Assessing Human Visual Inspection for Acceptance Testing: An Attribute Agreement Analysis Case Study

Christopher Drake
Lead Statistician, Small Caliber Munitions
QE&SA Statistical Methods & Analysis Group
• Background and Motivation
• Methodology and Rules of Thumb
• Metrics
• **Case Study**: Short Range Training Ammunition (SRTA) Trace Study
  – Test Objective
  – Test Setup & Execution
  – Data Analysis
  – Conclusions & Recommendations
• In the military, there are many applications which require visual assessment. Due to advances in technology, many automated computer visual inspection systems have been replacing traditional human visual inspection systems whenever possible for their superior performance with respect to accuracy, reliability, repeatability, and reproducibility.

• There are still some scenarios where human visual inspections are the only and best option due to certain constraints and limitations, so it is important to have a method for assessing the adequacy and effectiveness of these scenarios.

• Traditional Gauge R&R is a tool used to help quantify the inspection error and uncertainty in the system when continuous measurements are available.

• In the case of human observations with categorical responses, Traditional Gauge R&R are not applicable, and Attribute Agreement Analyses are used instead.

Figure 1: Conceptualizing Precision vs. Accuracy
• Quantifying ‘agreement’ and ‘effectiveness’ are two of the ways in which the Attribute Agreement Analysis assesses the categorical visual inspections in an attempt to understand inter-rater reliability.

• The analysis considers Kappa, a metric for quantifying agreement, as well as Prevalence and Bias when applicable.

• Contingency Analyses are also used to quantify misclassification rates.

• For sample size, traditional power and sample size tools can be leveraged based on signal-to-noise style assessments, but as a rule of thumb, more than 20 observations per observer is required (must also consider proportion defective).

• For the number of observers required, 2 is often considered a minimum, with diminishing returns on statistical power after 3 for most applications.

• For proportion defective guidelines, it is generally desired to have a balance of 50% good and 50% bad parts, but more broadly a 30-70% balance is acceptable (this proportion could affect sample size).

• When interpreting the magnitude of Kappa, general guidelines have been suggested as the following, although ultimately somewhat arbitrary:
  – $\kappa \leq 0 \rightarrow$ poor
  – $0.01 \leq \kappa \leq 0.20 \rightarrow$ slight
  – $0.21 \leq \kappa \leq 0.40 \rightarrow$ fair
  – $0.41 \leq \kappa \leq 0.60 \rightarrow$ moderate
  – $0.61 \leq \kappa \leq 0.80 \rightarrow$ substantial
  – $0.81 \leq \kappa \leq 1.0 \rightarrow$ almost perfect
• Kappa is stated to be a metric that measures “true agreement”, one in which takes into account the potential of agreement by pure chance, which is not true agreement.

• Range of possible values of Kappa are -1 to 1, though usually falling between 0 and 1, with a Kappa value of 1 representing perfect agreement, 0 representing agreement no better than would be expected by chance (simply guessing every rating), and negative values representing agreement worse than expected even by chance.

• Kappa values can be used for more than 2 level categorical responses, as well as ordinal responses. When dealing with ordinal responses, the weighted Kappa should be used.

• Constructing confidence intervals around Kappa is suggested, and hypothesis tests against some predetermined null hypothesis Kappa level can be used to produce valuable insights.

\[
\kappa = \frac{P_o - P_c}{1 - P_c}
\]

where \( P_o \) is the proportion of observed agreements and \( P_c \) is the proportion of agreements expected by chance.
• The ‘prevalence’ index looks to account for differences in proportions of *agreements* between classifying cases with ‘positive’ and ‘negative’ connotations.

• When the prevalence index is high, the proportion of chance agreement will also be high, resulting in lower kappa values. This effect is also greater for higher values of kappa.

• The ‘bias’ index is similar in nature to the prevalence index, except it looks at rater’s *disagreement* when classifying cases with ‘positive’ and ‘negative’ connotations.

• When bias values are larger, we expect to see higher kappa values. In contrast to prevalence, the effect of the bias index is greater when kappa is small.

• It is recommended to consider both prevalence and bias indices when assessing the magnitude of kappa values in the case when cases may be considered positive or negative.
The 7.62mm Short Range Training Ammunition (SRTA) Trace cartridge is a round that was designed for short range training scenarios. This round is used exclusively for training scenarios where there may be indoor ranges or limited ranges with a limited fan.

The 7.62mm SRTA Trace is a recently developed Army ammunition with the added trace capability.

- This round needed a method for assessing its trace performance for Lot Acceptance Testing (LAT) within budgetary constraints, so human visual inspection was chosen.

The purpose of this study:

- Characterize the baseline capability of the SRTA Trace visual inspection system.
- Use this study to make adjustments and revisions to the rating instructions.
- Compare results from a similar study with the revised instructions to validate improvement.

![Figure 2: Visual Summary of SRTA Trace Grading Instruction](image)
Many of the same experimental planning steps essential for any well designed are also applicable for these studies, especially with regard to randomization and replication.

Also leveraged during the design process were:
- recognition of and statement of the problem
- proper choice of varied factors and levels
- selection of the relevant response variable(s)
- appropriate choice of experimental design
- careful execution of the experiment
- proper data analysis

For the SRTA Trace AAA, 100 samples were fired with approximately 50% of the product being defective.

The observers were given detailed grading instructions for how to rate the trace events.
After the test was successfully completed and the data was gathered, it was appropriately formatted for analysis.

For each of the 100 total shots, the 5 raters either rated it as a pass or fail.

Subject matter experts also rated these same events and came to concurrence on a ‘standard’, or correct answer, and also took images of the events for future reflection.

The data was analyzed using JMP statistical software.

Figure 3: Raw Data Format in JMP Statistical Software
• From the data, we see clearly from the baseline study that overall kappa values are 78.72%.
• The overall effectiveness of 89.1% was also low compared to the desired 95% goal.
• There seemed to be some confusion with regard to specific images rated due to the low effectiveness.
• We can clearly see which images had the least agreement and effectiveness.

Figure 4: Attribute Agreement Analysis JMP Output for Baseline Study
• From the revised data, we see clearly that the overall kappa values are significantly improved, up from 78.72% to 85.97%.

• The overall effectiveness of 89.1% was also improved to 94.80, approximately meeting our goal of 95%.

• Many of the score for images which showed confusion among the raters were drastically improved after the new rating instructions.

• Still some marginal room to clarify these grading instructions based on a few of the images with lower agreement.

Figure 5: Attribute Agreement Analysis JMP Output for Revised Study
• Assessing systems which rely on human observation can be difficult due to the inherently noisy nature of the test environment and subjects.

• Attribute Agreement Analyses are an important tool to be able to more comprehensively and precisely quantify and assess the adequacy, agreement, and overall effectiveness of human observer dependent inspection systems.

• By identifying occurrences with the least agreement, we can iteratively adjust our instructions and system to better accommodate the required needs of the inspection system.

• In the case study detailed above, the initial baseline study showed several areas where lack of agreement was between observers and the standard was an issue. Using this information, adjustments to the rating instructions were made.

• After the second study was run, it was clear that this effort netted a large improvement across the board in % Agreement, Kappa value, and overall Effectiveness.

• With these improvements implemented, the lot acceptance methods for SRTA Trace ammunition are now more accurate and effective, decreasing the risk of providing defective ammunition to the Warfighter.

<table>
<thead>
<tr>
<th>Test</th>
<th>% Agreement</th>
<th>Kappa</th>
<th>Effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LCB</td>
<td>Mean</td>
<td>UCB</td>
</tr>
<tr>
<td>Baseline</td>
<td>61.5%</td>
<td>71%</td>
<td>79%</td>
</tr>
<tr>
<td>Revised Instructions</td>
<td>77.9%</td>
<td>86%</td>
<td>91.5%</td>
</tr>
</tbody>
</table>

Table 1: Summary of Attribute Agreement Analysis Metrics from Case Study


