STATISTICAL PROCEDURES FOR MEASUREMENT SYSTEMS
VERIFICATION AND VALIDATION

This guidance document describes the primary methods for validating measurement systems (test equipment, gauges, inspections or other analytical assessment tools or procedures) intended to assess the performance of materials or products for the purpose of accepting or rejecting the item(s) for continued use or shipment.

GENERAL

Specific details of calculations and interpretations can be found in the references listed at the end of this guidance document.

All measurements will be randomized and will require independent set-ups for repeated measurements. Whenever possible, the test operator will not be aware of the part number or serial number to avoid operator bias.

All experimental structures should ensure that no primary causal categories are confounded.

Functional, non-repeatable or destruct testing methods will require thoughtful determination of how the matched pairs will be created or selected for testing to minimize the effect of confounding product variation or use conditions with the measurement error.

Typical approaches are:
- Create large homogenous master samples for bulk materials (wet or dry)
- Take a larger sample and separate it into smaller samples
- Select sequential parts made under relatively the same conditions
- Functional tests are performed under constant ‘conditions of use’.

This may require the consultation of a subject matter expert in measurement systems analysis and/or testing methods. This structure will be recorded in the validation plan and final report.

All verification and validation plans must clearly state the acceptance criteria. This criteria is almost always situational and will be affected by the product requirements and business needs. The determination of acceptance criteria is collaborative effort among the effected stakeholders – operations, supply chain, R&D, Marketing, QA and Regulatory.

When validating new methods some considerations are:
- Supplier, IQA or early in-process testing:
  - The maximum false acceptance rate that can be tolerated based on the amount of scrap & rework or effect on capacity resulting from failures being detected downstream.
- The maximum false rejection rate that can be tolerated based on the amount of NCP and extra raw material that would have to be purchased as a result of rejecting acceptable material.

- In-process and Final Acceptance testing:
  - The maximum false acceptance rate that will not jeopardize label claims, product requirements and/or customer complaint goals.
  - The maximum false rejection rate that can be tolerated based on the amount of scrap & rework or effect on capacity

When validating replicate, replacement or alternate methods or operators the acceptance criteria will typically be evidence of equivalence or some level of improvement that is expected over the current system’s performance. This will typically result in a requirement for the maximum tolerable bias – or difference – between the new method and the current method that will not substantially affect the false acceptance or rejection rate.

Requirements for false acceptance & false rejection rates and equivalence must be determined prior to running the testing and analyzing the results. These requirements are necessary for determining the appropriate sample size and experimental structure.

Repeatability and reproducibility results are intended to be used as a guide in determining the final measurement protocol; there are no specific results that dictate usability. If a specific system requires a specific level of performance (e.g. inspector qualification) that requirement will be documented in the governing work instruction or SOP for the process.

Any screening test for input factors that is initiated in response to a functional failure must include correlation of the input factor to the functional failure mode.

Only invalid test results may be excluded from any analysis. They must still be recorded in the final validation report along with the rationale for claiming that the test was in fact invalid. (Instrument error, impossible value obtained, known error in sample prep, etc.)

All results must be reported with quantification of statistical significance and practical importance. All p values must be described as one tailed or two tailed with appropriate scientific rationale for either.
CONTINUOUS DATA

Repeatability

The standard sample size is 30 units measured twice randomly for each operator. Expensive, difficult or destruct testing may require smaller sample sizes and the rational for smaller sample sizes is to be documented in the validation plan.

Selected samples should span the full range of expected variation.

Repeatability of each test equipment should be established using a single operator.

Results are to be plotted on a Youden Plot and the measurement error, $\sigma_e$, the Intraclass Correlation Coefficient, $\rho (\rho_{ICC} = 1 - \sigma_e^2 / \sigma_T^2)$ and the Discrimination Ratio, $D_R$ ($D_R = \sqrt{(1 + \rho) / (1 - \rho)}$) are to be reported.

Specification Limits will be displayed on the plot to provide context for the measurement error and total observed variation in the study.

Youden Plot: The Youden plot is a square scatter diagram with equal X and Y axes; the axes are the same physical length and the same scale. The Youden plot also has a 45° line originating at zero. Measurement error, $\sigma_e$, is the scatter in the direction that is perpendicular to a 45° line. When used for graphical display of continuous measurement data, the plot may be specifically referred to as a “Measurement Discrimination Plot”.

The ideal measurement system has a small measurement error in relation to the specification range, if the units assessed in the study span the full range of the specification range, the Discrimination Ratio will be fairly large ($D_R \geq 7$). However, some measurement systems can still be valuable with lower discrimination ratios. Depending on the criticality of the system guard banding may be necessary. In these cases, consult a statistician or member of the Operational Excellence team to determine the best approach.
Measurement Stability

Each Youden Plot will also include a Measurement Stability analysis to determine if the repeated measurements were statistically stable. Unstable results are an indication that one or more test pairs *may* have an invalid reading causing a large difference between the pairs. These values should be investigated for possible exclusion and investigation of the measurement system.

Measurement stability is assessed with a basic Range Control chart assessing only the upper limit:  \( UCL_R = D_4 \bar{R} \) where \( D_4 = 3.267 \) (for \( n = 2 \))

Any values outside the upper control limit would indicate a statistically significant difference in measurement error for that pair. If any extreme values can be verified as being invalid (typographical error, physically impossible, or an invalid test condition) the pair may be removed from the analysis. Pairs may not be removed simply because they are outside of the control limits (extreme values); *pairs may only be removed if they are verified as being invalid results.*

Example of Youden Plot and Range Chart for Detecting Extreme Values:
Reproducibility

Reproducibility is assessed when more than one operator will be performing the test or when a new operator is assigned to the test as a qualification for performing the test. Reproducibility analysis will only be performed after each operator demonstrates adequate repeatability.

If the original R&R sample set is not available for future operators, reproducibility can be established by having one or more of the current operators included in the study to enable the necessary operator to operator comparison.

Reproducibility analysis will include a Youden plot with the first result for each operator. An Analysis Means (ANOM) approach will be taken to assess the statistical significance of operator to operator differences. The practical importance of any reproducibility difference is assessed qualitatively by comparison of the difference to the specification limits on the Youden Plot.

Example of Youden Approach for Assessing Reproducibility:
**Method Comparison**

Method comparison is performed when assessing a replicate, alternate or replacement measurement system. Method comparison is also appropriate when assessing the difference between systems or methods in different locations such as a vendor & IQA or IQA & Manufacturing.

If method comparison is required, repeatability and reproducibility will be established for each method individually then the two methods will be compared. Analysis will include Youden plots for repeatability of each system and either a Bias Youden or a Bland-Altman plot to display the size of the bias between the methods. In a Bias Youden plot the bias is the difference between the 45 degree line and a median line through the data cloud that is parallel with the 45 degree line. In a Bland Altman plot the average difference is plotted as the center line. Tolerance limits are then plotted about the average difference line. Typically the tolerance limits are \( \pm 2\sigma_d \), where \( \sigma_d \) is the standard deviation of the differences of each of the pairs. The plotted points are the differences of the individual pairs.

Example of Youden and Bland Altman Analyses for Method Comparison:

- **Youden Bias Plot**
- **Bland Altman Method Comparison Bias Plot**

Statistical significance of the bias will be assessed by a paired t-test. The required sample size will take into account the size of allowable bias or disagreement. The calculations performed to determine sample size will be included in the plan and final report.
CATEGORICAL DATA

Design and verification testing may be performed with a non-representative distribution of units, particularly marginally passing and failing units to optimize the test and set the cutoff limit(s).

All final validation testing for repeatability, reproducibility and method comparison must have a sample distribution of passing and failing units that is representative of the actual distribution of the process stream that is being tested. This is particularly important when assessing methods that are highly operator sensitive such as a visual inspection. If this distribution is not known the technique or assumptions utilized to estimate the distribution are to be documented in the validation plan and final report.

False Acceptance & False Rejection Rates

Each test must have a stated goal for acceptable (not to exceed) false acceptance and false rejection rates. These rates will be dependent on the purpose of the test and the consequences of a false result.

False acceptance and false rejection rates are determined by testing units with known pass/fail status (truth). If truth is unknown every effort should be made to establish truth. This may be done by utilizing a known ‘gold’ standard method, functional testing, or repeated testing to determine scientific and statistical consensus (most probable truth). In the case where human judgment is involved, especially when the criteria utilizes an operational definition and not a measureable physical characteristic, an expert panel may be utilized to determine ‘truth’. (e.g. visual inspection of cosmetic defects, functional anomalies or binning of customer complaints). The method for determining truth will be documented in the validation plan and final report.

The required sample size will take into account the necessary accuracy of the estimate of the false results. The calculations performed to determine the sample size will be included in the plan and final report.

The general format for the analysis is a 2x2 table. The critical statistics are the false acceptance rate and the false rejection rate.

Example Analysis of False Acceptance and Rejection Rates:

<table>
<thead>
<tr>
<th>Truth</th>
<th>Pass</th>
<th>Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pass</td>
<td>96</td>
<td>2</td>
</tr>
<tr>
<td>Fail</td>
<td>4</td>
<td>98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>False Acceptance Rate</th>
<th>False Rejection Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed</td>
<td>2%</td>
<td>4%</td>
</tr>
</tbody>
</table>

Beverly Daniels 7 of 16 9/14/15
Confidence intervals are to be calculated and reported against the goals for false results. Typically this calculation will require the Exact Binomial Confidence Limits as the rates will be fairly low and the Normal approximation to the Binomial will not be appropriate.

A final assessment of the false acceptance and rejection rates given the actual process distribution is to be included in the final report. This is particularly important if the validation test samples were not representatively distributed.

The estimated false rates that will be experienced are calculated by:

- **Expected False Acceptance Rate = Validation False Acceptance Rate x Proportion of Failing units in the Process Stream**
- **Expected False Rejection Rate = Validation False Rejection Rate x Proportion of Acceptable units in the Process Stream**

The false result rates have no predictive value unless the equipment and operators are also shown to be repeatable and reproducible.

**Repeatability & Reproducibility**

Each test system and operator for critical tests must be individually qualified by repeatability testing. Multiple operators and/or replicate test systems must have reproducibility studies.

The general statistic of interest is the Kappa test which provides a measure of agreement adjusted for the distribution of failing and passing units. The Kappa statistic will be reported with its corresponding confidence limits in the final validation report. In general a Kappa statistic of $\geq 0.95$ is required.

The sample set for the final validation Repeatability and Reproducibility study should be selected such that the distribution of passing and failing units is representative of the actual process distribution.

Repeatability and Reproducibility have no useful value if the system is not also accurate to truth.
Method Comparison

Method comparison is a special form of reproducibility that is performed when assessing a replicate, alternate or replacement measurement system. Method comparison is also appropriate when assessing the difference between systems or methods in different locations such as a vendor & IQA or IQA & Manufacturing.

If method comparison is required, repeatability and reproducibility will be established for each method individually then the two methods will be compared.

Statistical significance of any difference will be assessed by McNemar’s test: The Chi Square test can be used for n > 20. The exact Binomial p value will be used for smaller sample sizes, n ≤ 20.

Example Analysis of Statistical Significance of a Difference:

<table>
<thead>
<tr>
<th>Original Test Stand</th>
<th>Pass</th>
<th>Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Test Stand</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pass</td>
<td>a 25</td>
<td>b 5</td>
</tr>
<tr>
<td>Fail</td>
<td>c 15</td>
<td>d 55</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>one tailed</th>
<th>two tailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>p_BINOMIAL</td>
<td>0.00591</td>
<td>0.01182</td>
</tr>
<tr>
<td>p_χ²</td>
<td>0.02209</td>
<td>0.04417</td>
</tr>
</tbody>
</table>

If there is a statistically significant difference, the practical importance of the difference (bias) will be assessed by the estimated difference and its control limits vs. the pre-established acceptable difference. The required sample sizes, passing and failing unit distribution and the acceptable bias will be established prior to performing any testing and will be documented in the validation plan and final report.

Example Analysis of the Size of the Difference:

<table>
<thead>
<tr>
<th>Original Test Stand</th>
<th>Pass</th>
<th>Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Test Stand</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pass</td>
<td>a 25.0%</td>
<td>b 5.0%</td>
</tr>
<tr>
<td>Fail</td>
<td>c 15.0%</td>
<td>d 55.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>95% LCL</th>
<th>Δ</th>
<th>95% UCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Allowable Difference</td>
<td>0.015</td>
<td>0.1</td>
<td>0.185</td>
</tr>
</tbody>
</table>

Proportion Difference

Method Comparison: Original vs. New Test Stand

Beverly Daniels 9 of 16 9/14/15
Appendix: Formulas

CONTINUOUS DATA

Measurement Error, \( \sigma_e \):
The Range of measurements of each unit, \( R_i \):
\[ R_i = |X_A - X_B| \]

The Average Range of the sample set, \( \bar{R} \):
\[ \bar{R} = \frac{\sum R_i}{k} \]

The Repeatability or Measurement Error, \( \sigma_e \):
\[ \sigma_e = \frac{\bar{R}}{d_2} \]

Where \( k \) = the number of units in the sample set
\( d_2 = 1.128 \) (for \( n = 2 \))

Total Measured Variation, \( \sigma_T \):
The average value of each unit, \( \bar{X}_i \):
\[ \bar{X}_i = \frac{X_{Ai} + X_{Bi}}{2} \]

The product variation, \( \sigma_P \):
\[ \sigma_P = \sqrt{\frac{\sum (\bar{X} - \bar{X}_i)^2}{n-1}} \]

The total variation, \( \sigma_T \):
\[ \sigma_T = \sqrt{\sigma_P^2 + \frac{1}{2} \sigma_e^2} \]

Measurement Discrimination Formulas:

The Intraclass Correlation Coefficient, \( \rho \):
\[ \rho_{ICC} = 1 - \frac{\sigma_e^2}{\sigma_T^2} \]

The percentage of the total variation that is due to product variation is \( \rho \times 100 \).
\[ \%\sigma_p = \rho \times 100 \]

The percentage of the total variation that is due to measurement error is \((1-\rho) \times 100 \)
\[ \%\sigma_e = (1- \rho) \times 100 \]

The Discrimination Ratio is the number of distinct categories that the observed process variation can be divided into given the measurement error and the amount of observed variation.
\[ D_R = \sqrt{\frac{1+\rho}{1-\rho}} = \sqrt{\frac{2\sigma_T^2}{\sigma_e^2}} - 1 \]
Method Comparison Formulas, Continuous Data

Difference between replicates, $d_i$:

\[ d_i = b_i - a_i \]

(where $a$ and $b$ are the two respective readings of a single sample)

Mean of replicates, $\bar{X}_i$:

\[ \bar{X}_i = \frac{(a_i + b_i)}{2} \]

Centerline, $\bar{d}$:

\[ \bar{d} = \frac{\sum d_i}{n} \]

Standard deviation of the differences, $\sigma_d$:

\[ \sigma_d = \sqrt{\frac{\sum (\bar{d} - d_i)^2}{n-1}} \]

Upper and Lower Limits:

\[ \bar{d} \pm 2\sigma_d \]

The Bland Altman plot is created by plotting the mean of each pair, $\bar{X}_i$, on the X axis and the corresponding difference, $d_i$ on the Y axis for each pair of measurements.
CATEGORICAL DATA

False Acceptance & False Rejection Rates

The false rejection rate = \( \frac{\text{# falsely rejected units}}{\text{# true passing units}} \)

The false acceptance rate = \( \frac{\text{# falsely accepted units}}{\text{# true failing units}} \)

Kappa Score, single rater repeatability and two raters reproducibility with two categories:

<table>
<thead>
<tr>
<th>Measurement B</th>
<th>Measurement A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass</td>
<td>Fail</td>
</tr>
<tr>
<td>Pass</td>
<td>355</td>
</tr>
<tr>
<td>Fail</td>
<td>3</td>
</tr>
<tr>
<td>Observed</td>
<td>358</td>
</tr>
</tbody>
</table>

\[
P_{\text{Observed}} = P_{A\text{ Pass} \& B\text{ Pass}} + P_{A\text{ Fail} \& B\text{ Fail}}
\]

\[
P_{\text{Chance}} = (P_{A\text{ Pass}} \times P_{B\text{ Pass}}) + (P_{A\text{ Fail}} \times P_{B\text{ Fail}})
\]

\[
K = \frac{P_{\text{Observed}} - P_{\text{Chance}}}{1 - P_{\text{Chance}}}
\]
Kappa Score, multiple raters and multiple categories:

Overall Kappa Score:  
\[ K_{overall} = 1 - \frac{n(m^2) - \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}^2}{nm(m-1)\sum_{j=1}^{k} \bar{p}_j \bar{q}_j} \]

Kappa Score by Category:  
\[ K_{category} = 1 - \frac{\sum_{i=1}^{n} x_{ij}(m-x_{ij})}{nm(m-1)p_jq_j} \]

n = number of samples  
m = the number of judges, raters or inspectors  
k = number of categories  
x = an individual rating = 1  
p = \# ratings within a category/(nm)  
q = 1 - p
McNemar’s Test for Differences Between Two Correlated (Dependent) Results

Calculating p values for the statistical significance of the difference (the difference is between cells b and c)

One Tailed Exact Binomial (in Excel):

\[ p_{Binomial} = 1 - \text{BINOMDIST}(\text{max}(b,c), n_b + n_c, 0.5, \text{True}) \]

\( \text{max count}(b,c) \) is the larger of the two cell counts in b and c

\( n_b + n_c \) is the sum of the cell counts in b and c.

0.5 is Binomial probability given that if there were no real difference the count of b will equal the count of c.

\( \text{True} = \text{cumulative probability} \)

One Tailed Chi-Square Test (in Excel):

\[ p_{\chi^2} = \frac{\text{CHIDIST}(\chi^2, 1)}{2} \]

\[ \chi^2 = \frac{(|b - c| - 1)^2}{(b + c)} \]

degrees of freedom (df) = 1

Confidence Intervals for the Size of the Difference of Two Correlated Results

\( n = a + b + c + d \)

\( \Delta = |P_{A1} - P_{A2}| \)

\[ \sigma_{\Delta c} = \sqrt{\frac{P_{A1}(1 - P_{A1}) + P_{A2}(1 - P_{A2}) - 2(P_{A1}P_d - P_{A2}P_c)}{n}} \]

Confidence Limits = \( \Delta \pm z_{\alpha} \sigma_{\Delta c} \)

\( P_{A1} = \frac{(a + c)}{n}, \quad P_{A2} = \frac{(a + b)}{n} \)

\( P_a = \frac{a}{n}, \quad P_b = \frac{b}{n}, \quad P_c = \frac{c}{n}, \quad P_d = \frac{d}{n} \)
SAMPLE SIZE FORMULAS

Continuous Data

\[ n = \left( \frac{Z_{\alpha/2} \sigma}{\Delta} \right)^2 \] or \[ n = \left[ \frac{Z_{\alpha} + Z_{\beta}}{\Delta} \right]^2 \]

\( \sigma \): the standard deviation of the population you are trying to estimate
\( \Delta \): the amount of accuracy you want; this is the minimum amount of detectable difference from the true mean.

Note that \( \pm \Delta \) = the Confidence Interval.

\( \alpha \): alpha is the probability of detecting a difference that doesn’t exist, typically .05 or .01
\( \beta \): beta is the probability of not detecting a difference that does exist, typically 0.01

Categorical Data

Categorical data for the proportion of defective events generally follows a Binomial Distribution.

\[ n = \frac{Z_{\alpha/2} \sqrt{p_1(1-p_1)}}{\Delta^2} \] or \[ n = \left( \frac{Z_{\alpha} \sqrt{p_1(1-p_1)} + Z_{\beta} \sqrt{p_2(1-p_2)}}{\Delta} \right)^2 \]

<table>
<thead>
<tr>
<th>Confidence</th>
<th>80%</th>
<th>90%</th>
<th>95%</th>
<th>99.0%</th>
<th>99.9%</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>0.2</td>
<td>0.1</td>
<td>0.05</td>
<td>0.01</td>
<td>0.001</td>
<td>( z_{\beta} )</td>
</tr>
<tr>
<td>( Z_{\alpha} )</td>
<td>0.842</td>
<td>1.282</td>
<td>1.645</td>
<td>2.326</td>
<td>3.090</td>
<td>( Z_{\beta} )</td>
</tr>
<tr>
<td>( Z_{\alpha/2} )</td>
<td>1.282</td>
<td>1.645</td>
<td>1.96</td>
<td>2.576</td>
<td>3.291</td>
<td></td>
</tr>
</tbody>
</table>

Rare events (<5%) will require iteration using confidence intervals for the Exact Binomial. A starting point for the iteration is to use the Exact Binomial formula:

\[ n = \frac{\ln(1-P(Detection))}{\ln(1-p)} \]

\( \ln \): natural log
\( P(Detection) \): the probability of detecting a defect rate that is = p, typically .95 or .99
\( p \): the expected defect rate – or occurrence rate – of the process; may be the minimum rate which we want the system to reliably detect.
Appendix: References

Measurement Systems Analysis Manual, Internal IDEXX publication
Basic Statistics Manual, Internal IDEXX publication


Youden, William John, “Graphical Diagnosis of Interlaboratory Test Results”, Industrial Quality Control, May 1959, Vol. 15, No. 11


