Acceptance criteria for validation metrics in roadside safety based on repeated full-scale crash tests

Mario Mongiardini* and Malcolm H. Ray

Department of Civil and Environmental Engineering,
Worcester Polytechnic Institute,
100 Institute Road, Worcester,
MA 01609, USA
Email: mario@wpi.edu Email: mhray@wpi.edu
*Corresponding author

Marco Anghileri

Department of Aerospace Engineering,
Politecnico di Milano, via La Masa 34,
Milan 20156, Italy
Email: marco.anghileri@polimi.it

Abstract: This paper proposes acceptance criteria for quantitative comparison metrics to be applied in the Verification and Validation (V&V) process of computational models used in roadside safety. Typically, the degree of verification or validation of a numerical model is assessed by qualitatively comparing the shapes of two curves, but qualitative comparisons are subjective and open to interpretation. Using quantitative comparison metrics in the V&V process allows for an objective measure of the reliability of a numerical model. Two comparison metrics were selected from a group of 16 metrics found in the literature. Acceptance criteria suitable to the typical scatter of full-scale crash tests were established by comparing ten essentially identical vehicle redirectional crash tests. Since the tests were as identical as can be achieved experimentally, the values of the quantitative metrics represented the reasonable range for the metric corresponding to matched experiments. Typical residual errors expected in full-scale tests are also discussed.

Keywords: verification and validation; V&V; comparison metrics; full-scale crash tests; numerical simulations; roadside safety.

Reference to this paper should be made as follows: Mongiardini, M., Ray, M.H. and Anghileri, M. (2010) 'Acceptance criteria for validation metrics in roadside safety based on repeated full-scale crash tests', *Int. J. Reliability and Safety*, Vol. 4, No. 1, pp.69–88.

Biographical notes: Mario Mongiardini is a PhD candidate and a Research Assistant in Civil Engineering at Worcester Polytechnic Institute, USA. He holds his Master's degree in Mechanical Engineering from Politecnico di Milano, Italy. He previously participated in the European project 'ROBUST', during which he modelled full-scale crash tests between vehicles and barriers using the finite element method. His main research area is in computational mechanics, crashworthiness and roadside safety.

Malcolm H. Ray directs WPI's Impact Engineering Program and the Structural Mechanics Impact Laboratory. He is a Professor of Civil and Environmental Engineering at Worcester Polytechnic Institute, USA. He holds his BS, MS and PhD degrees in Civil Engineering from the University of Vermont, Carnegie Mellon University and Vanderbilt University, respectively. His research interests include using finite element methods to evaluate the crashworthiness of structures, in-service performance of traffic barriers and other roadside features, verification and validation methods in computational impact mechanics, and side-impact crash testing and evaluation methods. He has published nearly 80 articles, conference papers and reports and he holds patents on several roadside safety devices.

Marco Anghileri is an Assistant Professor in the Department of Aerospace Engineering at Politecnico di Milano, Italy. He holds his PhD degree in Aerospace Engineering from Politecnico di Milano. He is the director of the 'LAST' impact laboratory, and he has published several journal papers and conference articles in the field of crashworthiness for aeronautical structures and terrestrial vehicles. He has obtained several grants funded by the European community and has served as coordinator for the European project 'ROBUST' and the European Committee for standardisation on the numerical simulation of impacts between vehicles and barriers (CEN/TC226/TG1/CME).

1 Introduction

Assessing the degree of correspondence between two curves is a common task most engineers deal with daily. Often curves are compared either to verify or validate computer simulation models using data obtained from experimental tests or to assess the reproducibility of an experimental test (i.e. assess if different instances of the same test actually represent the same or similar physical events).

While in the past it was common to compare curves using subjective judgements, recently much attention has been given to using quantitative comparison metrics to measure how well two curves compare to each other (AIAA, 1998; DoD, 2003; ASME, 2006). As comparison metrics are mathematical measures of the agreement between two curves, they provide an objective way to compare a pair of curves. The many comparison metrics found in literature can be essentially grouped into two main categories: (1) deterministic and (2) stochastic metrics. Metrics within the deterministic group imply results are the same every time given the same input, while metrics of the stochastic group specifically address the statistical variation of experiments, calculation or both. In this paper, the behaviour of deterministic metrics only is investigated. Deterministic metrics can be further classified into three main types: (1) Magnitude-Phase-Composite (MPC) metrics, (2) single-value metrics and (3) Analysis of Variance (ANOVA) metrics.

MPC metrics treat the curve magnitude and phase separately using two different component metrics (i.e. M and P, respectively). The M and P component metrics are then combined together into a single-value comprehensive metric, C. In all MPC metrics, the phase component (P) should be insensitive to the magnitude differences but sensitive to differences in phasing or timing between the two curves. Similarly, the magnitude component (M) should be sensitive to differences in magnitude, but relatively insensitive to differences in phase. These characteristics of MPC metrics allow the analyst to identify the aspects of the curves that do not agree. For each component of the MPC metrics, zero indicates that the two curves are identical. The following five MPC metrics were found in literature: (1) Geers (1984), (2) Geers CSA (CSA, 1994), (3) Sprague-Geers (2003), (4) Russell (2006) and (5) Knowles-Gear (Schwer, 2007).

Each of the MPC metrics differs slightly in its mathematical formulation. The different variations of the MPC metrics are primarily distinguished in the way the phase metric is computed, how it is scaled with respect to the magnitude metric and how it deals with synchronising the phase. In particular, the Sprague-Geers metric (Sprague and Geers, 2003) uses the same phase component as the Russell metric (Russell, 2006). The magnitude component of the Russell metric is peculiar as it is based on a base-10 logarithm and it is the only MPC metrics among those found in literature to be symmetric (i.e. the order of the two curves is irrelevant). The Knowles-Gear metric (Schwer, 2007) is the most recent variation of MPC-type metrics. Unlike the previously discussed MPC metrics, it is based on a point-to-point comparison. In fact, this metric requires that the two compared curves are first synchronised in time based on the so-called Time of Arrival (TOA), which represents the time at which a curve reaches a certain percentage of the peak value. Typically, in literature the percentage of the peak value used to evaluate the TOA is 5%. Once the curves have been synchronised using the TOA, it is possible to evaluate the magnitude metric. Also, in order to avoid creating a gap between time histories characterised by a large magnitude and those characterised by a smaller one, the magnitude component M has to be normalised using the normalisation factor QS.

Single-value metrics give a single numerical value that represents the agreement between the two curves. The following single-value metrics were found in literature: (1) correlation coefficient metric (Cohen et al., 2003), (2) NARD correlation coefficient metric (Basu and Haghighi, 1988), (3) Zilliacus error metric (Whang et al., 1993), (4) RSS error metric (Whang et al., 1993), (5) Theil's inequality metric (Theil, 1975), (6) Whang's inequality metric (Whang et al., 1993) and (7) regression coefficient metric (Cohen et al., 2003). The first two metrics are based on integral comparisons, while the others are based on point-to-point comparisons (i.e. residuals at each time step are considered). The analytical formulation of the MPC and Single-value metrics previously cited is shown in the Appendix.

ANOVA metrics are based on the assumption that if two curves represent the same event then any differences between the curves must be attributable only to random experimental error (Ray, 1996; Oberkampf and Barone, 2006). The ANOVA is a standard statistical test that assesses whether the variance between two curves can be attributed to random error only. Conceptually, if two curves represent the same physical event, the mean residual error and the corresponding standard deviation should both be null; but, in practice, due to the presence of random experimental or numerical errors these two quantities are not exactly equal to zero. The conventional T-test represents an effective way to assess if the mean of the residual error is due to random errors only.

A review of the results and formulation of the metrics found in literature shows that there are really just three basic features of a shape comparison metric that are assessed: similarities in magnitude, similarities in phase and the shape of the residual error curve. The Sprague-Geers and the ANOVA metrics have been chosen as the best candidates in the Verification and Validation (V&V) process of computational models. In fact, while the Sprague-Geers metrics can assess the general similarity of magnitude or phase between the two curves, the ANOVA metric provides a direct assessment of the residual error, thereby supplying additional useful diagnostic information about the level of agreement.

In this paper, acceptance criteria for the two mentioned comparison metrics are proposed for the V&V process of computational models to be applied in the simulation of full-scale crash tests in roadside safety. In fact, although the use of comparison metrics allows an objective quantification of the reliability of the numerical model, developing acceptance criteria often relies on imprecise engineering judgement. Criteria suitable to represent the typical scatter of full-scale crash tests were established in this study by comparing ten essentially identical full-scale vehicle crash tests. Since the tests were as identical as can be achieved experimentally, the average values of the comparison metrics represented the reasonable range to asses a good match between two tests. Eventually, acceptance criteria for the two selected comparison metrics were proposed for the comparison of curves with such a noticeable scatter like in full-scale crash tests. The proposed criteria were also considered appropriate for comparing experimental and numerical results for the purpose of V&V of numerical models.

2 Repeated full-scale crash tests

A series of five crash tests (Set 1) with same new vehicles (i.e. 2000 Peugeot 106) and a rigid concrete barrier were performed as a part of the ROBUST project (ROBUST, 2006). The tests were independently carried out by five different test laboratories in Europe, herein called Labs #1 to #5, with the purpose of assessing the reproducibility of full-scale crash tests. As the main intent was to see whether experimental curves representing the same test resulted in similar responses, a rigid barrier was intentionally chosen in order to limit the scatter of the results (which is typically greater in the case of deformable barriers). In order to investigate the influence arising from different vehicle models on the reproducibility of crash tests, a second series of five tests (Set 2) was performed using the same barrier but with vehicles of different brands and models. All the vehicles used in the series, however, corresponded to the standard 900-kg small test vehicle specified by the European crash test standards, EN 1317 (CEN, 1998). In all cases, the three components of acceleration were measured at the vehicle centres of gravity. For reasons of conciseness and because the lateral response is generally thought to be the more critical in this type of redirectional tests, only lateral accelerations and velocities are discussed in this paper.

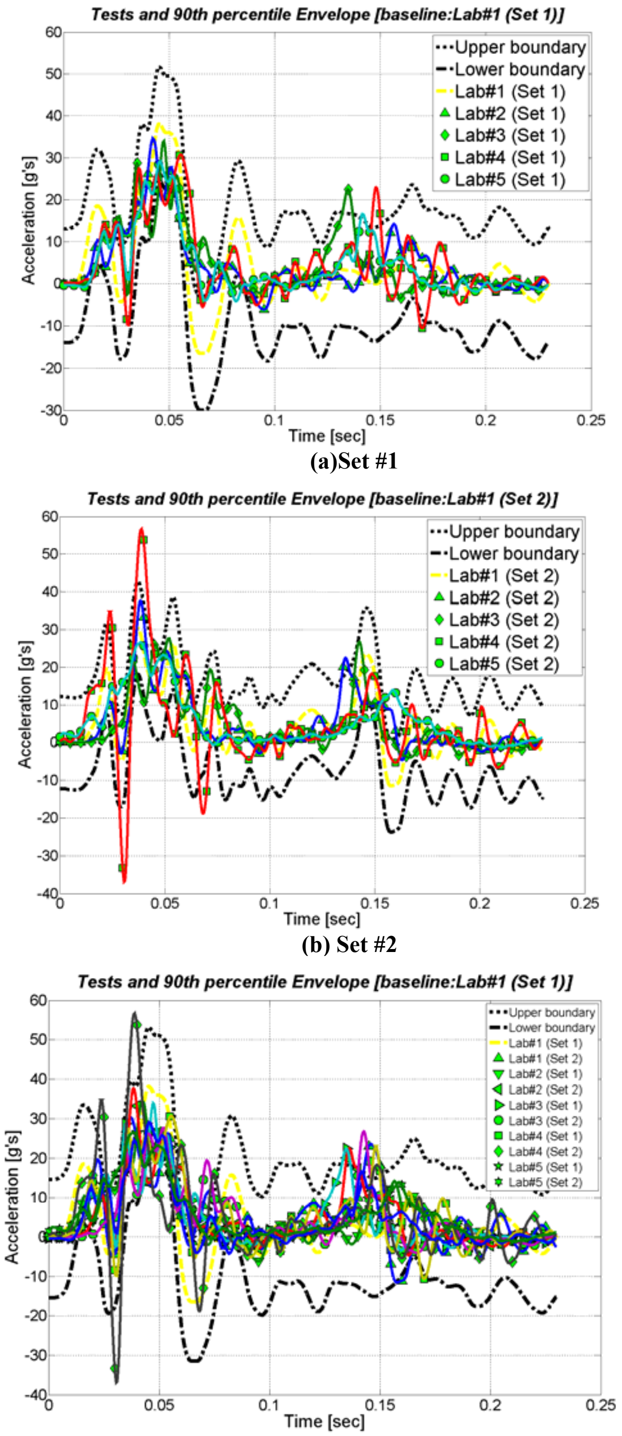
In order to compare the different time histories, it was necessary to prepare them by performing the following operations: (1) filtering, (2) re-sampling, (3) synchronising and (4) trimming. All these pre-processing operations as well as the following metrics evaluation were performed using a program written in Matlab® (MathWorks Inc., 2008).

One common method for assessing the validity of a simulation result or the reproducibility of multiple impact experiments in the biomechanics field is to develop response envelopes. If multiple experiments are available, the time histories for all the experiments can be plotted together. If the average response and standard deviation are calculated at each instant in time, the ±90th percentile envelope indicating the likely response corridor can be plotted. After the ten curves were pre-processed, the 90th percentile envelope for each of the two sets of tests, Set 1 (same new vehicle) and Set 2 (similar vehicles), was computed considering as the 'true' curve the response from the test of Lab #1 belonging to each set, respectively. The 90th percentile envelope for each set was evaluated by adding and subtracting to the respective 'true' curve the average of the standard deviations of the residuals for each specific set of tests multiplied by 1.6449 (i.e. the 90th percentile). Figure 1 shows the pre-processed curves and the respective envelopes for both Set 1 and Set 2. Also, all the ten tests from both sets were compared together considering the response of test Lab #1 from Set 1 as the 'true' curve and the results are shown in the bottom portion of Figure 1.

As expected, there is considerable scatter between the acceleration time histories shown in Figure 1 although there is also a clear trend. Any test response that falls within the response corridors shown in Figure 1 should be considered identical or at least equivalent impact events. As shown by the response envelopes, all ten experiments tend to remain inside the response envelopes although the test from Lab #4 in Set 2 has several peaks that are outside the response corridor. While calculating response corridors is a very useful technique, at least five experiments must be available before a corridor can be constructed and the level of confidence (i.e. the width of the corridor) will be wider the smaller the number of samples is. In roadside safety, the normal situation is that there is generally only one experiment. As a response corridor cannot be obtained from just one or two experiments, if an analyst desires to compare a single computational result to a single crash test experiment, the response corridor method is not a feasible option.

When comparing a computational result to an experiment, the analyst must decide what constitutes a reasonable acceptance criterion. While the metrics themselves are deterministic, a subjective judgement still has to be made about how close to zero (i.e. zero is a perfect match for all the metrics considered in this paper) is 'good enough'. Since all ten of the experiments discussed in this paper represent identical tests, the range of values observed should be an indicator of the acceptable range of scores for more or less identical tests. The purpose of this work, therefore, is to provide insight on acceptance criteria when using the proposed comparison metrics.

Figure 1 Ninetieth percentile envelope and acceleration time histories for (a) Set 1, (b) Set 2 and (c) Set 1 and Set 2 combined (see online version for colours)



(a)Set #1

(b) Set #2

(c) Sets #1 and #2 combined

3 Results using acceleration time histories

Once the time histories for the ten experiments were pre-processed, each was compared to the 'true' curve by evaluating the Sprague-Geers and ANOVA metrics using Matlab®. Initially, the two sets of tests, Set 1 with the same new vehicle and Set 2 with similar vehicles, were considered separately using the response from the Lab #1 test in each set as the 'true' curve. The choice of Lab #1 to represent the 'true' curve was arbitrary and it is reasonable to expect slightly different results if another test were used as the 'true' baseline test. Therefore, the question being evaluated in this paper is: 'Are the results from Lab #1 the same as those reported by Lab #2 to Lab #5?' The resulting metric values for Set 1, Set 2 and the combination of both sets are shown in Table 1 in the top, middle and bottom portions, respectively.

3.1 Sprague-Geers metric

3.1.1 Magnitude difference

The upper portion of Table 1 shows the values for the Sprague-Geers metric. The magnitude component of the metric is negative for all four of the comparison experiments indicating that the 'true' experiment generally experienced a higher magnitude. As shown by Schwer (2007), the amount of the magnitude score is roughly equal to the percent difference in magnitudes. In the case of Set 1, it varies between 14 and almost 26%. The last column in Table 1 shows possible acceptance criteria which are based on calculating the 90th percentile value of the observed metrics (i.e. the mean plus 1.67 times the standard deviation). As the sign of each comparison metric simply depends on which curve is greater and the focus of this paper is to define general acceptance criteria, the mean and the standard deviation for each set of data have been calculated using the absolute value of the metric results.

Even when the same make and model of vehicle is used, the acceleration time histories under identical impact conditions can vary as much as nearly 30% in magnitude. The results for Set 2, where different vehicle meeting the EN 1317 small car test vehicle criteria were used, are similar although the experiment from Lab #4 experienced a much higher magnitude score indicating that Labs #2, #3 and #5 tended to have smaller magnitudes than the Lab #1 ('true' test) and Lab #4 had a much higher magnitude. This is actually confirmed by the time history graphs in Figure 1 where the results for Lab #4 are clearly higher and even cross outside the response corridor. The large difference between Lab #4 and the other tests of this set of tests is reflected in the much larger absolute standard deviation of Set 2 compared to that of Set 1 (i.e. 14.74 vs. 4.85). It is not clear whether the differences between Sets 1 and Set 2 are due to the differences in the vehicles or the one unusual test from Lab #4 in Set 2.

When the magnitude component of the Sprague-Geers metric is combined for all ten tests, as shown in the bottom portion of Table 1, the mean absolute score is 20.13. The absolute standard deviation of the results is nearly 4.60. If the 90th percentile value were used to establish an acceptance criterion for the magnitude component, a value of 27.7 would be the result.

Table 1 Comparison metrics for Set 1, Set 2 and the combination of both sets

Metric	*Lab #2*	*Lab #3*	*Lab #4*	*Lab #5*	*Mean*[a]	*Std. dev.*[a]	*Possible acceptance criteria*
Dataset 1							
Sprague-Geers							
Magnitude	–23.0	–21.4	–14.4	–25.8	21.2	4.85	±29.2
Phase	21.8	28	25.8	22.9	24.6	2.81	±29.3
ANOVA							
Avg. residual error	0.00	–0.01	–0.01	0.00	0.005	0.006	±0.01
Std. dev. of residuals	0.19	0.24	0.23	0.20	0.22	0.02	0.25
T-score	–1.51	–1.87	–2.56	1.31			2.67
Dataset 2							
Sprague-Geers							
Magnitude	–2.6	–8.2	35.6	–9.3	13.92	14.74	±38.2
Phase	21.2	22.8	25.4	26.7	24.0	2.49	±28.2
ANOVA							
Avg. residual error	–0.01	0.00	0.00	–0.02	0.007	0.009	±0.02
Std. dev. of residuals	0.21	0.22	0.31	0.25	0.25	0.05	0.32
T-score	–2.3	0.66	–0.23	–6.27			2.67
Datasets 1 and 2 (combined)							
Sprague-Geers							
Magnitude	See Table 5 in the Appendix for individual scores				20.13	4.60	±27.7
Phase					27.2	3.60	33.2
ANOVA							
Avg. residual error					–0.01	0.01	±0.02
Std. dev. of residuals	See Table 5 in the Appendix for individual scores				0.24	0.04	0.31
T-score							2.67

[a] The mean and standard deviation are calculated based on the absolute value of the metrics.

3.1.2 Phase difference

The result for the phase component is similar to what obtained for the magnitude component as shown in Table 1. Due to the formulation (see the Appendix for details) the values for the phase component must always be positive, so it is not possible to determine from the metric value whether the test curve is leading or lagging the true curve in phase. For Set 1, the values varied from just below 22 to 28 with an absolute mean and standard deviation of 24.6 and 2.81, respectively. The phase component of the metric can be interpreted as being the percent out of phase of the signal.

The results for Set 2 were very similar and actually resulted in a smaller standard deviation than Set 1 possibly indicating that the difference between vehicles does not appear to play a major role at least in this type of test and with this type of vehicle. Combining both sets of data and calculating the 90th percentile indicates that a phase score of about 33 would be appropriate.

3.1.3 Proposed acceptance criteria

Based on the results of the ten essentially identical full-scale crash tests summarised in Table 1, an absolute upper bound value of 30 could be used as acceptance criteria for both the magnitude and phase components of the Sprague-Geers metric when evaluating acceleration time histories from full-scale crash tests. The values were rounded to the nearest rational number for convenience.

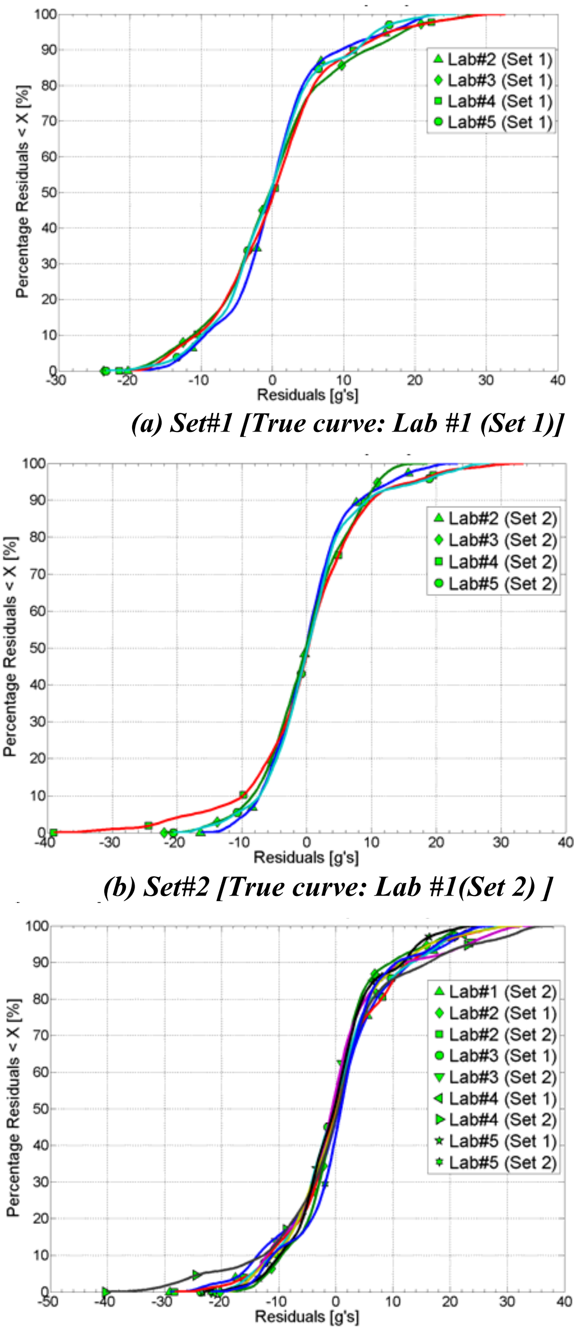
3.2 Analysis of variance metrics

While the Sprague-Geers metrics assess the magnitude and phase of two curves, the ANOVA examines the differences or residual errors between two curves. The average and standard deviation of the residuals were evaluated for each time history in both the two sets of data and the results are shown in Table 1. For all ten experiments, the average residual error was always close to zero as expected. The standard deviations of the residual errors were always under 32% and in all cases but one less than 25%. Since the time histories for all the crash tests represented essentially identical physical events, the residuals for each curve should be attributable only to random experimental error or noise. Statistically speaking, this means that the residuals should be normally distributed around a mean residual error equal to zero. As shown in the cumulative density function in Figure 2, the shape of the residual accelerations distribution is typical of a normal distribution for both the two sets of crash tests when taken separately or combined. Since the cumulative distribution is an ‘S’ shaped curve centred on zero, the distribution of the residuals is consistent with random experimental error as would be expected in these series of repeated crash tests. This is a very strong indicator that the ten tests are, in fact, similar impact events.

3.2.1 Average and standard deviation of residuals

In a previous study, Ray (1996) applied the ANOVA metrics to a set of six identical frontal rigid pole impacts with a small passenger vehicle and reported the results proposing as acceptance criteria: (1) a mean residual error less than or equal to 5% and (2) a standard deviation of the residual less than 20%. Since the tests used in this earlier study were of a type that is presumed to be highly repeatable (i.e. the same type of vehicle and the same crash test facility was used, the barrier was a rigid instrumented pole and the impact was a centre-on full-frontal impact), it was not known if these criteria would be reflective of more general roadside hardware crash tests performed under less ideal conditions. The data in Table 1 indicate that the mean residual error criterion of less than 5% appears to be adequate since none of the comparisons for the ten tests analysed in this paper resulted in a mean residual greater than 2%. The standard deviation of the residuals, however, was higher in these test series than in the one reported by Ray in 1996. The highest standard deviation of the residuals (i.e. 32%) was found for Lab #4 in Set 2, the same test that resulted in an unusually high magnitude score for the Sprague-Geers metric. With that exception, the standard deviations were generally between 20% and 25%, a little higher than Ray originally proposed.

Figure 2 Cumulative density function of the residual accelerations for (a) Set 1, (b) Set 2 and (c) the combination of Set 1 and Set 2 (see online version for colours)



(a) Set#1 [True curve: Lab #1 (Set 1)]

(b) Set#2 [True curve: Lab #1(Set 2)]

(c) All tests [True curve: Lab #1 (Set 1)]

3.2.2 T-test

The third component to the ANOVA is the T-test. The T-test is a score based on the standard deviation and mean of the residuals. At the 90th percentile confidence interval for experiments with a large number of samples (i.e. over 500 data points), the critical T-score is ±2.67 as can be found in any standard statistics textbooks. The T-score calculated for the eight comparisons were all acceptable except one, Lab #5 in Set 2 where the T-score was 6.27, well outside the acceptance range. The poor T-score in the case of Lab #5 in Set 2 strongly indicates there was some systematic error that caused the difference. Some of the T-scores, on the other hand, were less than unity indicating very high probability of representing the same physical event.

Note that Lab #4 in Set 2 has the lowest T-score but the highest Sprague-Geers and standard deviation scores. This apparent contradiction could be explained by looking at the definition of the T-test itself, which is inversely proportional to the standard deviation of the residuals. From a practical point of view, it seems that when the standard deviation becomes too large, the T-test loses its diagnostic meaning as a comparison metric. This is the reason a limit is proposed for the mean residual and the standard deviation. When the standard deviation becomes very large, the 90th percentile envelope is also wide meaning that more divergent curves can still fit in the appropriate envelope.

3.2.3 Proposed acceptance criteria

Based on these comparisons and Ray’s 1996 work, the average residual error should be less than 5% and the standard deviation of the error should be less than 25%. The resulting T-score should also be less than or equal to 2.67 in order to satisfy the ANOVA criteria.

4 Results using velocity time histories

In practice, velocity time histories are often used rather than acceleration time histories primarily because they are less noisy and the trends are more easily apparent. The acceleration histories for the ten experiments were integrated to obtain the lateral velocities. The velocity time histories and the 90th percentile response corridors for the tests of Set 1 are shown in Figure 3. The velocity corridor is much narrower and smoother than the corresponding acceleration time history response corridor shown in Figure 1.

4.1 Sprague-Geers metric

Following the same procedure used to compare two acceleration time histories, the Sprague-Geers metric was used also to compare two velocity time histories. Table 2 shows the values of the metrics calculated for Set 1. As shown in Table 2, the Sprague-Geers magnitude and phase metrics are much smaller for the velocity time history comparison than was the case for the acceleration time history comparison. The highest magnitude score was 5.1 and the maximum phase score was 3.5. Using these values to compute, the 90th percentile range results in an acceptance value of less than 10 for both magnitude and phase, much less than the value of 30 recommended for the acceleration time histories.

Figure 3 Ninetieth percentile envelope and velocity time histories for Set 1 (see online version for colours)

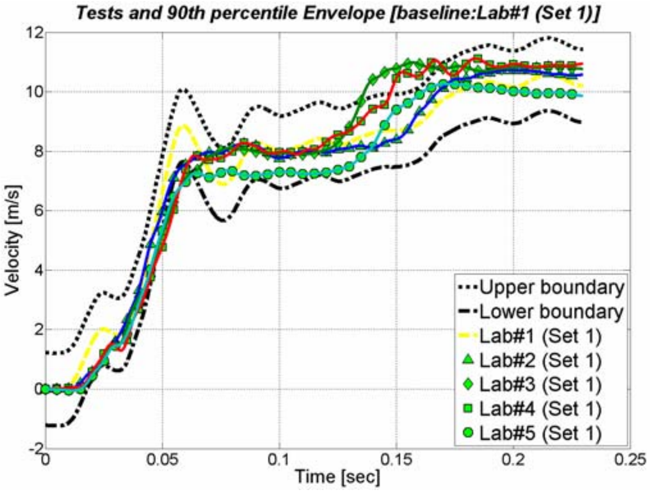


Table 2 Values of the comparison metrics using velocity time histories for Set 1

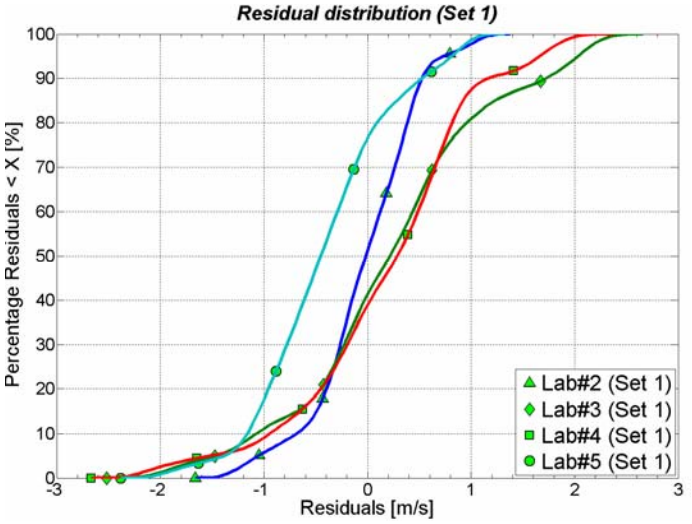
Metric	*Lab #2 (Set 1)*	*Lab #3 (Set 1)*	*Lab #4 (Set 1)*	*Lab #5 (Set 1)*	*Mean*[a]	*Std. dev.*[a]	*Upper bound acceptance*
Sprague-Geers							
Magnitude	0.5	5.1	4.5	–4.0	3.53	2.07	±3.53
Phase	2.0	3.5	3.1	2.8	2.9	0.64	±3.9
ANOVA							
Avg. residual error	0.0	–0.02	–0.02	0.04	0.02	0.02	±0.05
Std. dev. of residuals	0.05	0.09	0.08	0.06	0.07	0.02	0.10
T-score	4.99	–16.38	–14.17	44.0			

[a] The mean and standard deviation are calculated based on the absolute value of the metrics.

4.2 ANOVA

While the Sprague-Geers metric improved as the acceleration time history is integrated to a velocity time history, the ANOVA metrics became much worse. Although the average residual errors are still around zero and the standard deviation present small values, all the T-scores are unacceptably large with increases up to a factor of four. The reason for this poor performance with the velocity curves is that the integration process in essence accumulates the residual acceleration errors. When using an ANOVA technique, the evaluation of metrics should always be performed using time histories directly measured and not derived using either integration or differentiation. For example, if accelerations are measured experimentally, they should be the basis of the ANOVA comparison. Velocities and displacements, obtained by integrating the acceleration curve, will accumulate error at each subsequent integration step. This is shown graphically in Figure 4 where the distribution of the residuals is more spread out and the mean is not as close to zero as it was in the case of the directly measured acceleration time histories shown in Figure 2.

Figure 4 Cumulative density functions of the residual velocities for Set 1 (see online version for colours)



5 Discussion

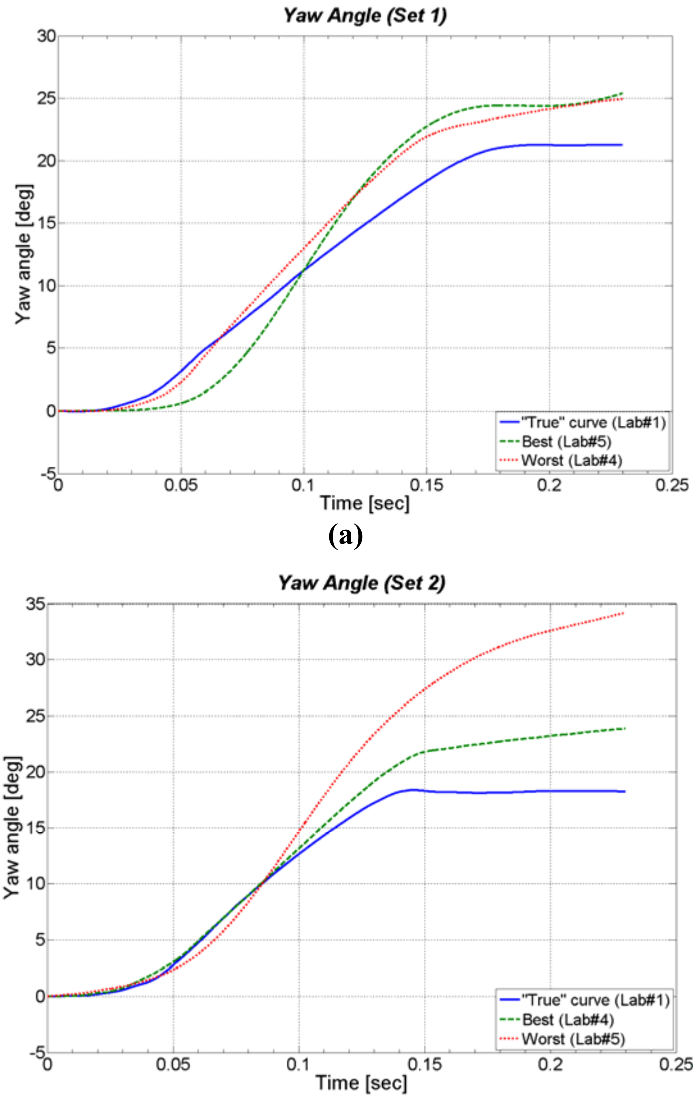
The purpose of the repeated crash test series was to explore how repeatable similar full-scale crash tests would be and to identify sources of possible discrepancies between test organisations. As discussed earlier, the results of the Sprague-Geers magnitude metric for Lab #4 in Set 2 represented a departure from most of the other test results and the ANOVA T-score for Lab #5 in Set 2 indicated a possible systematic error. During the ROBUST project, in fact, the investigation of all the test procedures and techniques used by the different test agencies actually identified several differences that could explain some of the discrepancies showed in this study. For example, it can be assumed that the different technique to mount the accelerometer block to the test vehicle used by each test agency had an effect on the acceleration time histories. In this case, a study performed during the Robust Project showed that a lightweight and more rigid block made of composite material can significantly improve the consistency of the testing results (ROBUST 2, 2006). This illustrates an important point: while two tests may be performed at the same impact conditions and use the same vehicle and barrier, the way data is collected and processed will also affect the results. The shape comparison metrics will be sensitive not only to differences in the impact conditions and test results but also in the way data was collected and processed.

In principle, the Sprague-Geers metric can be used to assess any type of shape: acceleration or velocity time histories. The data in Table 1 indicate that similar if not identical tests can still generate comparisons scores in the mid to upper 30s, whereas if the velocities are compared, the results for similar tests will generally be well under 10. The reason why the Sprague-Geers metrics give better results using velocities may be explained by the fact that the integration process acts like a filter on the original acceleration time histories, thus smoothing the high-frequency noise in the original curves. For this reason and because many analysts find them easier to interpret, it is recommended that velocity time histories be used for the evaluation of the Sprague-Geers metrics. If the magnitude and phase components result in values less than 10, the comparison can be considered valid.

For the ANOVA metrics, it is recommended that the results are computed based on the time histories collected in the physical experiments (e.g. accelerations if the original data in crash tests were collected with accelerometers). In evaluating the ANOVA metrics for a series of six identical frontal rigid pole impacts, Ray proposed an acceptance criterion of a mean residual error less than 5% of the peak and a standard deviation of less than 20% of the peak test acceleration. As shown in Table 1 and discussed above, this is probably a bit too restrictive and should be changed to an average residual error of less than 5% and a standard deviation of less than 25%. While Table 1 shows that the largest standard deviation of the residual error occurred in Lab #5 in Set 2 (i.e. 32%), the authors believe that the improved data collection and processing techniques previously recommended should eliminate such high values. This should, however, be checked when and if additional repeated crash tests become available. The T-score aspect of the ANOVA should be less than or equal to 2.67 to ensure 90th percentile confidence that the error distribution is normally distributed and can be attributed to experimental error.

In this study, one test exceeded the maximum T-score and two others were just below the limit of 2.67. A reason could be that the redirectional tests examined in this paper are much less repeatable than the frontal impact originally studied by Ray. The T-score criterion in some of these impacts can fail because the result of the test was, in actual fact, somewhat different even thought the initial impact conditions, vehicle and barrier were the same. For example, in a redirectional test, a slight difference in the suspension or steering system could cause a slightly different orientation of the front wheels which would in turn change the redirection angle and lateral forces. While the impact conditions may well be essentially identical, the result of the test may not be identical due to other uncontrollable variations in the experiment. Figure 5 shows the yaw-angle time histories of the two tests which gave the best and worst T-scores with the yaw-angle time history of the true curve for both Set 1 and Set 2. The yaw-angle time history in the best case is much closer to the yaw-angle history of the ‘true’ curve than in the worst case. In particular, the difference in the yaw-angle histories is more evident in Set 2, where the highest gap between the best and worst case occurred. The T-score component of the ANOVA metrics is, therefore, an extremely difficult test to pass. An acceptable T-score indicates not only that the impact conditions were similar but that the actual result of the test was very similar as well.

Figure 5 Yaw-angle time histories of the 'true' curve and the curves with the best and worst T-statistics for (a) Set 1 and (b) Set 2 (see online version for colours)



(a)

(b)

6 Conclusions

A comparison of ten repeated essentially identical crash tests was presented in this paper. Of the 16 different metrics found in literature by the authors, the magnitude and phase components of the Sprague-Geers metric were investigated to assess the similarity of magnitude and phase and the ANOVA metrics were investigated to examine the characteristics of the residual errors. The comparison metrics described in this paper were calculated using Matlab®.

The Sprague-Geers metric and the ANOVA metrics were used to quantitatively make comparisons between ten crash tests. Two sets of data were available, the first set of five tests used the same make, model and year of vehicle whereas the second set of five tests used different vehicles that met the requirements for the small car defined by EN 1317. The original raw time histories from the ten tests were filtered, re-sampled and synchronised in order generate accurate comparison results. The statistics derived from the analysis of the residuals confirmed the hypothesis that the errors were normally distributed and could, therefore, be attributed to normal random experimental error.

Using the data from these ten tests, recommendations for acceptance criteria when comparing repeated crash tests or comparing a crash test to a computation result were presented. Namely:

- When comparing two acceleration time histories from a crash test, the average residual error should be less than 5%, the standard deviation of the residual errors should be less than 25% and the T-score should be less than 2.67 for 90% confidence.
- When evaluating velocity time histories obtained by integrating an acceleration time history from a crash test, the Sprague-Geers magnitude and phase metrics should be strictly less than 10.

Roadside safety engineers and analysts should begin to use these comparison metrics to make objective decisions regarding the validity of computational solutions with respect to full-scale crash test results. The acceptance criteria suggested here may well change as the community begins to gain experience with using these types of quantitative metrics. The important point, however, is to use quantitative metrics in order that objective and mathematically precise values are the basis of comparisons and acceptance decisions.

Acknowledgements

The authors are grateful to the members of the ASME PTC-60 committee for providing helpful information about V&V metrics. Also, the authors are thankful to Dr. Chiara Silvestri for her help in finalising and revising the manuscript. This work was performed as part of National Cooperative Highway Research Program Project NCHRP 22-24 (2008).

References

AIAA (1998) *Guide for the Verification and Validation of Computational Fluid Dynamics Simulations*, American Institute of Aeronautics and Astronautics, Report AIAA G-077-1998, ISBN 1-56347-285-6, Reston, VA, USA.

ASME (2006) *Guide for Verification and Validation in Computational Solid Mechanics*, American Society of Mechanical Engineers (Performance Test Code Committee, PTC-60), American National Standard V&V 10-2006, ASME, New York, NY, USA.

Basu, S. and Haghighi, A. (1988) *Numerical Analysis of Roadside Design (NARD), Vol. III: Validation Procedure Manual*, Report No. FHWA-RD-88-213, FHWA, US Department of Transportation.

CEN (1998) 'European standard EN 1317-1 and EN 1317-2: road restraint systems', *European Standard*, European Committee of Standardization (CEN).

Cohen, J., Cohen, P., West, S.G. and Aiken, L.S. (2003) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed., Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

CSA (1994) *Comparative Shock Analysis (CSA) of Main Propulsion Unit (MPU)*, Technical Report, Validation and Shock Approval Plan, SEAWOLF Program, Contract No. N00024-90-C-2901, 9200/SER: 03/039.

DoD (2003) *DoD Modelling and Simulation (M&S)*, US Department of Defense, DoD Instruction 5000.61, Defense Modelling and Simulation Office, Washington, DC, USA.

Geers, T.L. (1984) 'An objective error measure for the comparison of calculated and measured transient response histories', *The Shock and Vibration Bulletin*, The Shock and Vibration Information Center, Naval Research Laboratory, Washington, DC, Bulletin 54, Part 2, pp.99–107.

MathWorks (2008) 'MATLAB® – high performance numeric computation and visualization software', *User Guide*, The MathWorks Inc., 3 Apple Hill Drive, Natick, MA, USA.

NCHRP (2008) *Recommended Procedures for Verification and Validation of Computer Simulations used for Roadside Safety Applications*, Interim Report, National Cooperative Highway Research Program (NCHRP) Project 22-24, Revision 1.2, October 2008. Available online at: http://civil-ws2.wpi.edu/Documents/Roadsafe/NCHRP22-24/RevisedInterimReport.pdf (accessed on 1 June 2008).

Oberkampf, W.L. and Barone, M.F. (2006) 'Measures of agreement between computation and experiment: validation metrics', *Journal of Computational Physics*, Vol. 217, No. 1 (Special Issue: Uncertainty Quantification in Simulation Science), pp.5–36.

Ray, M.H. (1996) 'Repeatability of full-scale crash tests and criteria for validating finite element simulations', *Transportation Research Record – Journal of the Transportation Research Board*, No. 1528, Transportation Research Board of the National academies, Washington, DC, USA, pp.155–160.

ROBUST (2006) *Road Barrier Upgrade of Standards (Robust) – Deliverable 4.1.1 – Full Scale Test Results – An Analysis*, Report document, ROBUST PROJECT, GRD-2002-70021. Available online at: http://www.ha-research.gov.uk/projects/projectdocuments.php?method=download&ID=100 (accessed on 1 April 2008).

ROBUST 2 (2006) *Road Barrier Upgrade of Standards (ROBUST) – Deliverable D.3.3 – New Lightweight Mounting Block – Preliminary Analysis of Data*, Report document, ROBUST PROJECT, GRD-2002-70021. Available online at: http://www.ha-research.gov.uk/projects/projectdocuments.php?method=download&ID=98 (accessed on 1 April 2008).

Russell, D.M. (2006) 'Error measures for comparing transient data: part I. Development of a comprehensive error measure', *Proceedings of the 68th Shock and Vibration Symposium*, pp.175–184.

Schwer, L.E. (2007) 'Validation metrics for response time histories: perspective and case studies', *Engineering with Computers*, Vol. 23, No. 4, pp.295–309.

Sprague, M.A. and Geers, T.L. (2003) 'Spectral elements and field separation for an acoustic fluid subject to cavitation', *Journal of Computational Physics*, Vol. 162, pp.149–184.

Theil, H. (1975) '*Economic Forecasts and Policy*', North-Holland Publishing Company, Amsterdam.

Whang, B., Gilbert, W.E. and Zilliacus, S. (1993) *Two Visually Meaningful Correlation Measures for Comparing Calculated and Measured Response Histories*, Report document, CARDEROCKDIV-U-SSM-67-93/15, Carderock Division, Naval Surface Warfare Center, Survivability, Structures and Materials Directorate, Research and Development, Bethesda, Maryland.

Appendix

A Analytical formulation of the MPC (Table 3) and single-value metrics (Table 4).

Table 3 Definition of MPC metrics

	Magnitude	*Phase*	*Comprehensive*
Integral comparison metrics			
Geers	$M_G = \sqrt{\frac{\sum c_i^2}{\sum m_i^2}} - 1$	$P_G = 1 - \frac{\sum c_i m_i}{\sqrt{\sum c_i^2 \sum m_i^2}}$	$\sqrt{M_G^2 + P_G^2}$
Geers CSA	$M_G = \sqrt{\frac{\sum c_i^2}{\sum m_i^2}} - 1$	$P_{CSA} = 1 - \frac{\lvert\sum c_i m_i\rvert}{\sqrt{\sum c_i^2 \sum m_i^2}}$	$\operatorname{sign}\left(\sum c_i m_i\right)\sqrt{M_{CSA}^2 + P_{CSA}^2}$
Sprague-Geers	$M_G = \sqrt{\frac{\sum c_i^2}{\sum m_i^2}} - 1$	$P_{SG} = \frac{1}{\pi}\cos^{-1}\frac{\sum c_i m_i}{\sqrt{\sum c_i^2 \sum m_i^2}}$	$\sqrt{M_{SG}^2 + P_{SG}^2}$
Russell	$M_R = \operatorname{sign}(m) \cdot \operatorname{Log}_{10}(1 + \lvert m \rvert)$ where $m = \frac{\left(\sum c_i^2 - \sum m_i^2\right)}{\sqrt{\sum c_i^2 \sum m_i^2}}$	$P_R = \frac{1}{\pi}\cos^{-1}\frac{\sum c_i m_i}{\sqrt{\sum c_i^2 \sum m_i^2}}$	$\sqrt{\frac{\pi}{4}(M_R^2 + P_R^2)}$
Point-to-point comparison metrics			
Knowles -Gear	$M_{KG}\sqrt{\frac{\sum\left(\frac{\lvert m_i \rvert}{m_{\max}}\right)^p (\tilde{c}_i - m_i)^2}{\sum\left(\frac{\lvert m_i \rvert}{m_{\max}}\right)^p (m_i)^2}}$ where $\tilde{c} = c(t - \tau)$ (with $\tau = TOA_c$)	$P_{KG} = \frac{\lvert TOA_c - TOA_m \rvert}{TOA_m}$	$\sqrt{\frac{10M_{KG}^2 + 2P_{KG}^2}{12}}$

Appendix (continued)

Table 4 Definition of single-value metrics

Integral comparison metrics	
Correlation coefficient	$\frac{n\sum c_i m_i - \sum c_i \sum m_i}{\sqrt{n\sum c_i^2 - (\sum c_i)^2}\sqrt{n\sum m_i^2 - (\sum m_i)^2}}$
Correlation coefficient (NARD)	$\frac{\sum c_i m_i}{\sqrt{\sum c_i^2}\sqrt{\sum m_i^2}}$
Weighted integrated factor	$\sqrt{\frac{\sum \max(m_i^2, c_i^2) \cdot \left(1 - \frac{\max\left(0, m_i \cdot c_i\right)}{\max(m_i^2, c_i^2)}\right)^2}{\sum \max\left(m_i^2, c_i^2\right)}}$
Point-to-point comparison metrics	
Zilliacus error	$\frac{\sum \lvert c_i - m_i \rvert}{\sum \lvert m_i \rvert}$
RSS error	$\frac{\sqrt{\sum (c_i - m_i)^2}}{\sqrt{\sum m_i^2}}$
Theil's inequality	$\frac{\sqrt{\sum (c_i - m_i)^2}}{\sqrt{\sum c_i^2} + \sqrt{\sum m_i^2}}$
Whang's inequality	$\frac{\sum \lvert c_i - m_i \rvert}{\sum \lvert c_i \rvert + \sum \lvert m_i \rvert}$
Regression coefficient	$\sqrt{1 - \frac{(n-1)\sum (c_i - m_i)^2}{n\sum (m_i - \overline{m})^2}}$

Appendix (continued)

B Values of the comparison metrics obtained by comparing the time histories from both Set 1 and Set 2 ('True' curve: Lab #1 in Set 1)

Table 5 Values of the comparison metrics considering the acceleration time histories of all the ten tests (Sets 1 and 2)

Metric	*Lab #1 (Set 2)*	*Lab #2 (Set 1)*	*Lab #2 (Set 2)*	*Lab #3 (Set 1)*	*Lab #3 (Set 2)*	*Lab #4 (Set 1)*	*Lab #4 (Set 2)*	*Lab #5 (Set 1)*	*Lab #5 (Set 2)*	*Mean*[a]	*Std. dev.*[a]	*Upper bound acceptance*
Sprague-Geers												
Magnitude	–16.7	–23	–18.9	–21.4	–23.5	–14.4	13	–25.8	–24.5	20.13	4.60	27.7
Phase	30	21.8	28.8	28	30.5	25.8	32.4	22.9	25	27.24	3.60	33.2
ANOVA												
Avg. residual error	–0.01	0	–0.01	–0.01	–0.01	–0.01	–0.01	0	–0.08	0.02	0.02	0.06
Std. dev. of residuals	0.26	0.19	0.25	0.24	0.26	0.23	0.32	0.2	0.21	0.24	0.04	0.31
T-score	–1.84	–1.51	–3.45	–1.87	–1.41	–2.56	–1.68	1.31	–7.99			

[a]The mean and standard deviation are calculated based on the absolute value of the metrics.