Summary of MAADCAP Statistics

Melissa Wong

March 3, 2020

1 Task Description

The MAADCAP task will crowd-source input from subject matter experts (SMEs) on the level of analytics required to detect specific cyber techniques in different kinds of data. The objective is to identify those areas where there is consensus across SMEs vs those areas where there is not consensus and thus suitable for further investment in additional research. Note that consensus is different from consistency, and the following definitions apply to the discussion in this paper:

- Consensus - Raters have exact agreement on how to apply various levels of the scoring rubric to the observed data.
  - Example statistics are Cohen's Kappa and Tastle's Consensus.
- Consistency - Raters do not have a common understanding of the rating scale, but are consistent in application of his/her own defintion of the rating scale.
  - Example statistics are Pearson's correlation and Spearman's correlation.

In other words, if consensus among the raters is high then one can infer the scoring rubric is an appropriate method for identifying areas in the defensive cyber space that require advanced analytics.

Table 1: Random Ratings

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Rater 1	4	3	1	5	5	2	2	5	5	1
Rater 2	4	4	4	2	2	5	4	1	1	3
Rater 3	4	5	3	5	1	1	1	3	2	5
Rater 4	5	1	2	3	1	3	3	4	2	1
Rater 5	2	3	4	1	2	4	2	3	5	4

2 Challenges

Non-response (both unit and item) occurred during MAADCAP Phase I. Given the large number of items (~3000) to be scored, not all SMEs will be asked to score all items during Phase II. Therefore, there will intentionally be items with missing data in addition to likely unintentional non-response. Thus, the statistics chosen must be able to handle missing data.

3 Recommended Statistics for Phase II

3.1 Krippendorf's alpha

Krippendorf's alpha (A. Hayes 2007) was initially developed for content analysis applications. Content analysis depends on the judgment of human observers, and so there is a need to measure reliability among the human observers. High reliability implies that the data from the human observers are exchangeable with the data from another set of observers (as opposed to being largely the result of individual idiosyncracies). Krippendorf's alpha has several advantages including that it can handle varying numbers of observers, different sample sizes and the presence or absence of missing data.

For Phase II, before assessing consensus on invidual items, Krippendorf's alpha should first be applied to the entire data set to determine if the data differ significantly from random data (K's $\alpha = 0$) or if there is systematic disagreement (K's $\alpha < 0$).

Table 1 illustrates why first assessing overall agreement is important. The table contains randomly generated scores. Question 1 appears to have a high degree of consensus/agreement among the raters, but that is misleading since we know this data was randomly generated. Krippendorf's alpha $= -0.14$ for Table 1 which indicates the data does not differ significantly from random data, and therefore further analysis of individual items is not recommended.

Similarly, if Krippendorf's alpha showed systematic disagreement across the data, that indicates an issue which needs to be resolved (e.g., inconsistent training on the scoring rubric) before attempting to do further analysis.

3.2 Tastle's Agreement

Assuming Krippendorf's alpha indicates the data in aggregate is not nearly random and there is not significant disagreement, the next step is to assess consensus on individual items. Phase I used Tastle's Consensus measure (Tastle 2009), but it has some known issues; namely that it is not monotonic with respect to the majority and is undefined at some edge cases (G. Beliakov 2014). A recommended alternative to Tastle's Consensus about the mean is Tastle's Agreement about the median since it does not have the same issues.

Tastle's Agreement about τ is defined as follows:

$$Agr(X,\tau) = 1 + \sum_{i=1}^{n} p_i log_2\Big(1 - \frac{|X_i - \tau|}{2d_x}\Big) \tag{1}$$

However, the range for Equation 1 is not [0,1] when $\tau =$ median. A scaled Agreement measure may be desirable for MAADCAP applications.

3.2.1 Scaled Agreement With Median

When there are an even number of scores which are evenly distributed, then the median (τ) is halfway between X_{min} and X_{max}):

$$\begin{aligned}
Agr(X,\tau) &= 1 + 0.5log_2\Big(1 - \frac{|X_{min} - \tau|}{2d_x}\Big) + 0.5log_2\Big(1 - \frac{|X_{max} - \tau|}{2d_x}\Big) \\
&= 1 + 0.5log_2\Big(1 - \frac{|\frac{d_x}{2}|}{2d_x}\Big) + 0.5log_2\Big(1 - \frac{|\frac{d_x}{2}|}{2d_x}\Big) \\
&= 1 + log_2\Big(\frac{3}{4}\Big) \\
&= log_2\Big(\frac{3}{2}\Big) \\
&= 0.58
\end{aligned}$$

When there are and odd number of scores with n scores at one end of the scale, and $n+1$ scores at the other end of the scale, then median (τ) is now the category with $n+1$ scores. In the example below, the $n+1$ scores are at X_{max} however we would get the same result for $n+1$ scores at X_{min}:

$$\begin{aligned}
Agr(X,\tau) &= 1 + \frac{n}{2n+1}log_2\Big(1 - \frac{|X_{min} - X_{max}|}{2d_x}\Big) + \frac{n+1}{2n+1}log_2\Big(1 - \frac{|X_{max} - X_{max}|}{2d_x}\Big) \\
&= 1 + \frac{n}{2n+1}log_2\Big(1 - \frac{d_x}{2d_x}\Big) + \frac{n+1}{2n+1}log_2\Big(1 - 0\Big) \\
&= 1 + \frac{n+1}{2n+1}log_2\Big(1 - \frac{1}{2}\Big) \\
&= 1 - \frac{n+1}{2n+1} \\
&= 0.5 \text{ as } n \to \infty
\end{aligned}$$

Agreement is maximum when all scores are in one category ($\tau = X_c =$ median):

$$\begin{aligned}
Agr(X,\tau) &= 1 + log_2\Big(1 - \frac{|X_c - \tau|}{2d_x}\Big) \\
&= 1 + log_2\Big(1 - 0\Big) \\
&= 1
\end{aligned}$$

Therefore the range of Agreement around the median is $[0.5, 1]$. Equation 1 when τ=median can be scaled to give a value between [0,1]. Since this is just a linear transformation of the equation, it will not affect the other properties of the equation (e.g. monotonicity with respect to majority).

$$\begin{aligned} Agr_{scaled}(X, \tau_{median}) &= \frac{Agr(X, \tau_{median}) - 0.5}{1 - 0.5} \\ &= 2Agr(X, \tau_{median}) - 1 \end{aligned}$$

3.2.2 Agreement With Other Values

Agreement can be calculated around other values of τ, such as the minimum score, maximum score or a specific category.

When τ = min, Agreement is a minimum when 1 score is in the minimum category and the remaining $n - 1$ scores are in the maximum category.

$$\begin{aligned} Agr(X, \tau_{min}) &= 1 + \frac{1}{n} log_2\left(1 - \frac{|X_{min} - \tau_{min}|}{2d_x}\right) + \frac{n-1}{n} log_2\left(1 - \frac{|X_{max} - \tau_{min}|}{2d_x}\right) \\ &= 1 + \frac{1}{n} log_2\left(1 - 0\right) + \frac{n-1}{n} log_2\left(1 - \frac{dx}{2d_x}\right) \\ &= 1 + \frac{n-1}{n} log_2\left(\frac{1}{2}\right) \\ &= 1 - \frac{n-1}{n} \\ &= \frac{1}{n} \\ &= 0 \text{ as } n \to \infty \end{aligned}$$

Maximum Agreement occurs when all scores are in the same category. Therefore the range of Agreement with the minimum is $[0, 1]$ and no transformation is needed. Similar results are obtained if Agreement is calculated with the maximum.

3.3 Confidence Intervals

The point estimates of Krippendorf's alpha and Tastle's Agreement give part of the picture. An example is easily illustrated with Tastle's Agreement with

$\tau = median$. In the table below, the proportion of scores in each category is the same, but the total number of raters differs. In all three cases, the Agreement score is 0.72 despite the fact that the total number of raters varies by factors of 10 to 100 between the three cases.

	1	2	3	4	5
small	1	0	1	0	3
large	10	0	10	0	30
huge	100	0	100	0	300

What is missing is some measure of uncertainty about the estimate of the Agreement score. As the number of raters increases, the uncertainty around the point estimate for the true population value of Agreement decreases. Calculating a confidence interval for the point estimate provides this additional information. The distribution of the Agreement statistic is unknown, so percentile and bias corrected bootstrap estimates of the confidence interval are calculated below (Efron and Narasimhan 2018). The bias correction has a significant effect for the small data set, and it becomes less significant as the sample size increases (i.e. the sampling distribution approaches a normal distribution).

	Percentile CI		BC CI	
	2.5%	98%	lower	upper
small	0.60	1.00	0.60	0.83
large	0.64	0.82	0.63	0.81
huge	0.68	0.75	0.68	0.75

3.4 Coefficient of Variation for Ordinal Categories

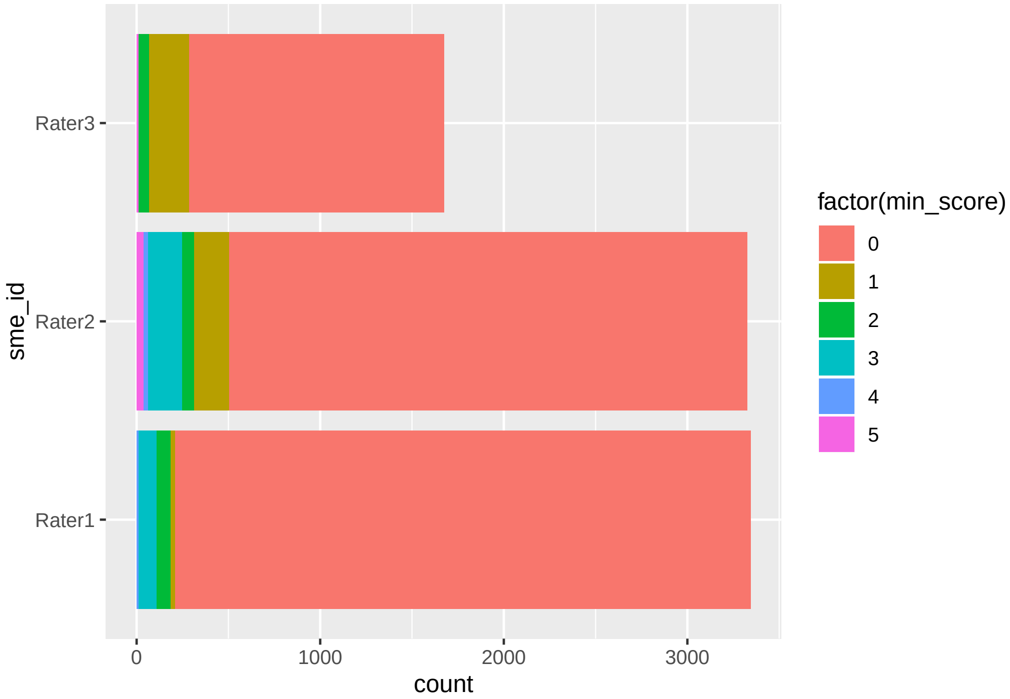
Another alternative to Tastle's Consensus is the Coefficient of Ordinal Variation (COV) (Kvalseth 1995a). This statistic is a measure of *dispersion*; $COV = 1$ when scores are evenly distributed between the two extreme categories and $COV = 0$ when all scores are in a single category. One strength of COV is that it is based on the cumulative probabilities of scores and does not depend on assigning numerical values to the categories which is arguably arbitrary (Kvalseth 1995b). Note that if we assign category values of $1, 2, ...k$ as is common practice, then the scaled version of Tastle's Agreement about the median and $1 - COV$ give similar results.

3.5 Rater Correlation

Since Phase II will assess Agreement with the minimum score (i.e., Tastle's Agreement where τ = minimum), it may be useful to first check Spearman's Correlation or Kendall's τ to compare the trend between an individual SME and the larger group which scored the same group of questions. Similar to using Krippendorf's α to identify systematic disagreement across the entire set of raters, Spearman's Correlation or Kendall's τ could be used to identify if one SME is systematically scoring differently from the rest of the group. That should be investigated further before proceeding with an Agreement measure about the minimum. For example, the individual may not understand the scoring rubric. Conversely, the individual may have unique experiences and insights which should not be discarded.

3.6 Rater Use of Scores

A simple bar chart as illustrated below can provide insight at a glance into how raters are using the scores. For example, in the chart below it is immediately apparent that a score of 0 is the predominant score used by all three raters. It is also easy to see that only one rater made full use of all scoring levels, and another rater scored only a subset of the questions. A simple graphic such as the one below may be easier for a user to interpret than a statistic in some cases.



4 Statistics for Nominal Data

Future versions of the tool may be applied to problems where the SMEs aren't using an ordinal rating scale, but rather are using nominal categories. In that case, Krippendorf's alpha is still applicable. However, Tastle's Agreement is only valid for ordinal/Likert scales, and thus an alternative measure of consensus is needed.

4.1 Mode and Mean Difference

Unlike ordinal data, the categories for nominal data are labels (e.g., eye color is blue, green, brown) and there is no scale or distance between the labels. The median does not exist for nominal data; instead, the mode should be reported. Note that unlike the median, the mode may not be unique.

Tastle's Agreement should not be used to assess the dispersion or variability in the scores since there is no distance measure. Instead, the mean difference (Wilcox 1973) can be used:

Table 2: Example - Nominal Data

	Q1	Q2	Q3
Rater 1	A	B	C
Rater 2	B	B	C
Rater 3	C	B	E
Rater 4	D	B	NA
Rater 5	E	B	E

$$MDA = 1 - \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} |f_i - f_j|}{N(K-1)}$$

where f_i = frequency of the ith category, N = number of cases and K = number of categories. The range for MDA is [0,1] where 0 occures when all cases fall in one category, and 1 occurs if and only if an identical number of cases occur in all categories.

	mda	mode
Q1.1	1	c("A", "B", "C", "D", "E")
Q2.1	0	B
Q3.1	0.25	c("C", "E")

Note that MDA is a measure of difference, not “agreement". Therefore, in the user interface it may be preferable to report $1 - MDA$ since that scale is consistent with Tastle's Agreement.

5 Statistics Considered, but Not Recommended

The following section summarizes other statistics considered and the reasons they are not recommended at this time for this application.

Metric	Purpose	Limitations
ICC (McGraw and Wong 1996)	Ten variations to handle different study designs (e.g., raters for each subject selected at random), measurement of interest (single rating or mean of several ratings) and type (agreement or consistency).	Strongly influenced by variance of the sample/population so may not be appropriate to compare different populations; uses list-wise deletion for missing elements so not suitable for fully-crossed designs with many missing elements; usual ANOVA assumptions apply (Normal distribution and homogeneity of variances).
Cohen's weighted κ	Measures observed agreement compared to agreement due to chance between two raters.	Value depends strongly on marginal distributions and prevalance of occurrence. Hard to interpret and compare.
Fleiss' kappa	Measures observed agreement compared to agreement due to chance for more than two raters.	Value depends strongly on prevalence of occurrence. Biased when there is missing data (A. Zapf 2016).
McNemar's Test	Test marginal homogeneity between two raters for two categories.	No generalization for multiple raters.
Stuart-Maxwell Test	Test marginal homogeneity between two raters for all categories simultaneously.	No generalization for multiple raters.
Cronbach's α	A measure of scale reliability (internal consistency), assumes questions are measuring one latent variable or dimension.	Not robust to missing data; sensitive to very large or very small number of test items.

6 Conclusion

For MAADCAP Phase II, we recommend using Krippendorf's alpha and Tastle's Agreement to analyze and prioritize the scoring data. Although there are some known issues with Tastle's Consensus, there is value in also using it for Phase II so that some comparisons can be made to the Phase I results. The guidelines for using these statistics are as follows:

1. First, confirm with Krippendorf's alpha that the rater scores in aggregate are not nearly random (i.e., the confidence interval includes zero) nor show systematic disagreement (i.e., the confidence interval includes negative numbers).

2. If the first condition is satistified, then the next step is to assess individual items. Using either the Agreement or Consensus statistic, each item can be placed into one of the following categories:

   a. Insufficient Data: Items do not have enough scores to draw conclusions (i.e., a confidence interval is wide or cannot be calculated at all); these items should be a high priority for additional SME scoring.

   b. Disagreement: Items with sufficient data to conclude there is systematic disagreement and further investigation is required (i.e., low Agreement/Consensus with a small confidence interval); these are items with significant differences in scores which may be difficult to reconcile.

   c. Mediation Candidate: Items with sufficient data to conclude that consensus building may be useful (i.e., moderate Agreement/Consensus with a small confidence interval).

   d. Agreement: Items with sufficient data to conclude there is systematic agreement (i.e., high Agreement/Consensus with a small confidence interval) and further investigation is not necessary; these items should be a low priority for additional SME scoring.

References

A. Hayes, K. Krippendorf. 2007. "Answering the Call for a Standard Reliability Measure for Coding Data." *Communication Methods and Measures* 1: 77–89.

A. Zapf, L. Morawietz, S. Castell. 2016. "Measuring Inter-Rater Relibability for Nominal Data-Which Coefficients and Confidence Intervals Are Appropriate." *BMC Medical Research Methodology* 16 (93).

Efron, B., and B. Narasimhan. 2018. "The Automatic Construction of Bootstrap Confidence Intervals." 2018-07. Stanford University.

G. Beliakov, S. James, T. Calvo. 2014. "Consensus Measures Constructed from Aggregation Functions and Fuzzy Implications." *Knowledge-Based Systems* 55: 1–8.

Kvalseth, T. O. 1995a. "Coefficient of Variation for Nominal and Ordinal Categorical Data." *Perceptual and Motor Skills* 80 (3): 843–47.

———. 1995b. "Comment on the Coefficient of Ordinal Variation." *Perceptual and Motor Skills* 81 (3): 621–22.

McGraw, K. O., and S. P. Wong. 1996. "Forming Inferences About Some Intraclass Correlation Coefficients." *Psychological Methods* 1 (1): 30–46.

Tastle, W. 2009. "Measuring Faculty Instructional Performance." *ISECON Proceedings* 26 (3).

Wilcox, A. R. 1973. "Indices of Qualitative Variation and Political Measurement." *The Western Political Quarterly* 26 (2): 325–43.