Calibrating Humans: Attribute Agreement Analysis

By Jd Marhevko,

VP Quality/Lean & EHS Systems, Accuride Corp
MBB, ASQ Fellow, CMQ/OE, CQE, CSSBB, Past-Chair ASQ QMD

Session: I3
Learning Objectives

Across this time we will...

- Learn about how to Calibrate Humans via Attribute Agreement Analysis (AAA)
- Understand the basic procedures involved in conducting an AAA
- Run an AAA simulation
- Review an example of AAA impact

WHAT YOU SHOULD WALK AWAY WITH...
When we measure or make an assessment of the goodness of an item, we need to be sure that our result is correct. If it is not correct, we take two risks:

- **Alpha $\alpha$ Risk:** We may inadvertently discard or rework a good item (Aw, darn)
- **Beta $\beta$ Risk:** We may inadvertently pass on a bad item (Boy, that was Bad)
Humans believe what they see. We do not generally question our expertise when assessing good or bad. There is generally **one type** of MSA of assessment conducted for Attributes to determine **HOW** right or wrong we are in our results.

This is an Attributes Agreement Analysis (AAA) for items we assess visually or by go/no go or needs to be categorized.
Why do We Need to Know?

We need to know how much error there is in our measurement processes for several reasons:

- Prevent $\alpha$ and $\beta$ errors…along with scrap/rework
- Reduce customer arguments over good/bad
- It is our **JOB** to ensure that our people are enabled to make the right pass/fail decision **EVERY** time
- It is a part of Risk Management (as well as within TS as a part of PPAP)
- **NOTE**: In TS, *every* measured or inspected item on a control plan is **REQUIRED** to have an MSA (or AAA) analysis conducted
Example: Even though a business has safety products with performance critical dimensions, ~99% of the product failures were for attribute reasons: Paint Color, Gloss, Scratches, etc.

• Several sites made the same PNs: There was inconsistency both across and within the various locations

• Internally, personnel were inconsistent on what was or wasn’t good; over-inspection resulted. This inadvertently increased the finished goods expectation (beyond the spec) where in-spec product was now no longer acceptable

• Product would be shipped back and forth for rework and/or replacement; a very expensive game of ping pong. Let’s change the game!
AAA Checks for the chances of 100% agreement on 3 features:

- **Within** “myself”; Did I repeatedly call it good or bad in a consistent manner (even if I was wrong)

- **Between** both me and “my peer”; Did both my peer and I repeatedly call it good or bad in a consistent manner (even if we were both wrong)

- Compared to **“Standard”**; Did I/we get it right each time

```
<table>
<thead>
<tr>
<th></th>
<th>Std</th>
<th>Op1</th>
<th>Op2</th>
<th>Op3</th>
<th>Y/N</th>
<th>Y/N</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>P/F</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>w/in</td>
</tr>
<tr>
<td>1</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>2</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>4</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>P</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>6</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
</tbody>
</table>
```
AAA Notes

- Pass/Fail “Samples”: Preferably 50 or more.
  NOTE: One unit might have several samples on it
- AAA: A check for accuracy in human performance
- The target for “Statistical Agreement” is $\geq 85\%$
- Another form of Agreement is called the Fleiss Kappa ($K$). AIAG calls out for $K \geq 75\%$
- An AAA should be performed to evaluate measurement system effectiveness and should be conducted BEFORE the process is “turned on”
- Also, like a “PM” activity, it should be used to keep human assessors “calibrated” on a regular basis
AAA: What It Looks Like

AAA Gives a series of graphs to show how the operators perform in general. The goal is to be in 100% agreement all of the time. While not feasible, 85% Statistical Agreement or a $K$ of 75% is acceptable.

Not an effective Statistical Agreement at < 85%... This team agrees with each other about 68% of the time. However, 95% of the time, they will range in agreement from 47% to 85%.
More AAA Notes

• **Fleiss Kappa** $K$ **Statistic**: The degree of **Absolute** agreement between ratings. Note: This is **not** the same as **Statistical** Agreement. 100% is ideal, 75% is advised by AIAG

• **AAA in Minitab**: Minitab runs both Absolute and Statistical AAA formats. Be sure to know which one you are running

• **Kendall’s Coefficient**: Use if the inspection data is ordinal (Ex. Good/better/best rating, 1-5 rating, etc.)

<table>
<thead>
<tr>
<th></th>
<th>Std</th>
<th>Op1</th>
<th>Op2</th>
<th>Op3</th>
<th>Y/N</th>
<th>Y/N</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>P/F</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>2</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>
Excel:
ASQ Statistics Division Model:
Enter operator data
Ensure correct calculations
% Statistical Agreement (≥ 85%)
Not Absolute Fleiss Kappa K

MiniTab 17 – Quick & Dirty…
Menu: Assistant
Note: Provides %Statistical Agreement (≥ 85%). (Not K) Quick and easy
Conduct a formal AAA to get K results
Measurement System Analysis
Attribute Agreement Worksheet
Create Table
Get/Populate Operator Results
Menu: Assistant
Attribute Agreement Analysis
Attribute Agreement Study
Selected data OK
How Easy Is It To Do?

• Let’s play with pennies
• Assess the pennies for pass/fail
• Enter the data
• Evaluate the process
• Do it again
The Game has Changed

• All sites went through 3 phases of AAA evaluation levels: Operator, Inspector, and site Management
• Pass/Fail samples were shared across the sites to refine internal inspection definitions
• All sites and engineering aligned on the product approval definitions (photos, videos, samples)
• A key customer was partnered with. The attribute product specs were re-verified. Teams were deployed to other customer locales for samples training

1. **Internal** rejects down by >30% across the business
2. **External** rejects down by 60-80% for key customers
3. **Premium freight down** across the business
Across this time we have...

- Learned how to Calibrate Humans via Attribute Agreement Analysis (AAA)
- Worked to understand the basic procedures involved in conducting an AAA
- Ran an AAA simulation
- Reviewed an example of AAA impact
Thank You!
Any Questions