Equivalency Testing of Bluehill® 3 and Bluehill® Universal – a Statistical Review

Authors: Elayne Gordonov and Rajiv Iyer

Introduction:

Regulated industries such as pharmaceutical and medical device must achieve a high level of product quality, as even the smallest product inconsistencies can have disastrous effects on patients. To achieve a high level of quality, laboratory equipment used to test highly regulated products must be validated. Validation often consists of calibrations, verifications, installation qualification (IQ), operation qualification (OQ), and performance qualification (PQ). One downside of system validation is that companies often avoid laboratory equipment updates or upgrades to avoid re-evaluating the equipment validation. This may lead to laboratories using unsupported operating systems, and working with outdated equipment and software which leads to missing out on product features that could improve accuracy and efficiency. Even worse, it is not uncommon for validated equipment to be operating on unsupported operating systems which leaves a lab vulnerable.

In an effort to help companies manage software validation, Instron offers equipment commissioning and qualification through an IQ/OQ service package. While this helps laboratories with the beginning steps of their validation process, many laboratories also conduct gauge repeatability and reliability (GR&R) studies or other equivalency tests as part of PQ.

With the 2017 introduction of Bluehill Universal Software, Instron conducted a series of tests to prove software equivalence between Bluehill 3 and Bluehill Universal.

Why Conduct Equivalency Testing:

Although GR&R is an important tool in validation of measurements from testing systems, when it comes to upgrading software versions, the technique of “equivalence testing” adds more value to the changes. In simple terms, GR&R is more suitable when a machine is being installed for the first time to validate the measurements from a testing system[1], however not essential for a change in product configuration such as software updates. In this case, equivalence testing provides more relevant information on the change in performance of the system by changing the software.

By definition, equivalence testing is a statistical technique used to determine the difference in means of the data sets. It is commonly applied to understand the data before and after configuration changes are made on a single system or when comparing two independent systems[1]. Further in many industrial applications small differences in the sample means of data sets might not be practically significant and therefore equivalence tests apply a user’s perspective on limits beyond which differences must be considered important[1].

In many scenarios, regulated companies still consider GR&R as an important step towards qualification of configuration changes, and in that case a type I gage study also known as P/T ratio is more applicable. P/T ratio, also known as Precision to Tolerance ratio, is a statistical method used to investigate the precision of a system within pre-defined specification limits[1]. This method purely analyzes the variations from the system and not from a part or operator. Performance of the testing system due to changes in the software configuration can be analyzed using this method.

Overall the basic expectations from regulated companies is to ensure the performance of the testing system is the same irrespective of changes in the software. The software should only enhance the performance and capability of the system and not introduce any variability. Therefore the above mentioned techniques will statistically validate the performance with different software configurations.

Test Methodology:

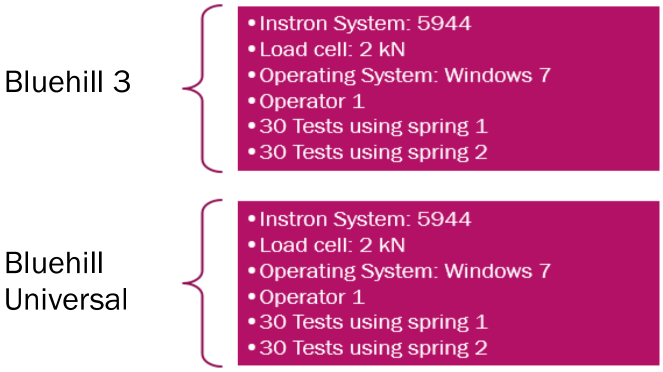
A 5944 single column table top model with a 2 kN load cell was used for testing. The computer was configured with Windows 7 to run Bluehill® 3 and Bluehill® Universal software. Test methodology for each statistical technique is detailed as follows:

Test Methodology: P/T Analysis

For the P/T analysis, a spring of known stiffness was used. The 5944 system was configured with circular compression platens to conduct the compression test. The test method was designed to run 30 times where the stiffness of the spring was measured for each specimen. The test ended every time at a fixed load and the slope of the response curve was used to automatically compute the final result. One part, one operator and one system was implemented to perform this P/T analysis.

Test Methodology: Equivalence Testing

For equivalence testing, the system performs the same compression test as detailed above. The first set of 30 specimens was collected with the Bluehill 3 and the second set of data was collected using Bluehill Universal. The entire procedure was followed for two springs with different stiffness to ensure the software capabilities are repeatable.



Bluehill 3
- Instron System: 5944
- Load cell: 2 kN
- Operating System: Windows 7
- Operator 1
- 30 Tests using spring 1
- 30 Tests using spring 2

Bluehill Universal
- Instron System: 5944
- Load cell: 2 kN
- Operating System: Windows 7
- Operator 1
- 30 Tests using spring 1
- 30 Tests using spring 2

Figure 1: Experimental Design

Results:

The analysis and results obtained from P/T analysis and equivalence testing are presented in this section.

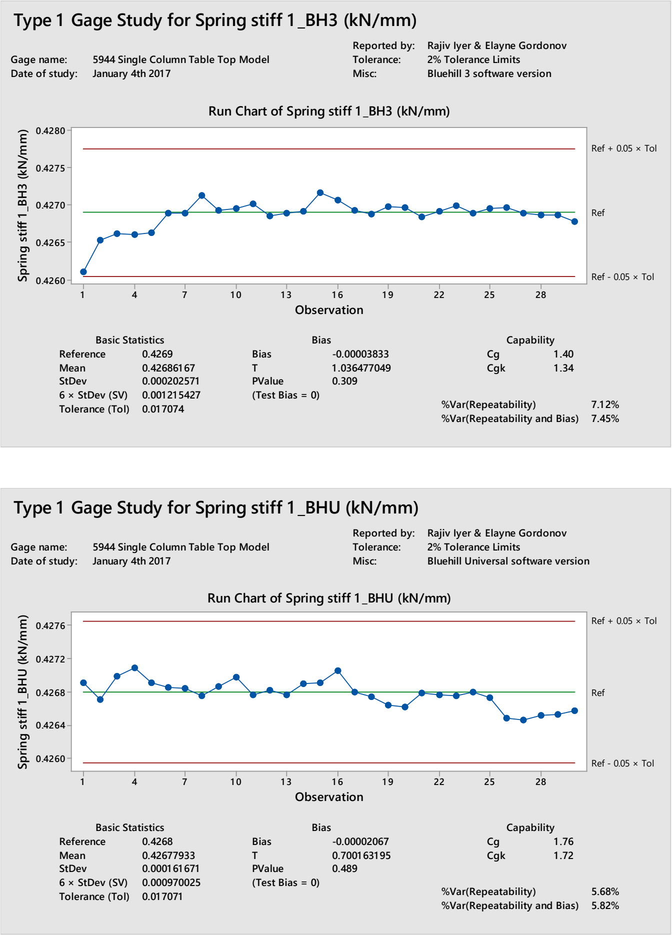
Analysis & Results: P/T Ratio

P/T ratio results for the spring stiffness data using Bluehill 3 and Bluehill Universal are presented in Table 1.

Table 1: P/T ratio results – BH3 v/s BHU

Machine	Software Version	P/T Ratio @ 2% Tolerance Limits
5944	Bluehill 3	7.12%
5944	Bluehill Universal	5.68%

Analysis on the 30 sample spring stiffness data using Bluehill 3 and Bluehill Universal indicate the testing system is significantly precise within the tight tolerance limits of 2%. The P/T ratio in Table 1 is well below 10% for both software versions indicating the testing system as gage capable according to AIAG guidelines[2]. This statistically validates the capabilities of measuring the same part using different software configurations on the same 5944 testing system. In other words, changing the software configuration from Bluehill 3 to Bluehill Universal does not introduce any variabilities in measurement and hence does not impact the system performance. Figure 2 presents a graphical summary of P/T analysis for spring stiffness data using both software versions.



Type 1 Gage Study for Spring stiff 1_BH3 (kN/mm)

Gage name: 5944 Single Column Table Top Model
Date of study: January 4th 2017
Reported by: Rajiv Iyer & Elayne Gordonov
Tolerance: 2% Tolerance Limits
Misc: Bluehill 3 software version

Basic Statistics
Reference 0.4269
Mean 0.42686167
StDev 0.000202571
6 × StDev (SV) 0.001215427
Tolerance (Tol) 0.017074

Bias
Bias -0.00003833
T 1.036477049
PValue 0.309
(Test Bias = 0)

Capability
Cg 1.40
Cgk 1.34

%Var(Repeatability) 7.12%
%Var(Repeatability and Bias) 7.45%

Type 1 Gage Study for Spring stiff 1_BHU (kN/mm)

Gage name: 5944 Single Column Table Top Model
Date of study: January 4th 2017
Reported by: Rajiv Iyer & Elayne Gordonov
Tolerance: 2% Tolerance Limits
Misc: Bluehill Universal software version

Basic Statistics
Reference 0.4268
Mean 0.42677933
StDev 0.000161671
6 × StDev (SV) 0.000970025
Tolerance (Tol) 0.017071

Bias
Bias -0.00002067
T 0.700163195
PValue 0.489
(Test Bias = 0)

Capability
Cg 1.76
Cgk 1.72

%Var(Repeatability) 5.68%
%Var(Repeatability and Bias) 5.82%

Figure 2: Graphical summary of P/T ratio for BH3 and BHU

Note: For all the testing and analysis conducted, 2% tolerance limits are considered as acceptable specification limits to investigate the system and software performance. This is based on empirical data and subject matter experts where the system performance within tight tolerance of 2% satisfies most of the applications intended with Instron testing systems. Therefore as observed in Figure 2, the precision of the system is investigated for 2% tolerance.

Analysis & Results: Equivalency Testing

The spring stiffness data from Bluehill 3® and Bluehill® Universal were applied to conduct the equivalency testing and analysis. The test involved collecting 30 stiffness tests from a single spring using Bluehill 3 and running the same test with Bluehill Universal. The data collected is used to estimate the sample mean or sample average.

Statistically, the means of the two data sets are compared where the analysis involves estimating the difference of means and computing the range of differences using 95% confidence interval. Limits of +/-0.0005 kN/mm are set as an acceptable range for difference of sample means.

P-value is estimated to compare the difference in the means with acceptable limits. The resulting p-value of 0.000 indicates the difference is insignificant and the means of the data sets are same. This further implies the stiffness data from Bluehill 3 and Bluehill Universal are statistically equivalent. Results from equivalence tests are presented as follows[3]:

Two-Sample Equivalence Test: Spring stiff 1_BHU (kN/mm), Spring stiff 1_BH3 (kN/mm)

```
Method
Test mean = mean of spring stiff 1_BHU (kN/mm)
Reference mean = mean of spring stiff 1_BH3 (kN/mm)

Equal variances were not assumed for the analysis.

Descriptive Statistics
Variable                    N     Mean       StDev      SE Mean
Spring stiff 1_BHU (kN/mm) 30  0.42678  0.00016167  0.000029517
Spring stiff 1_BH3 (kN/mm) 30  0.42686  0.00020257  0.000036984

Difference: Mean (Spring stiff 1_BHU (kN/mm)) - Mean (Spring stiff
1_BH3 (kN/mm))
                              95% CI for
Difference           SE      Equivalence    Equivalence Interval
-8.23333E-05  0.000047319  (-0.00016150, 0)    (-0.0005, 0.0005)

CI is within the equivalence interval. Can claim equivalence.

Test
Null hypothesis: Difference ≤ -0.0005 or Difference ≥ 0.0005
Alternative hypothesis: -0.0005 < Difference < 0.0005

α level: 0.05

Null Hypothesis       DF  T-Value  P-Value
Difference ≤ -0.0005  55   8.8266    0.000
Difference ≥ 0.0005   55  -12.307    0.000

The greater of the two P-Values is 0.000. Can claim equivalence.
```

Results indicate the p-value is 0.000 for both acceptable limits. This implies the difference of the stiffness sample means is insignificant. Therefore we can validate that the spring stiffness data from Bluehill 3 and Bluehill Universal are equivalent. Figure 3 presents a graphical summary of results from equivalence testing.

The study has been repeated using a second spring with different stiffness values. This is to ensure the performance is validated and verified for other parts.

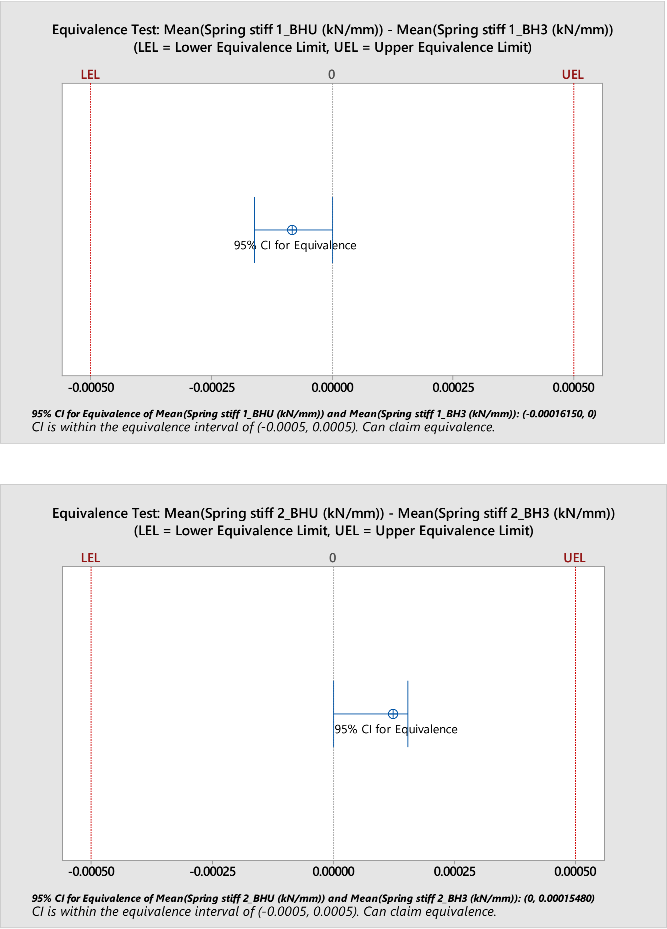


Figure 3: Graphical summary of equivalence testing using two different types of spring.

Conclusion:

As anticipated, updating Bluehill 3 software to Bluehill Universal has no effect on test results. A materials testing system outfitted with Bluehill 3 or Bluehill Universal is capable of producing equivalent test results with all other testing parameters constant. This has been statistically validated based on the tests performed.

References:

1. Montgomery, C. D., 2009, Introduction to Statistical Quality Control (6th ed.), John Wiley & Sons, USA.
2. Chrysler, Ford, and GM (2010), Measurement Systems Analysis Reference Manual (4th ed.), AIAG, Detroit, MI.
3. MINITAB 17, Statistical Software Package.