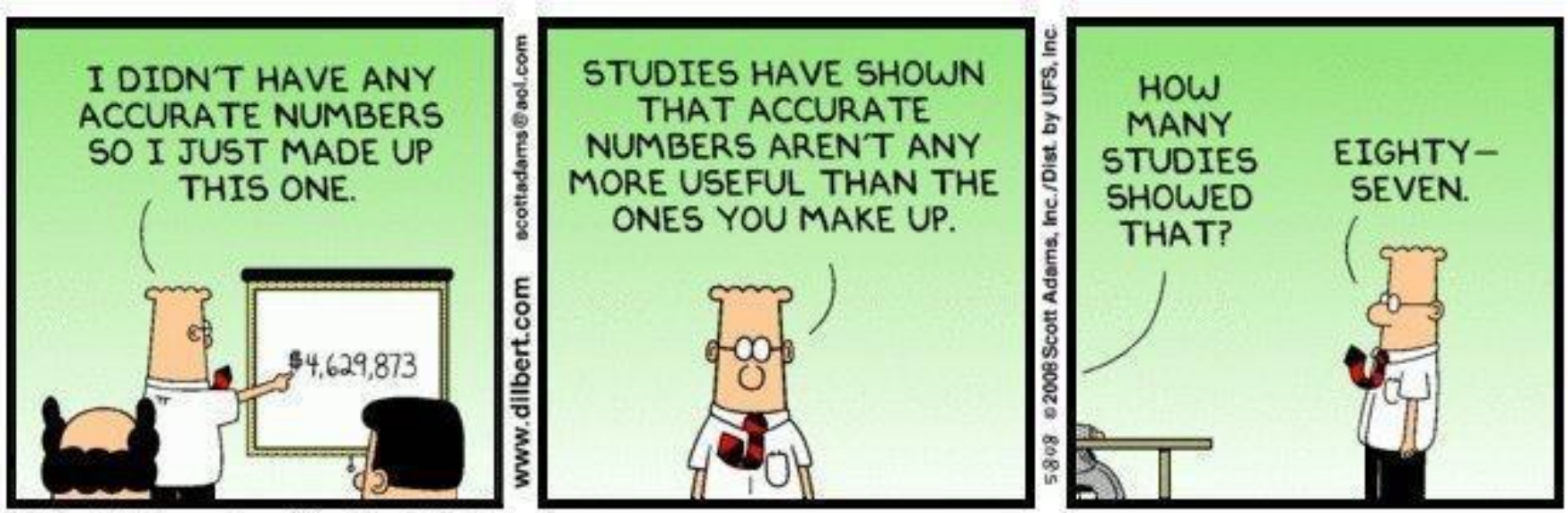
Gage Repeatability and Reproducibility (R&R) Studies

An Introduction to Measurement System Analysis (MSA)

Agenda

- Importance of data
- What is MSA?
- Measurement Error Sources of Variation
- Precision (Resolution, Repeatability, Reproducibility)
- Accuracy (Bias, Stability, Linearity)
- What is Gage R&R?
- Variable vs Binary Data
- Variable Gage R&R
  - Criteria for % of Tolerance
  - Type I and II error
- Attribute Agreement Analysis
  - Criteria for Kappa
- Key Points
- More Resources

Data


I DIDN'T HAVE ANY ACCURATE NUMBERS SO I JUST MADE UP THIS ONE.
$4,629,873
www.dilbert.com scottadams@aol.com
STUDIES HAVE SHOWN THAT ACCURATE NUMBERS AREN'T ANY MORE USEFUL THAN THE ONES YOU MAKE UP.
5-8-08 ©2008 Scott Adams, Inc./Dist. by UFS, Inc.
HOW MANY STUDIES SHOWED THAT?
EIGHTY-SEVEN.
© Scott Adams, Inc./Dist. by UFS, Inc.

Dealing with Data

- Making decisions based on data is critical in business, and in life
- "Garbage in, garbage out"
  - Need to ensure quality of data collected before analyzing or drawing conclusions
- How do you know if your data is "good"?

Measurement System Analysis (MSA)

What is MSA?

- Measurement System Analysis
- A controlled experiment where a sample of items are measured multiple times by different devices or people to separate the variation into specific sources
  - Gage Repeatability and Reproducibility (R&R) is a subset of MSA
- Provides estimate of "measurement error" to determine if variation is excessive or acceptable

Example of Gage R&R

Measuring thickness of a phone using calipers

If the thickness measuring process had no variation, then all measurements of each phone would be identical, regardless of who took the measurement, or which measurement device they used.

What does MSA evaluate?

- MSA can evaluate:
  - The process to setup and calibrate the measurement device
  - The technique used to setup the item prior to being measured
  - Whether different measurement devices (equipment and tools) or different versions of the same device influence the variation
  - The people who take the measurements
  - How the data is collected and recorded
  - The method for making a decision based on the data

MSA evaluates before, during and after the measurement is taken

Measurement Error

Measured Value = Actual Measurement + Measurement Error

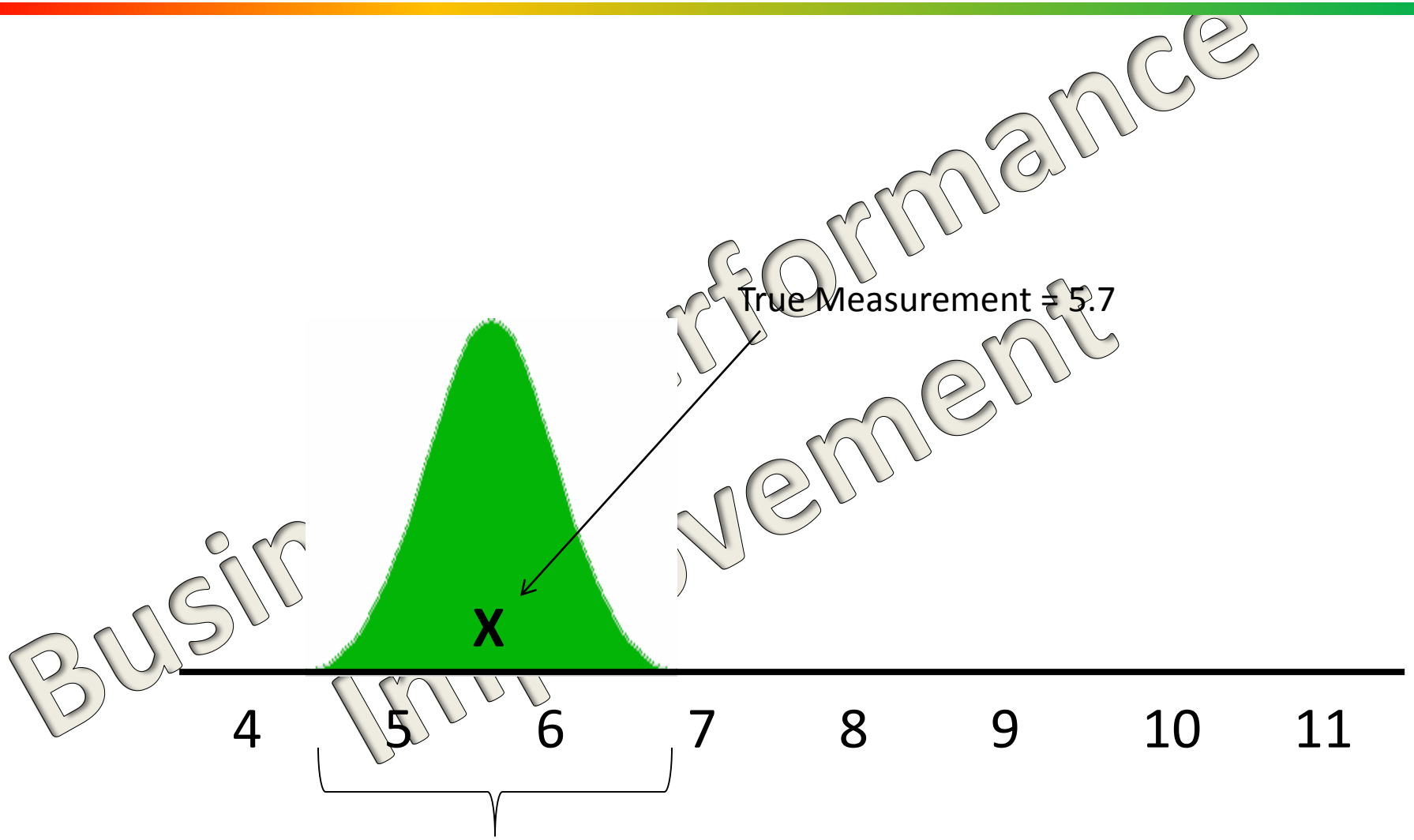
Example: Thermometer
Measured Value = **78.4 ° F**
Measurement Error = ??

What is true temperature?

- If measurement error is 1.5 ° F, then true temperature might be 74 - 83 ° F
- If measurement error is 0.1 ° F, then true temperature might be 78.2 – 78.6 ° F

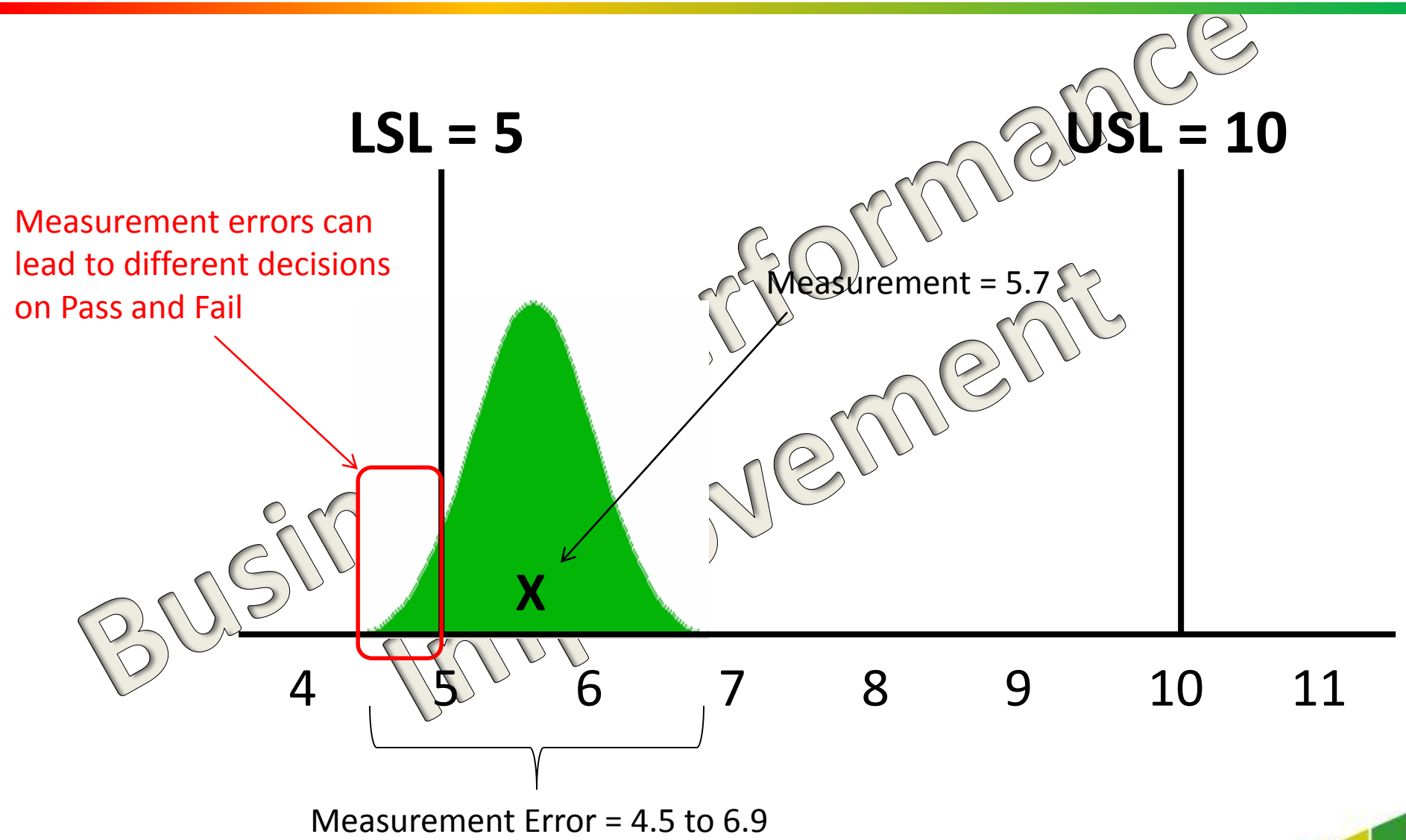
Must know measurement error to know the likely true value

What is Measurement Error?



Measurement Error for one item = 4.5 to 6.9

Is this a problem?



Real life MSA – Mortgage Loan

- Approval based on:
  - Credit score
  - Rental payment history
  - Previous home ownership
  - Job status and length
  - Income to debt ratio
  - Type of home
  - Familiarity with applicant and their references

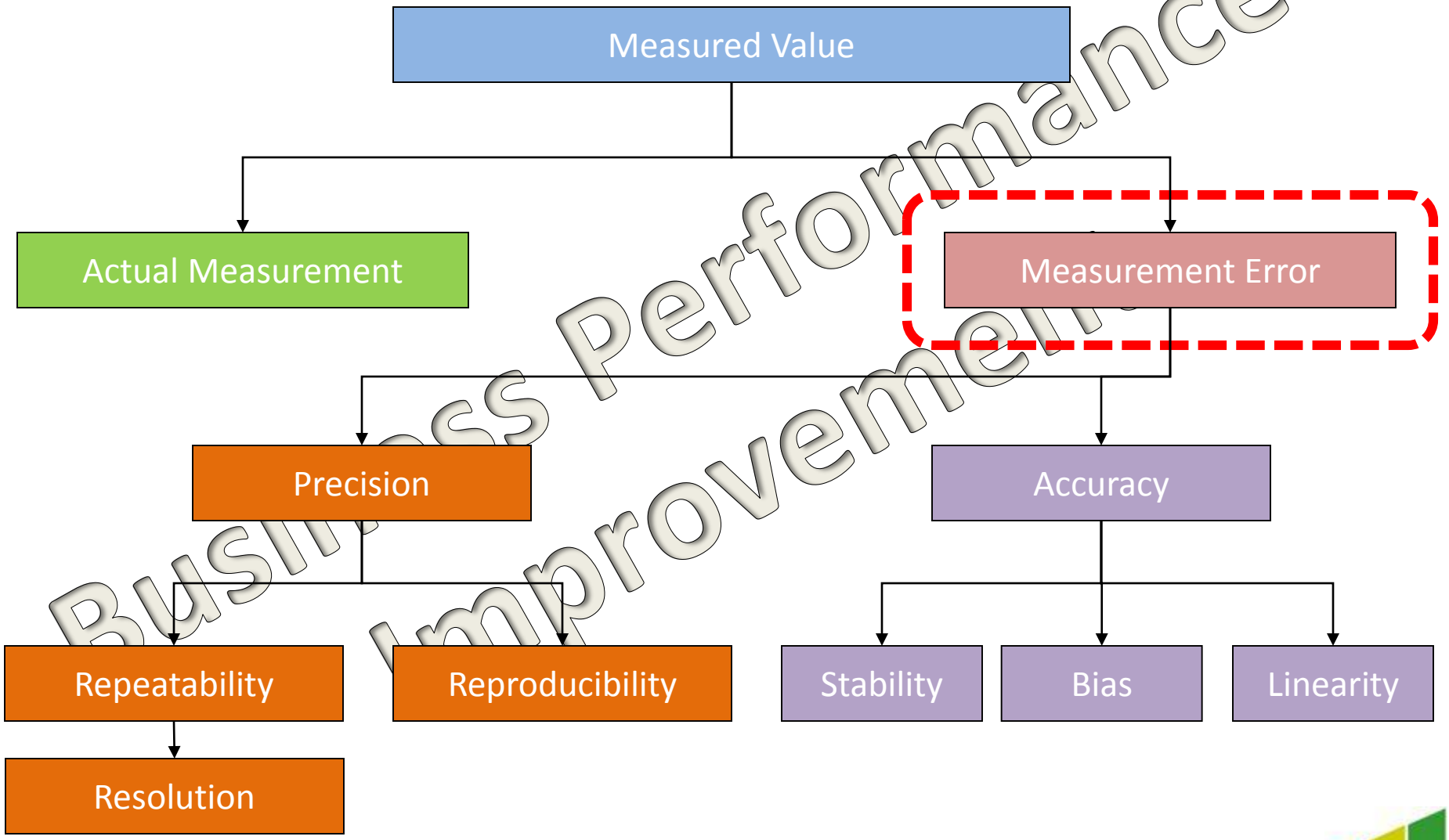
Why do we need a MSA?

- In order to make good decisions in business and in life, we need good data
- Without performing a MSA, we falsely assume the data is good
- If we are wrong and the data is not good, we might make an incorrect decision
- **MSA helps us determine if the data is good, so we can make the best decision possible**

Sources of Error

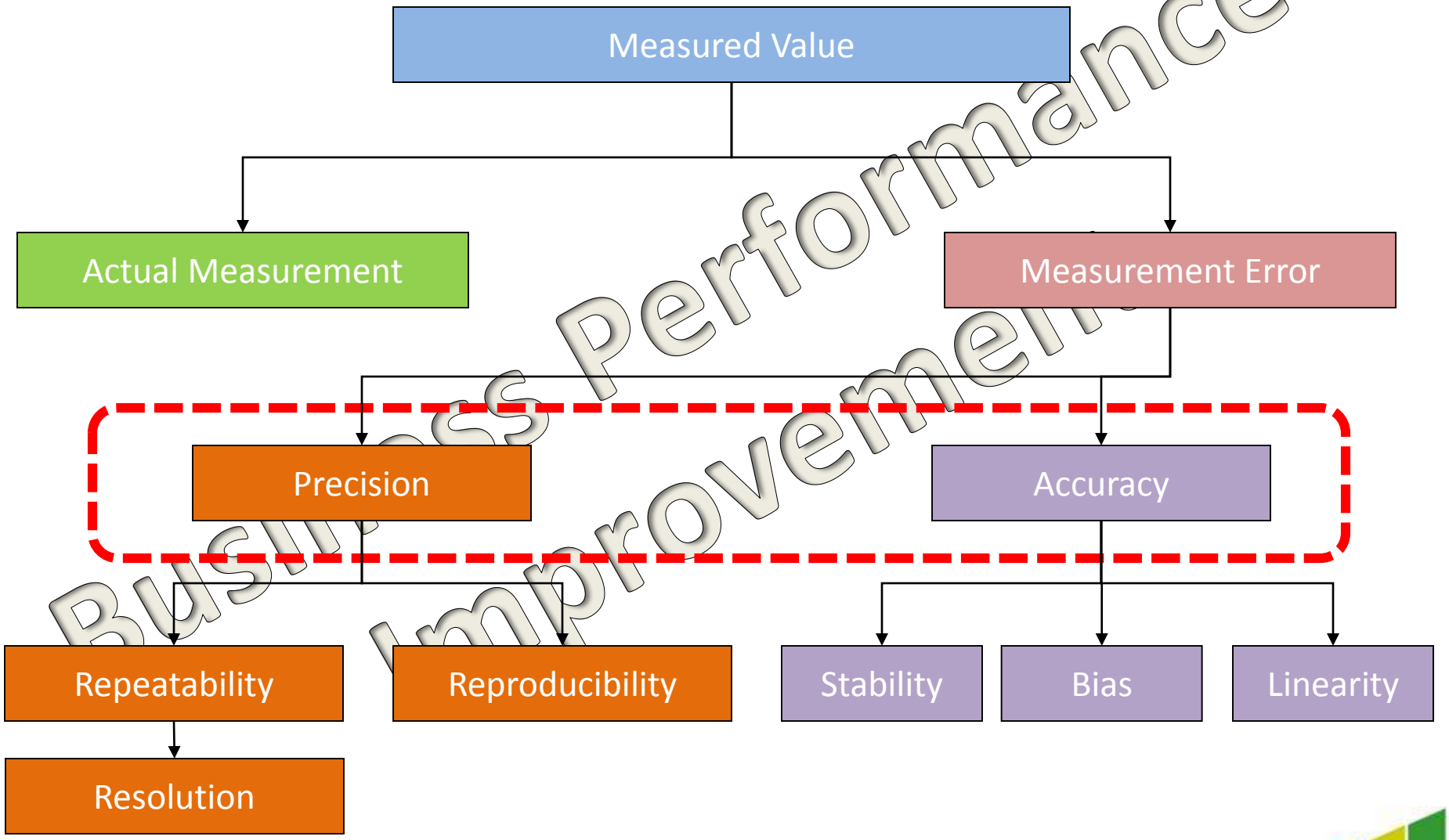
- These measurement sources can increase the measurement error
  - Repeatability
  - Reproducibility
  - Accuracy
  - Bias
  - Stability
  - Linearity
  - Resolution

Measurement Error



Measured Value = Actual Measurement + Measurement Error

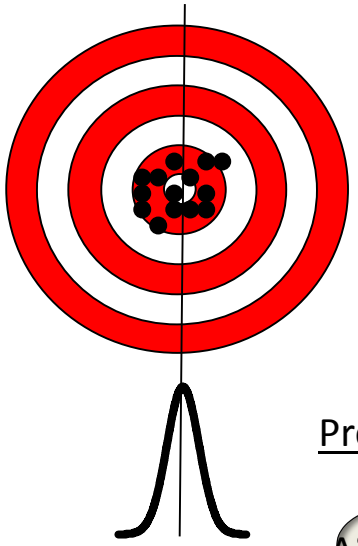
Measurement Error



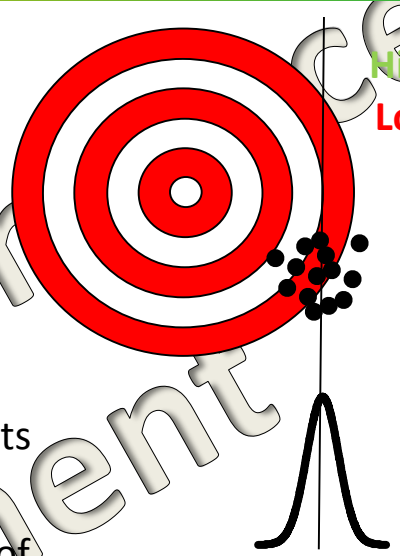
http://www.ncsu.edu/labwrite/Experimental%20Design/accuracyprecision.htm

Accuracy vs. Precision

High Precision
High Accuracy



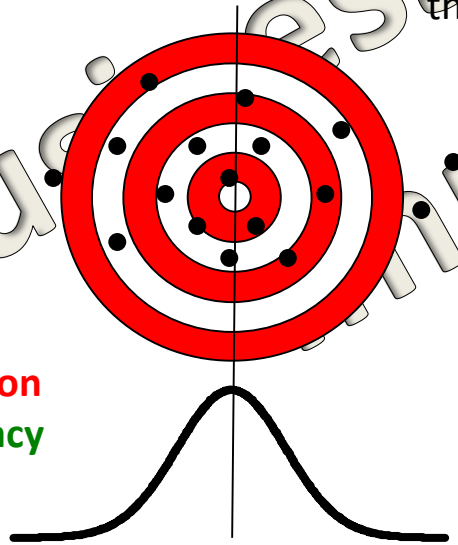
High Precision
Low Accuracy



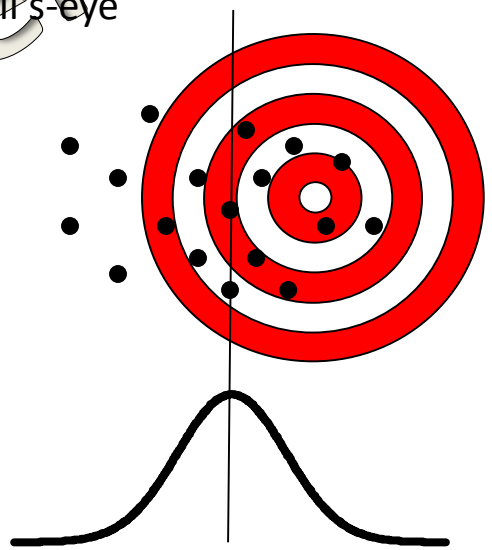
Precision: how spread out the shots are compared to each other

Accuracy: how close the average of the shots are to the bull's-eye

Low Precision
High Accuracy

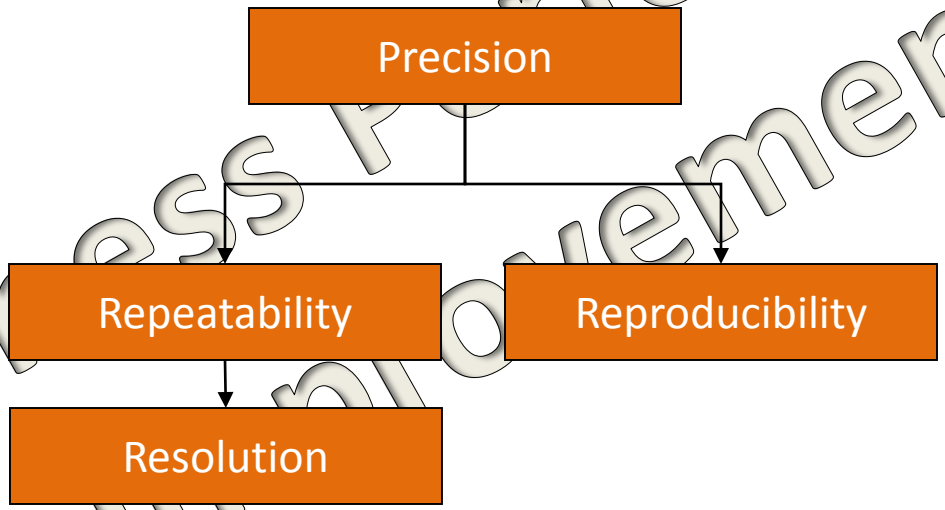


Low Precision
Low Accuracy



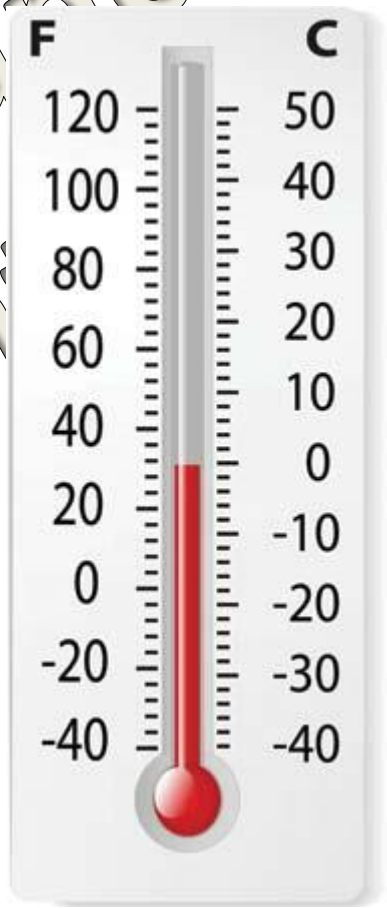
Precision

- Precision – how spread out are the measurements to each other



Repeatability

- The variation in measurements taken by a single person or instrument on the same item and under the same conditions
  - Ideally, the results should be identical
- Example: Thermometer fluctuates from 72 to 78 degrees every minute (not repeatable), but actual temperature is not changing



Example: Repeatability

SAME PART MEASURED OVER AND OVER AGAIN

REPEATABLE	NOT REPEATABLE
0.0036	0.0046
0.0037	0.0057
0.0035	0.0033
0.0036	0.0039
0.0036	0.0050
0.0037	0.0030
0.0036	0.0036
0.0035	0.0055

GOOD **BAD**

Reproducibility

- The variation induced when different operators, instruments, or laboratories measure the same or replicate items
  - Ideally, the average results between instruments or people should be identical
- Example: You think the thermometer shows 56 degrees C, but your neighbor thinks it shows 58 degrees C

Example: Reproducibility

COMPARE AVERAGES OF SAME PART TO EACH OTHER

REPRODUCIBLE

PERSON #1	PERSON #2
0.0046	0.0048
0.0057	0.0050
0.0032	0.0034
0.0039	0.0051
0.0050	0.0037
0.0030	0.0032
0.0036	0.0046
0.0056	0.0044
AVERAGE	AVERAGE
0.0043	**0.0043**

GOOD

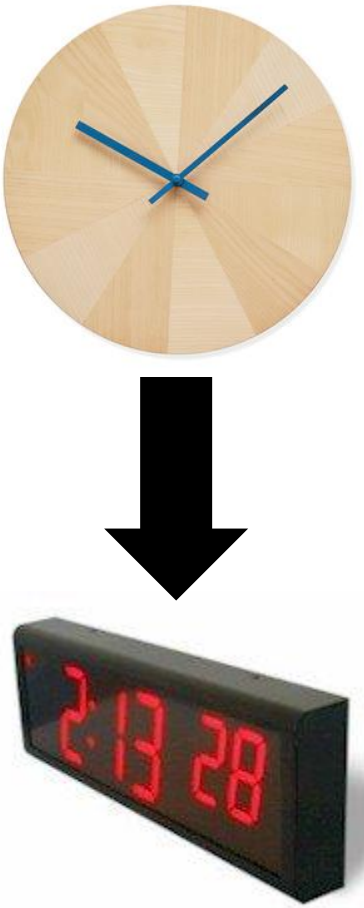
NOT REPRODUCIBLE

PERSON #1	PERSON #2
0.0043	0.0034
0.0052	0.0022
0.0031	0.0021
0.0033	0.0023
0.0045	0.0035
0.0034	0.0024
0.0039	0.0029
0.0052	0.0047
AVERAGE	AVERAGE
0.0041	**0.0029**

BAD

Resolution

- Ability of the measurement system to detect and indicate small changes
  - Ideally, the measurement can detect 10 or more values within likely range
- Each increment should be 10% or less of the range of values to be able to detect a change
- **Example**: Thermometer only displays in increments of 5 degrees (35, 40, 45, etc), unable to get readings between 35 and 40. Prefer to have readings like 35.4 degrees.



Example: Resolution

SAME PART MEASURED OVER AND OVER AGAIN

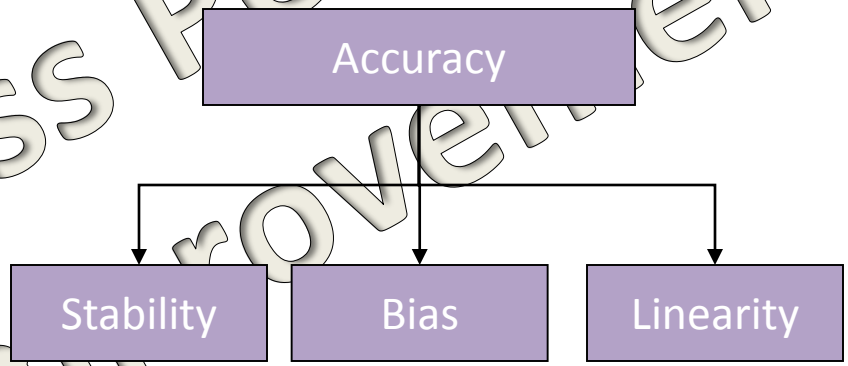
RESOLUTION	POOR RESOLUTION
0.0036	0.00
0.0037	0.01
0.0035	0.00
0.0036	0.00
0.0036	0.01
0.0037	0.00
0.0036	0.00
0.0035	0.01

GOOD

BAD

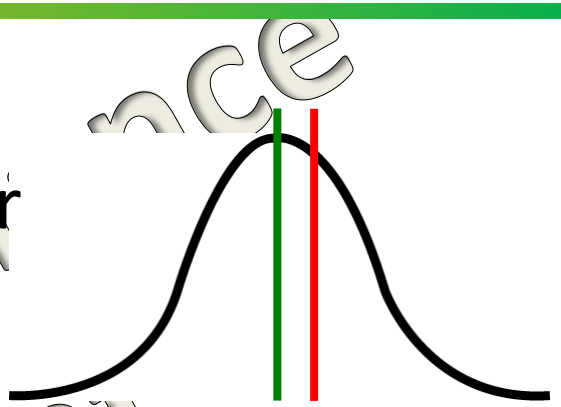
Accuracy

- Accuracy – how spread out the measurements are to each other, closeness to a reference value



Bias

- How well your measurements compare to a reference, standard or known value
  - Ideally, no difference between the measurement and the reference value
- Calibration is often performed to remove bias on a device or equipment
  - Only addresses one source of variation!
- **Example**: Thermometer is consistently 2 degrees higher than actual temperature



Example: Bias

THICKNESS OF PHONE IS KNOWN (REFERENCE) = 0.0041

BIAS

DEVICE

0.0046
0.0057
0.0038
0.0039
0.0050
0.0042
0.0036
0.0055

AVERAGE	STANDARD
0.0045	**0.0041**

BAD

NO BIAS

DEVICE

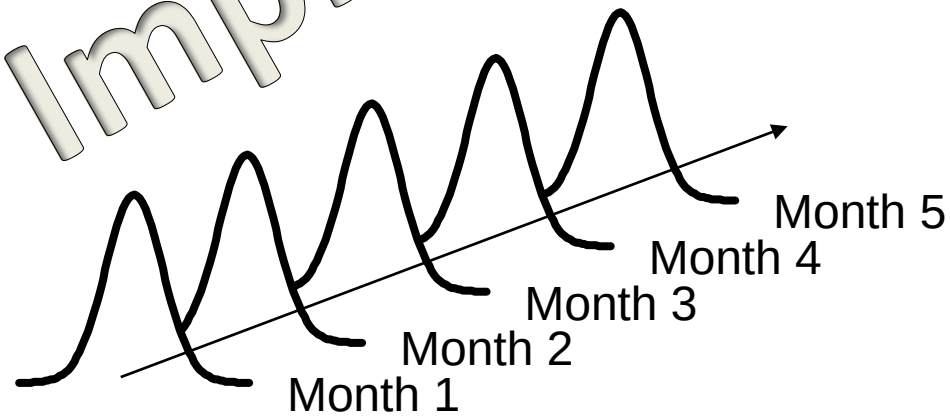
0.0043
0.0052
0.0031
0.0033
0.0045
0.0034
0.0039
0.0052

AVERAGE	STANDARD
0.0041	**0.0041**

GOOD

Stability

- The change in bias over time (drift)
  - Ideally, there should be no change in bias over time
  - Stability issues may increase or decrease the values over time
- Control charts are commonly used to track the stability of a measurement system over time
- **Example**: Thermometer performs well today, but gets progressively worse each month



Example: Stability

THICKNESS OF PHONE IS KNOWN (REFERENCE) = 0.0041

STABLE

DEVICE

Jan	0.0041
Feb	0.0041
Mar	0.0042
Apr	0.0041
May	0.0041
Jun	0.0042
Jul	0.0040
Aug	0.0041

GOOD

NOT STABLE

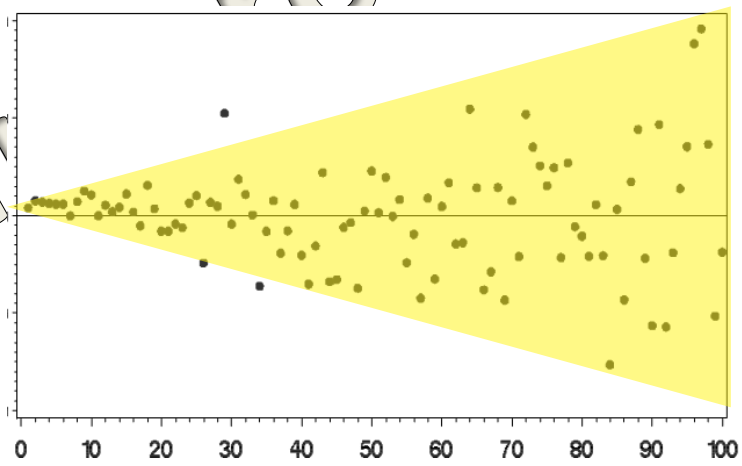
DEVICE

Jan	0.0041
Feb	0.0041
Mar	0.0042
Apr	0.0043
May	0.0045
Jun	0.0046
Jul	0.0047
Aug	0.0048

BAD

Linearity

- How accurate your measurements are through the expected range of measurements in which the device or instrument is intended to be used
  - Ideally, the measurement error will be the same across the range of likely values
  - Linearity often shows up as an increase in measurement error when measuring larger values
- **Example**: Thermometer is very good at low temperatures (around zero degrees C), but not as good near 100 degrees C or higher



Example: Linearity

COMPARE DIFFERENCE FROM STANDARD OVER RANGE OF VALUES

LINEAR

PART SIZE	DIFFERENCE FROM STANDARD
0.004	0.0001
0.005	0.0000
0.006	0.0002
0.007	0.0001
0.008	0.0001
0.009	0.0001
0.010	0.0000
0.015	0.0001

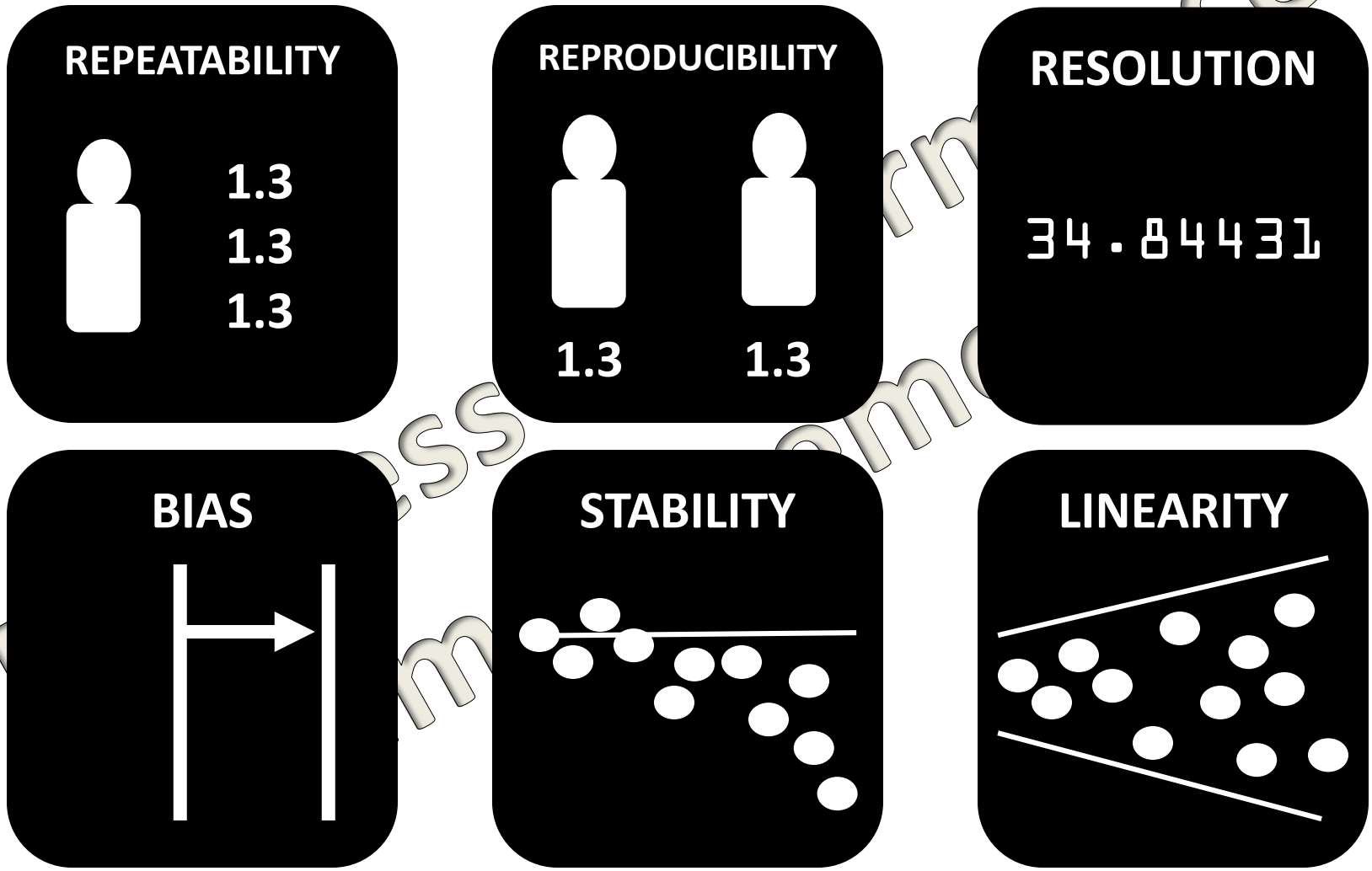
GOOD

NOT LINEAR

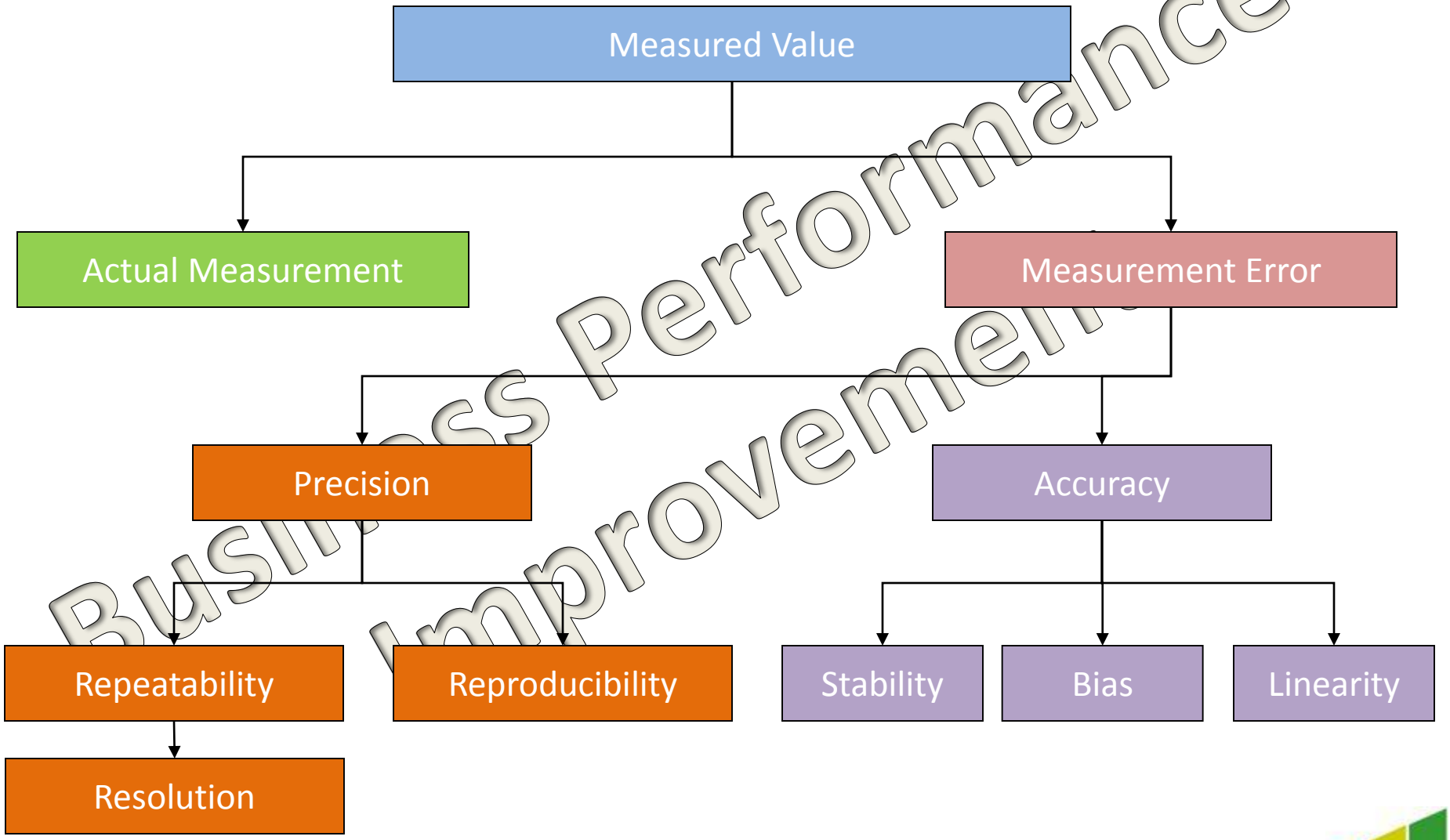
PART SIZE	DIFFERENCE FROM STANDARD
0.004	0.0001
0.005	0.0000
0.006	0.0002
0.007	0.0004
0.008	0.0005
0.009	0.0009
0.010	0.0010
0.015	0.0012

BAD

Summary of Variation Sources



Measurement Error



How to determine data validity?

- Lots of sources of measurement variation
- The most common drivers of measurement variation have been mentioned:
  - **Repeatability**
  - **Reproducibility**
  - Resolution
  - Bias
  - Stability
  - Linearity

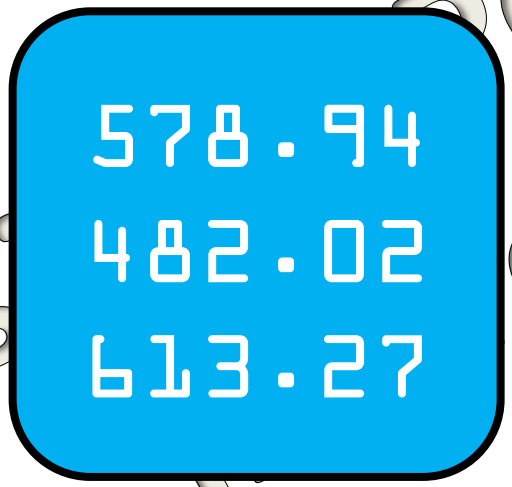
Gage R&R study

What is Gage R&R?

- Specialized experiment performed to check likely sources of measurement variation to determine whether the data is trustworthy
- R&R stands for Repeatability and Reproducibility
  - **Gage** = Process and devices used for collecting data
  - **Repeatability** = Differences between data points when you re-measure the same item
  - **Reproducibility** = Differences between people or devices when measuring the same item

Gage R&R depends on type of data

- In order to determine what type of Gage R&R to perform, need to know what type of data is being collected

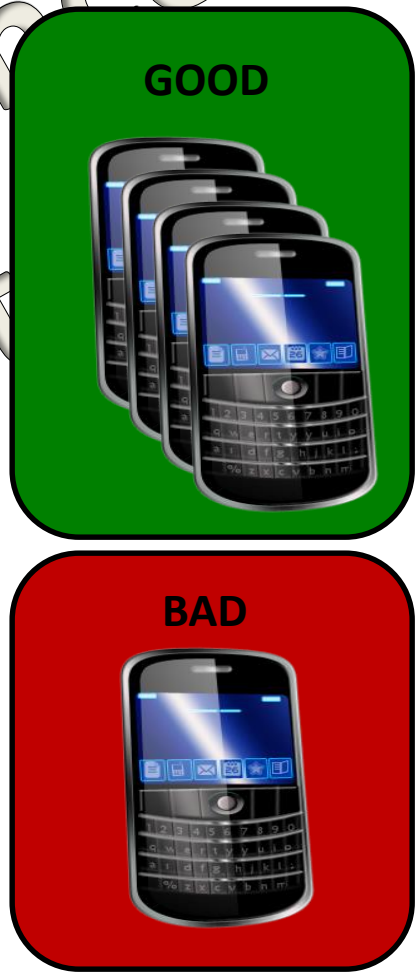


MEASUREMENTS
VARIABLE

BINARY
PASS/FAIL

Binary (Good/Bad)

- Based on individual decision whether something is acceptable or not (Go/No Go, Pass/Fail)
- Often expressed as a % of the total
  - **Delivery Success** (60%, where 12 were delivered on-time out of 20 total deliveries)
  - **Item Yields** (80%, where 4 were good out of 5 items tested)
  - **Categorization** (75%, where 3 out of 4 people recorded the item correctly)



Inspection Exercise

Count how often the 6th letter of the alphabet appears in the following paragraph:

The necessity of training farm hands for first class farms in the fatherly handling of farm live stock is foremost in the eyes of farm owners. Since the forefathers of the farm owners trained the farm hands for first class farms in the fatherly handling of farm live stock, the farm owners feel they should carry on with the family tradition of training farm hands of first class farmers in the fatherly handling of farm live stock because they believe it is the basis of good fundamental farm management.

Instructor: Answer available on hidden slide

Measurements (Variable)

- Usually requires a device in order to collect the data
- Can be expressed in decimal form
- Average and standard deviation can be calculated from results
  - **Temperature** (82.3 * C, thermometer)
  - **Speed** (72 MPH, speedometer)
  - **Weight** (35 kilograms, scale)
  - **Time** (4.4 seconds in 40-yard dash, stopwatch)
  - **Thickness** (13.55 cm, calipers)

Which one is best?

1. **Measurement**
   + Learn most with least amount of samples
   – Collecting data can take longer, cost of device
2. **Good /Bad**
   + Easiest to collect, better than no data
   – Requires lots of data points to understand results

Try to collect measurement data whenever possible!

Types of Gage R&R Studies

- **Variable**
  - Called “Variable Gage R&R”
  - Used with Measurement (variable) data
- **Attribute**
  - Called “Attribute Agreement Analysis” or “Attribute Gage R&R”
  - Used with Binary data

Variable Gage R&R Example

WHO WILL MEASURE

OR

ITEMS TO MEASURE

WHICH DEVICES

Exercise

Exercise: Measure how well you can estimate 10 seconds

1. Find a partner
2. Partner says “start” and starts stopwatch, tell partner “stop” when you think 10 seconds has elapsed. Repeat 6 times with each partner.
   - Make sure partner cannot see their results
3. Record all results, calculate average and standard deviation
   - What do you notice?

Gage R&R

- **Identify critical measurements**
  - Luminance of cell phone screen
- **Determine measurement method**
  - Photometer
- **Identify key variable (equipment, human, environment, etc)**
  - People can affect measurements more than equipment differences (use 3 in study)
- **Collect sample parts**
  - Identified 8 different cell phones to measure
  - **Goal is 30 or more observations**
- **Define repeatability plan**
  - Each person to measure each phone 2 times

Summary: 8 parts, 3 operators, 2 trials = 8 x 3 x 2 = 48 total measurements

Variable Gage R&R Example

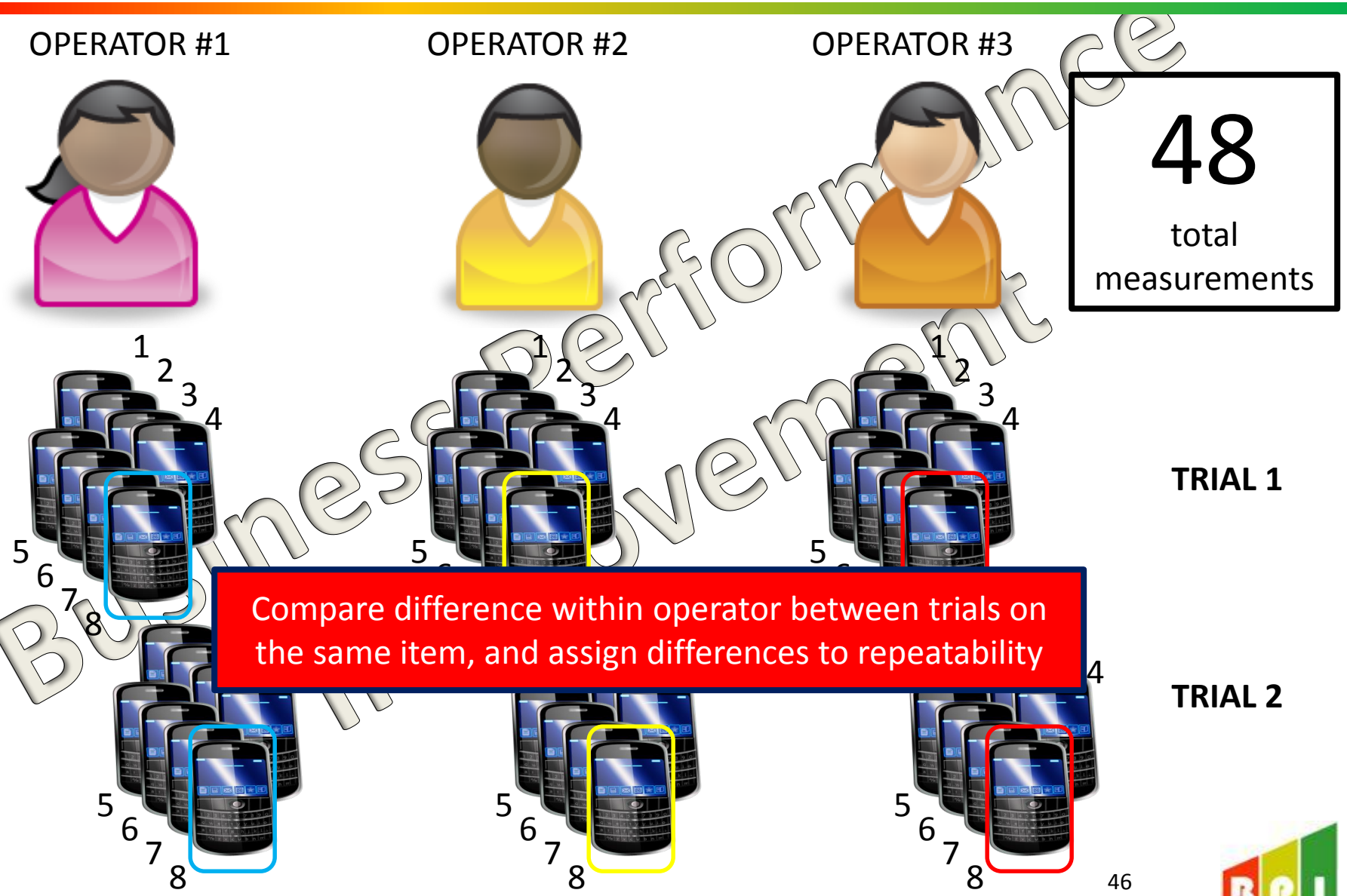
OPERATOR #1 | OPERATOR #2 | OPERATOR #3

48 total measurements

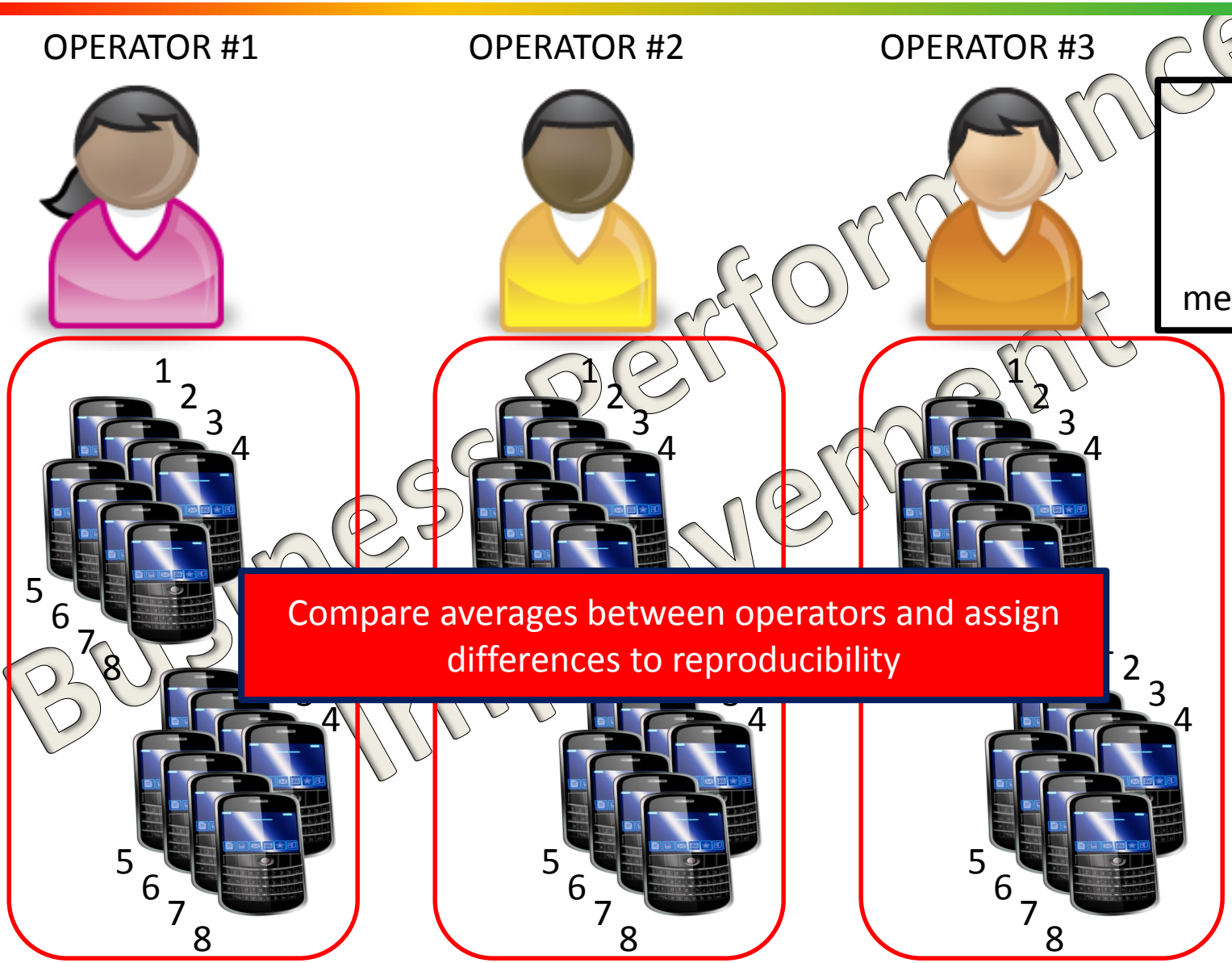
TRIAL 1

TRIAL 2

Repeatability



Reproducibility



Evaluation of Gage Study

- Two methods to evaluate Gage R&R studies
  - **% of study variation**
    - How much measurement variation is in the study
    - Helps determine if you can detect small changes
    - **Determines if we can detect trends of shifts in the process**
  - **% of tolerance**
    - How much measurement variation is in the study as a comparison to the tolerance width used in the process
    - Helps determine if the variation will impact the process
    - Only applies if the measurements have tolerances
    - **Contributes to risk of making correct pass/fail decision**

Results

Part	Repeat	OPERATOR		
		1	2	3
1	1	0.65	0.55	0.5
1	2	0.6	0.55	0.55
2	1	1	1.05	1.12
2	2	1	0.95	1
3	1	0.85	0.8	0.8
3	2	0.8	0.75	0.8
4	1	0.85	0.8	0.8
4	2	0.95	0.75	0.8
5	1	0.55	0.4	0.45
5	2	0.45	0.4	0.5
6	1	1	1	1
6	2	1	1.05	1.05
7	1	0.95	0.95	0.95
7	2	0.95	0.9	0.95
8	1	0.85	0.75	0.8
8	2	0.8	0.7	0.8

Tolerance

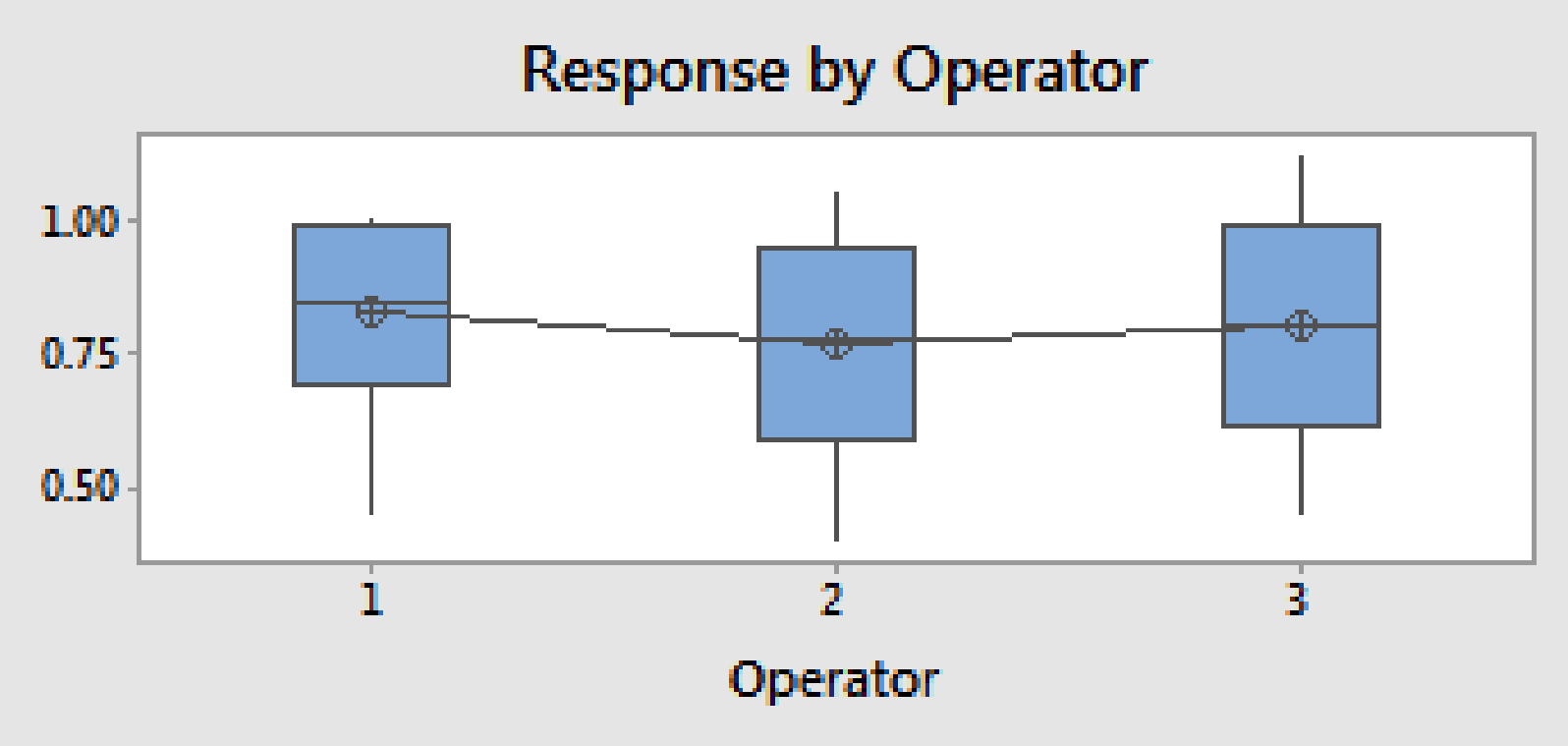
Lower Limit = 0.5

Upper Limit = 1.5

Review Result Averages

Part	Repeat	OPERATOR			Part Average
		1	2	3	
1	1	0.65	0.55	0.5	0.5667
1	2	0.6	0.55	0.55	
2	1	1	1.05	1.12	1.0200
2	2	1	0.95	1	
3	1	0.85	0.8	0.8	0.8000
3	2	0.8	0.75	0.8	
4	1	0.85	0.8	0.8	0.8250
4	2	0.95	0.75	0.8	
5	1	0.55	0.4	0.45	0.4583
5	2	0.45	0.4	0.5	
6	1	1	1	1	1.0167
6	2	1	1.05	1.05	
7	1	0.95	0.95	0.95	0.9417
7	2	0.95	0.9	0.95	
8	1	0.85	0.75	0.8	0.7833
8	2	0.8	0.7	0.8	
Operator Average		0.828125	0.771875	0.804375	**0.8015**

Reproducibility



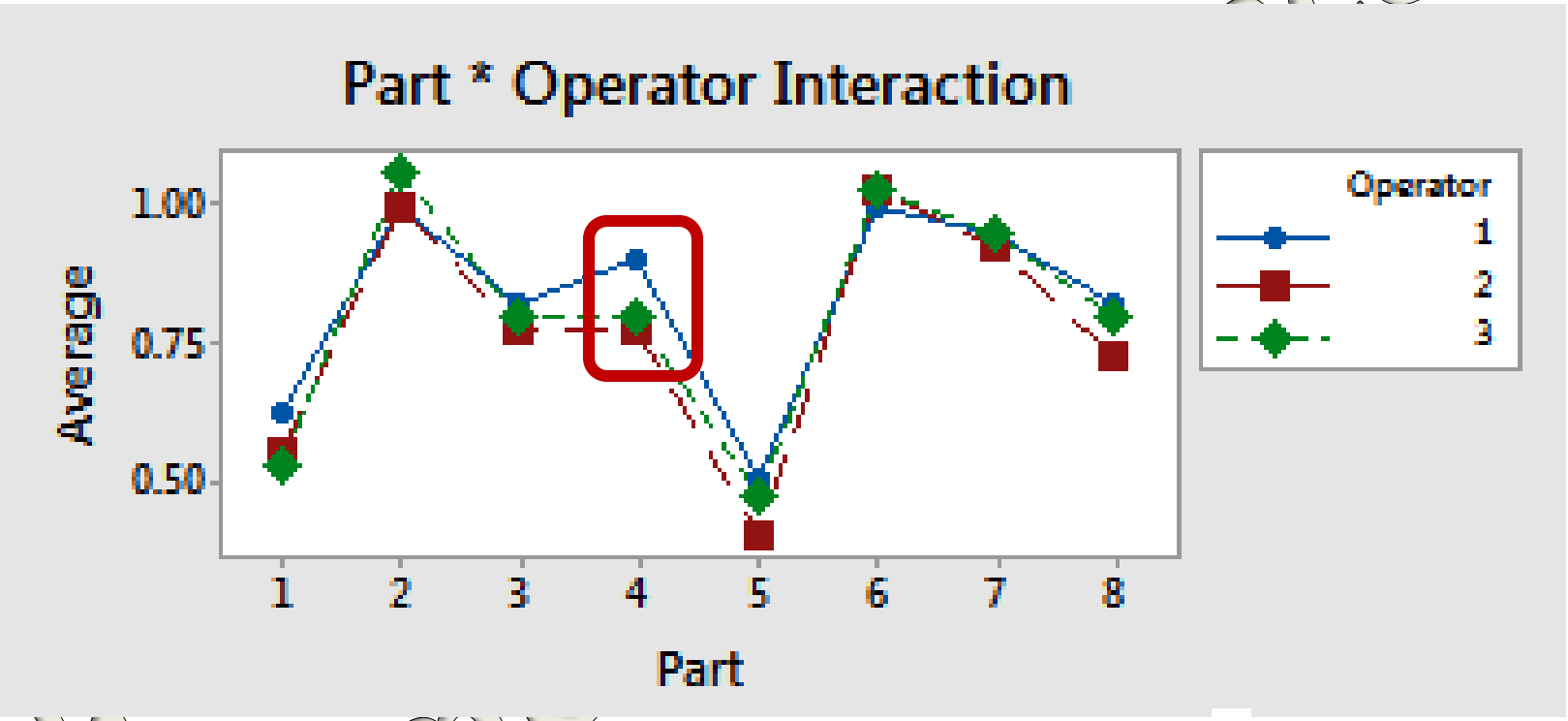
0.828125	0.771875	0.804375

Blue box represents spread of data (parts and repeatability)

Review Result Averages

Part	Repeat	OPERATOR			Part Average
		1	2	3	
1	1	0.65	0.55	0.5	0.5667
1	2	0.6	0.55	0.55	
2	1	1	1.05	1.12	1.0200
2	2	1	0.95	1	
3	1	0.85	0.8	0.8	0.8000
3	2	0.8	0.75	0.8	
4	1	0.85	0.8	0.8	0.8250
4	2	0.95	0.75	0.8	
5	1	0.55	0.4	0.45	0.4583
5	2	0.45	0.4	0.5	
6	1	1	1	1	1.0167
6	2	1	1.05	1.05	
7	1	0.95	0.95	0.95	0.9417
7	2	0.95	0.9	0.95	
8	1	0.85	0.75	0.8	0.7833
8	2	0.8	0.7	0.8	
Operator Average		0.828125	0.771875	0.804375	**0.8015**

Reproducibility



0.567	1.02	0.8	0.825	0.458	1.017	0.942	0.783

Start to see patterns and differences in results

Standard Deviation

Source	StdDev (SD)	Study Var (6 × SD)	%Study Var (%SV)	%Tolerance (SV/Toler)
Total Gage R&R	0.050096	0.30057	24.08	30.06
Repeatability	0.042735	0.25641	20.54	25.64
Reproducibility	0.026139	0.15683	12.57	15.68
Operator	0.026139	0.15683	12.57	15.68
Part-To-Part	0.201901	1.21141	97.06	121.14
Total Variation	0.208024	1.24814	100.00	124.81

Calculated from data, based on repeat, parts and operator

Study Variation

Source	StdDev (SD)	Study Var (6 × SD)	%Study Var (%SV)	%Tolerance (SV/Toler)
Total Gage R&R	0.050096	0.30057	24.08	30.06
Repeatability	0.042735	0.25641	20.54	25.64
Reproducibility	0.026139	0.15683	12.57	15.68
Operator	0.026139	0.15683	12.57	15.68
Part-To-Part	0.201901	1.21141	97.06	121.14
Total Variation	0.208024	1.24814	100.00	124.81

Multiply by 6 to estimate spread of distribution (+/- 3 std devs)

% Study Variation

Source	StdDev (SD)	Study Var (6 x SD)	%Study Var (%SV)	%Tolerance (SV/Toler)
Total Gage R&R	0.050096	0.30057	24.08	30.06
Repeatability	0.042735	0.25641	20.54	25.64
Reproducibility	0.026139	0.15683	12.57	15.68
Operator	0.026139	0.15683	12.57	15.68
Part-To-Part	0.201901	1.21141	97.06	121.14
Total Variation	0.208024	1.24814	100.00	124.81

Total Gage R&R = 0.30057 / 1.24814 = 0.2408 x 100% = **24.08%**

Dividing each "Study Var" by "Total Variation"

% Tolerance

Source	StdDev (SD)	Study Var (6 × SD)	%Study Var (%SV)	%Tolerance (SV/Toler)
Total Gage R&R	0.050096	0.30057	24.08	30.06
Repeatability	0.042735	0.25641	20.54	25.64
Reproducibility	0.026139	0.15683	12.57	15.68
Operator	0.026139	0.15683	12.57	15.68
Part-To-Part	0.201901	1.21141	97.06	121.14
Total Variation	0.208024	1.24814	100.00	124.81

% of Tolerance = Study Var / Spread of Tolerance

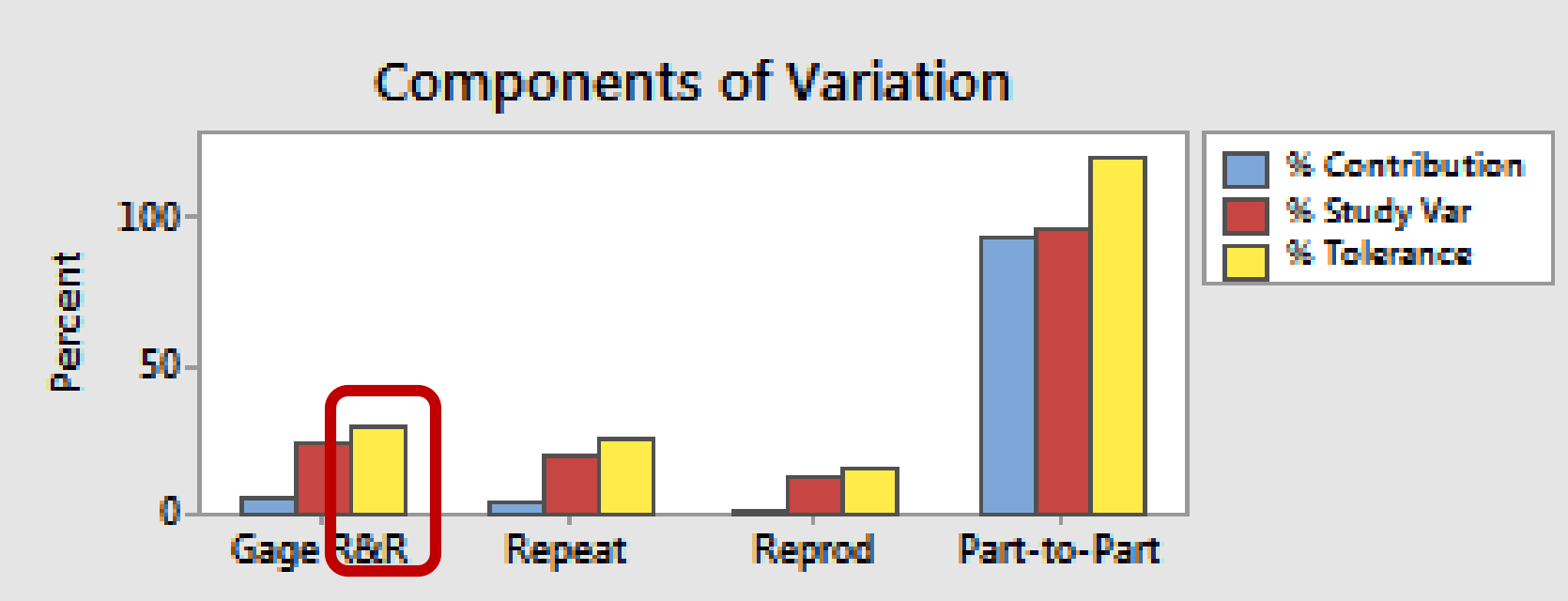
Tolerance is 0.5 to 1.5 (provided on part requirement sheet)

Spread of tolerance = 1.5 – 0.5 = **1.0**

Total Gage R&R % of Tolerance = 0.25641 / 1.0 = 0.2564 = **25.64%**

Dividing each "Study Var" by "Spread of Tolerance"

Summary of Variation %



Breakdown	Gage R&R	Repeat	Reprod	Part-to-Part
% Study Var	24.08%	20.54%	12.57%	97.06%
% Tolerance	**30.06%**	25.64%	15.68%	121.14%

Criteria for Tolerance/Study Variation %

EXCELLENT

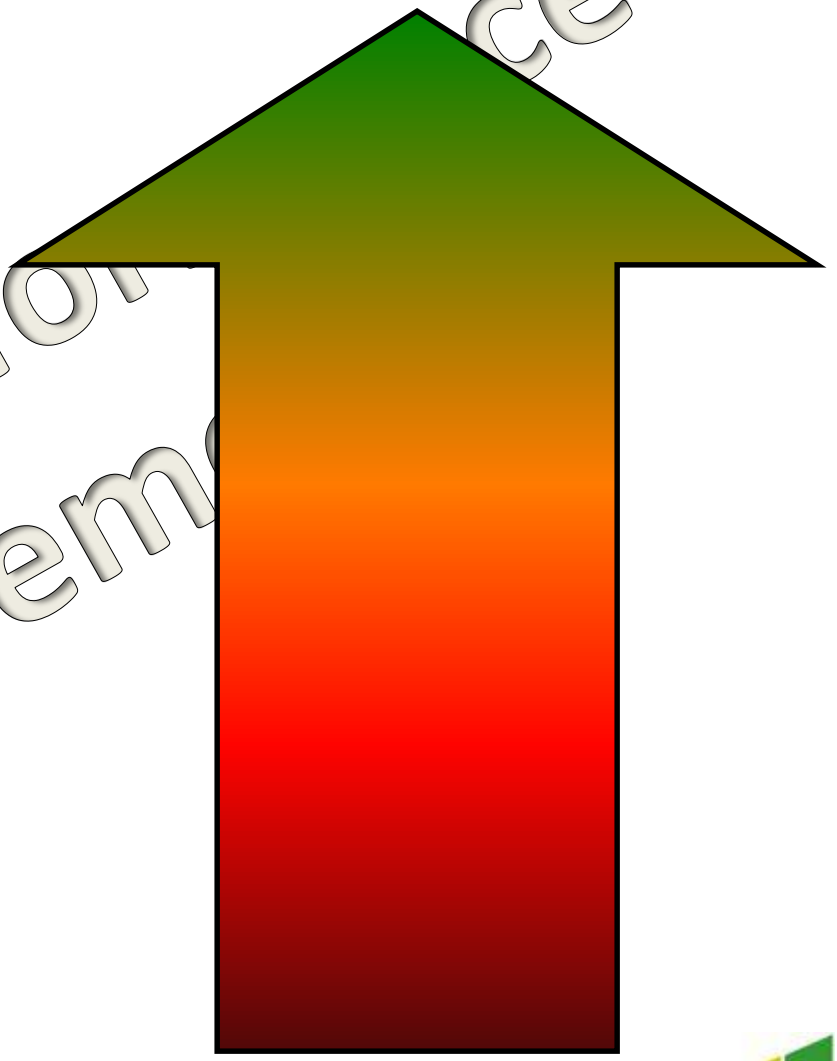
Less than 10%

MARGINAL

Between 10% and 30%

POOR

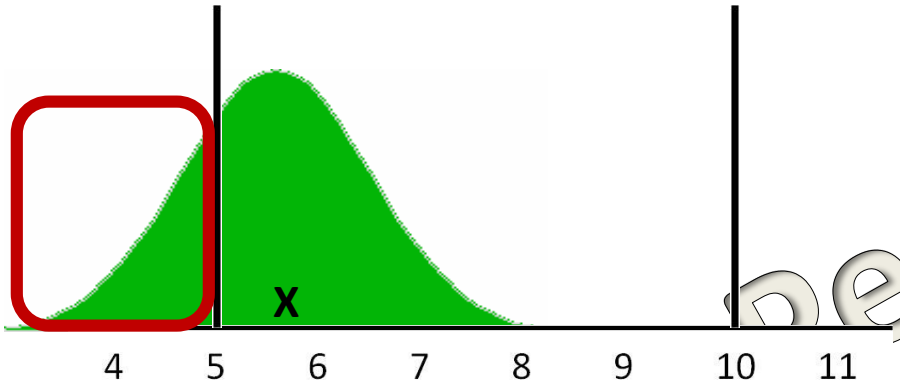
Greater than 30%



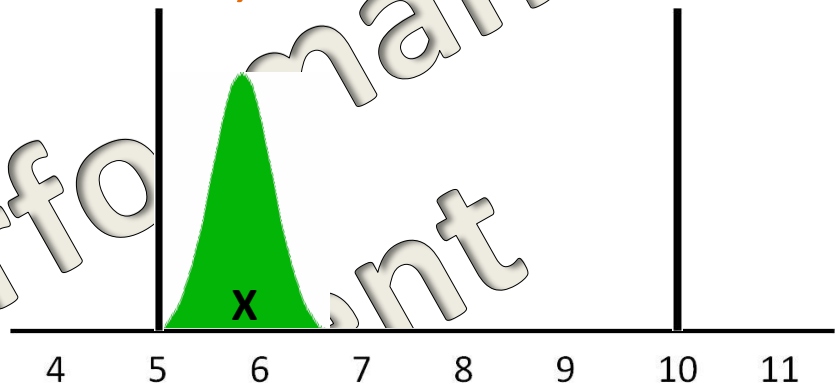
Applies to “% of Study Variation” and “% of Tolerance”

Is this a problem?

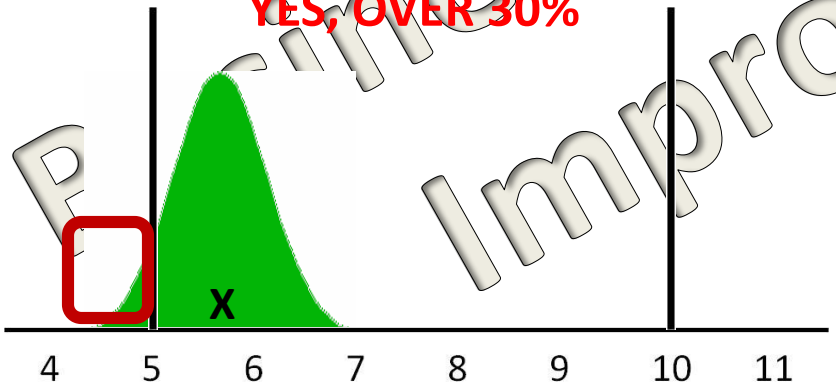
100%
YES, OVER 30%



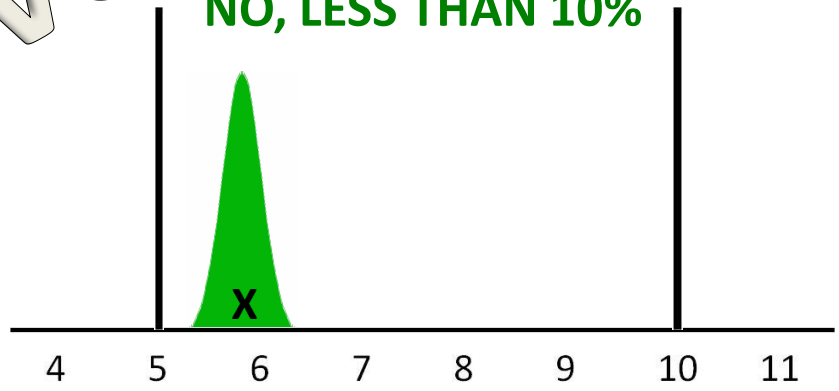
20%
MAYBE, BETWEEN 10 AND 30%



50%
YES, OVER 30%

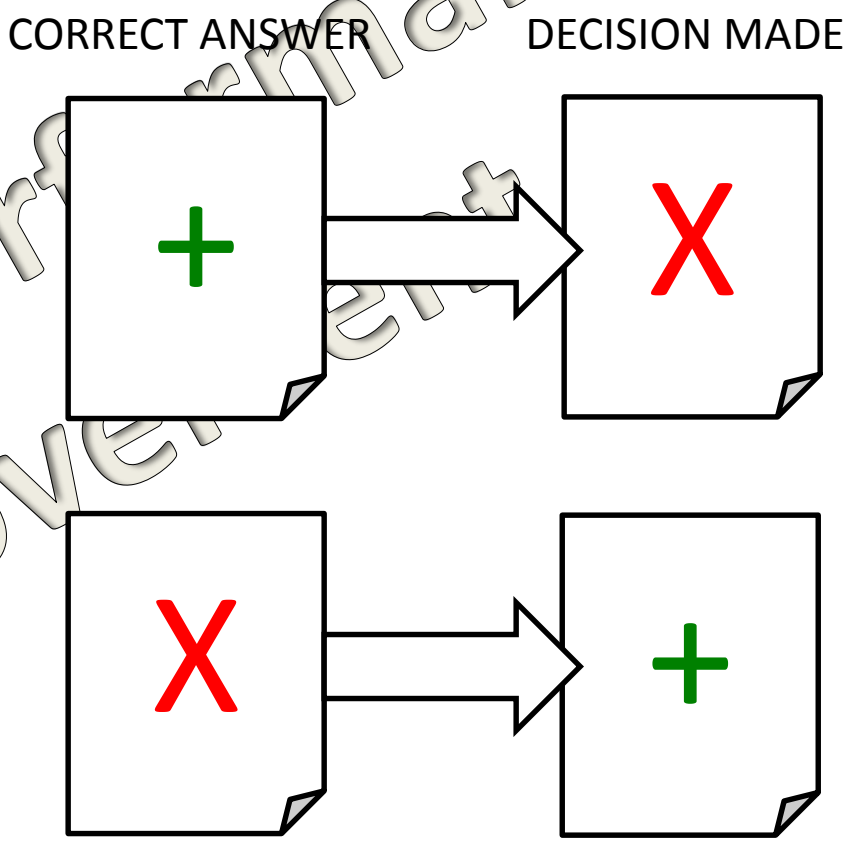


5%
NO, LESS THAN 10%



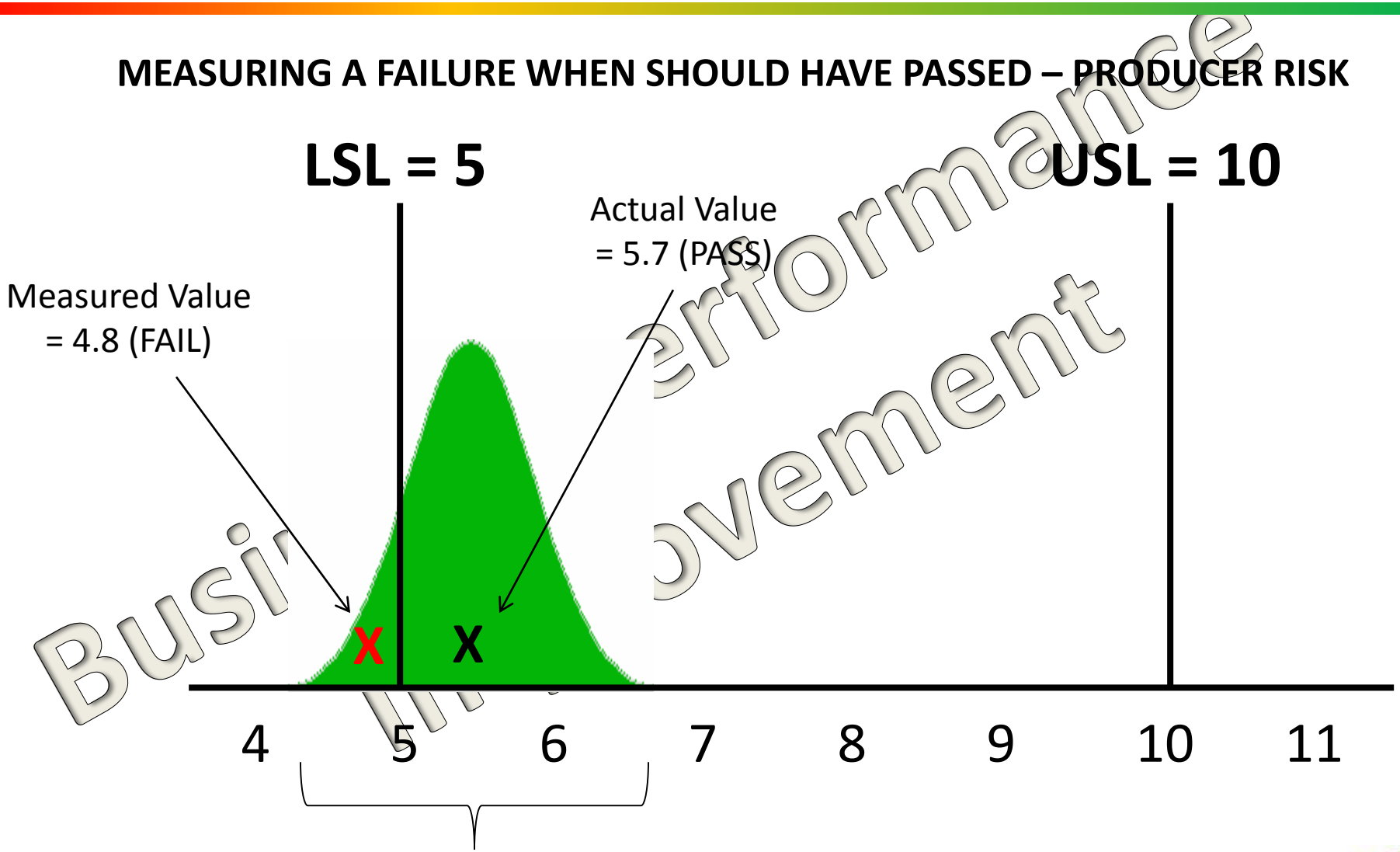
Incorrect Decisions

- Two types of mistakes
  - **Type I error**
    - Producer Risk
    - Good item called bad
  - **Type II error**
    - Consumer Risk
    - Bad item called good



Type I Error

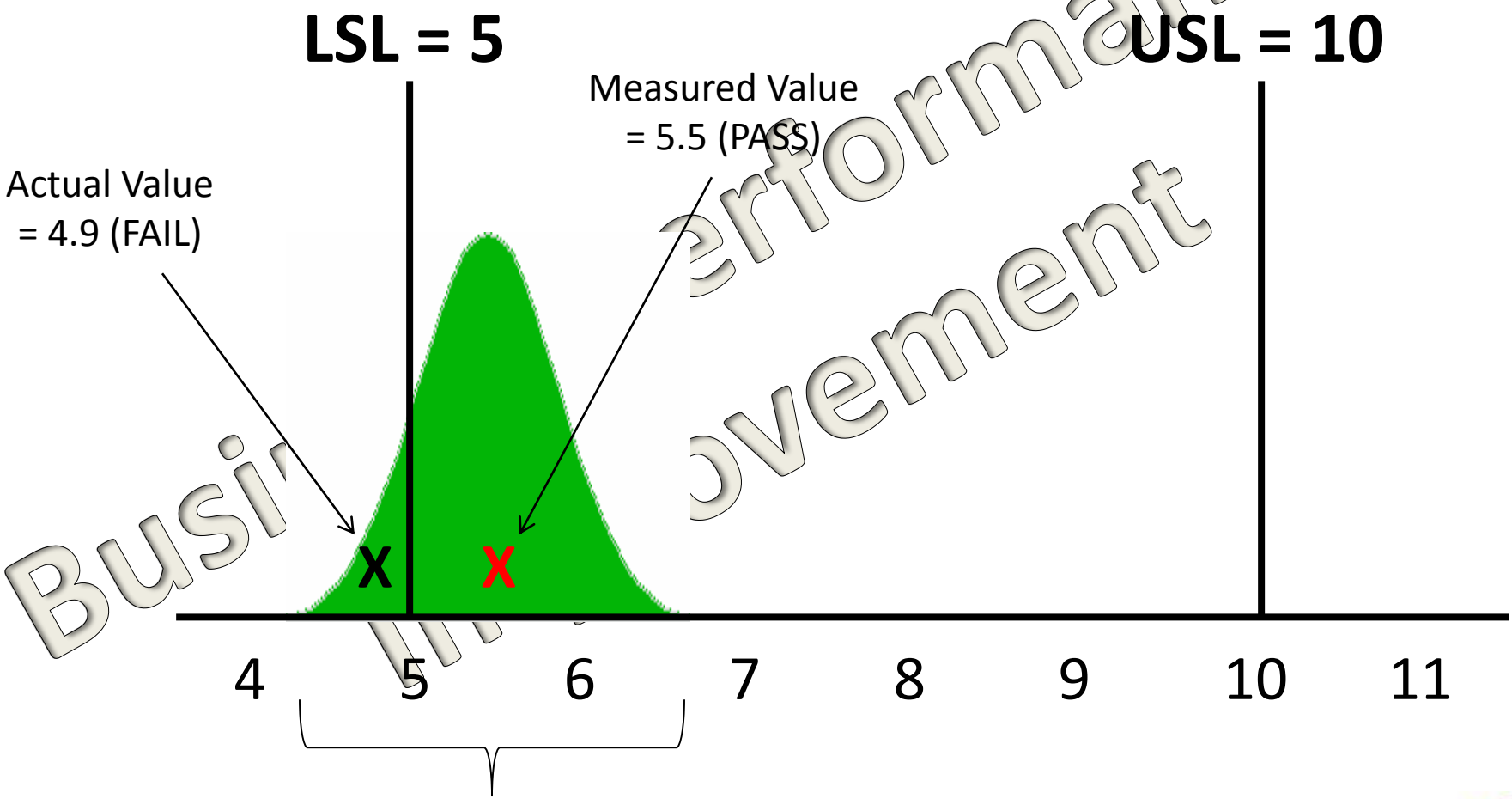
MEASURING A FAILURE WHEN SHOULD HAVE PASSED – PRODUCER RISK



33% measurement error

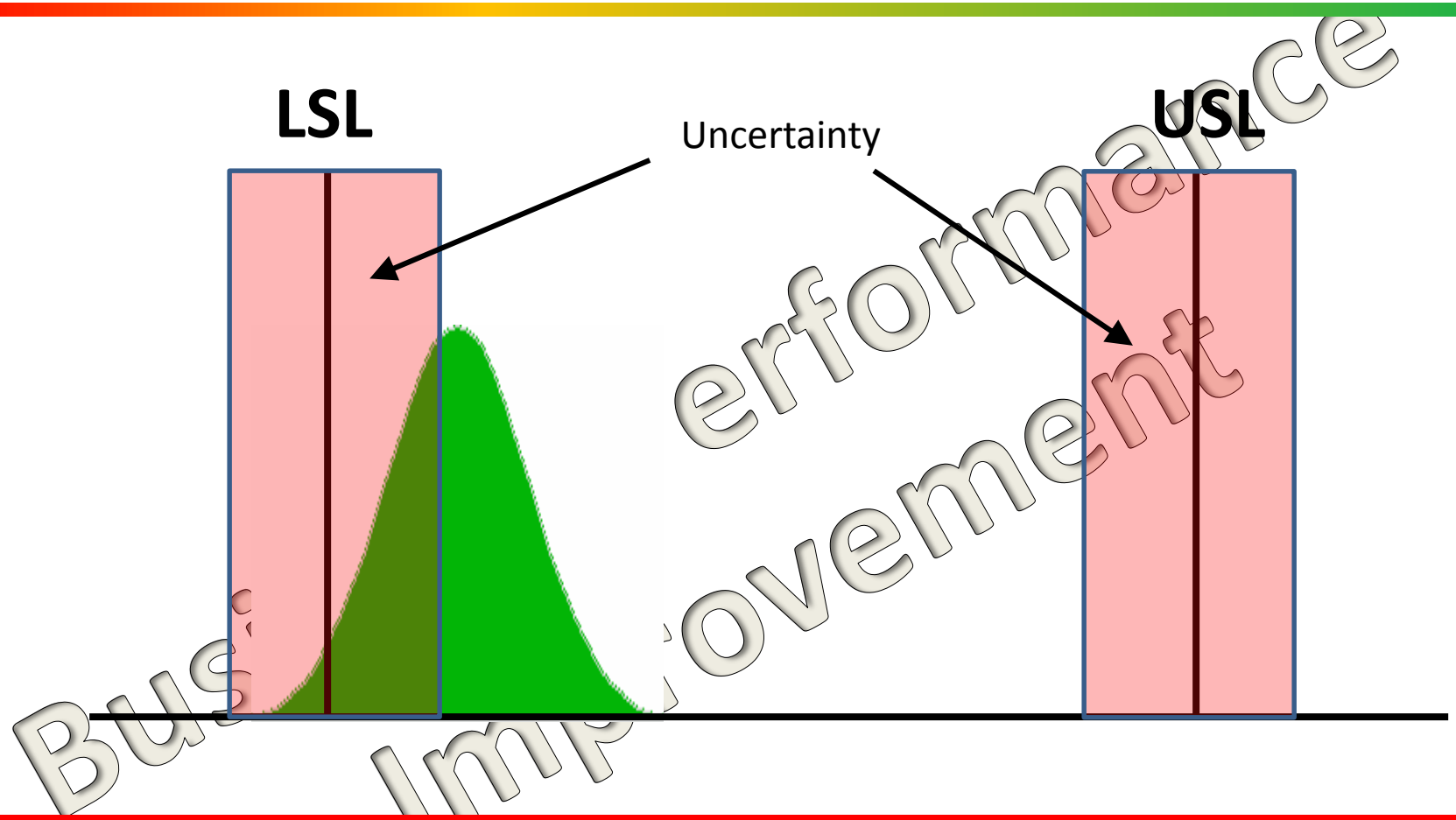
Type II Error

MEASURING A PASS WHEN SHOULD HAVE FAILED – CONSUMER RISK



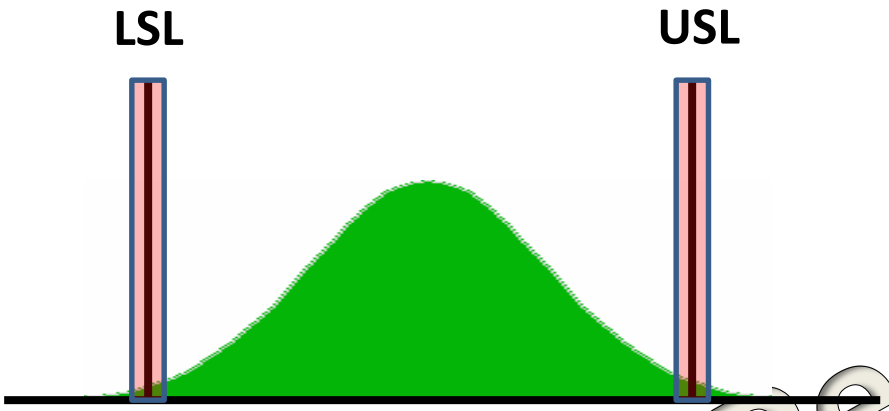
33% measurement error

Measurement Error vs Spec Limits



Measurements that falls within the uncertainty area could lead to Type I or II errors

Measurement Error vs Uncertainty

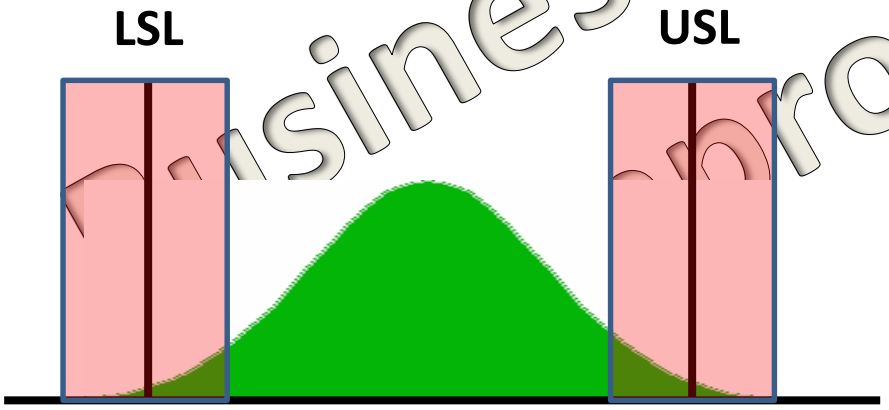


Gage R&R% of Tolerance

1%

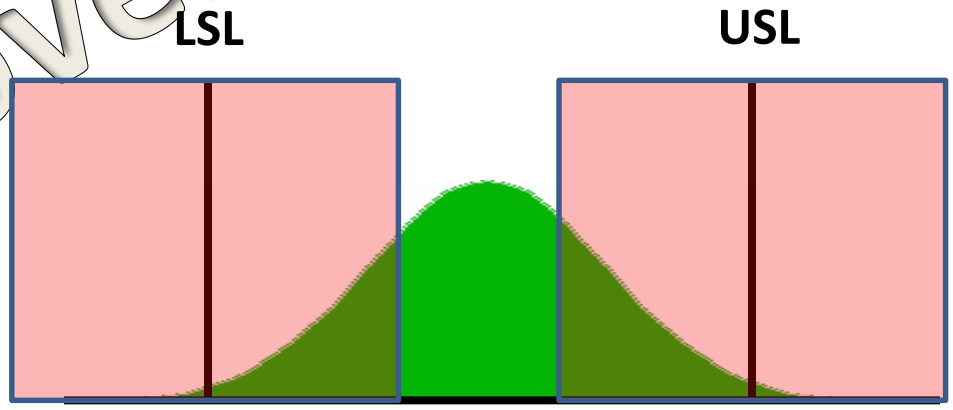
Gage R&R% of Tolerance

10%



Gage R&R% of Tolerance

30%



Gage R&R% of Tolerance

80%

What if it is “Marginal”?

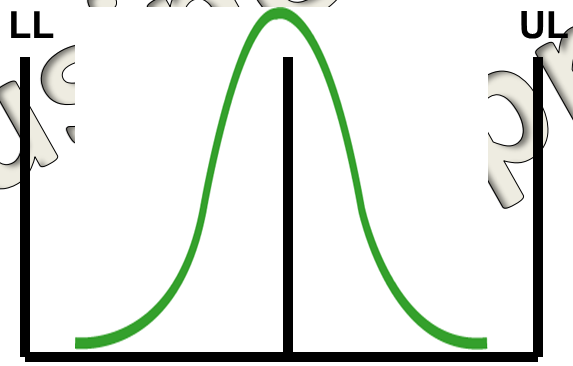
If you are “marginal”, there are two options

- Improve it, because it is not yet “Excellent”

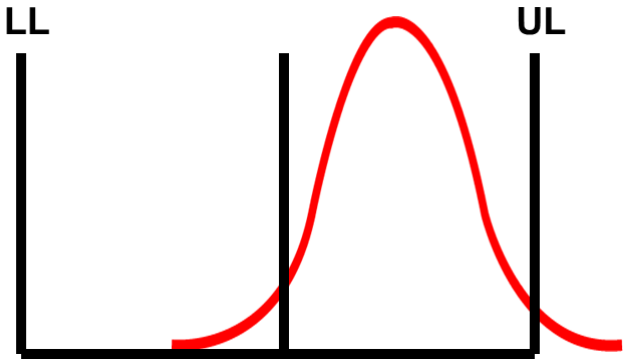
OR

- Decide based on capability analysis

If capability (Cpk/Ppk) > 1.33, then you **might not** have to improve it

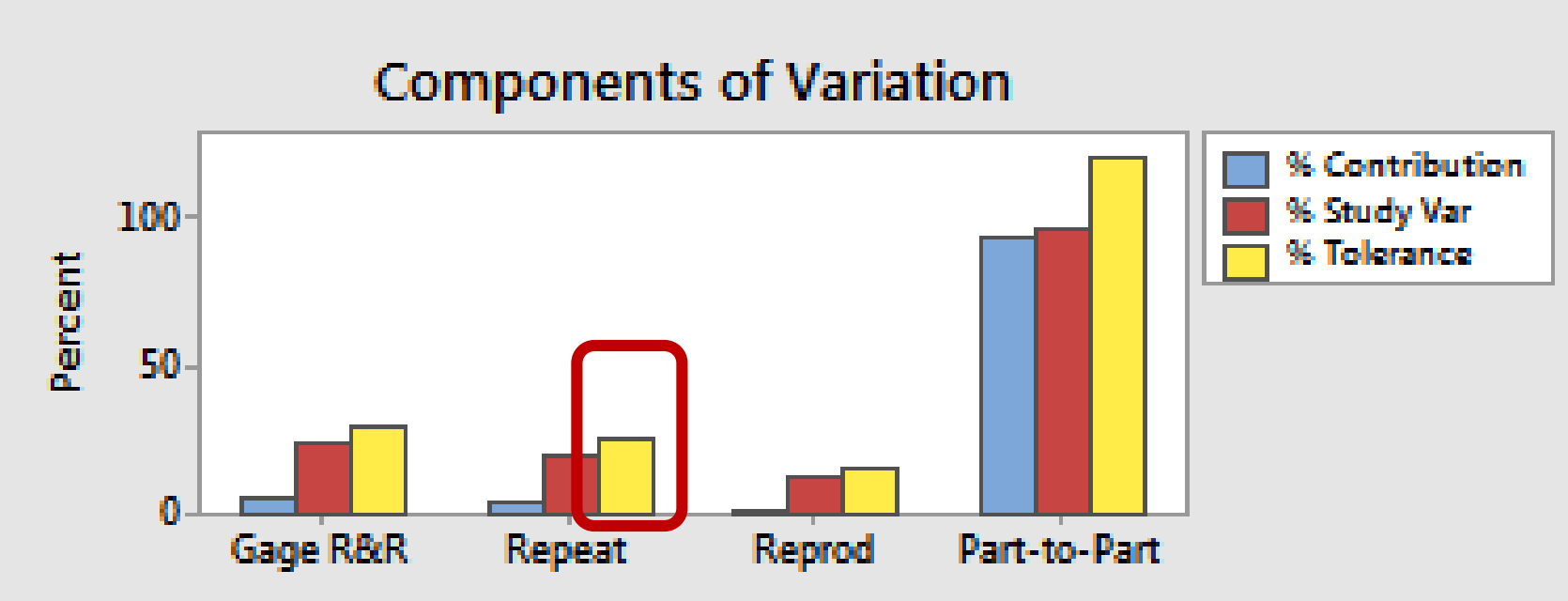


If capability (Cpk/Ppk) < 1.33, then you **should** improve it



[Learn more about Capability in our Powerpoint course >>>]

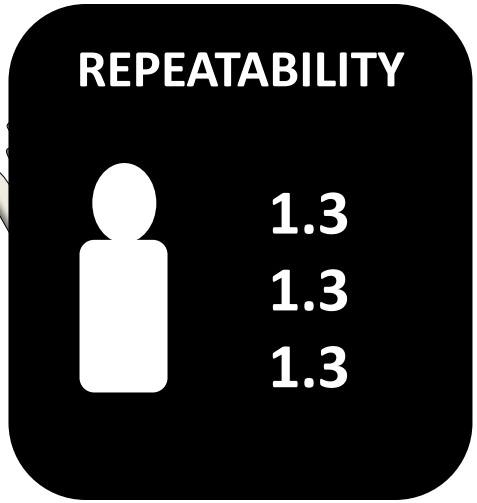
Repeatability or Reproducibility?



Breakdown	Gage R&R	Repeat	Reprod	Part-to-Part
% Study Var	24.08%	20.54%	12.57%	97.06%
% Tolerance	**30.06%**	25.64%	15.68%	121.14%

How to Improve Repeatability

- **Repeatability**
  - Repeat measurements multiple times and use the average result (not the individual results)
  - Utilize measurement devices with less measurement variation
  - Standardize measurement process and documentation for the individual
  - Standardize location of measurements being taken on part, circuitry, angle of measurement, document, etc.



How to Improve Reproducibility

- **Reproducibility**
  - Standardize devices allowable for use (or reduce options)
  - Re-calibrate assessment devices to make certain they are not biased, or misaligned, and implement PM or calibration schedule to prevent future issues
  - Standardize measurement process and documentation, and ensure all are trained properly
  - Standardize location of measurements being taken on part, circuitry, angle of measurement, document, etc.

How many runs do I need?

- Minimum numbers recommended for running a Gage R&R (for **variable** measurements – not attribute Gage R&R)

Parts	Operators	Repetitions	Reference Total
8	2	2	>30
5	3	2	30
5	2	3	30

- Minimum numbers recommended for running a Gage R (**no reproducibility**)

Parts	Operators	Repetitions	Reference Total
10	1	3	30
8	1	4	>30

Attribute Gage R&R

- Used for binary (pass/fail) data
- Also called “Attribute Agreement Analysis”
- Examples:
  - Do operators select the correct defect code?
  - Do analysts find the same errors in the file?
  - Do bankers accept or reject the same applications?
  - Do managers give the same assessment score to candidates after interview?
  - Can doctors identify a known disease?
- **What other examples can you think of?**

Attribute Agreement Analysis

WHO WILL MEASURE

OR

ITEMS TO MEASURE

WHICH DEVICES

Attribute Agreement Analysis

How does it work?

- Identify good/bad criteria or count criteria
  - Binary: Does phone power up correctly?
  - Count: How many defects were found?
- Determine measurement method
  - Binary: Visual
  - Count: Magnification scope
- Identify key variable (equipment, human, environment, etc)
  - Inspectors can affect outcomes more than magnification scope (use 3 inspectors in study).
- Collect sample parts
  - Identified 8 different phones to inspect
- Define repeatability plan
  - Each operator to inspect each phone 2 times

- Summary: 8 parts, 3 operators, 2 trials = 8 x 3 x 2 = 48 total assessments

Attribute Agreement Analysis

Part	Repeat	OPERATOR			Part Agreement %
		1	2	3	
1	1	GOOD	GOOD	GOOD	83%
1	2	BAD		GOOD	
2	1	BAD		GOOD	67%
2	2	BAD	BAD	GOOD	
3	1	BAD	BAD	BAD	100%
3	2	BAD	BAD	BAD	
4	1	GOOD	GOOD	GOOD	100%
4	2	GOOD		GOOD	
5	1	BAD		GOOD	50%
5	2	GOOD	GOOD	BAD	
6	1	BAD	BAD	BAD	83%
6	2	BAD	BAD	GOOD	
7	1	GOOD	GOOD	GOOD	83%
7	2	GOOD	GOOD	BAD	
8	1	BAD			00%
8	2	BAD			
Operator Agreement %		75%	88%	63%	**38%**

NOT AGREE

AGREE

6 agree, 2 not agree = 75%

Attribute Agreement Analysis

Part	Repeat	OPERATOR 1	OPERATOR 2	OPERATOR 3	Part Agreement %
1	1	GOOD	GOOD	GOOD	83%
1	2	BAD	GOOD	GOOD	
2	1	BAD	BAD	GOOD	67%
2	2	BAD	BAD	GOOD	
3	1	BAD	BAD	BAD	100%
3	2	BAD	BAD	BAD	
4	1	GOOD	GOOD	GOOD	100%
4	2	GOOD	GOOD	GOOD	
5	1	BAD	BAD	GOOD	50%
5	2	GOOD	GOOD	BAD	
6	1	BAD	BAD	BAD	83%
6	2	BAD	BAD	GOOD	
7	1	GOOD	GOOD	GOOD	83%
7	2	GOOD	GOOD	BAD	
8	1	BAD	BAD	BAD	100%
8	2	BAD	BAD	BAD	
Operator Agreement %		75%	88%	63%	**38%**

5 out of 6 assessments agree = 83%

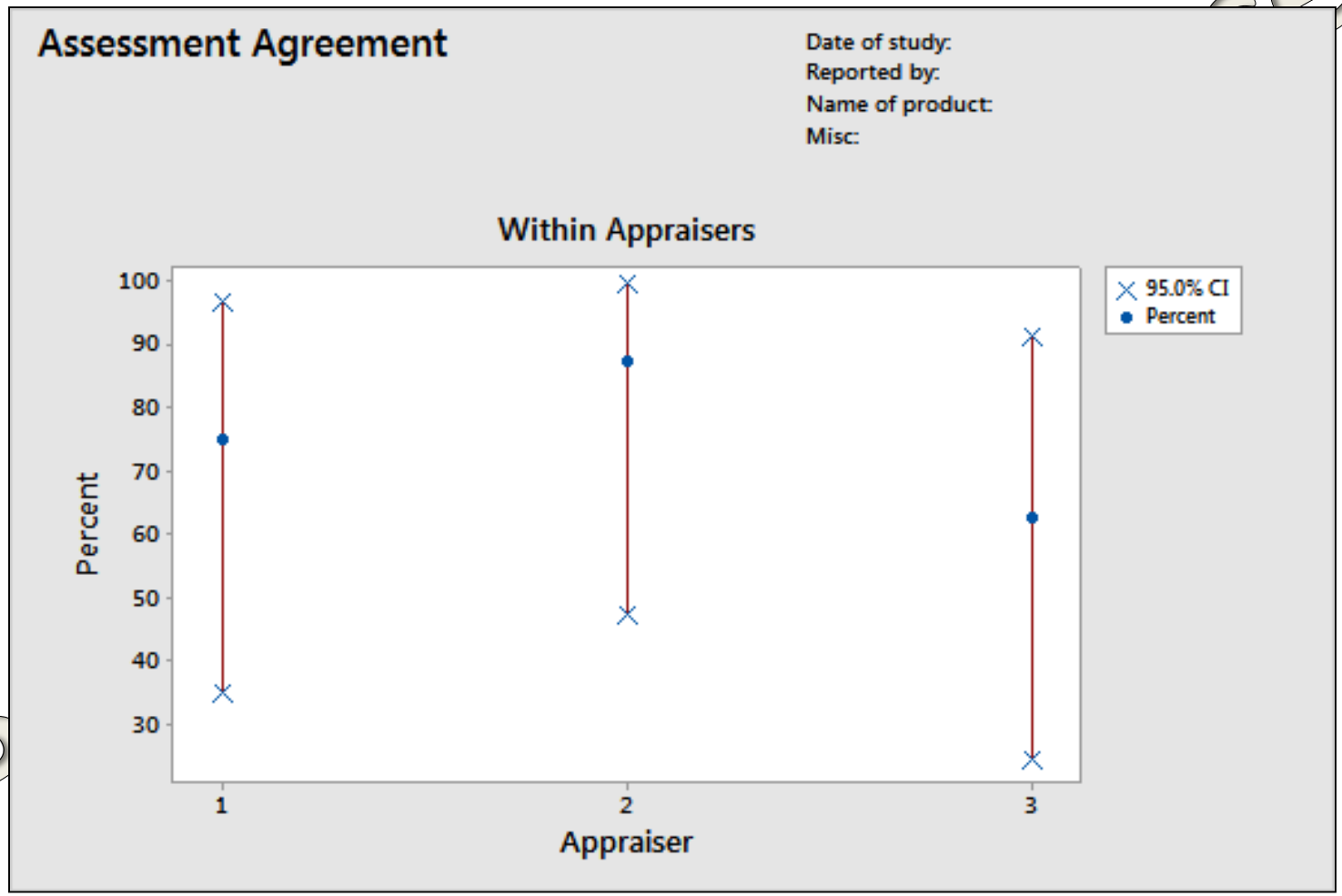
6 out of 6 assessments agree = 100%

Attribute Agreement Analysis

Part	Repeat	OPERATOR			Part Agreement %
		1	2	3	
1	1	GOOD	GOOD	GOOD	83%
1	2	BAD	GOOD	GOOD	
2	1	BAD	BAD	GOOD	67%
2	2	BAD	BAD	GOOD	
3	1	BAD	BAD	BAD	100%
3	2	BAD	BAD	BAD	
4	1	GOOD	GOOD	GOOD	100%
4	2	GOOD	GOOD	GOOD	
5	1	BAD	BAD	GOOD	50%
5	2	GOOD	GOOD	BAD	
6	1	BAD	BAD	BAD	83%
6	2	BAD	BAD	GOOD	
7	1	GOOD	G	D	83%
7	2	GOOD	G		
8	1	BAD			100%
8	2	BAD			
Operator Agreement %		75%	88%	63%	**38%**

3 out of 8 assessments completely agree = 38%

% Agreement



75%	88%	63%

Kappa Values

- Kappa value = the degree of agreement made by multiple appraisers when assessing the same samples/parts
- How much better is your assessment compared to guessing
  - If you flipped a coin and you guessed heads or tails, you would be right about 50% of the time by chance
    - **Kappa = 0** means that you were equal to random chance (50%)
    - **Kappa < 0** means that you were worse than random chance (less than 50% correct)
    - **Kappa > 0** means that you were better than random chance (more than 50% correct)
    - **Kappa = 1** means that you were correct 100% of the time

Kappa Values

EXCELLENT

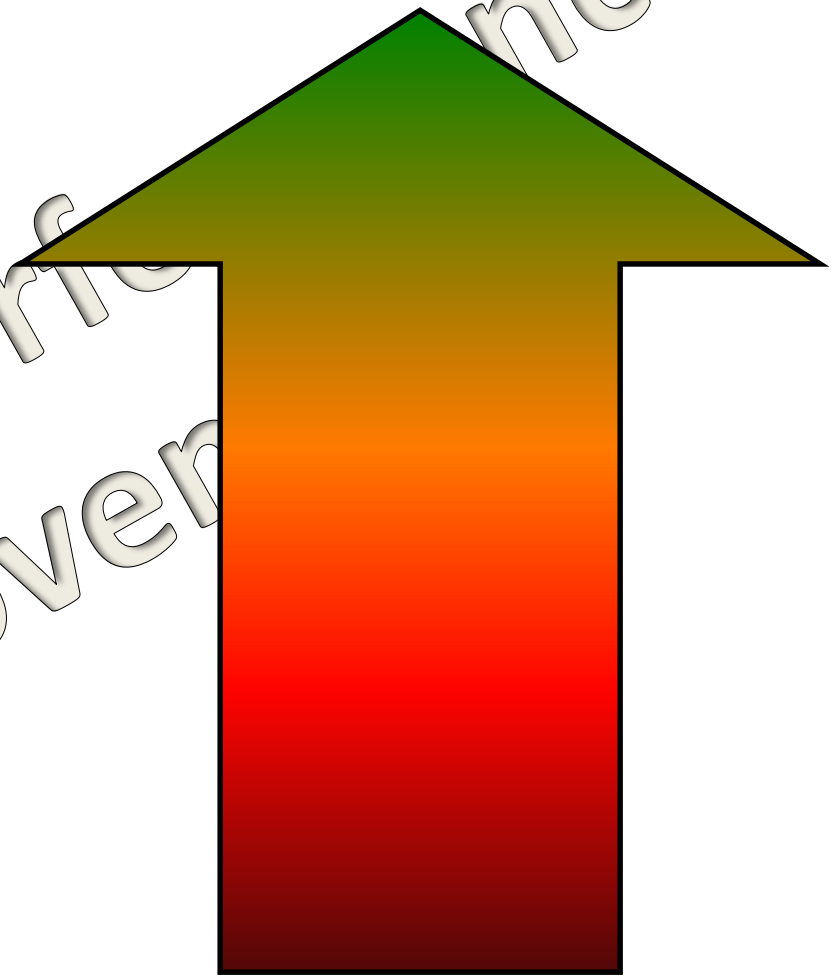
Greater than 0.9

MARGINAL

Between 0.7 and 0.9

POOR

Less than 0.7



http://www.sigmaxl.com/AttributeMSABinary.shtml

Calculating Kappa

- The formula for kappa is:

$$\frac{P_o - P_e}{1 - P_e}$$

P_o = Probability observed

P_e = Probability expected

Calculating Kappa for Coin Flips

Ex: Flipping a coin 100 times

- 45 heads and 55 tails
- $P_o = 45/100 = .45$
- $P_e = 50/100 = .50$

$$\frac{P_o - P_e}{1 - P_e}$$

Kappa = (0.45 – 0.50) / (1 – 0.5)

Kappa = -0.05 / 0.5 = -0.1

Kappa is near zero, so it matches close enough to our expectations

Kappa Values

Fleiss' Kappa Statistics

Appraiser	Response	Kappa	SE Kappa	Z	P(vs > 0)
1	BAD	0.466667	0.353553	1.31993	0.0934
	GOOD	0.466667	0.353553	1.31993	0.0934
2	BAD	0.746032	0.353553	2.11010	0.0174
	GOOD	0.746032	0.353553	2.11010	0.0174
3	BAD	0.238095	0.353553	0.67344	0.2503
	GOOD	0.238095	0.353553	0.67344	0.2503

- Appraiser 2 is marginal at separating GOOD vs BAD
- Appraiser 1 is poor at separating GOOD vs BAD
- Appraiser 3 is poor at separating GOOD vs BAD

http://www.sigmaxl.com/AttributeMSABinary.shtml

Back to Results

Part	Repeat	OPERATOR			Part Agreement %
		1	2	3	
1	1	GOOD	GOOD	GOOD	83%
1	2	BAD	GOOD	GOOD	
2	1	BAD	BAD	GOOD	67%
2	2	BAD	BAD	GOOD	
3	1	BAD	BAD	BAD	100%
3	2	BAD	BAD	BAD	
4	1	GOOD	GOOD	GOOD	100%
4	2	GOOD	GOOD	GOOD	
5	1	BAD	BAD	GOOD	50%
5	2	GOOD	GOOD	BAD	
6	1	BAD	BAD	BAD	83%
6	2	BAD	BAD	GOOD	
7	1	GOOD	GOOD	GOOD	83%
7	2	GOOD	GOOD	BAD	
8	1	BAD	BAD	BAD	100%
8	2	BAD	BAD	BAD	
Operator Agreement %		75%	88%	63%	**38%**

Kappa Values

Fleiss' Kappa Statistics

Response	Kappa	SE Kappa	Z	P(vs > 0)
BAD	0.462937	0.0912871	5.07122	0.0000
GOOD	0.462937	0.0912871	5.07122	0.0000

Overall system needs improvement!

Key Points

- Without performing a MSA, you cannot prove that your measurements are valid
- Calibration only removes bias, not other sources of measurement variation
- Gather 30 or more measurements for your gage study (ex: 5 parts, 3 people/devices and 3 repeats)
- Evaluate your gage results as you go, and stop the study if you can already identify problems
- Randomize data collection and keep part identifiers hidden from person

Key Points (cont'd)

- Clearly mark or control items, but don't make markings visible to operators (blind study)
- Don't let operators watch each other, so true behavior can be captured (depends on purpose of study)
- Use typical items seen in the process
- Measure to as finite a number as possible. Do not round
- Make detailed observations as the parts are being measured
- Treat each measurement as a new item (full setup and break down each time)

Criteria Summary

Variable Gage R&R

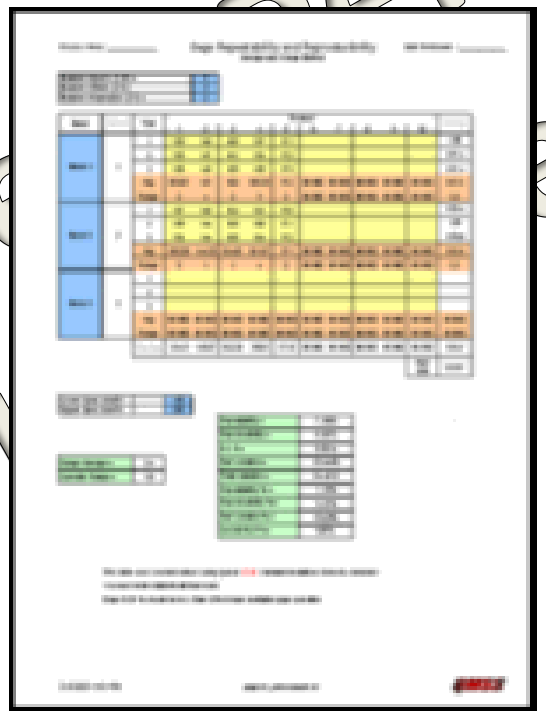
- **% of Study Variation or % of Tolerance**
  - Excellent: 10% or less
  - Marginal: 10-30%
  - Poor: Over 30%

Attribute Gage R&R (Agreement Analysis)

- Kappa value
  - Excellent: 0.90 or greater
  - Marginal: 0.70 - 0.90
  - Poor: Less than 0.70

Get Free Gage R&R Template

- Get an Excel Gage R&R Template and enter code “ ” for your free copy
- http://biz-pi.com/product.asp?id=10



Contact

- For more training materials and resources
  - Capability Analysis
  - Control Charts
  - Lean Six Sigma Overview
  - Root Cause Analysis
  - 5S
  - Cost of Poor Quality
  - Templates and Diagrams
- Visit Business Performance Improvement at: **http://biz-pi.com/store.asp**

Additional Resources

Business Performance Improvement

http://www.biz-pi.com