This contents of this paper were presented at the Materials and Processes for Medical Devices Conference (MPMD) September 23-25, 2007 Desert Springs Resort & Spa, Palm Desert, CA. The manuscript will appear in the conference proceedings (in press).

GR&R: Understanding Sources of Error in Mechanical Testing Results

M. Viveiros, J. Ritchey
Instron, Norwood, MA, USA

Abstract

The need to ensure the repeatability and reproducibility of mechanical test results between individual test systems, whether those systems are located in the same lab or different labs or a comparison is being made between multiple suppliers, has recently surfaced as a critical concern throughout the medical device industry. This is not surprising since product quality is critical and quality assurance depends on the ability of testing systems to provide accurate results. Further, as many industries expand R&D and manufacturing operations into different parts of the world, data comparisons have become increasingly more important and complex. Gage repeatability and reproducibility, also known as GR&R, is a type of statistical analysis that is often performed by quality and product engineers as a method of test equipment validation and verification. However, in cases where GR&R values are higher than expected, it is necessary to investigate and resolve or at least minimize sources of variation. The purpose of this paper is to highlight a variety of error sources and provide suggestions and guidelines for conducting a successful GR&R study.

Introduction

A GR&R study is a thorough investigation that provides a statistical approximation of the variation and percent of process variation for a test measurement system. Such studies are recommended by the Automotive Industry Action Group (AIAG), Six Sigma and ISO 9000 quality plans, and suggest that quantification of the repeatability and reproducibility of a test system is required in order to determine how much of the observed variability is a product of the test system versus part-to-part variation or process changes. The term repeatability defines how well the system can produce a known result over multiple tests. Reproducibility is the ability of another operator to produce the same results from similar parts with the same level of consistency. The output of a GR&R study is a quantitative result by which a test system can be measured. Statistical norms for GR&R values fall under three different categories. A GR&R value less than 10% is ideal for most measurement systems. This value suggests that the variability in the test system is negligible and the results can be used to identify variability between parts or differences in production processes. A GR&R value between 10% and 30% suggests that the variability in the system is not negligible but may be acceptable for evaluating part variability. The performance of the test system should be evaluated for areas of improvement to decrease inconsistencies. Finally, a GR&R value greater than 30% suggests that the error in the system is too great and will prevent differentiation between system error and part variation.

It is important to note that these value ranges were originally developed for strict regulation of manufacturers in the automotive industry and most of those studies were using non-destructive test methodologies. A non-destructive test uses a single specimen between multiple operators to eliminate, or at least minimize, the error generated from part variability in the statistical analysis. However, for materials testing systems, a destructive GR&R, where the actual parts pulled from the production line are tested, is more important. Because the statistical analysis can become significantly more complex in destructive GR&R, there is question as to whether these ranges are appropriate.

In many quality assurance labs, the purpose of mechanical testing is to ensure that products meet or exceed requirements, to identify changes in process that affect the critical requirements of the part and to evaluate the consistencies of those processes for reducing part variation. In the medical device industry, there is a trend towards using GR&R studies as a method for evaluating a test system’s ability to perform these tasks, the standard operating procedures and the operators who run the systems.

However, the danger in using GR&R exclusively as a methodology for evaluating a mechanical testing system is that GR&R does not address accuracy. It is possible to have very low GR&R values and test results which are wrong and not truly representative of the material or the product tested. The ability of a test system to provide accurate results depends not only on the quality of the test system but also on errors that can be introduced into the system. A fishbone diagram, shown in Figure 1, shows the major categories by which a test system should be evaluated for error sources that can affect both the accuracy of a test system and the ability of the gage to produce repeatable and reliable results. The major categories include the following: method, measurement, operator, material, machine, and environment. Within these major categories are sub-category sources of error that must be specifically addressed. This is not an exhaustive list of error sources, but rather common sources that should be considered.

Therefore, when trying to determine and understand errors in mechanical testing results, one must examine the testing system in both a qualitative, physical manner and a quantitative manner. For the former, all elements of the

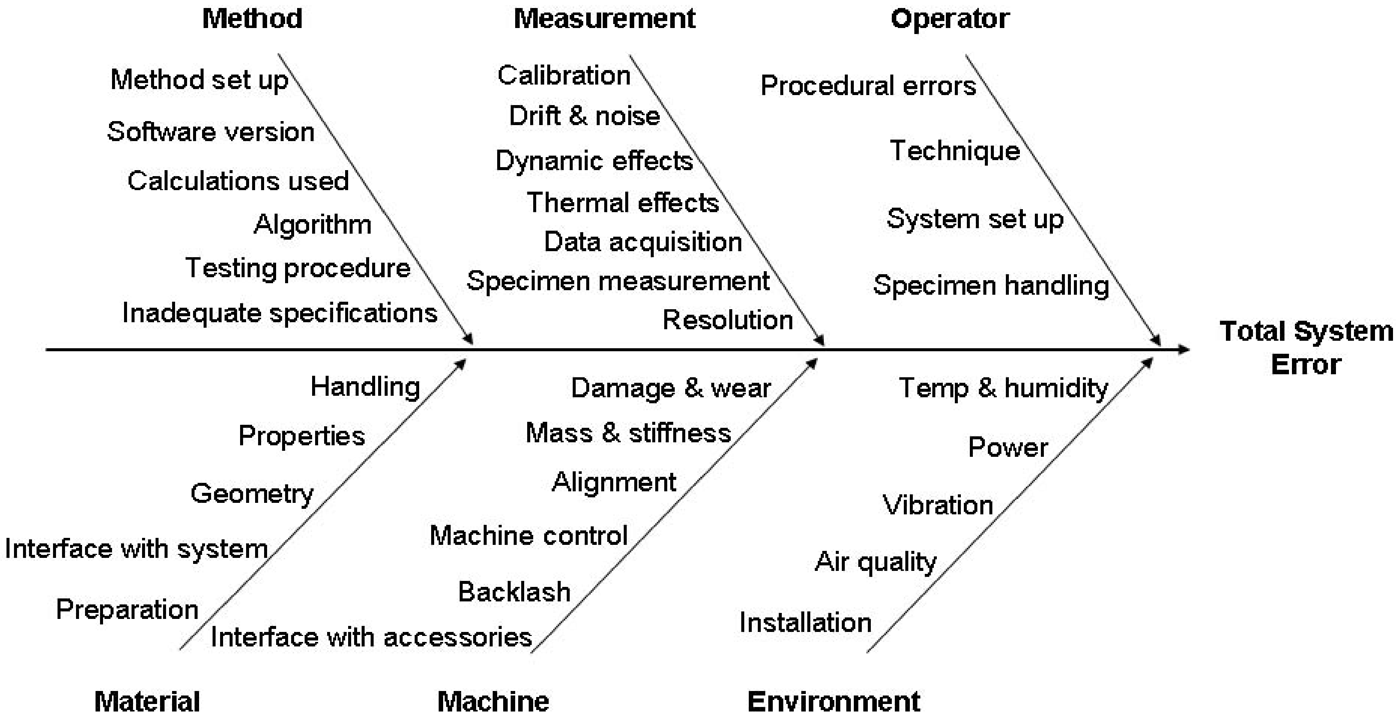


Figure 1: Fishbone diagram shows the major categories, along with sub-categories, by which a test system should be evaluated for error sources that can affect the accuracy of a test system. Note that this is not an exhaustive list of error sources, but rather common sources.

fishbone diagram must be considered and evaluated. For the latter, results must be calcualted and analyzed typically through a GR&R study. In order to demonstrate these theories in practice, a non-destructive study was conducted on four different materials test systems from four different manufacturers. The study included both a qualitative examination to evaluate sources of potential error in each system and a quantitative analysis, non-destructive GR&R, to compare mean peak load values and determine a non-destructive GR&R value. The ultimate output was to provide recommendations for improving destructive test results.

Methods

Apparatus: Four different electromechanical materials test systems, from four different manufacturers, were evaluated in this study. All four systems were configured with 6-inch compression platens and a load cell that was appropriate for the expected maximum load values. Adhesive markers were placed on the lower platen to aid in operator placement of the specimen repeatabily. All but one of the systems, System #4, used a software program for test control and output of results. System #4 used a touch panel that allowed for test control and results were manually recorded. ASTM E-04 verification of calibration was performed on all systems' load weighing transducers. Although all of the systems were not located in the same test lab, temperature and humidity readings were taken at each location and other gerneral environmental observations were made to ensure that environmental differences would not affect results.

Specimen Preparation: Three different springs were used as specimens and labelled as Spring 1, Spring 2 and Spring 3. The stiffness of the springs varied such that under the same amount of compression, the peak loads varied by approximately 5 to 10 pounds. All three springs had a special mechanical fixture designed to allow for a single point of compression on the spring and therefore, minimize the effects of system alignment on the results. The moving parts on the mechanical fixture had marks to ensure that the moving parts were consistently aligned for every test.

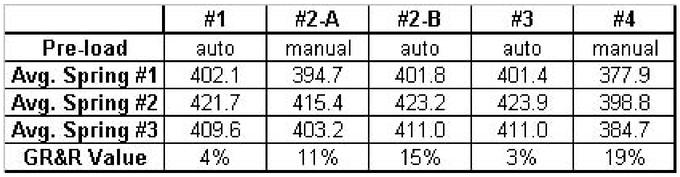
Procedure: A single standard operating procedure was created for all four test systems, and is summarized as follows: balance the load cell; insert the specimen on the center of the platen with careful attention to the mechanical fixture alignment marks; apply a 5-pound pre-load to the specimen; compression the specimen to 0.25-inches; take a load reading at the 0.25-inch point; repeat for the next specimen. Systems #1, #2 and #3 all allowed for an automatic pre-load setting to be configured in the method, rather than requiring the operator to manually adjust the position of the crosshead to achieve the pre-load value. System #4 did not have an automatic pre-load feature, therefore, the manual method was required. For System #2, the primary users of the equipment did not use the automatic preload feature, despite its benefits for increased productivity, repeatabilty and ease-of-use. Therefore, for System #2, both manual and automatic pre-load settings were used, labeled as System #2-A and System #2-B, respectively.

Each operator tested each spring 10 times in a predetermined order that allowed the time between tests on each specimen to remain consistent. The same two operators were used throughout the study. The GR&R values were calculated using a proprietary analysis program that was validated using Minitab Software. Minitab or an equivalent software program that allows for the generation of GR&R values could be used to generated similar results.

Results and Discussion

A summary of the average peak load values reported for all five test configurations and the resulting GR&R values are summarized in Table 1. The first and most interesting point of discussion, is the fact that in this relatively basic compression test, which was designed to overly simplify the test parameters of the actual tests on sporting equipment, a range of 24 to 26-pounds is seen for each spring between systems. The differences in these mean values are the result of both the poor gage repeatabilty and reproducibility as well as sources of error in the test systems.

Table 1: Summary of results from a GR&R study conducted on 4 different materials testing(systems using springs. The table shows the average value in pounds for each of the springs tested and the resulting GR&R value for a specific test frame.



	#1	#2-A	#2-B	#3	#4
Pre-load	auto	manual	auto	auto	manual
Avg. Spring #1	402.1	394.7	401.8	401.4	377.9
Avg. Spring #2	421.7	415.4	423.2	423.9	398.8
Avg. Spring #3	409.6	403.2	411.0	411.0	384.7
GR&R Value	4%	11%	15%	3%	19%

Based on the results generated, the known system evaluations and the expected results from the springs selected, we can assume that Systems #1 and #3 are repeatabile and reproducible since the GR&R values are less than 10%. The key to repeatable and reproducible results for these test systems was the quality of the manufacture of the test system frame and control electronics, and the automatic pre-load feature, which eliminates unnecessary procedural steps for the operator, therefore, reducing the potential for operator error.

For System #2, two different methodologies were used, automatic pre-load and manual pre-load. The standard operating procedures for the current users of this test system were to use the manual pre-load setting because, through investigation, they discovered that the system would always overshoot the automatic pre-load value. Therefore, despite the added effort, the users felt more confident with the manual pre-load method in its abilty to achieve the desired pre-load value. However, when using the manual pre-load method, the users would balance (or zero) both the extension and load transducers. Balancing of the load after the pre-load has been set will result in an approximate 5-pound net lowering in results. The System #2-A results correspond with this suggestion.

In regards to the higher GR&R values for System #2-A, #2-B and #4, focus needs to be shifted to the qualitative review of the systems and setups. Possible reasons for the higher GR&R values can be attributed to several different sources of error as mentioned in the fishbone diagram (Figure 1). In the case of System #2-A and #2-B, the data rate on this system has a maximum setting of 5 points per second. Because the test only runs for approximately 15 seconds, there are only about 75 data points that characterize the load-extension data. For comparison, System #1 has a data rate of 100 points per second and, therefore, 1500 data points to characterize the load-extension curve. When looking for the load value at a specific point (e.g. 0.25-inches of compression) and there is no exact data point corresponding to that specified point, the software will interpolate the data or pick the next closest point. The more data points that are available, the more repeatable the system will report the correct value.

Another important source of error to examine is speed accuracy. When reviewing the system service records and when looking at the raw data generated from System #2, it was not clear that the accuracy of the speed had been verified according to ASTM standards. The actual time to complete the test did not correspond with calculated values derived from speed and displacement. Although speed accuracy may not be completely important for testing springs, it is definitely a significant factor to consider when testing strain-sensitive materials. It is an important qualitative step to address before completing any comparative study between systems.

One last qualitative issue revolved around the accuracy associated with the control electronics on the test systems. Control electronics and the speed at which they can respond can often be vital to succussful test results. In this test method, a pre-load was required to insure repeatable test results. Unfortunately, System #2 tests contained a large amount of overshoot and error generated when both automatically and manually setting preload values. Although application dependent, these types of system characteristics can have an impact on both accuracy as well as repeatability.

Looking in more detail at System #4, several factors emerged as issues and sources of error. Most importantly, was the system compliance. System compliance incorporates the mechanical compliance, or system slack, that exists in the test frame, the load cell and the accessories and has a significant implication on the total stiffness of the system. A test system with low stiffness used in this type of compression application, will consistently yield lower peak load results than a similar system with a higher stiffness. Because of the slack in the frame with the lower stiffness, the actual distance travelled is lower than expected and therefore, the load values reported will be lower. It is important to note that low stiffness can affect the accuracy of readings. It is possible for the system to provide consistent, yet inaccurate readings, which would not lead to high GR&R values. Validating the accuracy of the system for a specific test is therefore necessary before conducting a GR&R evaluation.

Similar to the issues associated with System #2, data rate, control electronics and speed accuracy are all issues that must also be addressed in System #4. Additionally, because System #4 does not use software and requires the operator to go through a series of repetitive manual steps for each test, it is likely that the operator variability is much higher as compared with the other systems. In our previous work, it has been found that, under conditions where total system error is low, operator error is typically the greatest source of error, as compared with the test system and the material. Therefore, it is important to have very detailed operating procedures, regularly scheduled training for operators and test methodologies that limit the number of steps required by an operator to minimize the chance for error. It is helpful when conducting a GR&R to have an observer present during all testing to compare different operator procedures and to document errors and actions that may be the cause for differences in test results.

Conclusions

The sources of error that were revealed and discussed in this study are just a sample of the types of errors that can be identified with both quantitative and qualitative evaluations of a materials test systems. Every test sytem is unique and test setups and configurations can vary widely. In the case of a non-destructive test with a spring, the specimen is overly simplified to limit the errors that can complicate the analysis. When evaluating a destructive test with real test specimens, everything from the environment from which the materials originated, to how they were prepared, to the condition and environment in which they were tested needs to be considered. These qualitative measures are important to evaluate because they can lead to highly variable results which, when quantified, lead to high GR&R values. Once the qualitative issues are addressed, error sources minimized, and system configuration reviewed, a solid GR&R study can be implemented to explore the potential variance in the parts that are being produced.

A non-destructive GR&R using a single controlled specimen, like that which was described in this study, is a good way to evaluate a test system’s basic functionality. If the test sytsem is not able to successfully pass the non-destructive study, with results under 10%, it will not be possible to get acceptable destructive GR&R results. When conducting a destructive GR&R, the methodology described above for identifying and resolving sources of error is critical to ensuring accurate, repeatable and reproducible results.