evaluators 7, 12, 22, 25, 29, 36, and 40 as tending to score significantly higher than average and evaluators 17, 19, and 30 as tending to score significantly lower than average. In any environment where human judgement is involved, variability in that judgement is expected and welcome. The interest for this problem lies in whether these more extreme evaluators have too much influence on the determination of which applicant programs are eventually selected for funding, or if the evaluators that tended to give higher scores were simply assigned to applications that were stronger, and vice versa. This element is investigated further in Section 6 of this report, as it is addressed by the final research question proposed by the client.

5 Was the number of evaluators enough for each application that it normalized the possible consistently low or high scoring evaluators?

5.1 Exploratory Data Analysis

The client is interested in determining if the number of evaluators assigned to each application prevents individual evaluators from having an undue influence on each application's final average score. Ten evaluators were identified as giving scores significantly different from average in Section 4, so there is reason to believe that in some cases these evaluators who consistently score higher or lower than the average evaluator may have an exceptionally high impact on an application's final score. The client identified an undue impact of a single evaluator as being greater than a five point impact on score.

The assignment of evaluators to applications is not symmetrical. As outlined in Section 1.3, programs seeking a Community Impact Grant can apply in one of three Impact Areas: Basic Needs, Early Childhood Success, or Workforce Development. Evaluators are also assigned exclusively to grading applications in one of these sections. Table 4 shows the number of programs that applied in each category, the number of evaluators assigned to each category, and the number of significant evaluators identified in each category.

Table 4: Distribution of Applicant Programs and Evaluators to Each Application Category

	Number of Applicant Programs	Number of Evaluators Assigned	Number of High Scoring Evaluators	Number of Low Scoring Evaluators
Basic Needs	12	14	2	0
Early Childhood Success	7	15	3	2
Workforce Development	12	14	2	1
Total	31	43	7	3

Even within these categories, each evaluator does not grade the same number of applications, and every application does not receive the same number of evaluations. In the Basic Needs application category, seven applicants received four evaluations, four applicants received three evaluations, and one applicant received only two evaluations. In the Early Childhood Success application category, all seven applicants received six evaluations. In the Workforce Development application category, six applicants received four evaluations, five applicants received three evaluations, and one applicant received two evaluations. This asymmetry may not be ideal for fair comparisons between applicants but it does allow the impact an evaluator has on score to be compared across several levels.

To determine if the number of evaluators assigned to each application was sufficient to mitigate the effects of the evaluators who graded significantly higher or lower than average, the impact of individual evaluators on average score should be compared across the number of evaluators assigned to each application.

5.2 Methods

To determine if the number of evaluators assigned to each application was sufficient to mitigate the effects of the evaluators who graded significantly higher or lower than average, the impact each evaluator had on the scores of the applications to which they were assigned must first be quantified. A common method for quantifying the impact of an outlier is taking the mean of the dataset with the outlier and the mean of the dataset without the outlier, then subtracting the former from the latter [3]. The difference is considered the impact of the outlier on the dataset. Since the score for each application is calculated as the average of the scores given by different evaluators, we can quantify the impact of a particular evaluator by comparing the average score for an application including the score given by the evaluator of interest to the average score without the score given by the evaluator of interest. In line with the outlier method described, the difference in these two average scores can be considered the impact a particular evaluator has on an application's score. Since applications have different numbers of evaluators assigned to them, the range of impact for an evaluator for applications with a certain number of evaluators, can be compared across all available numbers of evaluators assigned to an application. By this method, the number of evaluators that is enough to contain the impact of the significantly harsh or generous evaluators can be determined.

5.3 Analysis

To determine whether the number of evaluators assigned to each application was enough to mitigate the effects of the extreme evaluators, the impact of each evaluator on each application was calculated by subtracting the average application score without the evaluator of interest from the final average score. These impact scores are then sorted by the number of evaluators assigned to the application. Figure 7 contains parallel boxplots of the distribution of individual evaluator impact on final average score by the number of evaluators assigned to the application on which that impact was calculated.

As seen in Figure 7, all these distributions are centered at zero. This is because the sum of all the impacts on an individual application is zero, since they are all subtracted from the average score. Therefore, the interest lies in the range of each of these distributions. As the number of evaluators assigned to each application increases, the range of individual evaluator impact generally decreases. Extreme scoring evaluators, identified in Section 4, occur on applications with all numbers of evaluators, but the impact of these significant evaluators is clearly more controlled as the number of evaluators increases. The client specified that it would be preferred if any particular evaluator did not impact an evaluators final average score by greater than five points. The

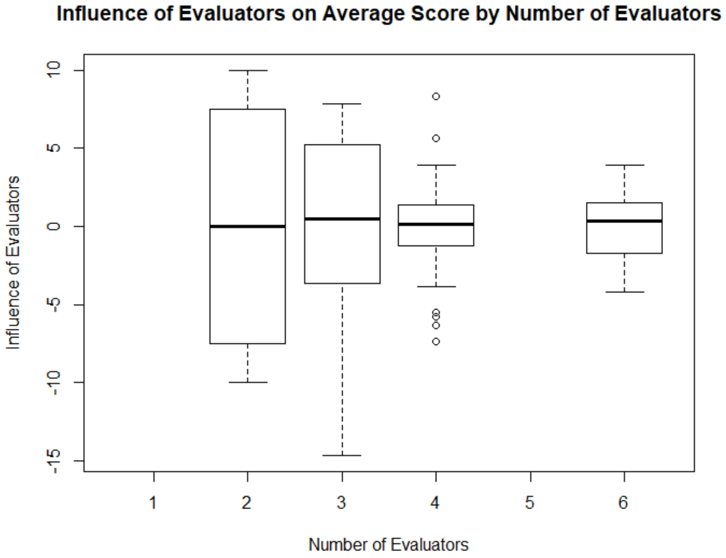


Figure 7: Parallel Boxplots of the Distribution of Evaluator Impact on Application Final Score

only distribution where this occurs is the distribution of evaluator impact for applications scored by six evaluators.

5.4 Results

In Section 5.3, the distributions of the impact of evaluators on applications with different numbers were compared to determine if the number of evaluators assigned to each application was sufficient to mitigate the effects of extreme scoring evaluators. The client stated that the maximum tolerable impact of a single evaluator on an application's final average score is five points. As seen from Figure 7, the only distribution in which evaluator impact was contained to five points or less is the distribution corresponding to applications scored by six evaluators. As explained in Section 5.1, the only application category in which applications received six evaluations is the Early Childhood Success application category. This indicates that applications in Basic Needs and Workforce Development are more susceptible to the influence of extreme evaluators. This could be an unfair advantage if they are assigned to consistently higher-scoring evaluators, and could be an unfair disadvantage if they are assigned to consistently lower-scoring evaluators.

6 Were the evaluation prompts adequate for evaluators to make consistent judgements?

6.1 Exploratory Data Analysis

To begin to evaluate whether the application and the following evaluation prompts were adequate for evaluators to make consistent judgements, inconsistencies in judgement must first be identified. To this end, the variance in scores given on each prompt on each individual application were calculated. This directly quantifies the difference between how evaluators graded the exact same response on the exact same application. These individual variance scores are then totaled for each prompt. These totals give us a measure of consistency in scoring for each prompt. Table 5 contains the modified variance scores given to each prompt. The full prompt names corresponding to the variable names used in the following figures and discussion can be found in Table A.1 in the Appendix. As can be seen from Table 5, most of the scores are very similar, with one clear exception. This drastically higher modified variance score corresponds to the only prompt in the Collaboration section of the application. To definitively identify the outliers in the modified variance score, a Rosner's Generalized Extreme Studentized Deviate (ESD) Test can be conducted. The prompts corresponding to the outliers formally identified by this test should be investigated further to determine potential causes of aberrant variability.

Table 5: Modified Variance Scores for Each Evaluation Prompt

Prompt	Modified Variance
PDNS_PAAM	14.46
PDNS_APAIA	17.36
PDNS_TP	16.05
PDNS_PNIA	17.45
PDNS_SCO	6.32
C_CPKR	**95.37**
PPM_OpM	16.75
PPM_AEOpM	13.97
PPM_CIAM	13.52
PPM_AEOcM	14.11
PPM_AETM	5.94
PBR_DFS	20.73
PBR_APE	15.73
PBR_EcS	18.71
PBR_PBSD	19.39
AFI_AUF	6.11
AFI_ACD	4.29
AFI_LB	7.44
AFI_RSD	18.34

Another way in which prompts may be considered inadequate, thus preventing evaluators from making consistent judgements, is if they are redundant. In order to identify prompts which may be evaluating the same information as another, the correlation between prompt scores can be examined. A correlation matrix between all prompt scores in the data set was generated to find any instances of potential redundancy. Table 6 shows the correlation values between the nineteen prompts in all five sections of the application. As can be seen from Table 6, there are three noticeably high between-prompt correlation values. Two variables from the Program Description and Need For Service section of the application, PDNS_PAAM and PDNS_APAIA, have a correlation value of 0.75. Two variables from the Program Budget Request Section, PBR_APE and PBR_EcS, have a correlation value of 0.76. Finally, two variables from different application sections, PBR_PBSD and AFI_RSD, have a correlation value of 0.78. To definitively identify the outliers in correlation values, a Rosner's Generalized Extreme Studentized Deviate (ESD) Test can be conducted [5]. The pairs of prompts corresponding to the outliers formally identified by this test should be investigated further to determine potential causes of extreme correlation values.

Table 6: Between-Prompt-Correlation for All Evaluation Prompts

	PDNS	**PDNS**	PDNS	PDNS	PDNS	C	PPM	PPM	PPM	PPM	PPM	PBR	PBR	**PBR**	PBR	AFI	AFI	AFI	**AFI**
	PAAM	**APAIA**	TP	PNIA	SCO	CPKR	OpM	AEOpM	CIAM	AEOcM	AETM	DFS	APE	**EcS**	PBSD	AUF	ACD	LB	**RSD**
PDNS_PAAM		**0.75**	0.42	0.47	0.31	0.11	0.37	0.40	0.59	0.34	0.26	0.03	0.27	0.28	0.22	0.24	0.25	0.18	0.15
PDNS_APAIA			0.29	0.53	0.25	0.05	0.41	0.48	0.59	0.47	0.18	0.04	0.22	0.13	0.11	0.13	0.23	0.17	0.10
PDNS_TP				0.56	0.29	0.02	0.37	0.28	0.29	0.42	0.41	0.16	0.33	0.36	0.24	0.35	0.32	0.16	0.25
PDNS_PNIA					0.37	0.01	0.50	0.35	0.47	0.47	0.40	0.19	0.34	0.35	0.30	0.22	0.33	0.20	0.31
PDNS_SCO						0.08	0.29	0.34	0.29	0.23	0.24	0.02	0.17	0.26	0.19	0.19	0.22	0.18	0.15
C_CPKR							-0.08	0.04	0.12	0.01	0.04	0.21	0.13	0.10	0.21	0.02	0.03	-0.00	0.13
PPM_OpM								0.58	0.64	0.63	0.44	0.15	0.35	0.46	0.38	0.35	0.38	0.25	0.38
PPM_AEOpM									0.51	0.62	0.37	0.18	0.42	0.41	0.44	0.38	0.36	0.26	0.45
PPM_CIAM										0.54	0.33	0.18	0.32	0.35	0.29	0.26	0.39	0.20	0.31
PPM_AEOcM											0.46	0.10	0.47	0.50	0.30	0.33	0.31	0.16	0.36
PPM_AETM												0.12	0.44	0.52	0.29	0.43	0.33	0.27	0.34
PBR_DFS													0.27	0.32	0.30	0.18	0.27	0.26	0.26
PBR_APE														**0.76**	0.51	0.43	0.21	0.28	0.47
PBR_EcS															0.48	0.49	0.28	0.30	0.46
PBR_PBSD																0.43	0.17	0.23	**0.78**
AFI_AUF																	0.29	0.24	0.55
AFI_ACD																		0.16	0.29
AFI_LB																			0.31
AFI_RSD																			

6.2 Methods

The final research question outlined by the client requires the identification of inconsistencies in evaluator judgement to identify any insufficient prompts. In the previous section, two quantities were generated as indicators of potential inconsistencies in prompt scoring. Outlying values of these two metrics can be considered indications of inconsistencies in evaluator judgement. Rosner's Generalized Extreme Studentized Deviate Test will definitively identify outliers in modified variance and correlation values.

Rosner's Generalized ESD Test is used to detect one or more outliers in an approximately normal univariate dataset [5]. The Generalized ESD test is more flexible than other methods for determining outliers in the fact that it only requires an upper bound for the suspected number of outliers rather than an exact amount. For this specified upper bound k, the Generalized ESD Test essentially performs k separate tests: the first for a single outlier, the second for two outliers, up to a kth test for k outliers. The test operates under the null hypothesis that there are no outliers in the dataset, with the alternative hypothesis that there are up to k outliers in the dataset. Rosner indicates in his own simulation studies that the test is very accurate for samples that have 25 or more observations and reasonably accurate for samples that have 15 or more observations [5].

The R package *EnvStats* contains the function *rosnerTest* which performs the Rosner's Generalized ESD Test [4]. To identify potential inconsistencies in prompt scoring, the *rosnerTest* function in R will be used to perform the Rosner's Generalized ESD Test on the modified variance values and the between-prompt-correlation values. The prompts corresponding to the outliers identified by this process will be considered further to determine potential causes of the inconsistencies in judgement indicated by their abnormal values of either of these two metrics.

6.3 Analysis

In Section 6.1, two quantities were generated as indicators of potential inconsistencies in prompt scoring: modified variance and correlation. The modified variance scores for each prompt are displayed in Table 5. The second quantity examined as an indicator of potential inconsistencies is between-prompt correlation. These correlation values can indicate areas of potential redundancy in scoring or prompt content. The between-prompt correlation values are shown in Table 6. Outlying values of these two metrics can be considered indications of inconsistencies in evaluator judgement. While potential outliers were already noted in Section 6.1, the Rosner's Generalized Extreme Studentized Deviate Test, as outlined in Section 6.2, will definitively identify outliers in modified variance and correlation values.

The Rosner's Generalized ESD Test operates under the assumption of normality. This assumption can be checked via a QQ plot of the data. Figure 8 contains the Normal QQ Plot for the modified variance scores obtained in Sec-

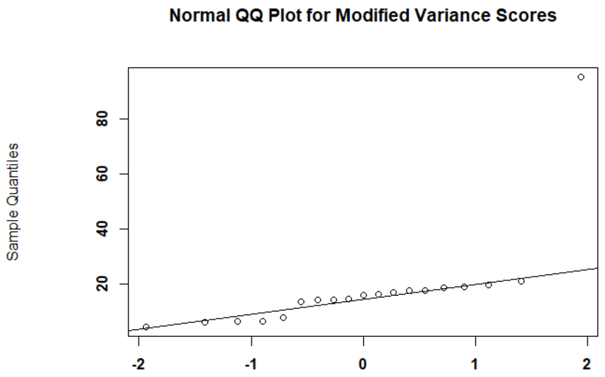


Figure 8: Normal QQ Plot for Modified Variance Scores for Each Prompt

tion 6.1. As seen in Figure 8, aside from one data point, the data falls along the QQ line, and can therefore be reasonably approximated by a normal distribution. Figure 9 contains the Normal QQ Plot for the between-prompt correlation scores obtained in Section 6.1. As can be seen from Figure 9, aside from slight deviations at the tails of the QQ Plot, the data falls along the QQ line, and can therefore be reasonably approximated by a normal distribution. Since both of these variables meet the normality assumption, it is safe to conduct the Rosner's Generalized ESD Test.

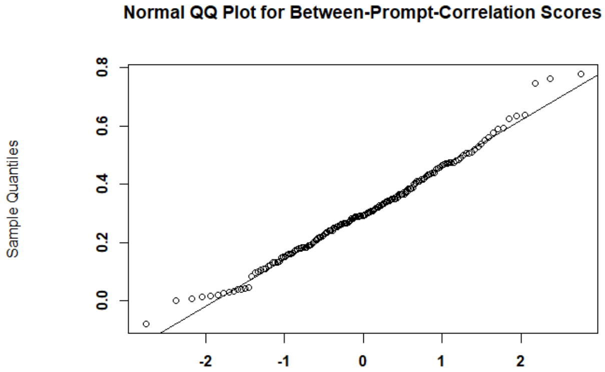


Figure 9: Normal QQ Plot for Between-Prompt-Correlation Scores for Each Pair of Prompts

For the modified variance scores, one suspected outlier was identified in Section 6.1. Since the Rosner's Generalized ESD Test tests for up to k outliers where k is less than 10, the test is run with $k=4$ outliers. This cushion is to make sure there are no additional outliers, besides the one suspected. Using the *rosnertest* function in the R package *EnvStats*, the Rosner's Generalized ESD Test was performed for the distribution of modified variance scores for up to four outliers. Figure A.4 in the Appendix contains the R output results of this testing procedure. The test identified one outlier, corresponding to the C_CPKR variable, at a significance level of 0.05.

For the between-prompt-correlation scores, three suspected outliers were identified in Section 6.1. Using the *rosnertest* function in the R package *EnvStats*, the Rosner's Generalized ESD Test was performed for the distribution of between-prompt correlation scores for up to 10 outliers. Figure A.5 in the Appendix contains the R output results of this testing procedure. The test identified three outliers corresponding to the correlations between the variables PDNS_PAAM and PDNS_APAIA, PBR_APE and PBR_EcS, and PBR_PDSD and AFI_RSD.

6.4 Results

In Section 6.3, a single outlier was identified from the distribution of the modified variance scores calculated as a measure of score consistency for each prompt. The variable with the outlying modified variance score is C_CPKR, which corresponds to the scores given on the only prompt in the Collaboration Section of the Application. Not every applicant program has an element of collaboration in their Community Impact Proposal. Therefore this prompt is only meant to be addressed by evaluators if applicants have included an element of collaboration in their application. The collaboration prompt is presented to evaluators as the following question: "Collaboration (if applicable): How confident are you that each collaborator involved plays a key role in the program that leads to participant success?" Evaluators may select from the answers "Very Confident," "Confident," "Moderately Confident," "Slightly Confident," "Not at All Confident," or leave the answer blank, to indicate that the element of collaboration is not relevant to the application being scored. A selection of one of the answers corresponds to a score of one-to-five, while leaving the answer blank corresponds to a score of zero. The modified variance score for this prompt is far higher than any other application prompt. This indicates that there is significant inconsistency in the way evaluators are judging this prompt. A closer look at this variable reveals that on many applications, one evaluator would give a score of five for collaboration, while the other evaluators would give the same application a score of zero for collaboration, indicating that collaboration was not an applicable element of this application. These vast discrepancies are an indication of a prompt that is not sufficiently adequate for evaluators to make consistent judgements.

Furthermore, three outliers were identified from the distribution of between-prompt correlation scores in Section 6.3. The first pair of variables with an outlying high correlation score is PDNS_PAAM and PDNS_APAIA. The variable PDNS_PAAM corresponds to the scores given on the first prompt in the Program Description and Need for Service section of the application, which is meant to address Program Alignment with Agency Mission. The prompt is presented to evaluators as the following question: "How well does the program proposed align with the overall agency's mission?" The variable PDNS_APAIA corresponds to the scores given on the second prompt in the Program Description and Need for Service section of the application, which is meant to address whether Activities Purpose Align with Impact Area. The prompt is presented to evaluators as the following question: "How well do the activities and the purpose of the program align with the impact area chosen?" The scores of these two prompts were found to be significantly highly correlated in Section 6.3. This could simply be because the subjects of these prompts are related, it could be a coincidence, or evaluators could potentially be unclear about the difference between these two questions. After all, program alignment with agency mission and program activities and purpose alignment with impact area are concepts that are highly intertwined. Evaluators may be confused about which parts of the application to look at for each prompt. Regardless of the cause, an extremely high correlation between two prompts such as these is an indication that the same information may be being captured twice in the scoring process.

Additionally, the variables PBR_APE and PBR_EcS have an exceptionally high between-prompt correlation score that was found to be an outlier in Section 6.3. The variable PBR_APE corresponds to the scores given on the second prompt in the Program Budget Request section of the application, which is meant to address Appropriate Program Expenses. The prompt is presented to evaluators as the following question: "Are program expenses appropriate considering program description?" The variable PBR_EcS corresponds to the scores given on the third prompt in the Program Budget Request Section of the application, which is meant to address Expenses versus Number Served. The prompt is presented to evaluators as the following question: "Are program expenses appropriate for number of persons served?" The scores of these two prompts were found to be significantly highly correlated in Section 6.3. This could be because applications that have well-developed budget requests simply do well on both questions, or it could be coincidental. It could be that there is confusion between the two prompts considering they both address the appropriateness of program expenses. Evaluators may be missing the distinction between program description and number served in terms of budget appropriateness. Again, regardless of what the cause of the correlation is, the final application score weights could be misleading if two prompts are capturing the same information.

Finally, the last outlier in between-prompt-correlation scores is the correlation between the variables PBR_PDSD and AFI_RSD. The variable PBR_PDSD corresponds to the scores given on the third prompt in the Program Budget Request section of the application, which is meant to address the Program Budget Surplus and Deficit. The prompt is presented to evaluators as the following question: "Comparing prior year actuals to the proposed budget, has the organization responded to deficits or significant surpluses appropriately?" The variable AFI_RSD corresponds to the scores given on the fourth prompt in the Agency Financial Information section of the application, which is meant to address Response to Surplus and Deficit. The prompt is presented to evaluators as the following question: “Comparing prior year Completed Financial Statements to the Current Year Budget, has the organization responded to deficits or significant surpluses appropriately?” Clearly, both questions are worded very similarly with regard to surplus and deficits. Additionally, the documents which are meant to be compared for each question sound quite similar as well. It is possible that evaluators struggle to distinguish between the “proposed budget” and the “Current Year Budget” or even “prior year actuals” and “prior year Completed Financial Statements”. While it is also possible that the correlation between the scores on these two prompts is a coincidence, or simply the result of a better financial situation of the applicant, it seems much more likely that evaluators find it difficult to distinguish the content addressed in these two prompts.

7 Conclusion

The Community Impact Program of United Way of Northeast Georgia is interested in improving the Community Impact Grant Application Evaluation Process. The primary goal is to ensure that the application process is designed in a way that gives every applicant the ability to present the best application possible, and every evaluator the ability to score each question as accurately as possible. The Director of Community Impact, Mark Madison, proposed five research questions to guide the statistical analysis that attempts to identify any potential areas of improvement in the application and subsequent evaluation process. The culmination of this research results in the following recommendations.

The client is interested in whether or not applicant responses were consistent enough to imply a common understanding of application prompts. In Section 2, it was found that, with the exception of one confirmed outlier, the average sentiment scores for each application follow an approximately normal distribution. This indicates reasonably consistent sentiment among applicants and therefore consistent understanding of the application prompts. The confirmed outlier has the lowest average sentiment score among all programs, but has one of the highest average total evaluator scores among all programs. This indicates that this program understood the application prompts thoroughly, even though they had the only statistical inconsistency in sentiment among programs. However, their exceptional score could also be due to evaluator bias. This program is arguably the most well-known program in this sample, so it is possible that evaluators felt an inherent trust of this applicant despite the difference in application sentiment. While this application’s average sentiment score was an anomaly, it is not an indication of lack of understanding of the application prompts. Therefore, the applicant responses were consistent enough to indicate general understanding of application prompts. The only recommendation to the client is to consider omitting application names when evaluators are grading in order to limit inherent biases.

The client is also interested in which prompts, if any, were consistently scored very poorly or very well. The results described in Section 3 identified one prompt that scored significantly lower than average, and four prompts that scored significantly higher than average. The prompt that was identified as scoring significantly lower than average is the only prompt from the Collaboration section of the application. As previously mentioned, the Collaboration section of the application is an optional component, as not all applicants have an element of collaboration in their proposals. Though the collaboration prompt is one of the prompts scored on a one-to-five point scale, if an application does not have a collaboration element, a score of zero is assigned. Many applicant programs did not have a collaboration element in their applications, therefore it is not surprising that this prompt would have a significantly lower score relative to the other prompts.

Of the four prompts that were identified as receiving scores significantly higher than average, three are from the first application section, Program Description and Need for Service. The significant prompts from this section address Program Alignment with Agency Mission, Activities Purpose Align With Impact Area, and Target Population. This indicates that the content meant to be addressed by this application section is most likely very clear to both applicants and evaluators, as applicants were able to do a good job explaining their programs. However, as mentioned previously, the scores available in this dataset came from applications that made it to the second round of evaluations already. Therefore, this trend could also indicate that Program Description and Need for Service was a more important determining factor in whether applications would move to the next round, relative to other application sections. Finally, the last prompt identified as being scored significantly higher than average is the third prompt in the Program Performance Measures section of the application, which addresses whether performance measures Contribute to Impact Area Measures. Similar to the other prompts, this indicates that the content meant to be addressed by this prompt was very clear to both applicants and evaluators, as applicants gave clear descriptions of how their performance measures contribute to impact area measures. Again, this could also indicate that this particular component of the application was a key determining factor in whether applications would proceed to this round of evaluations. Therefore the only recommendation to the client regarding these prompts scored significantly higher than others is to give special consideration in the future to what elements of the application are of the highest priority when advancing applications to the evaluation round and which prompts are given the highest weight in the final score, to assure that their priorities are appropriately reflected.

Additionally, the client is interested in which evaluators, if any, tended to give consistently low or high scores relative to the others. The results described in Section 4 indicate that the following seven evaluators tended to give significantly higher than average scores: Evaluator 7, Evaluator 12, Evaluator 22, Evaluator 25, Evaluator 29, Evaluator 36, and Evaluator 40. Likewise, the following three evaluators tended to give significantly low scores relative to the others: Evaluator 17, Evaluator 19, and Evaluator 30. The significance of these results is potentially weakened by the fact that each evaluator did not evaluate all applications, and in fact did not evaluate the same number of applications. Therefore, it could be the case that the evaluators tending to give higher scores, were assigned to genuinely better applications, and vice versa for those tending to give lower scores. However, the significance level for these tests was set very low, to ensure that only evaluators with highly significantly different score distributions were identified by the T-tests. The importance of these significant evaluators is whether they have undue influence on the applications they score. Therefore, the impact of evaluators on the applications they graded is considered further, to identify if there is cause for concern.

Accordingly, the client is interested in whether the number of evaluators assigned to each application is sufficient to normalize the possible consistently low or high scoring evaluators. The results described in Section 5 indicate that for most of the applications, the number of evaluators assigned to each application was not sufficient to limit the impact of evaluators to a tolerable level. Only applications that received six evaluations were consistently impacted by individual evaluators by less than five points in their final average score. The only application category in which applicants received six evaluations is the Early Childhood Success category. Furthermore, every application in this category received exactly six evaluations, therefore it is the only application category in which applications can be fairly compared. To ensure that the impact of consistently high or low scoring evaluators is contained to a reasonable level, it is recommended that every applicant receive at least six evaluations. Additionally, to ensure that all applicants receive fair consideration for funding, all applicants, regardless of application category should receive the same number of evaluations. At the very least, applications within the same category should receive the same number of evaluations, if the categories are considered for funding separately. Without this symmetry, applications are not given fair and equal consideration.

The client's final inquiry is whether the evaluation prompts were adequate for evaluators to make consistent judgements. The results described in Section 6 indicate four instances where inadequacies may exist. The first is with the evaluation prompt addressing a collaboration element in an application, if it is present. The significantly high variability in evaluator scoring on this prompt suggests a need for clarity. It is recommended that the client review the Collaboration prompt, in an attempt to provide further clarity to evaluators that the question should only be answered if applicants specifically address an element of collaboration in their application. Additionally, the client may consider redesigning the application so that the Collaboration section is a permanent feature of the application document, and applicants write in "Not Applicable" if the section does not apply to them. This could prevent confusion for the evaluators.

The additional instances of potential inadequacy are indicated by the significantly high correlations of scores between the following pairs of prompts:

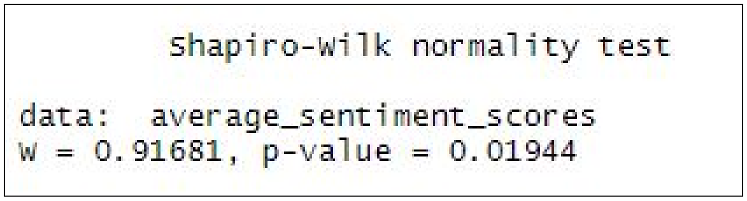
the first and second prompts in the Program Description and Need for Service section of the application, which address Program Alignment with Agency Mission and Activities Purpose Align with Impact Area, respectively; the second and third prompts in the Program Budget Request section of the application, which address Appropriate Program Expenses and Expenses versus Number Served, respectively; and finally the third prompt in the Program Budget Request section of the application, and the fourth prompt in the Agency Financial Information section of the application, which address Program Budget Surplus and Deficit and Response to Surplus and Deficit, respectively. As noted in Section 6, these exceptionally high correlations could be due to an overlap in content addressed by the two prompts in each pair. If evaluators are unable to distinguish between the content addressed by two separate prompts, this results in the content being evaluated twice, which in turn leads to that content receiving a higher weight in the final score than intended in the original design of the evaluation. Therefore, it is recommended that the client review these three pairs of prompts and evaluate whether additional distinction between the content intended to be addressed by each prompt in each respective pair can be provided. Should additional distinction prove impossible to provide, the client may consider addressing the weighting of scores corresponding to these prompts in the final score. Along these lines, the client may also consider scoring all prompts on a single scale, providing the same number of options for evaluators to select from, and then scaling these scores based on the importance of the content to the delegation of funding. This would prevent differences in the number of options from influencing evaluator decisions.

Overall, United Way of Northeast Georgia has a strong Community Impact Grant application and evaluation process. This process could be strengthened further with the recommendations offered to the client above, in order to ensure that the application process is aligned with the Director's Goal to give every applicant the ability to present the best application possible, and every evaluator the ability to score each question as accurately as possible.

8 Appendix

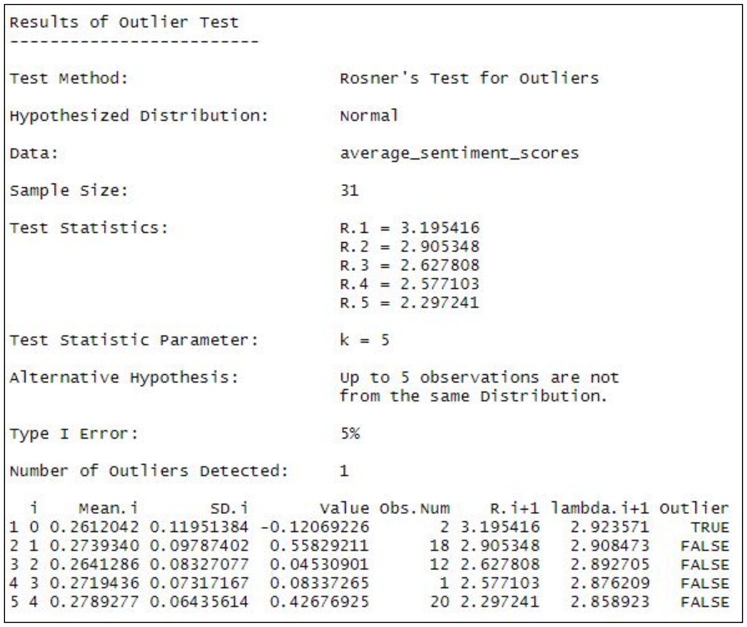
Table A.1: Variable Names and Prompt Definitions

Variable Name	Prompt Definition
PDNS_PAAM	Program Description and Need for Service: Program Alignment with Agency Mission
PDNS_APAIA	Program Description and Need for Service: Activities and Purpose Align With Impact Area
PDNS_TP	Program Description and Need for Service: Target Population
PDNS_PNIA	Program Description and Need for Service: Program Need In Impact Area
PDNS_SCO	Program Description and Need for Service: Similar County Outcomes
C_CPKR	Collaboration: Collaborator Plays Key Role
PPM_OpM	Program Performance Measures: Output Measure
PPM_AEOpM	Program Performance Measures: Achieves Estimated Output Measures
PPM_CIAM	Program Performance Measures: Contribute to Impact Area Measures
PPM_AEOcM	Program Performance Measures: Achieves Estimated Outcomes Measure
PPM_AETM	Program Performance Measures: Ability to Effectively Track Measures
PBR_DFS	Program Budget Request: Diverse Funding Sources
PBR_APE	Program Budget Request: Appropriate Program Expenses
PBR_EcS	Program Budget Request: Expenses vs. Served
PBR_PBSD	Program Budget Request: Program Budget Surplus / Deficit
AFI_AUF	Agency Financial Information: Ability to Utilize Funds
AFI_ACD	Agency Financial Information: Agency Capacity to Deliver
AFI_LB	Agency Financial Information: Local Board
AFI_RSD	Agency Financial Information: Response to Surplus / Deficit



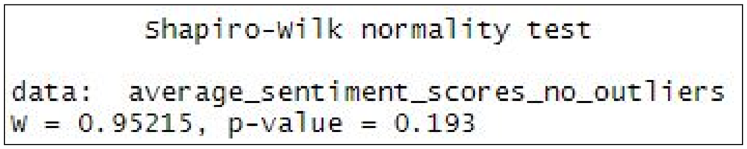
```
        Shapiro-Wilk normality test

data:  average_sentiment_scores
W = 0.91681, p-value = 0.01944
```

Figure A.1: R Output for Shapiro-Wilk Test for Normality on All Average Sentiment Scores



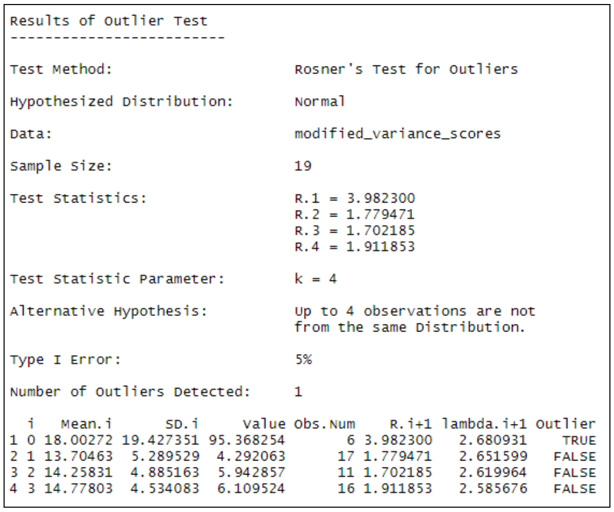
```
Results of Outlier Test
-------------------------

Test Method:                     Rosner's Test for Outliers

Hypothesized Distribution:       Normal

Data:                            average_sentiment_scores

Sample Size:                     31

Test Statistics:                 R.1 = 3.195416
                                 R.2 = 2.905348
                                 R.3 = 2.627808
                                 R.4 = 2.577103
                                 R.5 = 2.297241

Test Statistic Parameter:        k = 5

Alternative Hypothesis:          Up to 5 observations are not
                                 from the same Distribution.

Type I Error:                    5%

Number of Outliers Detected:     1

  i    Mean.i       SD.i       Value Obs.Num    R.i+1 lambda.i+1 Outlier
1 0 0.2612042 0.11951384 -0.12069226       2 3.195416   2.923571    TRUE
2 1 0.2739340 0.09787402  0.55829211      18 2.905348   2.908473   FALSE
3 2 0.2641286 0.08327077  0.04530901      12 2.627808   2.892705   FALSE
4 3 0.2719436 0.07317167  0.08337265       1 2.577103   2.876209   FALSE
5 4 0.2789277 0.06435614  0.42676925      20 2.297241   2.858923   FALSE
```

Figure A.2: R Output for Rosner's ESD Test for Outliers on All Average Sentiment Scores



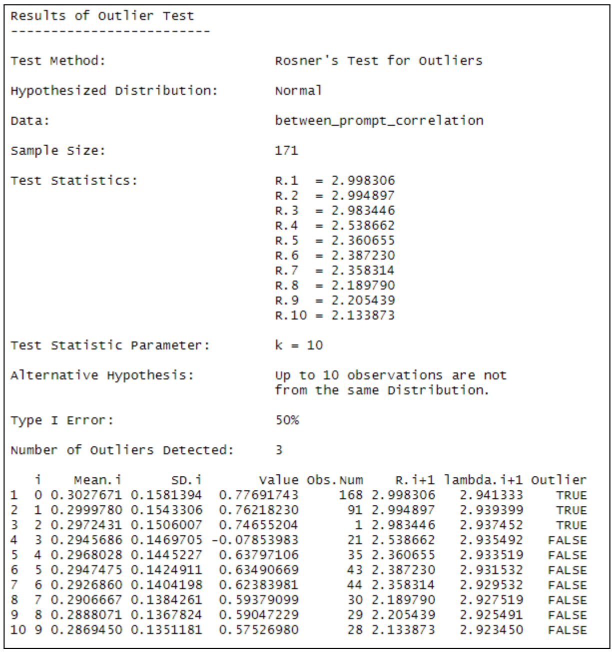
```
        Shapiro-Wilk normality test

data:  average_sentiment_scores_no_outliers
W = 0.95215, p-value = 0.193
```

Figure A.3: R Output for Shapiro-Wilk Test for Normality on Average Sentiment Scores without the Identified Outlier



```
Results of Outlier Test
-------------------------

Test Method:                     Rosner's Test for Outliers

Hypothesized Distribution:       Normal

Data:                            modified_variance_scores

Sample Size:                     19

Test Statistics:                 R.1 = 3.982300
                                 R.2 = 1.779471
                                 R.3 = 1.702185
                                 R.4 = 1.911853

Test Statistic Parameter:        k = 4

Alternative Hypothesis:          Up to 4 observations are not
                                 from the same Distribution.

Type I Error:                    5%

Number of Outliers Detected:     1

  i   Mean.i      SD.i     Value Obs.Num    R.i+1 lambda.i+1 Outlier
1 0 18.00272 19.427351 95.368254       6 3.982300   2.680931    TRUE
2 1 13.70463  5.289529  4.292063      17 1.779471   2.651599   FALSE
3 2 14.25831  4.885163  5.942857      11 1.702185   2.619964   FALSE
4 3 14.77803  4.534083  6.109524      16 1.911853   2.585676   FALSE
```

Figure A.4: R Output for Rosner's ESD Test for Outliers on Modified Variance Scores



```
Results of Outlier Test
-------------------------

Test Method:                     Rosner's Test for Outliers

Hypothesized Distribution:       Normal

Data:                            between_prompt_correlation

Sample Size:                     171

Test Statistics:                 R.1  = 2.998306
                                 R.2  = 2.994897
                                 R.3  = 2.983446
                                 R.4  = 2.538662
                                 R.5  = 2.360655
                                 R.6  = 2.387230
                                 R.7  = 2.358314
                                 R.8  = 2.189790
                                 R.9  = 2.205439
                                 R.10 = 2.133873

Test Statistic Parameter:        k = 10

Alternative Hypothesis:          Up to 10 observations are not
                                 from the same Distribution.

Type I Error:                    50%

Number of Outliers Detected:     3

   i    Mean.i      SD.i       Value Obs.Num    R.i+1 lambda.i+1 Outlier
1  0 0.3027671 0.1581394  0.77691743     168 2.998306   2.941333    TRUE
2  1 0.2999780 0.1543306  0.76218230      91 2.994897   2.939399    TRUE
3  2 0.2972431 0.1506007  0.74655204       1 2.983446   2.937452    TRUE
4  3 0.2945686 0.1469705 -0.07853983      21 2.538662   2.935492   FALSE
5  4 0.2968028 0.1445227  0.63797106      35 2.360655   2.933519   FALSE
6  5 0.2947475 0.1424911  0.63490669      43 2.387230   2.931532   FALSE
7  6 0.2926860 0.1404198  0.62383981      44 2.358314   2.929532   FALSE
8  7 0.2906667 0.1384261  0.59379099      30 2.189790   2.927519   FALSE
9  8 0.2888071 0.1367824  0.59047229      29 2.205439   2.925491   FALSE
10 9 0.2869450 0.1351181  0.57526980      28 2.133873   2.923450   FALSE
```

Figure A.5: R Output for Rosner's ESD Test for Outliers on Between-Prompt-Correlation Scores

9 References

[1] Charfuelan, Marcela and Schroder, Marc. Language Technology Laboratory, DFKI GmbH, Correlation Analysis of Sentiment Analysis Scores and Acoustic Features in Audiobook Narratives,www.dfki.de/fileadmin/user_upload/import/6463_correlationAnalysis.pdf.

[2] de Winter, J. F.C. and Dodou, D. (2010) ”Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon (Addendum added October 2012),” Practical Assessment, Research, and Evaluation: Vol. 15 , Article 11.

[3] Frost, Jim. “5 Ways to Find Outliers in Your Data.” Statistics By Jim, 7 July 2020, statisticsbyjim.com/basics/outliers/.

[4] Millard SP (2013). EnvStats: An R Package for Environmental Statistics. Springer, NewYork. ISBN 978-1-4614-8455-4, https://www.springer.com.

[5] ”NIST/SEMATECH e-Handbook of Statistical Methods”, http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h3.htm

[6] Rinker, T. W. (2019). sentimentr: Calculate Text Polarity Sentiment version 2.7.1. http://github.com/trinker/sentimentr

[7] Silge J, Robinson D (2016). “tidytext: Text Mining and Analysis Using Tidy DataPrinciples in R.” JOSS, 1(3). https://doi.org/10.21105/joss.00037

[8] Thode, Henry C. Testing for Normality. Dekker, 2002. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3693611/: :text=The%20main%20tests%20for %20the,test%20(7)%2C%20and%20the

[9] United Way of Northeast Georgia, 15 November 2020, https://www.unitedwaynega.org/