

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

Guy S. Nusholtz, Zine Aoun, Laura Di Domenico, Timothy Hsu, Manuel A. Gracián and
Jesús A. Prado
Chrysler Group LLC

ABSTRACT

Reliable testing of a mechanical system requires the procedures used for the evaluation to be repeatable and reproducible. However, it is never possible to exactly repeat or reproduce the tests that are used for evaluation. To overcome this limitation, a statistical evaluation procedure can generally be used. However, most of the statistical procedures use scalar values as input without the ability to handle vectors or time-histories. To overcome these limitations, two numerical/statistical methods for determining if the impact time-history response of a mechanical system is repeatable or reproducible are evaluated and elaborated upon. Such a system could be a vehicle, a biological human surrogate, an 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 those response time histories. The two statistical procedures use the t and T2-tests. One uses a statistical comparison between the average time history of a set (Representative Curve (RC)) and the individual time histories of that set or other sets. The other procedure uses a statistical comparison of sets of time histories.

CITATION: Nusholtz, G., Aoun, Z., Di Domenico, L., Hsu, T. et al., "Statistical Considerations for Evaluating Biofidelity, Repeatability, and Reproducibility of ATDs," *SAE Int. J. Trans. Safety* 1(1):2013, doi:10.4271/2013-01-1249.

INTRODUCTION

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 an ATD, this operation has two basic goals: one, to directly compare an ATD with itself, repeatability; two, to compare one ATD to one or more 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 anatomical structure can generate a collection (set) of response time histories as a result of the impact. 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 with every other time-history in the set. For reproducibility, the time histories in one set are compared with 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 with each value in that set. Reproducibility compares the average of one set with 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 time histories of accelerations, velocities, forces, etc., standard statistics to study R & R may not apply.

One possible approach for 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; finally add, for each system, the multiple tests together obtaining the system averages. Statistics can

then be performed on the averages. However, if the “underlying mechanical system” of each sample is not the same, a similar response under impact is not produced with the only difference that it carries noise contamination, then it is not altogether clear what the average represents. The average 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 from sets of vehicle acceleration time-histories, and then creating sets of shape comparisons among and between sets of vehicles. The shape coefficient values were first transformed and then analyzed using the ANalysis Of VAriance (ANOVA) procedure. This allowed them to study 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 are identically distributed. The noise is considered to be independent of the underlying response signal, normally distributed, stationary and, for small sample sizes, insignificant with respect to the signals. If these conditions are not met then addition of the signals may become 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 the meaning of the average of several time-histories from different samples is. Beyond the question of what the average means, there is the question about 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](#).

Although the statistical methods in this paper are aimed at biomechanical processes 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 in this paper is repeatability and reproducibility of a system.

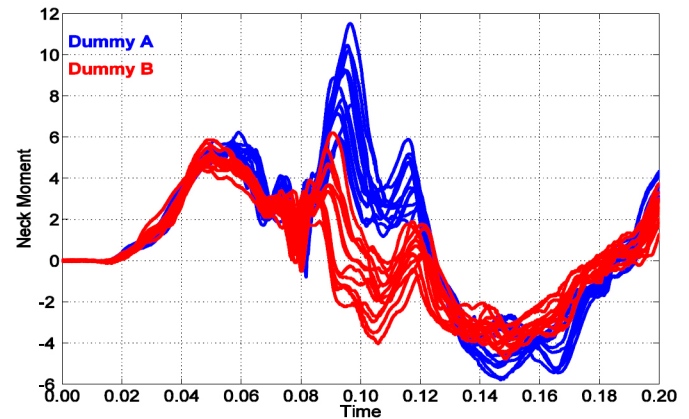


Figure 1. Dummy A and Dummy B - Neck Moment Signals

BACKGROUND FOUNDATION

Magnitude

The magnitude uses the norm of a vector defined by:

$$\|x\| := \sqrt{\int x^2(t) dt} \quad (1)$$

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. \(1\)](#). For ease of comparison, the magnitude from [\(Eq. \(1\)\)](#) 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 two dummies with twelve tests each are used, and we used the average magnitude of all two dummies for the normalization. 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. It is important to be careful in order to ensure that the analysis can be done using the relative magnitudes only.

Shape

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

$$p(x, y) := p(x, y, \tau) \quad (2)$$

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

$$p(x, y, h) = \frac{\int x(t)y(t+h) dt}{\sqrt{\int x^2(t) dt \int y^2(t) dt}} \quad (3)$$

Where h is a time lag on the autocorrelation.

Creation of the Representative Curve

A collection of a 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, a set of those measurements is formed. In general, the measurements will not be the same and an average curve will be a good Representative Curve (RC) of the set.

An RC is derived from a group of signals by averaging, 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, a detailed example of MTM application is reported in [Appendix C](#).

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, which is not optimum. In [Fig. 2](#) an example of RC for PMHS data, obtained from Maltese et al (2002) is presented. The RCs, for the signals of [Fig. 1](#), are shown below in [Figs. 3](#) and [4](#). RCA and RCB are the representative curves for dummies A and B, respectively.

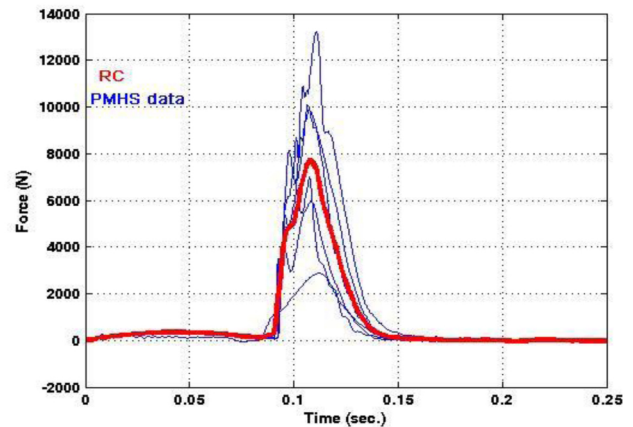


Figure 2. RC for PMHS data

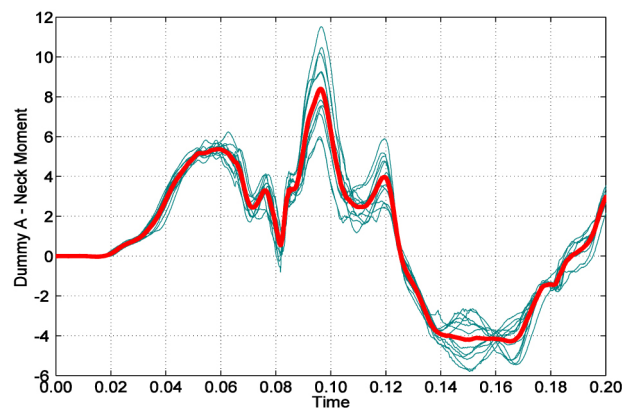


Figure 3. RCA for Dummy A

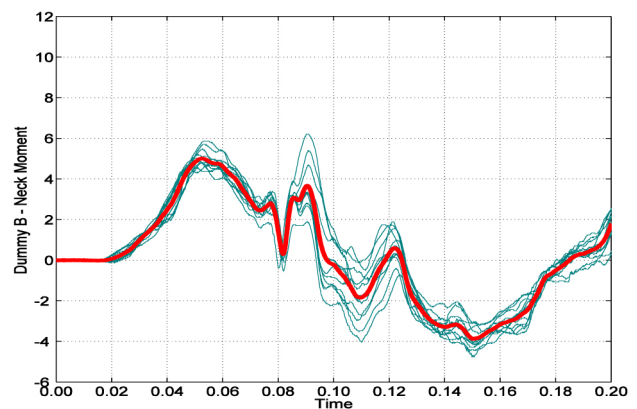


Figure 4. RCB for Dummy B

Average Shape Coefficient and RC

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 with: the given time-history is compared with all other time-histories in the set using the

shape function. 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 statements are true, (Nusholtz et al. 2010):

i. 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.

ii. 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.

iii. 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.

iv. When the cases i, ii or iii are not applicable, it is possible that the average signal will not have the best shape coefficient.

v. The cases i, ii, iii and iv are in agreement with the Central Limit theorem for random vectors.

Statement iv 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 with 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.

REPEATABILITY ANALYSIS

Repeatability is a considerable concern in the use of ATDs in crash testing and the analysis of a biomechanical data set. In particular, it should be observed whether the histograms of the magnitude and the correlation shape (between signals of the same set) show a uni-modal distribution with mean value above a specified threshold, the possible presence of outliers should be checked and the distribution and the variability of the data set should be evaluated.

A plot of the correlation shape values between all the possible couples of time history replications of the same mechanical system provides an easy visual assessment of potential outliers. A signal that has a shape that does not

resemble the shape of the other signals of the set under examination will present a poor correlation shape value with all the other signals in the set, so that from a visual inspection of the cross correlation shape values it can be easily detected. Time histories that present a significant low correlation shape with all the other repeated time histories of the same mechanical system should be further analyzed to verify their validity. For the example presented here, we have twelve (12) repeated tests for each dummy, thus for the repeatability analysis the correlation shape of sixty six (66) pairs of combination for the data of each dummy is evaluated. Box-plots of the magnitude and the correlation shape values are presented in Fig. 5, Fig. 6, Fig. 7, Fig. 8.

An empirical analysis based on data of an extensive number of ATD tests across a wide range of ATDs (Hybrid III, THOR, EuroSID, EuroSID-re, WorldSID50, SID-IIs, BioRID, etc.) has shown that, in general, a set of signals representing a single ATD can indicate an “acceptable” repeatability if the correlation shape distribution is uni-modal with a mean shape value that is above a maximum threshold, for example, it has been observed that a value sufficiently high such as, for example, either 95% or 97% should provide sufficient confidence to ensure the repeatability necessary for most ATD use in the evaluation of impact response.

Fig. 5 below shows the box plots for the correlation shape values for the two dummies used in this example. The median of the shape coefficient correlation for dummy B repeatability is less than 0.95, which suggests that is not repeatable for many impact response evaluations.

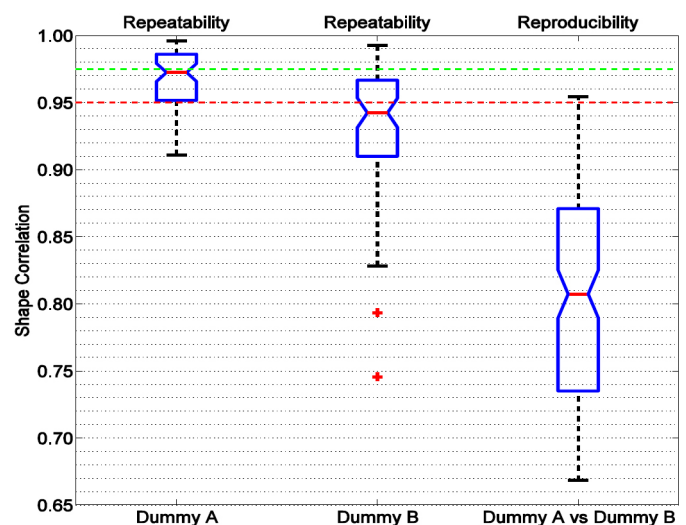


Figure 5. Shape Comparison - Dummy A versus Dummy B

The normalized magnitudes from the signals of Fig. 1 are shown in Fig. 6. The notches around the median values represent about 1.7 standard deviations (see McGill et al. 1978 for further details about the box plots). The notches for both dummies are smaller than 0.07 indicating standard deviations of less than 0.05 which suggests that both

dummies could be considered repeatable with respect to magnitude using this assessment. The above approach to assess repeatability does not use the RC. It should be observed that when the mechanical system is not repeatable then the representative value of the RC could be affected either by the possible presence of influential outliers among the time histories or by the lack of repeatability (Nusholtz et al. 2010). Thus the repeatability should be checked using the correlation shape values among the time histories instead of using the RC.

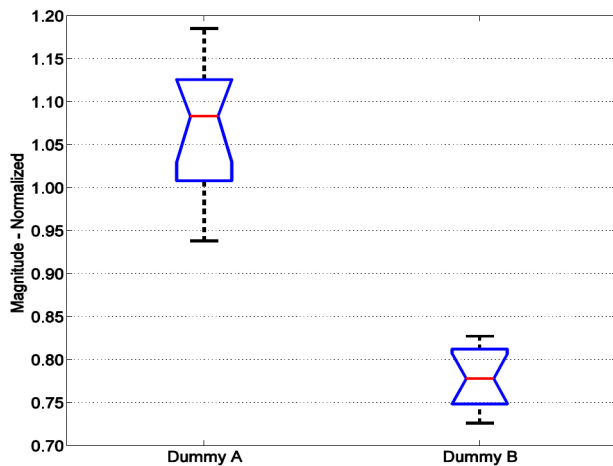


Figure 6. Magnitude analysis - Dummy A versus Dummy B

In the following, as an example, is presented a possible approach of repeatability analysis using the Representative Curve. In this case, once the RC has been evaluated, the RC is used within its own set to determine the shape correlation between the RC and each of the time-histories of its own set. The average shape coefficient with the shape coefficient distributions can then be used to estimate the repeatability of the set of time-histories.

The shape (RC) correlations for A and B are shown in the first two columns of Fig. 7. The Green and Red horizontal lines could be used as one aspect of the measure of repeatability for the dummy in consideration.

For example, an acceptability criterion of 97.5%, i.e. the Green level in Fig. 7, or an acceptability criterion of 95%, i.e. level the Red horizontal dotted line in Fig. 7 could be used. For example, since both medians are above the Red line, this method of comparison indicates that both dummies are repeatable at the 95% level, using the RC-correlation shape approach.

However, it should be noted that the two most extreme cases noted using the correlation shape between signals) are completely confounded when the repeatability is analyzed using the RC (second box-plot from the left in Fig. 7).

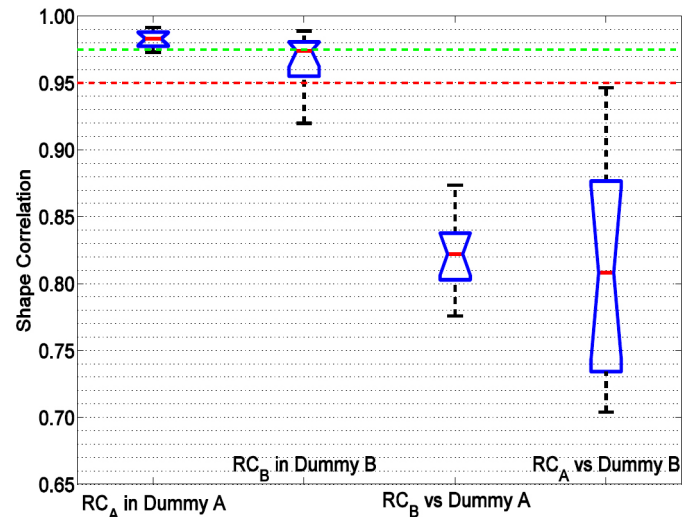


Figure 7. RC-Box Plots - Dummy A versus Dummy B

In the following section two approaches are introduced to study the mechanical system reproducibility: the Representative Curve Test and the Cross-correlation Test. However, when the mechanical system is found non-repeatable then there is no reason to examine reproducibility (see the flow-chart presented in Appendix B).

REPRODUCIBILITY ANALYSIS

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, 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 and magnitude of the signal set.

In other words, the RC is a “valid” summary of the signal shape and magnitude for the signal set only when all the signals belonging to the same set are independent replications of the same phenomenon. Consequently, the shape and magnitude are functionally the same for each signal, the only difference 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 so that the RC will be a valid and informative curve for analyzing the shape and magnitude of the signal set.

For example, when the average correlation 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 may not be a valid representation of the average signals: any analysis based on that RC could lead to misleading and unreliable conclusions.

Before defining the RC-Test, it is necessary to introduce some of the notations used. Given the set of signals $\mathbf{S}_A = \{x_1, x_2, \dots, x_M\}$ with representative curve RC_A then:

- Magnitude values of the signals in $\mathbf{S}_A$

$$\|x_1\|, \|x_2\|, \dots, \|x_M\|$$

- Mean magnitude in set $\mathbf{S}_A$

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

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

$$p(x_1, RC_A), p(x_2, RC_A), \dots, p(x_M, RC_A)$$

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

$$p(x_1, RC_B), p(x_2, RC_B), \dots, p(x_M, RC_B)$$

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

$$\overline{p(\mathbf{S}_A, RC_A)} = \overline{p(RC_A, \mathbf{S}_A)} = \frac{1}{M} \sum_{i=1}^M p(x_i, RC_A) \quad (4)$$

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

$$\overline{p(\mathbf{S}_A, RC_B)} = \overline{p(RC_B, \mathbf{S}_A)} = \frac{1}{M} \sum_{i=1}^M p(x_i, RC_B) \quad (5)$$

Similar notations for set $\mathbf{S}_B$ of signals $\{y_1, y_2, \dots, y_N\}$ 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 three of the following equalities hold:

$$\overline{\|\mathbf{S}_A\|} = \overline{\|\mathbf{S}_B\|}$$

$$\overline{p(\mathbf{S}_A, RC_A)} = \overline{p(\mathbf{S}_B, RC_A)}$$

$$\overline{p(\mathbf{S}_A, RC_B)} = \overline{p(\mathbf{S}_B, RC_B)}$$

The RC-test is based on the multivariate two-sample T2-Test (Casella and Berger (2001)). The T2-Test relies on the assumptions that the data has a multivariate normal distribution and that the covariance matrices of the two sets of data are the same. In our evaluation of the method we have noticed that these assumptions could be reasonably accepted for all the time histories that we have analyzed even-tough a transformation of the correlation data would improve the normality of the distribution of the data (see [Appendix B](#) and [Appendix D](#)).

When the T2-Test leads to the conclusion 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 . Thus, there is evidence to conclude that they are reproducible. On the other hand, when the T2-Test rejects the hypothesis of similarity for the two sets of signals ($\mathbf{S}_A$ and $\mathbf{S}_B$), then the implication is that at least one signal characteristic among magnitude and shape correlations with respect to either RC_A or 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 $\mathbf{S}_B$ with respect to RC_A and one for the average shape relationship of set $\mathbf{S}_A$ with respect to RC_B). The t-tests would provide evidence to determine which of the three factors considered (i.e. magnitude, shape with respect to RC_A and shape with respect to RC_B) would lead to the difference observed between the two sets of signals. The RC-Test compares the shape of the two sets of signals ($\mathbf{S}_A$ and $\mathbf{S}_B$) using the representative curve of both the two sets; it should not be expected that both comparisons would always lead to the same conclusion.

In general, when the variability of the signals in $\mathbf{S}_A$ is similar to the variability observed among the signals of $\mathbf{S}_B$, then most likely 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 RCs when: for example, the shape standard deviation among pairs of signals in $\mathbf{S}_A$ is significantly lower than the shape standard deviation among pairs of signals in $\mathbf{S}_B$, then using the comparison set $\mathbf{S}_B$ to RC_A could be more relevant in the