INJURY BIOMECHANICS RESEARCH
Proceedings of the Thirty-Ninth International Workshop

Statistical Considerations for Evaluating Biofidelity, Repeatability, and Reproducibility of ATDs

G. Nusholtz, Z. Ben Aoun, L. Di Domenico, T. Hsu, M. A. G. Luna, and J. A. P. De La Mora

Chrysler

This paper has not been screened for accuracy nor refereed by any body of scientific peers and should not be referenced in the open literature.

ABSTRACT

Presented in this paper are two numerical/statistical methods for determining if the impact time-history response of a mechanical system is repeatable or reproducible; such a system could be a vehicle, a biological human surrogate, Anthropometric Test Device (ATD or dummy), etc. The responses could be sets of time-histories of accelerations, forces, moments, etc., of a component or of the system. The example system evaluated is the BioRID II rear impact dummy. The evaluation begins by transforming the sets of time-histories into sets of relative-shapes and magnitudes of the response time histories. The two statistical procedures use the t and T^2-tests. One uses a statistical comparison of the average time history of a set (Representative Curve (RC)) to the individual time histories of that set or other sets. The other procedure uses a statistical comparison of sets of time histories. The statistical analysis is then performed on the sets of the relative-shapes and magnitudes. Both methods indicate the BioRID II is neither repeatable nor reproducible for the neck moment.

INTRODUCTION

Although the statistical methods in this paper are aimed at biomechanical process that are used to define Anthropometric Test Devices (ATDs or dummies) they are general enough to be applicable to a wide range of systems. The primary focus of the statistical evaluation is repeatability and reproducibility of a system.

Determining Repeatability and Reproducibility (R&R) of a mechanical system from a parameter measurement time history, like force, moment, acceleration, or displacement, contains at its core the comparison of time histories, processed or raw, to determine similarities or differences. In the case of ATDs this operation has two basic goals: One, to directly compare an ATD to itself, repeatability; Two, to compare one ATD to one or more other ATDs of the same design, reproducibility. For repeatability, the object of investigation is subjected to repeat nondestructive impacts, and the time-history responses are obtained; each area of interest forming a collection (set) of response time histories. Reproducibility requires that two or more copies of the system are subjected to the same/similar inputs and the set of responses from each system is obtained. For repeatability each time history can be compared numerically/statistically to every other time-history in the set. For reproducibility the time histories in one set are companied to the time histories in every other set.

A second approach to determine if the mechanical system is repeatable or reproducible is to use the average of a given set of time-histories. For repeatability the average of a given set is compared to each value in that set. Reproducibility compares the average of one set to each time-history in a second set.

In general, when experiments of this type are run the responses are contaminated by sample-to-sample variation, test-to-test variability, random noise, instrumentation noise, and noise from unknown sources. This noise is generally addressed statistically. If the responses are single valued functions (scalars) then standard statistics can be used to determine the level of repeatability or reproducibility. However, if the responses are time-histories, such as accelerations, velocities, forces, etc., standard statistics may not apply. One possible approach of addressing the noise in the system is to obtain the underlying response by running multiple tests on the same sample for repeatability and on different samples that represent the same system for reproducibility and add them together obtaining an average. Statistics can then be performed on the averages. However, if the "fundamental response" of each sample is not the same, then it is not altogether clear what the average represents. It may not capture the underlying physics.

Nusholtz et al. (2010) presented procedures for comparing the time histories of different classes of vehicles. They took the approach of first creating an average signal and then creating sets of shape comparisons. The shape coefficient values were first transformed and then analyzed using the ANalysis Of VAriance (ANOVA) procedure. This allowed for studying whether there was significant difference among the average shape coefficients for the different vehicle groups considered. Another method for comparing time histories is CORrelation Analysis (CORA) that includes the forming of an average curve as part of a method of comparing Post Mortem Human Surrogate (PMHS) and/or ATD time-histories. This was introduced in 2009 by Gehre et al.

Functionally, the methods above rely on the Multidimensional, or Vector, Central Limit Theorem (CLT), which when applied to a time-history response assumes that the random variables to be identically distributed. The noise is considered to be independent of the underlying response signal, normally distributed, stationary and, for small sample size, insignificant with respect to the signals. Addition of the signals becomes problematic and the resulting time-history may represent a characteristic system that is different from any of the test samples and may not represent the response of the underlying structure. For example, Nusholtz et al. (2010) created a simple linear system that consisted of a spring and a mass with a half-sine like solution. The mass and the spring were varied to produce time-histories with different periods. The average curve that was created from the addition of the different time-histories produced a characteristic response that looked like a non-linear response resembling a soliton (Dodd et al. (1984)).

The question remains as to what is the meaning of the average of several time-histories from different samples. Beyond the question of what the average means, there is the question as to what is meant by average. In this paper, we expand the use of the average curve, Representative Curve or RC, (un-normalized average) developed by Nusholtz et al. (2009) to address R&R. In addition, methods are presented to perform similar analysis without the creation of an RC. The examples presented utilize BioRID II data obtained from the Partnership for Dummy technology and Biomechanics (PDB) for tests performed from December 2008 to January 2009, see Bortenschlager et al. (2009). A sample of the data is shown in Fig. 1.

BACKGROUND FOUNDATION

Shape

The shape coefficient (relative shape coefficient) (Nusholtz et al. (2007)) is not a property of any one signal but is a relationship between two signals that defines how close, in shape, is one signal to another and is defined as follows:

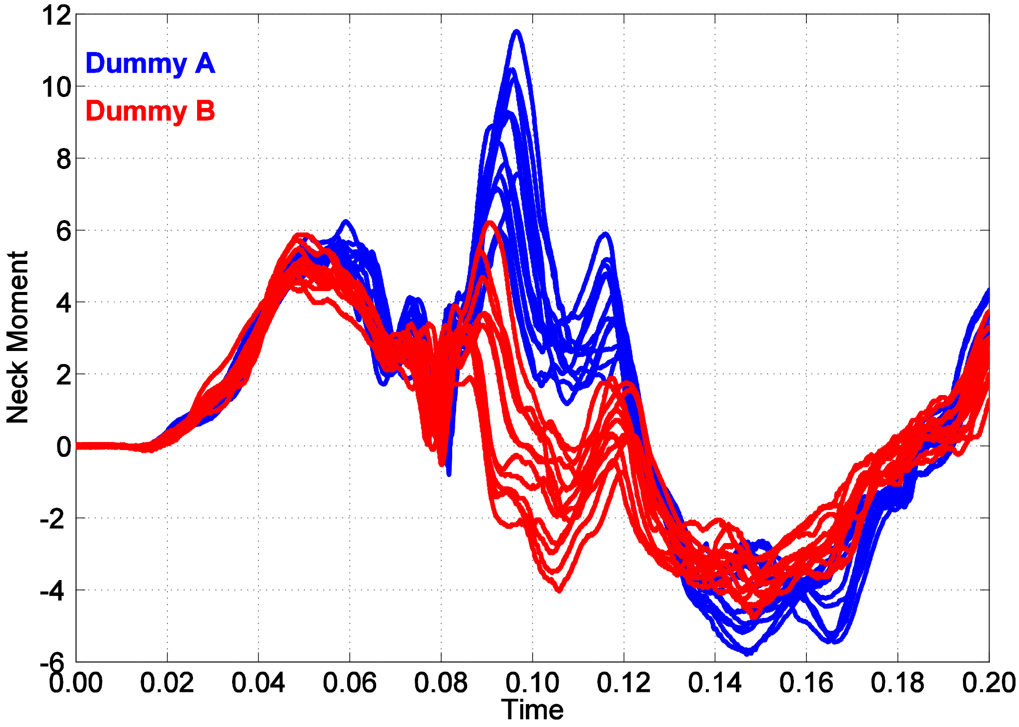

Figure 1: Dummy A and Dummy B – Neck Moment Signals

$$p(x,y) \coloneqq p(x,y,\tau) \tag{1}$$

where τ is the amount of time alignment that maximizes the shape function, or shape operator defined by:

$$p(x,y,h) = \frac{\int x(t)y(t+h)\,\mathrm{d}t}{\sqrt{\int x^2(t)\,\mathrm{d}t \int y^2(t)\,\mathrm{d}t}} \tag{2}$$

Creation of the Representative Curve (RC)

Repeated scalar measurement, such as weight, height, length, etc. on a given object, forms a set of those measurements. In general the measurements will not be the same and an average value will be a good represented value of the set. In a similar manner, given a set of repeated vector measurements, such as time-histories of forces, accelerations, moments, displacements, etc. of a given object, such as an ATD, forms a set of those measurements. In general the measurements will not be the same and an average curve will be a good Represented Curve (RC) of the set.

An RC is derived from a group of signals by adding, point wise, all the signals in the group. Since each signal represents a time-history, or vector, it is important to determine how to configure each signal/vector before adding them together. In the case of time-histories that have a well-defined constant sampling rate, configuring the signals consists of aligning the signals in time such that the global least-square difference is minimal.

Two methods for least-square alignment have been presented by Nusholtz et al. (2009). The first is a direct least square approach in which all the signals are aligned simultaneously. The second approach, Mean-To-Mean (MTM), is a procedure in which the signals are aligned in stages and eventually added together. In general, both methods produce similar results. However, the MTM is computationally more efficient for large sample sizes and consequently can handle larger data sets easier. The MTM procedure is used in this paper to generate the RC.

The MTM method starts with an initial curve. In theory this could be any initial curve; however, if the curve is not carefully chosen then it is possible that a local minimum will be determined by the procedure that is not optimum. One possible approach is to choose any of the signals from the set that will be used to form the RC. In this study, we used the method outlined in Nusholtz et al. (2009). The resulting RCs, from the signals of Fig. 1, are shown below in Figs. 2 and 3. RC_A and RC_B are the representative curves for dummies A and B, respectively.

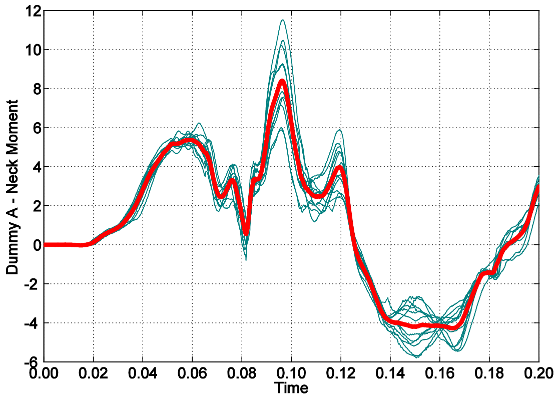

Figure 2: RCA for Dummy A

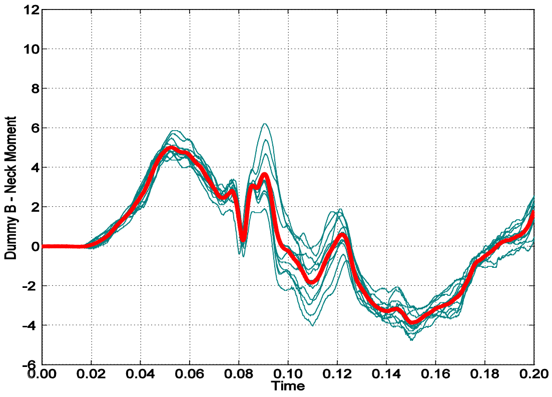

Figure 3: RCB for Dummy B

Average Shape Coefficient

To determine the average shape coefficient for a given time-history that is being compared to a set of time-histories, a shape coefficient is determined for all possible pairs that include the given time history and every other time history in the set that it is being compared to: The given time-history is compared, using the shape function to all other time-histories in the set. All of these numerical estimates of the shape coefficients obtained from the shape function are then used to form the average and the distribution. The average shape coefficient and the distribution of shape coefficients for an RC are determined in a similar manner. In the case where the RC is compared to the set that was used to create it, the following is true, Nusholtz et al. (2010):

1. It can be shown that for two signals the aligned average signal will always provide the best average shape coefficient. However this does not always hold for more than two signals.

2. When there are more than two signals: If all the signals have the same magnitude (i.e. the integral of the squared signal), irrespective of the signal shapes, the average signal will provide the best average shape coefficient.

3. When there are more than two signals: If all the signal parts have the same shape coefficient, irrespective of the signal magnitudes, the average signal will provide the best average shape coefficient.

4. When the above cases (1, 2 or 3) are not satisfied, it is possible that the average signal will not have the best shape coefficient.

5. 1, 2, 3 and 4 are in agreement with the Central Limit theorem for random vectors.

Statement 4 above does not indicate how different the shapes or magnitudes have to be for the RC not to be the best representation of the shape. To estimate what level of difference is needed, in the magnitude and/or shape, see Nusholtz et al. (2010).

The RC average shape coefficient and its distribution of shape coefficients obtained are used to determine the shape quality of the RC with respect to the signals that were used to form the RC. If the RC produces a higher average shape coefficient value when compared to the average shape coefficient value using any of the time histories in the set, then it is considered a better estimate of the shape of the set of signals than any of the base signals. It is the best estimate of the shape.

Magnitude

The magnitude uses the norm of a vector defined by:

$$\|x\| \coloneqq \sqrt{\int x^2(t)\,\mathrm{d}t} \tag{3}$$

The definition above differs from the one originally proposed by Xu et al. (2000) where the squared value was used for consistency with the shape function definition. However, for presentation purposes, in this paper, it has been found convenient to define the magnitude in the norm form as shown above in Eq. (3). For ease of comparison, the magnitude from (Eq. (3)) is scaled by the overall average, of the magnitudes of all the time-histories forming the sets, to yield a normalized magnitude. In this paper we used data from the test series run by PDB where eight dummies are used and in the normalization we used the average magnitude of all eight dummies. The normalization process distorts the information content by eliminating the actual absolute value of each magnitude and leaving only the relative values between the magnitude coefficients. Care must be taken to ensure that the analysis can be done using the relative magnitudes only.

The normalized magnitudes from the signals of Fig. 1 are shown in Fig. 4. The notches around the median values represent about 1.7 standard deviation, see McGill et al. (1978) for further details about the box plots. The notches for both dummies are smaller than 0.07 resulting to standard deviations of less than 0.05 which suggest that both dummies could be considered repeatable using this assessment. On the other hand, the notches around the two medians do not overlap which indicate that the medians are significantly different at at-least the 95% confidence level. A t-test yields the same conclusion. This suggests that the dummy is not reproducible.

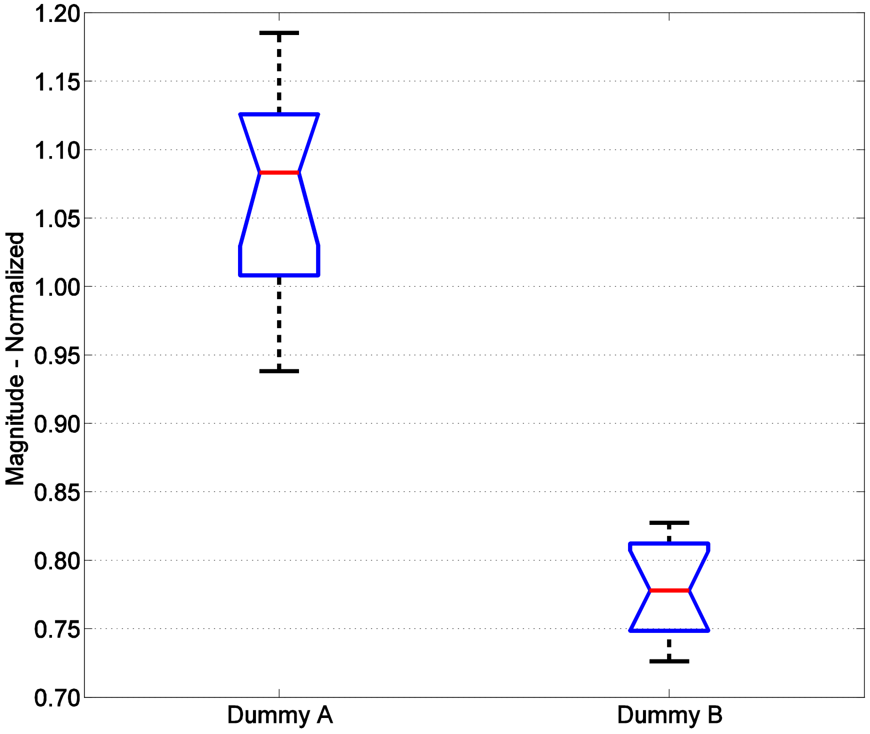

Figure 4: Magnitude Comparison – Dummy A versus Dummy B

SET EVALUATION USING AVERAGE SHAPE AND MAGNITUDE

Once the RC has been obtained the distributions of shape and magnitude of a set with respect to the average shape and magnitude of that set can be found. The RC is compared, through the shape function, directly to the members of the set of time-histories to obtain a set of shape coefficients. The average shape coefficient and the average magnitude along with the distributions can then be used to estimate the repeatability of the set of time-histories.

Representative Curve Test

In general the RC is an average time-history of a set of signals that represents the set. Therefore the RC from one set can be compared to the RC from a second set through a statistical test. The statistical Representative Curve-Test (RC-Test) incorporates a comparison between two sets of signals based on both the average signal shape correlation and its distribution as well as the average signal magnitude and its distribution. Using this testing procedure, two sets of signals will be considered to be similar only when both the signal shape and the signal magnitude can be concluded to be statistically similar.

The RC-Test is based on the fundamental assumption that the representative curve of both of the two sets under examination is a reliable and informative summary curve that is able to convey all the necessary and sufficient information to study the shape of the signal set.

In other words, the RC is a “valid” summary of the signal shape for the signal set only when all the signals belonging to the same set are independent replications of the same phenomenon. Consequently, the shape is functionally the same for each signal the only difference in shape among the signals is assumed to be to random variation (Nusholtz et al. 2010).

For this reason before performing any statistical analysis based on the RC it is essential to check that the RC comes from a set of signals that arise from the same phenomenon, mechanical or general system. One heuristic approach to estimate weather the RC is appropriate is provided by the evaluation of the correlation shape between any pair of signals in the set and then by the definition of the histogram of the correlation shape values: if the histogram shows a uni-modal distribution with mean value above a specified threshold value then the set of signals could be considered to arise from the same phenomenon and the RC should be a valid and informative curve for analyzing the shape of the signal set.

Conversely when the average correlations among the signals is below the specified threshold and/or the histogram of the correlation shape values presents more than one peak then the RC should not be evaluated because it is not a valid representation of the average signals and any analysis based on that RC could lead to misleading and unreliable conclusions that have no value in the field.

Repeatability and reproducibility are a considerable concern in the use of ATDs in crash testing. In this regard an analysis, based on empirical data, of an extensive number of ATD tests across a wide range of ATDs, Hybrid III, THOR, EuroSID, WorldSID, SID-IIs, BioRID, etc., has shown that, in general, a set of signals representing a single ATD can indicate an "acceptable" repeatability if the mean shape value is above 95%. Two ATDs of the same design that have undergone similar testing can be considered to be reproducible it the mean value of the shape descriptor formed by the cross comparison of the two sets is also above 95%.

An analysis of an extensive number of PMHS and computationally generated signals set has shown that a set of signals can be considered to be generated from an independent replication of the same phenomenon when they show mean correlation shape values above 75%. As the mean value of the shape gets lower the confidence of an independent replication of the same phenomenon goes down. This has implications for PMHS testing, in that if the average shape value for a set of PMHS tests is not 75% or above the RC may not be valid. Nonetheless, because of the intrinsic limitations and problematic nature of biological data, the many factors of variability (age, sex, medical conditions, bone conditions, etc.), and the very limited number of tests, that may not be achievable.

Thus, it is appropriate to evaluate/compare sets using statistical tests based on the RC if and only if the sets are derived from the same phenomenon (i.e. the sets satisfy the conditions stated above).

Before defining the RC-Test it is necessary to introduce some of the notations used. Given the set of signals $\mathbf{S}_\mathrm{A} = \{x_1, x_2, \cdots, x_\mathrm{M}\}$ with representative curve RC_A then:

- Magnitude values of the signals in $\mathbf{S}_\mathrm{A}$

$$\|x_1\|, \|x_2\|, \cdots, \|x_\mathrm{M}\|$$

- Mean magnitude in set $\mathbf{S}_\mathrm{A}$

$$\overline{\|\mathbf{S}_\mathrm{A}\|} = \frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}} \|x_i\|$$

- Shape correlations between the signals of set $\mathbf{S}_\mathrm{A}$ and its representative curve RC_A

$$p(x_1, \mathrm{RC}_\mathrm{A}), p(x_2, \mathrm{RC}_\mathrm{A}), \cdots, p(x_\mathrm{M}, \mathrm{RC}_\mathrm{A})$$

- Shape correlations between the signals of set $\mathbf{S}_\mathrm{A}$ and the other set's representative curve RC_B

$$p(x_1, \mathrm{RC}_\mathrm{B}), p(x_2, \mathrm{RC}_\mathrm{B}), \cdots, p(x_\mathrm{M}, \mathrm{RC}_\mathrm{B})$$

- Mean shape correlation between set $\mathbf{S}_\mathrm{A}$ and its representative curve RC_A

$$\overline{p(\mathbf{S}_\mathrm{A}, \mathrm{RC}_\mathrm{A})} = \overline{p(\mathrm{RC}_\mathrm{A}, \mathbf{S}_\mathrm{A})} = \frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}} p(x_i, \mathrm{RC}_\mathrm{A}) \tag{4}$$

- Mean shape correlation between set $\mathbf{S}_\mathrm{A}$ and the second set's representative curve RC_B

$$\overline{p(\mathbf{S}_\mathrm{A},\mathrm{RC}_\mathrm{B})}=\overline{p(\mathrm{RC}_\mathrm{B},\mathbf{S}_\mathrm{A})}=\frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}}p(x_i,\mathrm{RC}_\mathrm{B}) \tag{5}$$

Similar notations for set $\mathbf{S}_\mathrm{B}$ of signals {y_1, y_2, ⋯, y_N} with representative curve RC_B could be constructed from the above definitions.

Using the RC-Test two sets of signals are considered to be similar when there is evidence to conclude that all the three following equalities hold:

$$\overline{\|\mathbf{S}_\mathrm{A}\|}=\overline{\|\mathbf{S}_\mathrm{B}\|}$$

$$\overline{p(\mathbf{S}_\mathrm{A},\mathrm{RC}_\mathrm{A})}=\overline{p(\mathbf{S}_\mathrm{B},\mathrm{RC}_\mathrm{A})}$$

$$\overline{p(\mathbf{S}_\mathrm{A},\mathrm{RC}_\mathrm{B})}=\overline{p(\mathbf{S}_\mathrm{B},\mathrm{RC}_\mathrm{B})}$$

The RC-test is based on the multivariate T^2-Test (Casella and Berger (2001)). When the T^2-Test leads to conclude that the two sets of signals are similar, there is evidence that the two sets have similar magnitude and similar shape with respect to both RC_A and RC_B. On the other hand, when the T^2-Test rejects the hypothesis of similarity for the two sets of signals ($\mathbf{S}_\mathrm{A}$ and $\mathbf{S}_\mathrm{B}$) then the implication is that at least one signal characteristic among magnitude, shape correlation, with respect to either RC_A and RC_B, is not similar. The difference could be driven by RC_A or RC_B shape correlation with respect to the set that it was not derived from, or a combination. It could also be that the average magnitudes are not similar for the two sets of signals. When this occurs, three t-tests Casella and Berger (2001) should be performed on the three individual characteristics (i.e. one for the average magnitude, one for the average shape relationship of set $\mathbf{S}_\mathrm{B}$ with respect to RC_A and one for the average shape relationship of set $\mathbf{S}_\mathrm{A}$ with respect to RC_B). The t-tests would provide evidence about which of the three factors considered (i.e. magnitude, shape with respect to RC_A and shape with respect to RC_B) would lead the difference observed between the two sets of signals. In the following section all the possible RC-Test conclusions and their implications are analyzed.

It should be observed that the RC-Test compares the shape of the two sets of signals ($\mathbf{S}_\mathrm{A}$ and $\mathbf{S}_\mathrm{B}$) using the representative curve of both the two sets; it should not be expected that both comparison would always lead to the same conclusion. In general, when the variability of the signals in $\mathbf{S}_\mathrm{A}$ is similar to the variability observed among the signals of $\mathbf{S}_\mathrm{B}$ then it is irrelevant which RC is used because both the RCs will lead to the same conclusion. There could be a difference in the RC-Test conclusion using different RC when: for example, if the shape standard deviation among pairs of signals in $\mathbf{S}_\mathrm{A}$ is significantly lower than the shape standard deviation among pairs of signals in $\mathbf{S}_\mathrm{B}$ then using the comparison of set $\mathbf{S}_\mathrm{B}$ to RC_A should be more relevant in the discrimination between $\mathbf{S}_\mathrm{A}$ and $\mathbf{S}_\mathrm{B}$ than using the comparison of set $\mathbf{S}_\mathrm{A}$ to RC_B. Heuristically, the fact that the signals in $\mathbf{S}_\mathrm{A}$ have a shape standard deviation among pair of them lower than the one observed for the signals in $\mathbf{S}_\mathrm{B}$, implies that the signals in $\mathbf{S}_\mathrm{A}$ are more "similar" to each other (i.e. smaller random error): they have more power to discriminate themselves from another set of (potentially different) signals.

Using Eq. (4) for set A and substituting A with B to get the same analysis for set B, each RC (RC_A and RC_B) is used within its own set to determine the shape coefficient correlation. A visual assessment is presented in fig 5. The shape correlations for A and B are shown in the first two columns of Fig. 5. They could be used as one aspect of the measure of repeatability of the dummy in consideration. The Green and Red horizontal lines indicate the acceptability criteria. The Green level acceptability criterion amounts to 97.5%. The 95% level is indicated by the Red horizontal dotted line. Since both medians are above the Red line, this method of comparison indicates that both dummies are repeatable at the 95% level.

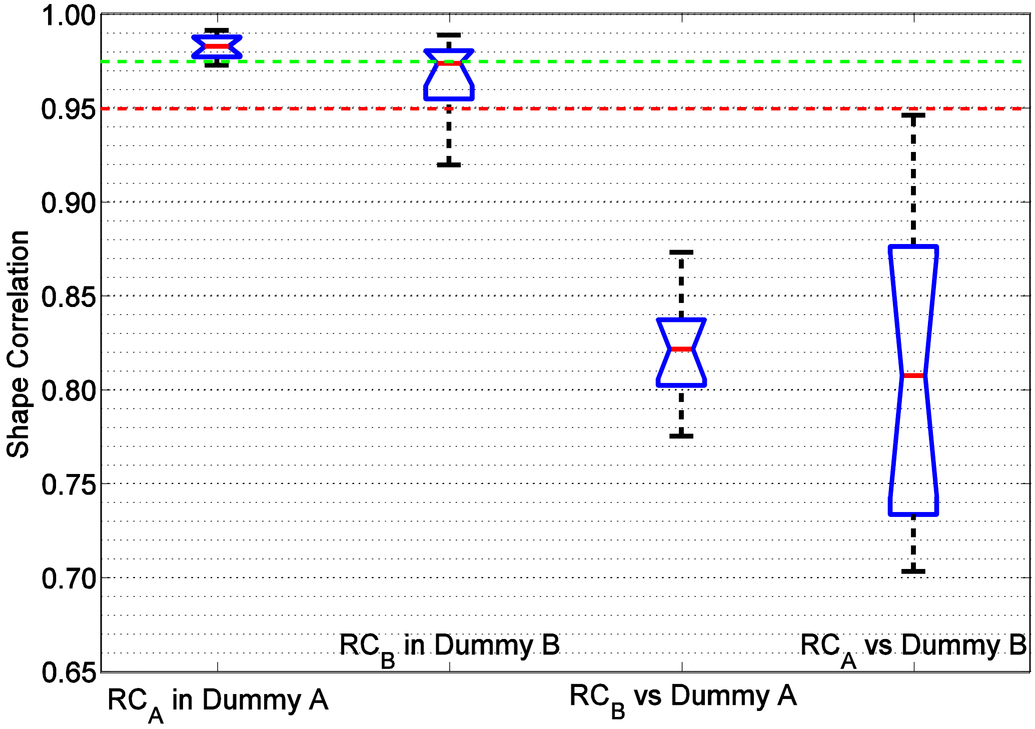

Figure 5: RC-Test – Dummy A versus Dummy B

The analysis across the sets is performed using Eq. (5) and its mirror equation by switching the subscripts A and B and performing a formal RC-Test as described above. Namely $\mathrm{RC_A}$ to the signals in the set of Dummy B or conversely $\mathrm{RC_B}$ to the signals in the set of Dummy A, would assess the reproducibility of the dummy. Box plots of these shape correlations are shown in columns 3 and 4 in Fig. 5. The medians of the box plots are much lower than the 95% acceptable limit and thus the dummy is considered not reproducible.

Interpretation of RC-Test

The RC-Test allows performing a comparison between two sets of signals using simultaneously the signal magnitude and shape information. Considering that the test with respect to the magnitude is straightforward in this section we will focus only on the shape factor comparison implications.

The RC-Test between two sets of signals (for example, $\mathbf{S}_\mathrm{A}$ and $\mathbf{S}_\mathrm{B}$) can lead to one of the following possible conclusions with respect to the shape correlation factor. There are a total of nine (9) cases. They are grouped, per their relevance and association, in three separate sets as follows:

Set One: Valid Comparisons

Set 1 - Case 1: $\overline{p(\mathbf{S}_\mathrm{A}, \mathrm{RC_A})} > \overline{p(\mathbf{S}_\mathrm{B}, \mathrm{RC_A})}$ and $\overline{p(\mathbf{S}_\mathrm{B}, \mathrm{RC_B})} > \overline{p(\mathbf{S}_\mathrm{A}, \mathrm{RC_B})}$

Set 1 - Case 2: $\overline{p(\mathbf{S}_\mathrm{A}, \mathrm{RC_A})} = \overline{p(\mathbf{S}_\mathrm{B}, \mathrm{RC_A})}$ and $\overline{p(\mathbf{S}_\mathrm{B}, \mathrm{RC_B})} = \overline{p(\mathbf{S}_\mathrm{A}, \mathrm{RC_B})}$

*First Set - First Case***.** In the first case the Representative Curve of the first set (i.e. $\mathrm{RC_A}$) has a significantly higher average shape correlation with the signals of $\mathbf{S}_\mathrm{A}$ than with the signals of $\mathbf{S}_\mathrm{B}$. Also, $\mathrm{RC_B}$ has a significantly higher average shape correlation with the signals of $\mathbf{S}_\mathrm{B}$ than with the signals of $\mathbf{S}_\mathrm{A}$. Thus, each one of the two RCs has a significantly higher average shape with the set of signals used to generate it than with the other set of signals. Heuristically speaking, in this first case, each RC is "drastically more similar" in shape with the signals of the sets of which it is the representative curve than with the signals of the other set. Consequently there is evidence to conclude that the two sets of signals have a different signal shape.

First Set - Second Case. In the second case there is evidence to conclude that the two sets of signals have similar shape: observationally, either of the two RCs, when used within its own set or the other set would yield the same mean shape coefficient.

The next two sets (set two and set three) of RC test results are uninformative and do not allow the determination of the similarities or differences between the two RCs. There are two possible reasons for this loss of statistical power: significant noise contamination and one or more of the RCs is invalid. In the first of these two sets (set two), it is possible that the RC is valid and additional statistical work or increased data will allow a determination of the difference or similarity of the two RCs. In the second set of these two sets (set three), it is unlikely that additional statistical work or increased data will be beneficial because the RC is most likely invalid.

Set Two: Significant Noise Contamination

Set 2 - Case 1: $\overline{p(\mathbf{S}_{\mathrm{A}}, \mathrm{RC}_{\mathrm{A}})} > \overline{p(\mathbf{S}_{\mathrm{B}}, \mathrm{RC}_{\mathrm{A}})}$ and $\overline{p(\mathbf{S}_{\mathrm{B}}, \mathrm{RC}_{\mathrm{B}})} = \overline{p(\mathbf{S}_{\mathrm{A}}, \mathrm{RC}_{\mathrm{B}})}$

Set 2 - Case 2: $\overline{p(\mathbf{S}_{\mathrm{A}}, \mathrm{RC}_{\mathrm{A}})} = \overline{p(\mathbf{S}_{\mathrm{B}}, \mathrm{RC}_{\mathrm{A}})}$ and $\overline{p(\mathbf{S}_{\mathrm{B}}, \mathrm{RC}_{\mathrm{B}})} > \overline{p(\mathbf{S}_{\mathrm{A}}, \mathrm{RC}_{\mathrm{B}})}$

Set 2 - Case 3: $\overline{p(\mathbf{S}_{\mathrm{A}}, \mathrm{RC}_{\mathrm{A}})} < \overline{p(\mathbf{S}_{\mathrm{B}}, \mathrm{RC}_{\mathrm{A}})}$ and $\overline{p(\mathbf{S}_{\mathrm{B}}, \mathrm{RC}_{\mathrm{B}})} > \overline{p(\mathbf{S}_{\mathrm{A}}, \mathrm{RC}_{\mathrm{B}})}$

Set 2 - Case 4: $\overline{p(\mathbf{S}_{\mathrm{A}}, \mathrm{RC}_{\mathrm{A}})} > \overline{p(\mathbf{S}_{\mathrm{B}}, \mathrm{RC}_{\mathrm{A}})}$ and $\overline{p(\mathbf{S}_{\mathrm{B}}, \mathrm{RC}_{\mathrm{B}})} < \overline{p(\mathbf{S}_{\mathrm{A}}, \mathrm{RC}_{\mathrm{B}})}$

Set Two – First and Second Cases. These cases have been generated using valid RCs and Gaussian random noise. In general they are borderline situations in which the data do not provide enough evidence to make a definitive conclusion about the two sets under examination: The two sets may present a difference in shape that it is not detectable simultaneously by the two RCs. As an example, consider case 1 and case 2 which are mirror images of each other by switching RC_{A} with $\mathrm{RC}_{\mathrm{B.}}$ In the first case, RC_{A} is a good discriminator between the two sets and the partial conclusion is that the two sets are different. On the other hand, RC_{B} does not discriminate between the two sets and thus we do not have enough evidence for a definitive resolution on the shape similarity of the two sets.

This situation may arise when the difference between the two sets is comparable to the difference generated by the percentage of random noise. Therefore, if there is a difference the RC of one set may be able to pick up the difference in shape between the two sets while the RC of the other set could be confounded by the noise. Additional information or analysis is needed to help to settle the comparison conclusion.

Set Two - Third and Fourth Cases. These cases have been generated using valid RCs and Gaussian random noise and can be generated when there is a significant difference in random noise between the two sets. In general, it is expected that the following holds for any two sets of signals, $\mathbf{S}_{\mathrm{i}}$ and $\mathbf{S}_{\mathrm{j}}$, which have a similar percentage of random noise:

$$\overline{p(\mathrm{RC}_{\mathrm{i}}, \mathbf{S}_{\mathrm{i}})} \geq \overline{p(\mathrm{RC}_{\mathrm{i}}, \mathbf{S}_{\mathrm{j}})} \tag{6}$$

Equation (6) states that the mean shape coefficient correlation between a Representative Curve, RC_{i}, and set $\mathbf{S}_{\mathrm{i}}$, of which it is the representative curve of, is higher than the mean shape coefficient when correlated with any other set of signals. The inequality in (Eq. (6)) holds as long as the sets under investigation have a similar level of random noise and there is significant difference in the underlying signal in which case the comparison test will lead to one of the first four possible conclusions listed above, namely Set One case 1 and 2 and Set Two case 1 and 2.

When there is a relevant difference of random noise between the sets of signals under investigation then the relationship in Eq. (6) could be reversed. Performing a comparison test may lead to either of the cases listed above. An example showing this facet has been constructed for two sets, $\mathbf{S}_{\mathrm{A}}$ and $\mathbf{S}_{\mathrm{B}}$, such that $\mathbf{S}_{\mathrm{A}}$ includes 20 uni-modal (single hump Gaussian-like) signals with 10% Gaussian noise while $\mathbf{S}_{\mathrm{B}}$ includes 20 single hump signals, the same as $\mathbf{S}_{A,}$ with 20% Gaussian noise. Thus, the base signals are the same for both sets and the only difference is the level of random noise: 10% for $\mathbf{S}_A$ and 20% for $\mathbf{S}_B$. This implies that theoretically RC_A and RC_B should be essentially the same curve because the two sets are generated from the same base signals. However, RC_B will have a higher correlation shape with $\mathbf{S}_A$ rather than with $\mathbf{S}_B$ since $\mathbf{S}_A$ has a smaller random noise level. These cases could also emerge when the two sets, under investigation, present both a relevant difference in random noise and a difference in signal shape.

However, it should be noted that, in general, the case with a high differences in noise levels between the sets may be inherently challenging to any techniques using the RC.

Set Three: Possible Invalid RC

Set 3 - Case 1: $\overline{p(\mathbf{S}_A, RC_A)} = \overline{p(\mathbf{S}_B, RC_A)}$ and $\overline{p(\mathbf{S}_B, RC_B)} < \overline{p(\mathbf{S}_A, RC_B)}$

Set 3 - Case 2: $\overline{p(\mathbf{S}_A, RC_A)} < \overline{p(\mathbf{S}_B, RC_A)}$ and $\overline{p(\mathbf{S}_B, RC_B)} = \overline{p(\mathbf{S}_A, RC_B)}$

Set 3 - Case 3: $\overline{p(\mathbf{S}_A, RC_A)} < \overline{p(\mathbf{S}_B, RC_A)}$ and $\overline{p(\mathbf{S}_B, RC_B)} < \overline{p(\mathbf{S}_A, RC_B)}$

We were unable to generate the above cases using valid RCs and Gaussian random noise. Most likely, there has been a violation of the necessary conditions to define a valid RC. Consequently when these conditions occur, they give evidence that the RC is not valid and therefore the statistics using the RC may not be valid.

The third case can be easily recognized as an unlikely case. The first inequality of the third case implies that the noise level of S_A is significantly higher than the noise level of S_B. On the other hand the second inequality implies the reverse, i.e. that the noise level of S_B is significantly higher than the noise level of S_A. Thus the contradictory statements imply the third case is not likely under these conditions. Similarly under the same assumptions and example, it can be shown that cases 1 and 2 are unlikely. For example, consider case 2, the first inequality implies that the noise level of S_A is significantly higher than the noise level of S_B; On the other hand the second equality implies that the noise level of S_A must be less than the noise level of S_B reaching a contradiction.

Cross Correlation Test

The Cross-Correlation test compares two full sets of the signals shape coefficients and thus the RC curve is not needed for this comparison technique. Using this approach, any two sets of signals will be considered to be similar when they present similar average shape correlation among the pairs of signals. In particular, this procedure is based on the comparison between the average shape correlation among pairs of signals belonging to the same set and the average shape correlation among pairs of signals formed taking one signal from each set under examination. It is important to recognize that this procedure is "valid" only if the signals belonging to the same set are independent replications of the same phenomenon.

The technical details of this approach are reported in the following. If $\mathbf{S}_A$ and $\mathbf{S}_B$ are two sets of signals, such that $\mathbf{S}_A = \{x_1, x_2, \cdots, x_M\}$ and $\mathbf{S}_B = \{y_1, y_2, \cdots, y_N\}$, then using the cross-correlation approach the two sets are considered to have the same shape when it can be concluded that the following two hypotheses hold:

$$\begin{aligned} &\overline{p(\mathbf{S}_A, \mathbf{S}_A)} = \overline{p(\mathbf{S}_A, \mathbf{S}_B)} \\ \text{and} \quad &\overline{p(\mathbf{S}_B, \mathbf{S}_B)} = \overline{p(\mathbf{S}_A, \mathbf{S}_B)} \end{aligned} \tag{7}$$

Where,

$$\overline{p(\mathbf{S}_A, \mathbf{S}_A)} = \left(\frac{1}{M \times (M-1)/2}\right) \sum_{i=1}^{M} \sum_{j>i}^{M} p(x_i, x_j) \tag{8}$$

is the average of the shape coefficients for the signals in $\mathbf{S}_A$. Assuming that this set is a collection of responses of some mechanical system subject to the same repeated input, then this average shape coefficient would be a measure of the repeatability of this mechanical system "A."

$$\overline{p(\mathbf{S}_\mathrm{B}, \mathbf{S}_\mathrm{B})} = \left(\frac{1}{\mathrm{N} \times (\mathrm{N}-1)/2}\right) \sum_{i=1}^{\mathrm{N}} \sum_{j>i}^{\mathrm{N}} p(y_i, y_j) \tag{9}$$

is the average of the shape coefficients for the signals in $\mathbf{S}_B$. Similarly, if the data collected for this set are responses to the same repeated input, then this average shape coefficient would be considered a measure of the repeatability of this mechanical system "B."

$$\overline{p(\mathbf{S}_\mathrm{A}, \mathbf{S}_\mathrm{B})} = \left(\frac{1}{\mathrm{M} \times \mathrm{N}}\right) \sum_{i=1}^{\mathrm{M}} \sum_{j=1}^{\mathrm{N}} p(x_i, y_j) \tag{10}$$

is the average of the shape coefficients between the signals in $\mathbf{S}_A$ and $\mathbf{S}_B$. The correlation here being across the two mechanical systems, this average shape coefficient would be a measure of the reproducibility of the mechanical system.

The hypotheses (Eq. (7)) above can be studied using, for example, two t-tests, one for each equality. When there is evidence that $\overline{p(\mathbf{S}_\mathrm{A}, \mathbf{S}_\mathrm{A})} > \overline{p(\mathbf{S}_\mathrm{A}, \mathbf{S}_\mathrm{B})}$ and $\overline{p(\mathbf{S}_\mathrm{B}, \mathbf{S}_\mathrm{B})} > \overline{p(\mathbf{S}_\mathrm{A}, \mathbf{S}_\mathrm{B})}$ then it can be concluded that the signals of the two sets, $\mathbf{S}_A$ and $\mathbf{S}_B$, have different shapes. On the other hand, when there is evidence that $\overline{p(\mathbf{S}_\mathrm{A}, \mathbf{S}_\mathrm{A})} = \overline{p(\mathbf{S}_\mathrm{A}, \mathbf{S}_\mathrm{B})}$ and that $\overline{p(\mathbf{S}_\mathrm{B}, \mathbf{S}_\mathrm{B})} = \overline{p(\mathbf{S}_\mathrm{A}, \mathbf{S}_\mathrm{B})}$ then it is appropriate to conclude that the two sets of signals, $\mathbf{S}_A$ and $\mathbf{S}_B$, have similar shapes. For all the other possible relationships among the three means, defined in Eqs. (8) to (10), it is not possible to provide a clear interpretation of the shape relationship between the two sets. However, it has been observed that the cases for which it is not possible to provide a clear interpretation usually arise when either there is a relevant difference in the noise level between the two signal sets under comparison or when the noise level of the two signal sets is comparable with the signal shape difference between the two sets.

For the examples presented here, we have twelve (12) repeat tests for each dummy. For repeatability analysis, the analysis within each set, we have sixty six (66) pairs of combination for the data. Similarly, when studying the reproducibility of the dummy, one would seek the shape coefficients from one set against the other(s) which lands us one hundred forty four (144) combinations. After collecting these numerical estimates, one could use standard statistics tools to evaluate a variety of measures such as averages, standard deviations, means, box plots, etc.

Figure 6 below shows the repeatability and the reproducibility box plots for the two dummies used in this example. The neck moment is shown here and the plots highlight the distribution of all of the shape coefficients in a concise manner. The median of the shape coefficient correlation for dummy B repeatability is less than 0.95 which suggests that the dummy is not repeatable. The median of the shape coefficient correlation, when comparing the two sets, is lower than 0.81, and is much lower than either of the medians for each dummy, which is a clear indication that the response of the dummy is not reproducible.

It is important to notice that when there is a relevant difference in the error level between the two signal sets under comparison then it is challenging to compare the two sets with any statistical methodology including the two methodologies considered in this study.

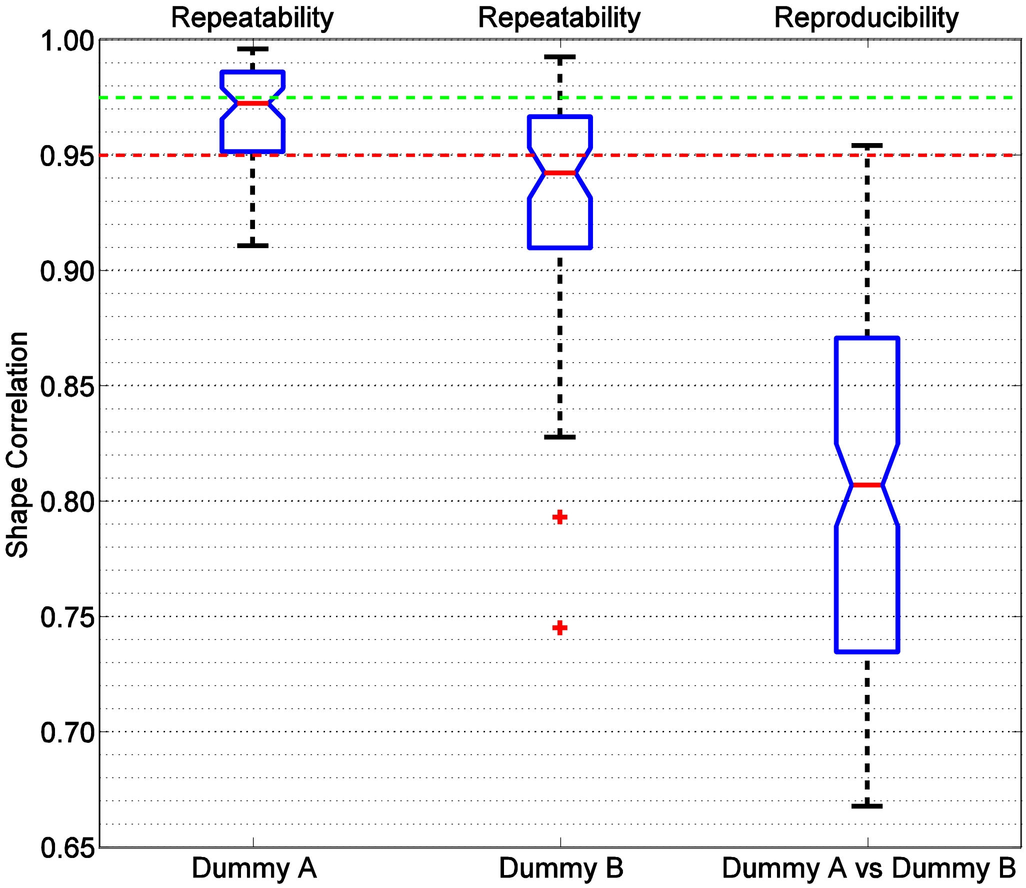

Figure 6: Shape Comparison – Dummy A versus Dummy B

DISCUSSION

The two approaches proposed are simple, reliable and consistent. Both methods can be used to address repeatability and reproducibility of mechanical devices such as ATDs. The differences between the two approaches are: The cross-correlation analysis determines if members of one set are similar to each other, repeatability, and that one set is similar to a second set, reproducibility. The RC-test determines if the RC from one set is similar to the time-histories in the set that it represents, repeatability and that the RC from one set is similar to the RC from a second set, reproducibility. Although in many cases the results from either test or set of tests will produce equivalent results, it is not always the case that one test can be substituted for the other: Which test is chosen will depend on the objects of the study.

In general, if the object of the study is to compare the two sets of time-histories from a mechanical system to determine if they come from the same population and are similar, then it would be appropriate to use the cross-correlation approach. In many cases, however, a visual representation of the average response is desired and an average curve (RC) is generated. Nevertheless, attention should be paid in the RC generation process, because when the signal set under examination presents extreme signal noise or it includes signals that cannot be considered independent samples of the same phenomenon then the RC is not informative for the purpose of studying the shape of the signals in the set (Nusholtz et al. 2010).

The objective is not to determine that the two sets are the same but that the average response is the same. To determine if the average responses of two systems are the same or if the response of one system is similar to a standard, the RC-test can be used. For example, assume a RC has been generated for a set of biomechanical data generated from a series of tests using Post Mortem Human Surrogates (PHMS).

Determining which ATD design is a better estimate of the PMHS response can be done through the use of RC-testing.

In addition, there are some minor operational differences. For example the cross-correlation test requires more operations than the RC-Test method after the RC has been created. In fact the within set comparison of the cross-correlation analysis involves $\mathrm{M} \times (\mathrm{M}-1)/2$ shape calculations, see Eq. (8), compared to just M shape calculations for the RC-Test analysis, see Eq. (4). Similarly, the analysis for across the sets, the shape correlations count for the cross-correlation analysis is $\mathrm{M} \times \mathrm{M}$, assuming the same number of tests in each set, compared to just M operations for the RC-Test. The RC-Test method obviously requires the evaluation of the Representative Curves (RCs) which are not required for the cross-correlation analysis. In the example used in this document, we have twelve (12) repeat tests per set. We performed sixty six (66) shape comparison calculations for the cross-correlation analysis with the sets compared to a similar analysis of twelve (12) operations for the RC-Test. For the comparison across the sets, the counts is one hundred forty four (144) versus twelve (12), for the cross-correlation tests as compared to the RC-Test, respectively.

The reproducibility box plots for both the cross-correlation and the representative curve analyses clearly indicate that the dummy is not reproducible as shown in Fig. 7 columns five through seven. On the other hand, the two methods do not suggest the same conclusion for repeatability. In fact, the cross-correlation analysis suggests that the dummy is not repeatable as the median of the third box plot of Fig. 7 is below the cutoff 0.95 whereas the medians of both the second and fourth box plots showing RC-Test analysis are above the cutoff and thus considered repeatable.

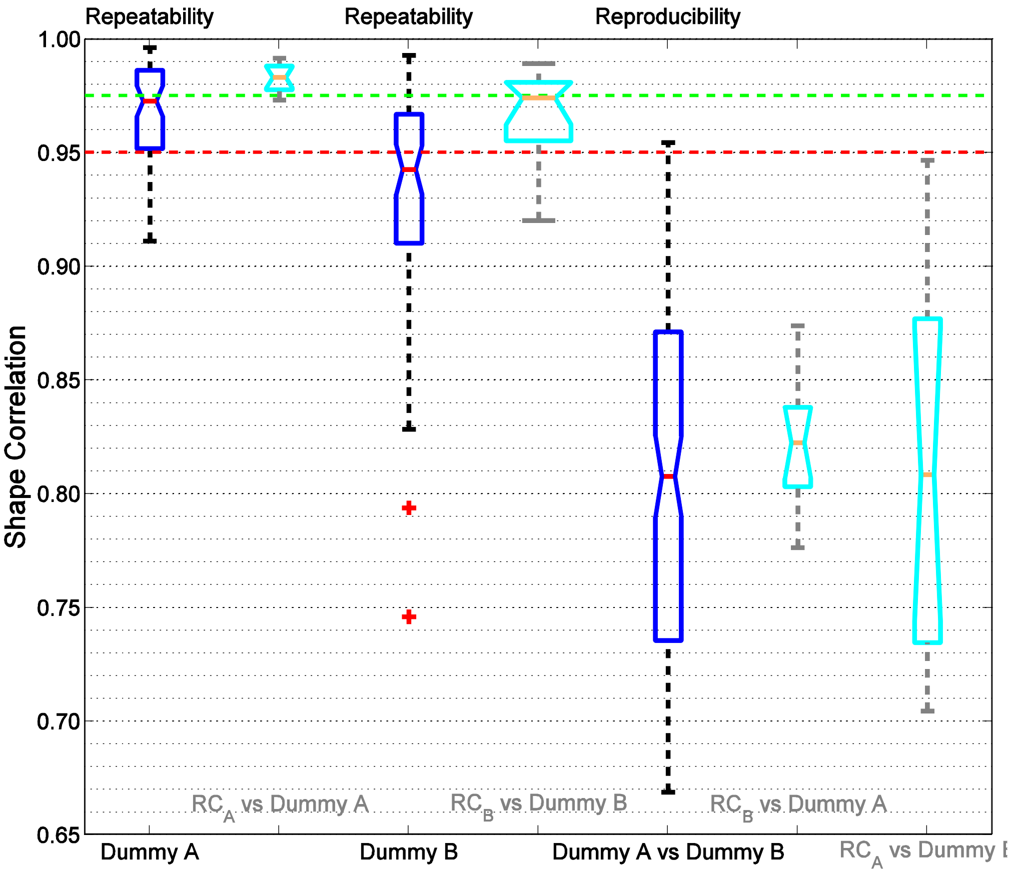

Figure 7: Cross-Correlation/RC-Test – Dummy A versus Dummy B

CONCLUSION

Two statistical methods have been developed to determine if a mechanical system is repeatable and reproducible. The mechanical system used as an example was an ATD, the BioRID II.

In both methods, sets of the mechanical response (time-histories) of a mechanical system resulting from impacts are formed, the set of neck moments time histories in the BioRID II for example. In the case of repeatability the impacts would be repeated nondestructive impacts on a single sample of the mechanical system. In the case of reproducibility more than one sample of the mechanical system is impacted and the impact can be either destructive or nondestructive. In the examples used in this paper all the impacts were non-destructive.

In the first method, the RC-test, a representative curve is generated for each set of time-histories. Repeatability is determined by statistically evaluating, using t and T^2-tests, the similarity of the RC in terms of shape and magnitude to members of the set that it was derived from. Reproducibility is determined also by statistical evaluations using t and T^2-tests. In this case the statistical test evaluates how close the RC obtained from the first mechanical system is in terms of shape and magnitude to the data of the second mechanical system and how close the second RC is in terms of shape and magnitude to the data of the first mechanical system.

The procedure to evaluate the RC has, as its purpose, the goal of providing an estimate of what the underlying signal without the noise component should be. This is in general, informative and valuable; however, when the RC is not a valid representation of the underlying signal then the analysis based on the RC is not representative of the set that was used to create it and the conclusions could be completely misleading. Attention should be paid to the signals before generating an RC for a signal set.

In the second method no RC is formed. In the case of repeatability each member of a set of time-histories obtained from a set of impacts to a single system is compared to every other member in terms of shape and magnitude. For reproducibility each member of a set of time-histories resulting from a set of nondestructive impacts is compared to every member of a set of nondestructive impacts to a second system in terms of shape and magnitude. Once this comparative set is obtained the shape and magnitude in terms of the average and distribution can be statistically compared, using t-test, to the average and distribution used to form the two repeatability average and distributions.

Fundamentally the two methods are not testing the same conditions. In the first method (RC-test) repeatability is defined as to how close the RC is to the data used to form it and reproducibility is testing how close two RCs are. In the second method repeatability is testing how close each time-history in a set is to the other time histories and reproducibility is testing how close the member in one set of time-histories obtained from one sample of a mechanical system is to a second set of time-histories obtained from a second sample of a mechanical system.

ACKNOWLEDGEMENTS

We are greatly indebted to the Partnership for Dummy Biomechanics (PDB) for their gracious sharing of their test data for the example analyses provided in this paper.

REFERENCES

BORTENSCHLAGER, K., HARTLIEB, M., HIRTH, A., KRAMBERGER, D., and STAHLSCHMIDT, S. (2009). Detailed Analysis of BioRID-II Response Variations in Hardware and Simulation. 21st International Technical Conference on the Enhanced Safety of Vehicles (ESV), Stuttgart, Germany. Paper Number 09–0492

CASELLA, G. and BERGER, R. L. (2001). Statistical Inference. Duxbury Press, Second edition.

Dodd W.G., Eilbeck J.C., Gibbon J.D., and Morris H.C. Solition and Nonlinear Wave Equations. Academic Press, London, 1984.

GEHRE, C., GADES, H., and WERNICKE, P. (2009). Objective Rating of Signals Using Test and Simulation Responses. 21st International Technical Conference on the Enhanced Safety of Vehicles (ESV), Stuttgart, Germany. Paper Number 09–0407.

McGill, R., TUKEY, J. W., and LAREN, W. A. (1978). Variations of Box Plots. The American Statistician, Vol. 32, No. 1, pp. 12–16.

NUSHOLTZ, G. S., HSU, T. P., and BYERS, L. C (2007). A Proposed Side Impact ATD Bio-Fidelity Evaluation Scheme Using Cross-Correlation Approach. 20th International Technical Conference on the Enhanced Safety of Vehicles (ESV), Lyon, France. Paper Number 07–0399.

NUSHOLTZ, G. S., HSU, T. P., SHI, Y., BABAII KOCHEKSERAII, S., and GRACIAN LUMA, M. A. (2009). Creating Representative Curves from Multiple Time Histories of Vehicle, ATD and Biomechanics Tests. 21st International Technical Conference on the Enhanced Safety of Vehicles (ESV), Stuttgart, Germany. Paper Number 09–0249

NUSHOLTZ, G. S., HSH, T. P., GRACIAN LUMA, M. A., DI DOMENICO, L., and BABAII KOCHEKSERAII, S. (2010). The Consequence of Average Curve Generation: Implications for Biomechanics Data. Stapp Car Crash Journal, Vol. 54.

XU, L., AGARAM, V., ROUHANA, S., HULTMAN, R. W., KOSTYNIUK, G. W., McCLEARY, J., MERTZ, H., NUSHOLTZ, G. S., and SCHERER, R. (2000). Repeatability Evaluation of the Pre-Prototype NHTSA Advanced Dummy Compared to the HYBrid III. SAE Technical Paper Series 2000-01-0165.