DATAWorks 2018 - March 21, 2018

Always a Step Ahead

ARDEC

ARMAMENTS

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

QE&SA

UNPARALLELED COMMITMENT & SOLUTIONS

U.S. ARMY

U.S. ARMY RDECOM®

U.S. ARMY ARMAMENT RESEARCH, DEVELOPMENT & ENGINEERING CENTER

AGENDA

- **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

Background and Motivation

- 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.

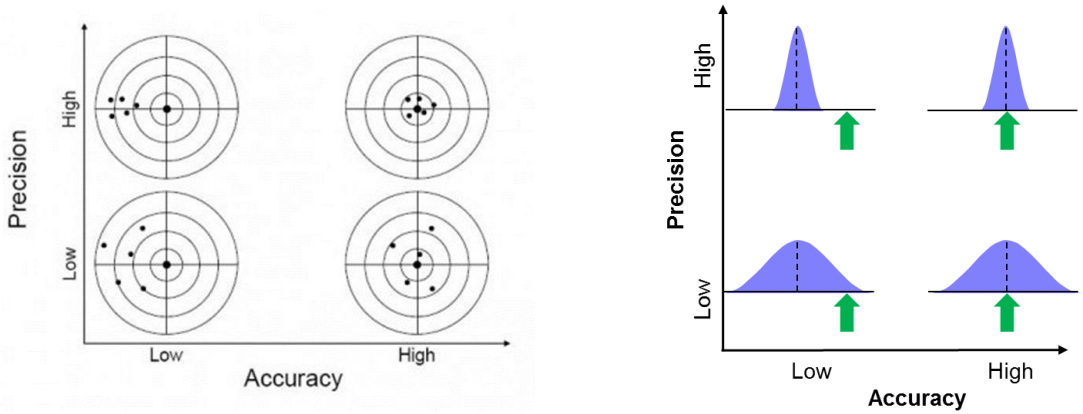


Figure 1: Conceptualizing Precision vs. Accuracy

METHODOLOGY & RULES OF THUMB

- 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

METRICS: *KAPPA*

- 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 pre-determined 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.

SITUATIONAL METRICS: *PREVALENCE & BIAS*

- 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.

CASE STUDY: *BACKGROUND*

- 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 it's 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.

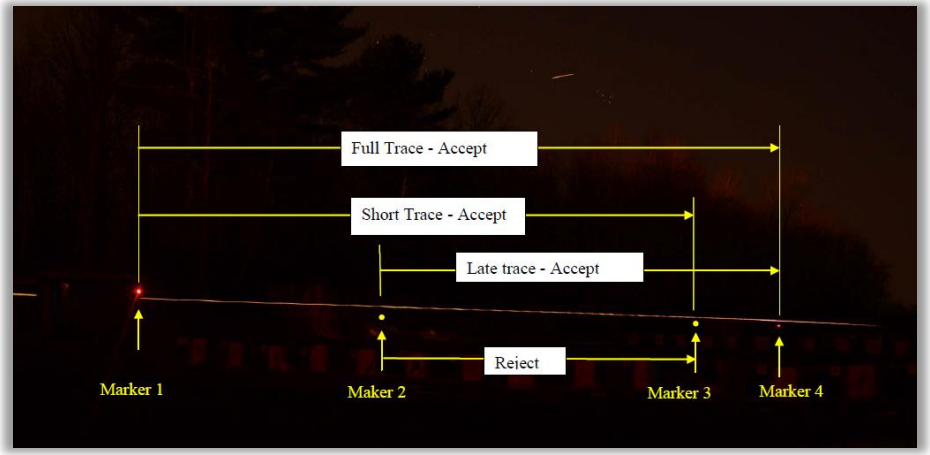


Figure 2: Visual Summary of SRTA Trace Grading Instruction

CASE STUDY: *TEST DESIGN*

- 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.

CASE STUDY: *DATA ANALYSIS*

- 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.

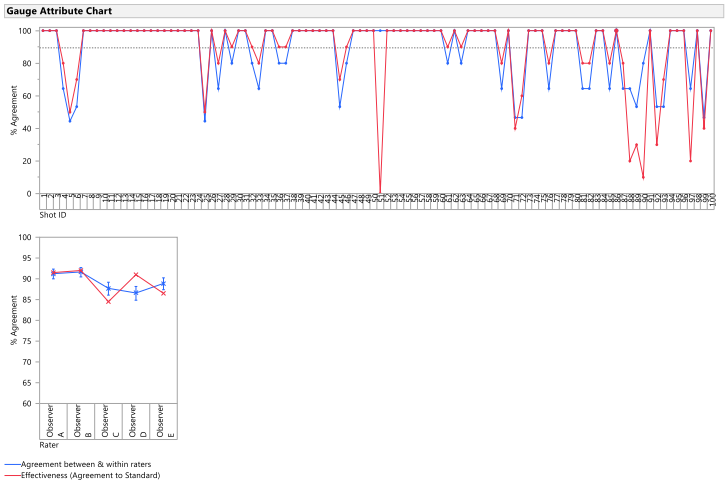
- The data was analyzed using JMP statistical software.

	Shot ID	Observer A	Observer B	Observer C	Observer D	Observer E	Standard
1	1	0	0	0	0	0	0
2	2	0	0	0	0	0	0
3	3	0	0	0	0	0	0
4	4	0	0	0	0	0	0
5	5	0	1	1	1	0	1
6	6	0	0	0	0	0	0
7	7	1	1	1	1	1	1
8	8	1	1	1			
9	9	1					

Figure 3: Raw Data Format in JMP Statistical Software

CASE STUDY: DATA ANALYSIS CONT.

- 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



Agreement Report

Rater	% Agreement	95% Lower CI	95% Upper CI
Observer A	91.7500	90.5896	92.7787
Observer B	91.5000	90.3087	92.5569
Observer C	90.0000	88.6280	91.2228
Observer D	85.7500	83.9111	87.4103
Observer E	87.0000	85.2917	88.5366

Number Inspected	Number Matched	% Agreement	95% Lower CI	95% Upper CI
100	77	77.000	67.846	84.157

Effectiveness Report

Agreement Counts

Rater	Correct(0)	Correct(1)	Total Correct	Incorrect(0)	Incorrect(1)	Grand Total
Observer A	86	97	183	12	5	200
Observer B	89	95	184	9	7	200
Observer C	70	99	169	28	3	200
Observer D	96	86	182	2	16	200
Observer E	77	96	173	21	6	200

Effectiveness

Rater	Effectiveness	95% Lower CI	95% Upper CI	Error rate
Observer A	91.5000	86.8104	94.6254	0.0850
Observer B	92.0000	87.4011	95.0159	0.0800
Observer C	84.5000	78.8393	88.8604	0.1550
Observer D	91.0000	86.2234	94.2313	0.0900
Observer E	86.5000	81.0709	90.5534	0.1350
Overall	89.1000	87.0168	90.8840	0.1090

Misclassifications

Standard Level	0	1
0	.	37
1	72	.
Other	0	0

Conformance Report

Rater	P(False Alarms)	P(Misses)
Observer A	0.0490	0.1224
Observer B	0.0686	0.0918
Observer C	0.0294	0.2857
Observer D	0.1569	0.0204
Observer E	0.0588	0.2143

Assumptions	
NonConform =	0
Conform =	1

Agreement across Categories

Category	Kappa	.2 .4 .6 .8	Standard Error
0	0.7872		0.0149
1	0.7872		0.0149
Overall	0.7872		0.0149

Agreement Comparisons

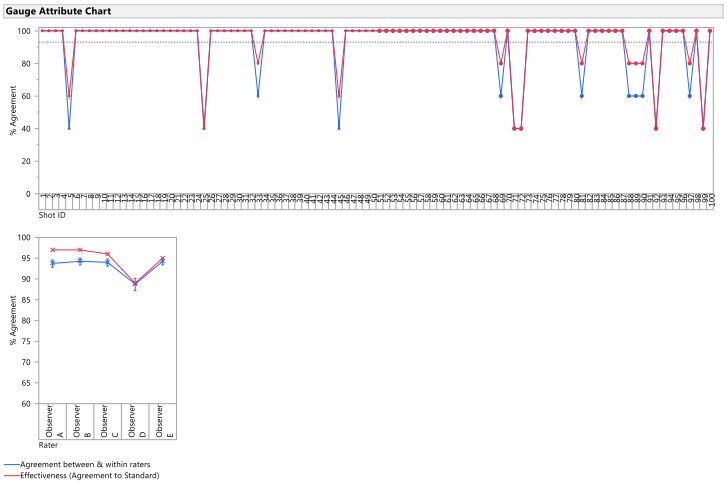
Rater	Compared with Rater	Kappa	.2 .4 .6 .8	Standard Error
Observer A	Observer B	0.8695		0.0349
Observer A	Observer C	0.8155		0.0408
Observer A	Observer D	0.7725		0.0437
Observer A	Observer E	0.8172		0.0410
Observer B	Observer C	0.7675		0.0444
Observer B	Observer D	0.8209		0.0398
Observer B	Observer E	0.8087		0.0414
Observer C	Observer D	0.6222		0.0503
Observer C	Observer E	0.7694		0.0461
Observer D	Observer E	0.6962		0.0481

Rater	Compared with Standard	Kappa	.2 .4 .6 .8	Standard Error
Observer A	Standard	0.8297		0.0394
Observer B	Standard	0.8399		0.0384
Observer C	Standard	0.6883		0.0498
Observer D	Standard	0.8204		0.0400
Observer E	Standard	0.7291		0.0479

Figure 4: Attribute Agreement Analysis JMP Output for Baseline Study

CASE STUDY: DATA ANALYSIS CONT.

- 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.



Agreement Report

Rater	% Agreement	95% Lower CI	95% Upper CI
Observer A	93.7500	92.8166	94.5692
Observer B	94.2500	93.3852	95.0078
Observer C	94.0000	93.1008	94.7886
Observer D	88.7500	87.1853	90.1453
Observer E	94.2500	93.3852	95.0078

Number Inspected	Number Matched	% Agreement	95% Lower CI	95% Upper CI
100	86	86.000	77.863	91.474

Effectiveness Report

Agreement Counts

Rater	Correct(0)	Correct(1)	Total Correct	Incorrect(0)	Incorrect(1)	Grand Total
Observer A	43	54	97	2	1	100
Observer B	43	54	97	2	1	100
Observer C	45	51	96	0	4	100
Observer D	45	44	89	0	11	100
Observer E	43	52	95	2	3	100

Effectiveness

Rater	Effectiveness	95% Lower CI	95% Upper CI	Error rate
Observer A	97.0000	91.5481	98.9745	0.0300
Observer B	97.0000	91.5481	98.9745	0.0300
Observer C	96.0000	90.1629	98.4337	0.0400
Observer D	89.0000	81.3687	93.7458	0.1100
Observer E	95.0000	88.8250	97.8456	0.0500
Overall	94.8000	92.4899	96.4270	0.0520

Misclassifications

Standard Level	0	1
0	.	20
1	6	.
Other	0	0

Conformance Report

Rater	P(False Alarms)	P(Misses)
Observer A	0.0182	0.0444
Observer B	0.0182	0.0444
Observer C	0.0727	0.0000
Observer D	0.2000	0.0000
Observer E	0.0545	0.0444

Assumptions	
NonConform =	0
Conform =	1

Agreement across Categories

Category	Kappa	.2	.4	.6	.8	Standard Error
0	0.8597					0.0316
1	0.8597					0.0316
Overall	0.8597					0.0316

Agreement Comparisons

Rater	Compared with Rater	Kappa	.2	.4	.6	.8	Standard Error
Observer A	Observer B	0.9594					0.0284
Observer A	Observer C	0.8597					0.0509
Observer A	Observer D	0.7240					0.0665
Observer A	Observer E	0.9596					0.0282
Observer B	Observer C	0.8998					0.0435
Observer B	Observer D	0.7634					0.0623
Observer B	Observer E	0.9192					0.0395
Observer C	Observer D	0.8603					0.0504
Observer C	Observer E	0.8998					0.0436
Observer D	Observer E	0.7623					0.0631

Rater	Compared with Standard	Kappa	.2	.4	.6	.8	Standard Error
Observer A	Standard	0.9393					0.0345
Observer B	Standard	0.9393					0.0345
Observer C	Standard	0.9198					0.0391
Observer D	Standard	0.7826					0.0604
Observer E	Standard	0.8992					0.0439

Figure 5: Attribute Agreement Analysis JMP Output for Revised Study

CASE STUDY: *CONCLUSIONS*

- 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 asses 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 1: Summary of Attribute Agreement Analysis Metrics from Case Study

Test	% Agreement			Kappa			Effectiveness		
	LCB	Mean	UCB	LCB	Mean	UCB	LCB	Mean	UCB
Baseline	61.5%	71%	79%	0.758	0.787	0.816	87%	89.1%	90.9%
Revised Instructions	77.9%	86%	91.5%	0.798	0.86	0.922	92.5%	94.8%	96.4%

REFERENCES

1. Sim, J. and Wright, C. C. (2005) **"The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements",** in *Physical Therapy*. Vol. 85, No. 3, pp. 257–268

2. Fleiss, J.L. (1971) **“Measuring Nominal Scale Agreement Among Many Raters”,** *Psychological Bulletin*, Vol. 76, pp 378-382.

3. Landis JR, Koch GG. (1977) **“The Measurement of Observer Agreement for Categorical Data”,** *Biometrics.* 33:159 –174.

4. Montgomery, D. (2013) **“Introduction to Statistical Quality Control”,** Seventh Edition, Wiley

QUESTIONS