Engineering with Computers (2007) 23:295–309
DOI 10.1007/s00366-007-0070-1

ORIGINAL ARTICLE

Validation metrics for response histories: perspectives and case studies

Leonard E. Schwer

Received: 15 November 2006 / Accepted: 20 January 2007 / Published online: 12 June 2007

Abstract The quantified comparison of transient response results are useful to analysts as they seek to improve and evaluate numerical models. Traditionally, comparisons of time histories on a graph have been used to make subjective engineering judgments as to how well the histories agree or disagree. Recently, there has been an interest in quantifying such comparisons with the intent of minimizing the subjectivity, while still maintaining a correlation with expert opinion. This increased interest has arisen from the evolving formalism of validation assessment where experimental and computational results are compared to assess computational model accuracy. The computable measures that quantify these comparisons are usually referred to as validation metrics. In the present work, two recently developed metrics are presented, and their wave form comparative quantification is demonstrated through application to analytical wave forms, measured and computed free-field velocity histories, and comparison with Subject Matter Expert opinion.

Keywords Validation · Metric · Wave form · Magnitude · Phase

Abbreviations

TOA	Time-of-arrival
V&V	Verification and validation
SME	Subject Matter Experts

L. E. Schwer (✉)
Schwer Engineering and Consulting Service,
6122 Aaron Court, Windsor, CA 95492, USA
e-mail: Len@Schwer.net

1 Introduction

The comparison of experimental measurements and simulation results is a daily activity for most analysts. The types, and methods, of these comparisons vary widely, as do the qualifiers describing the agreement between the results. Most often the description of the agreement is subjective as illustrated by terms such as ‘fair,’ ‘good’, or ‘excellent’ agreement. The most common quantified comparison is the relative error between the measurement and simulation results, typically stated as a percent difference between the experiment and simulation. Relative error is probably the most widely known example of a metric—a mathematical measure of the difference between two items with the general properties of distance, such that the measure is zero only if the two elements are identical.

Part of a validation assessment, of a numerical simulation capability, comparisons between simulation results and laboratory results (measurements) are performed. Based on these comparisons, assessments are made about the applicability of the simulation capability for making predictive simulations of a similar nature. There are three basic elements of a validation assessment:

- The items to be compared, i.e. system response quantities,
- The manner in which to make the comparison, i.e. computable metric,
- Determination of the accuracy of the comparison, i.e. metric evaluation.

The focus of the present work is on the latter two items, i.e. selection and evaluation of a validation metrics. It is hoped that by illustrating the application of simple validation metrics, the utility of such metrics in the assessment activity will be made evident. This work also hopes to encourage the use of metric-based comparisons on a wider basis in the computational mechanics community.

A simple metric such as relative error works well for point-to-point comparisons, e.g. maximum deflection of a cantilever beam. However when comparisons involve time or spatial variations, e.g. velocity history at a point or deflection along a beam, then the application of a simple metric like relative error becomes sensitive to inaccuracies in the time and space dimensions as well as the system response quantity. Consider two identical wave forms, however one is shifted in time, i.e. different time of arrival. A point-to-point comparison of the shifted wave forms will indicate a disagreement, while aligning the wave forms will produce perfect agreement.

Consider the example of comparing velocity histories, both the shape of the wave form (magnitude) and time-of-arrival (phase) contribute to the assessment of the agreement.

Validation metrics for response histories can be grouped into two broad categories: those based on a point-to-point comparison, and those that use other bases of comparison. Both categories quantify the shape of the wave form (magnitude) and time-of-arrival (phase) contribute to the overall evaluation of the agreement. Typically separate magnitude and phase metrics are developed, based on either normalized time integrations of the wave forms, or point-wise differences between a measurement and its corresponding computed value. Then the magnitude and phase metrics are combined to provide a single value which combines the magnitude differences with the corresponding phase differences. Small values of the combined metric indicate good agreement, and larger numbers indicate poorer agreement.

Perhaps the earliest reference to a metric for comparing wave forms is the 1984 publication of Geers [1]. Integrals of the two wave forms to be compared are computed and used to evaluate the difference in the magnitude and phase of the wave forms. The phase form of this metric was improved by Russell [2, 3]. The new phase form was combined with the 1984 metric by Sprague and Geers [4]. The latest version of this metric is presented and discussed in the present work.

The 2002 work of Oberkampf and Trucano [5] is their earliest work in a series of wave form metrics based on the point-to-point evaluation method based on relative error. The most recent work being that of Oberkampf and Barone [6] where they account for experimental uncertainty when multiple (repeat) experiments are available for comparison with deterministic (single) simulation results. This reference also provides an important section on ‘‘Recommended Features of Validation Metrics,’’ which is suggested reading for those interested in developing such metrics.

The main criticisms of the use of relative error based metrics are the need to account for phase differences, e.g. different times-of-arrival, and more importantly the treatment of relative errors when the base wave form is near zero. The relative error is typically expressed as

$$\mathrm{RE} = \frac{c - m}{m}$$

where c is the computed and m is the measured system response quantity. When the measured response is small, i.e. crosses the time axis in a transient event, the relative error is unbounded. An unpublished 2004 work by Knowles and Gear [7] addresses the criticisms of using relative error while still basing the metric on point-to-point evaluations. The Knowles and Gear metric is presented and discussed in the present work.

In the sections that follow, two validation metrics are presented and illustrated by application to analytical wave forms for assessing magnitude and phase differences; additional analytical wave form comparisons, suggested by Geers [1], are presented in an appendix. Next a more germane demonstration of the validation metrics is presented by application to a velocity history wave form obtained from a well instrumented, and characterized, wave propagation experiment (Groethe and Gran [8]) and the corresponding numerical simulation result by three independent simulation groups. In a subsequent subsection, the two metrics are compared with Subject Matter Expert (SME) opinions to assess the ability of the metrics to mimic this important aspect of validation assessment.

2 Validation metrics

In this section two validation metrics for comparing measured and simulated response histories are presented. The selected metrics, Sprague and Geers [4] and Knowles and Gear [7], are representative of the two categories of wave form metrics:

1. An integral comparison where (time) integration of the wave forms are combined in the metric.
2. A point-to-point comparison where the error at each discrete (time) point is used in the metric.

Also, these two metrics are relatively recent developments.

2.1 Sprague and Geers metric

There have been several modifications by Geers [1] of his proposed error measures, e.g. Sprague and Geers [9], but only the most recent Sprague and Geers [4] error measures are presented.

If $m(t)$ is the measured history and $c(t)$ is the corresponding computed history, then the following time integrals are defined

$$\begin{aligned} \vartheta_{mm} &= (t_2 - t_1)^{-1} \int_{t_1}^{t_2} m^2(t)\mathrm{d}t \\ \vartheta_{cc} &= (t_2 - t_1)^{-1} \int_{t_1}^{t_2} c^2(t)\mathrm{d}t \\ \vartheta_{mc} &= (t_2 - t_1)^{-1} \int_{t_1}^{t_2} m(t)c(t)\mathrm{d}t \end{aligned} \tag{1}$$

where $t_1 < t < t_2$ is the time span of interest for the response history. The error in magnitude is given by

$$M_{\mathrm{SG}} = \sqrt{\vartheta_{cc}/\vartheta_{mm}} - 1 \tag{2}$$

which is insensitive to phase discrepancies, as it is based upon the area under the squared response histories. Equation (2) is represents the ratio of the area under the squared computed and measured response histories, with the –1 providing a zero metric value when the two areas are identical.

The phase error by

$$P = \frac{1}{\pi}\cos^{-1}\left(\vartheta_{mc}/\sqrt{\vartheta_{mm}\vartheta_{cc}}\right) \tag{3}$$

which is insensitive to magnitude differences; this version of the phase error was proposed by Russell,[1] in companion papers [2, 3], and adopted by Sprague and Geers in their most recent work. If $c(t) = m(t)$ it is easy to see that $P = 0$, for the case where $c(t) = m(t - \tau)$, i.e. a phase shift or time-of-arrival difference, it will be demonstrated subsequently that the evaluation of P is sensitive to τ.

Finally, Geers original idea of a Comprehensive Error Factor, given by

$$C_{\mathrm{SG}} = \sqrt{M_{\mathrm{SG}}^2 + P^2} \tag{4}$$

is retained by Sprague and Geers. The idea is that one number represents the combined magnitude and phase differences. Geers makes no claim this is the only, or even the best, way to combine magnitude and phase metrics, just the combination he found useful. In the next section, an alternative method of combining metrics is presented.

2.2 Knowles and Gear metric

The Knowles and Gear [7] metric treats the wave form magnitude and time-of-arrival (TOA) characteristics separately. Then, in a manner similar to the Sprague and Geer metric, combines these two component metrics to provide a single measure of agreement between two wave forms.

The magnitude portion of the metric is based on a weighted sum-of-squared differences between the wave forms, i.e.

$$M_{\mathrm{KG}} = \sqrt{\sum_{i=1}^{N}\left[Q_i \bullet (\tilde{c}_i - m_i)^2\right]/\mathrm{QS}} \tag{5}$$

As before, $m(t)$ is the measured history and $\tilde{c}(t) = c(t - \tau)$ is the time-of-arrival shifted computed history. If TOA_m and TOA_c are the times-of-arrival of the measured and calculated wave forms, then for $\mathrm{TOA}_c > \mathrm{TOA}_m \rightarrow \tau = \mathrm{TOA}_c - \mathrm{TOA}_m$. Shifting the computed wave form, i.e. using $\tilde{c}(t)$ rather than $c(t)$, focus this metric on the wave form differences (magnitude) rather than time-of-arrival (phase) differences.

In (5) Q_i is a weighting factor, QS is a normalization factor, and N is the number of discrete time measurements. The weighting factor is designed to scale the sum-of-squares differences to the maximum value of the measurement, $m_{\max} = \max_i(|m_i|)$, i.e.

$$Q_i = \left(\frac{|m_i|}{m_{\max}}\right)^p (t_{i+1} - t_{i-1}) \tag{6}$$

where the value $p = 1$ is recommend to place more weight on the large values of $m(t)$.

The normalization factor QS is chosen to provide a value of unity when the magnitude of the two wave forms differs by 100%, i.e. $\tilde{c}(t) = 2m(t)$, and is given by

$$\mathrm{QS} = \sum_{i=1}^{N}\left(\frac{|m_i|}{m_{\max}}\right)^p (m_i)^2 (t_{i+1} - t_{i-1}) \tag{7}$$

Some form of magnitude normalization is required, else self-similar wave forms with large magnitudes will have larger values of the metric than the corresponding smaller magnitude wave forms.

Since the Knowles and Gear metric is a point-to-point metric, i.e. at each discrete time both the measurement and calculation wave form values are compared in (5), some form of time-based interpolation is required. The current practice is to use the time line of the measurement as the base and interpolate the calculation results to that base time line. Alternatively, both time lines could be interpolated to a independent base time line.

[1] The interested reader is urged to review the work by Russell for a nice derivation of this metric [2] and an impressive evaluation of 11 validation metrics [3].

An important limitation of the weighted sum-of-squared differences magnitude metric is that it cannot differentiated between an under or over prediction of the measured and computed wave forms. In a case where there are several measurements, it is desirable to know if the corresponding computations under or over predict the measurement and for which measurements.

In place of a phase metric Knowles and Gear use a time-of-arrival (TOA) relative error metric. Recall TOA_m and TOA_c are the times-of-arrival of the measured and calculated wave forms, respectively, then the TOA metric is

$$M_{\mathrm{TOA}} = \frac{|\mathrm{TOA}_c - \mathrm{TOA}_m|}{\mathrm{TOA}_m} \tag{8}$$

The time-of-arrival for each wave form is determined by the time at which the wave form attains some percentage of the maximum wave form value, a range of 5–10% is recommended for wave forms with relative fast rise times; this percentage may need to be changed if wave forms contain slow, rather than fast, rise times.

The Knowles and Gear magnitude and TOA metrics are combined, using what they term *importance factors*. They assign an importance factor of 10 to their magnitude metric, M_{KG}, and an importance factor of 2 to their time of arrival metric, M_{TOA}, and then form the following weighted average:

$$C_{\mathrm{KG}} = \sqrt{\frac{10M_{\mathrm{KG}}^2 + 2M_{\mathrm{TOA}}^2}{12}} \tag{9}$$

This is similar to the Sprague and Geers comprehensive error factor given by (4), but Sprague and Geers use importance factors of unity for the magnitude and phase components of their combined metric, and they do not form the average.

Equation (9) is a simplification of a more general proposal by Knowles and Gear for combining metric values for different system response quantities. For each of K system response quantities there are M_{Ki} metric evaluations, i.e. one evaluation for each ith occurrence of the Kth system response quantity. Next a weighted average of the M_{Ki} metric evaluations is formed i.e.

$$\mathrm{M}_K = \sqrt{\sum_{i=1}^{n} Q_i M_{Ki}^2 / \mathrm{Q}} \tag{10}$$

where n is the number of metric evaluations for system response quantity K, e.g. magnitude metric for $n = 18$ wave forms. Here the *quality factors* Q_i are used to combine multiple metric evaluations for the same system response quantity, e.g. multiple magnitude comparisons from the same experiment, and form a weighted average with the average factor given by:

$$\mathrm{Q} = \sum_{i=1}^{n} Q_i \tag{11}$$

An example of a quality factor Q_j is given by

$$Q_j = \sum_{i=1}^{N} \left(\frac{|m_i|}{m_{\max}}\right)^p (t_{i+1} - t_{i-1}) \tag{12}$$

which is the time sum of all the weighting factors given previously in (6). Note these quality factors are based solely on the *measured* data, i.e. they are intended to reflect the quality of the measured data. At present, the same quality factor is recommended for the magnitude and time-of-arrival system response quantities, i.e.

$$Q_{\mathrm{MAG}} = Q_{\mathrm{TOA}} = \sum_{i=1}^{N} \left(\frac{|m_i|}{m_{\max}}\right)^p (t_{i+1} - t_{i-1}) \tag{13}$$

Finally, an *importance factor* IMP_K is assigned to each system response quantity and a weighted average of all metric evaluations for all system response quantities is formed:

$$M = \frac{\mathrm{IMP}_K \mathrm{Q}_K \mathrm{M}_K + \mathrm{IMP}_L \mathrm{Q}_L \mathrm{M}_L + \ldots}{\mathrm{IMP}_K \mathrm{Q}_K + \mathrm{IMP}_L \mathrm{Q}_L + \ldots} \tag{14}$$

In this way different measurements in a single experiment, say velocity history at a point, maximum displacement at another point, and strain along a length, could be evaluated with different metrics but combined to provide a single measure of comparison, M.

2.3 Application to idealized wave forms

Geers [1] provides an idealized measured response history in the form of an analytical wave form given by the expression

$$m(t) = \mathrm{e}^{-(t-\tau)} \sin 2\pi(t-\tau) \tag{15}$$

where the parameter τ is used to adjust the time of arrival (TOA) of the wave form. To illustrate the magnitude and phase error factors, Geers offers three variations on (15) for comparison; Russell [3] expanded the list of comparisons to 20, but does not include adjustments to the time of arrival. In the Appendix the two validation metrics, i.e. Sprague and Geers and Knowles and Gear, are applied to the comparisons suggested by Geers; perhaps these analytical function comparisons, or others like Russell's, will be used by other validation metric investigators as benchmark cases.

Two simple cases are examined here: a difference only in magnitude and a difference only in time of arrival; neither Geers [1] nor Russell [3] considered these cases.

2.3.1 *Magnitude difference*

The magnitude only case serves as both a demonstration of the metrics and as a verification of the implementation of the metrics. The wave form taken as the measurement, $m(t)$, is an exponentially decaying sine wave given by (15). Where τ is used to adjust the time-of-arrival. For this case, the computed wave form, c_{20} (t), has the same analytical representation as the measured wave form, but with a 20% increase in magnitude

$$c_{20}(t) = 1.2\mathrm{e}^{-(t-\tau_{20})}\sin 2\pi(t-\tau_{20}) \qquad (16)$$

The times-of-arrival in this case are identical, i.e. $\tau = 0.14$ and $\tau_{20} = 0.14$. Because the metric algorithms are implemented to operate with discrete data, rather than the continuous analytical form used in this example case, the above measured and calculated wave forms were sampled using a time increments of $\Delta t = 0.02$ over the time interval of $0 \le t \le 2$. Figure 1 shows the sampled measured and computed wave forms obtained from (15) and (16).

The values of the components of the two metrics for this case are summarized in Table 1. The notation for the metric components in this table, and in the remainder of the text, is as follows:

- S&G—Sprague and Geers: magnitude, phase and combined, Eqs. (2), (3) and (4),
- K&G—Knowles and Gear: magnitude, TOA and combined, Eqs. (5), (8) and (9).

Both the Sprague and Geers and Knowles and Gear metrics produce identical results for the magnitude metric, i.e. 20% as expected, and zero for the Phase or TOA component. The Knowles and Gear combined metric value differs slightly from the Sprague and Geers combined metric due to the different weighting (importance factors) and forms of the combined metric formulas, see (4) and (9).

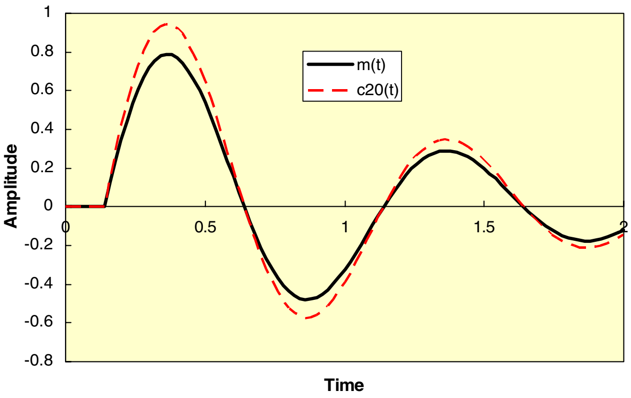

Fig. 1 Idealized response histories for illustrating 20% magnitude difference

Table 1 Metric values for the 20% magnitude increase example

	S&G	K&G
Magnitude	20%	20%
Phase or TOA	0	0
Combined	20%	18%

In this particular case, the two combined metric values could be made equal if the Knowles and Gear importance factor for the time-of-arrival metric component was zero.

2.3.2 *Time of arrival differences*

The other sample case not reported by Geers [1], nor Russell [3], is that of comparing shifts in time-of-arrival (TOA) only. Time-of-arrival differences are an *important* consideration in any comparison of measured and computed response histories, since this severs as a check on the speed of propagation in the media. As was done for the preceding magnitude-only case, an *approximate* 20% phase error factor was 'manufactured' for comparison with the measured response, $m(t)$ from (15), by shifting the TOA of the calculated response:

$$c_{p20}(t) = \mathrm{e}^{-\left(t-\tau_{p20}\right)}\sin 2\pi\left(t-\tau_{p20}\right) \qquad (17)$$

where $\tau_{p20} = 0.24$ to provide for an early TOA, and $\tau_{p20} = 0.04$ for a later TOA, relative to the measured response given by (15) with $\tau = 0.14$. Note: both time shifts, τ_{p20}, were obtained by trail-and-error to generate phase and TOA metric values of about 20%, i.e. a metric value comparable with that used in magnitude only example. The same duration interval of $0 \le t \le 2$, and time increment $\Delta t = 0.02$, was used. The two TOA shifted wave forms, and the measured wave form, are shown in Fig. 2.

The values of the various metric components for this ±TOA case are summarized in Table 2. By design, the phase metric value for the Sprague and Geers metric, i.e. (3), is 20% for the selected values of the time shift parameters τ, and the metric value is the same for negative or positive (relative) TOA's. The Sprague and Geers comprehensive metric value is also 20%, for both TOA's, since the magnitude error component is *essentially* zero. Note the magnitude error is not *exactly* zero because the extent, or range, of the time shifted functions is different from the baseline measured function, and thus the metric values, over this fixed time interval, are different. The numeric value of a metric is dependent on the time interval over which they are evaluated.

The Knowles and Gear TOA metric value is 63% for both TOA cases. For the Knowles and Gear metric the wave form's time-of-arrival is calculated at 15% of the maximum ordinate value. Thus these calculated TOA's are

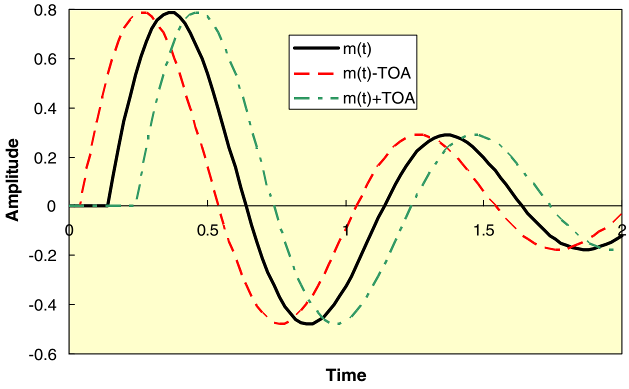

Fig. 2 Idealized response histories for illustrating TOA phase error of 20%

Table 2 Metric values for the phase and time-of-arrival example

		S&G	K&G
m(*t*) + TOA	Magnitude	–0.5%	1%
	Phase or TOA	20%	63%
	Combined	20%	26%
m(*t*) – TOA	Magnitude	0.1%	0
	Phase or TOA	20%	63%
	Combined	20%	26%

0.06, 0.16, and 0.26 for the early arrival (–TOA), baseline (measurement), and late arrival (+TOA) cases, respectively. The form of the Knowles and Gear TOA metric is quite different from the Sprague and Geers phase metric, and thus the two metrics should not be expected produce similar metric values.

The present example also indicates the combined metric value for Sprague and Geers and Knowles and Gear metrics are comparable at 20 and 26%, respectively. This is an accident of the particular wave forms selected for comparison. As an example, had the Knowles and Gear combined metric used equal important factors for magnitude and TOA, then the combined metric in this case would be 44% rather than the 26% listed in the table.

2.3.3 Evaluation

As part of a validation assessment, the entity requesting the modeling effort and resulting simulations will need to establish acceptable values of the validation metrics used to asses the model accuracy. As an illustration, a metric value of say 20% for a given system response quantity might be deemed as the upper limit on acceptable accuracy when comparing simulation to experiment. However, guidance in assigning this upper limit on acceptable accuracy is an open topic in the V&V community.

Hopefully with more wide spread use of various validation metrics, and their improvement, some community-based guidelines for acceptability will evolve. As an example, in the post-presentation question-and-answer period, documented in the 1984 Geers paper [1], when asked about the *acceptable* values of the metrics, Geers provided the following rule-of-thumb guidance on values for his combined metric:

> My personal reaction has been that anything below about 20% is really good. When you get to around 20–30%, it is getting fair. When you get above the 30–40% range, that is rather poor.

However, at this point it is worth emphasizing a point well made by Oberkampf and Trucano [5] about the *value* any validation metric:

> However, we do not believe the quantitative value of any metric is important in an absolute sense. By this we mean that the functional form of the metric is not absolute or unique. It only measures the agreement between the computational results and the experimental data in such a way that positive and negative errors cannot cancel.

2.4 Application to velocity wave forms

While it is instructive to apply validation metrics to analytical wave forms, application to experimental data and simulations is the goal. Also, such applications can further illustrate the strength and weakness of metrics, which might be overlooked when ‘manufacturing’ wave forms. In this section the application of validation metrics to the velocity wave forms shown in Fig. 3 will be used to compare measurement and simulations. These wave forms represent radial velocity in a geological medium due to a nearby energetic source. In addition to the experimental velocity

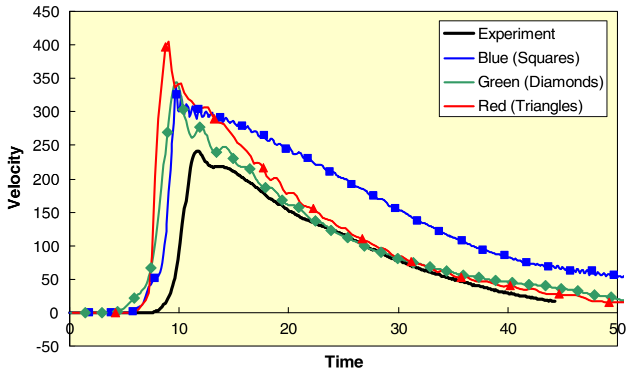

Fig. 3 Comparison of measured velocity wave form with three simulation results

history, three independent simulations of the wave form are presented for comparison. The simulations were performed by three different organizations each using a different numerical algorithm (code), spatial discretization, and non-linear material model for the geological medium. Further, the simulation results were collected *before* the experimental results were made available to those providing the simulations, i.e. a ‘blind’ (predictive) validation study.

2.4.1 Verification and errors bars

While the focus of the present work is on validation metrics, it is worth mentioning at this point that verification of the algorithms (codes) and calculations, e.g. grid convergence, is a necessary prelude to any validation assessment. Here it is assumed that such verification efforts have been completed, and the validation focuses on the non-linear material models used to represent the geological medium.

Both the experimental results and the numerical simulations should include assessments of errors and uncertainties, i.e. error bars. Even when only one measurement is provided, the experimentalist should provide an estimate of the measurement error. Similarly, the aforementioned calculation verification would provide numerical error estimates on the simulations results. For the purpose of the present illustration, it is assumed these errors, experimental and numerical, are negligibly small.

2.4.2 Traditional validation assessment

Figure 3 is a fairly typical graphic for presenting time varying results. The observer of such a graphic, in say a technical report or during a presentation of results, would likely read, or hear, accompanying text providing a *subjective* view of the relative merit of each simulation result compared with the measurement. Details such as time-of-arrival (TOA), maximum velocity, velocity decay rates, and perhaps integration of the velocity histories to provide permanent displacements, would be discussed. In the end however, the observer of the graphic uses their own *engineering judgment* to assess the relative overall ‘goodness of agreement’ (accuracy) of the simulation results, and makes an assessment of each simulation’s ability to accurately reproduce the experimental results.

This type of personal, or group based, assessment of accuracy of the numerical results, based on viewing a chart[2] of the results, can be sufficient if the person, or group, is experienced in making such assessments, i.e. Subject Matter Experts. However, when non-experts are presented with the same chart and asked to make an assessment, as often happens when presenting such results to clients, or more importantly assessment or licensing groups, then the assessments may be quite different from those of the Subject Matter Experts.

The following text is presented as a plausible assessment of the simulation adequacy of the results presented in Fig. 3 by a non-expert. These observations could be made:

- All of the simulations over predict the measurement, i.e. the measurement appears to be a lower bound on the velocity predicted by all the simulations.
- All of the simulations predict a TOA that is earlier than the measured TOA.
- Two of the three simulations provide a *fairly good* representations of the velocity decay.
- Based on these observations, the non-expert, and perhaps expert, probably would be willing to rank the Blue (Squares) simulation result as *less accurate* compared to the Green (Diamonds) and Red (Triangles) results.

Eliminating the Blue (Squares) simulation, next consider the remaining two simulations and their comparison with the measured velocity. Now the evaluators, expert and non-expert, are asked to select which of these two simulation results *best* represents the measured velocity wave form. It is suggested that most evaluators would select the Green (Diamonds) simulation result, as in *some sense* it lies between the Red (Triangles) simulation result and the measurement. Thus, in this plausible validation assessment, the Green (Diamonds) simulation result is deemed *better* than the other two simulations results.

2.4.3 Validation metric assessment

Next the validation metrics proposed by Sprague and Geers and Knowles and Gear are applied to the simulation results shown previously in Fig. 3. A comparison of the components of the two metrics is shown in Table 3. The magnitude metric comparison indicates that both metrics provide very similar values, and further rank the three simulation results in the same order, i.e. Green (Diamonds), Red (Triangles) and Blue (Squares) from best-to-worst agreement with the experimental wave form. This ranking of the wave form is also in agreement with the above hypothetical expert traditional assessment of the wave form comparisons. More will be said about how these validation metrics and Subject Matter Expert opinion align in the last section.

2.5 Observations

These validation metrics are quite simple to implement, e.g. a spreadsheet application was used for the present

[2] A traditional validation assessment, based on viewing a chart, a.k.a viewgraph, is known in the validation vernacular the ‘Viewgraph Norm’ implying the assessment is both qualitative and subjective.

Table 3 Metric components of wave forms shown in Fig. 3

	Sprague and Geers			Knowles and Gear		
	Magnitude	Phase	Combined	Magnitude	TOA	Combined
Blue (Squares)	0.60	0.08	0.61	0.54	0.17	0.50
Green (Diamonds)	0.26	0.13	0.29	0.27	0.23	0.26
Red (Triangles)	0.45	0.15	0.47	0.48	0.21	0.45

calculations and comparison. The metrics require the measured and calculated responses to have the same sampling time lines to perform the required point-to-point evaluations; these evaluations are further simplified by the use of a uniform sampling interval.

The separation of wave form metrics into magnitude and phase (time-of-arrival) components is an advantage when more detailed investigation of the comparisons are needed. Typically the comparative emphasis is on the magnitude error. The ability of any metric to represent magnitude differences exactly, i.e. as in the preceding section 20% magnitude difference example, is likely to be considered a requirement by evaluation groups.

Although the Knowles and Gear metric provides values comparable to the Sprague and Geers metric, its inability to indicate the sign of the magnitude component, i.e. an over or under prediction, is a significant limitation. Similarly, the emphasis on time-of-arrival, rather than phase differences, is considered a limitation. The most promising of the ideas suggested by Knowles and Gear is that of the possibility of a weighting function to emphasize what are deemed important aspects of the wave forms being compared, i.e. the power p in (6). Also, the concept of what they term 'importance factors' for combining metric components and metrics has potential when validation assessments consider different types of measurements.

3 Subject Matter Expert opinion and validation metrics

The result of applying a validation metric to a pair of waveforms is a mathematical quantification of the comparison. Oberkampf and Trucano [5] correctly stress that the quantitative value of any metric does not imply adequacy of the comparison. However, the underlying intent of any validation metric should be consistency with Subject Matter Expert opinion, i.e. provide quantified mimetic expert evaluations.

This section describes a quantified assessment by a group of experts who were asked to compare five pairs of velocity waveforms. The SME assessment is compared with two validation metrics designed to quantifying waveform differences. The comparisons presented in this manuscript should be interpreted as anecdotal evidence, as the author is unskilled in expert opinion solicitation, which is essential to establishing valid SME assessments.

3.1 Summary of expert opinion

The Subject Matter Experts were provided with five charts showing pairings of experimental velocity waveforms. Only experimental wave forms where used in this case study that included pairs of replicate wave forms, e.g. from repeat tests or symmetric measurements from a single test. All the experts had previous *equal* access to prior knowledge of the pedigree of the waveforms, e.g. the data were obtained from archived field measurements, the quality of the data had been openly discussed, and a metric evaluation of the data had been previously presented to the experts.

The experts were asked to respond to the following statement:

> The community of experts is asked to provide a zero-to-one assessment of how poorly or well the paired wave forms agree, with zero poor agreement and one good agreement.

All but one expert responded in a direct manner to this statement. A few experts included unsolicited qualification statements, and one expert responded in a manner that required a subjective evaluation to transform the expert's opinion into the requested 0-to-1 value. Details of the responses, quantifying comments and subjective evaluation are presented in a subsequent section.

A summary of the waveforms and corresponding expert opinions are presented in Figs. 4, 5, 6, 7 and 8. The velocity waveforms are shown on the left and the SME assessments by the 11 experts are presented on the right, preceded by the mean of the SME evaluations.

3.2 Validation metrics

Both Sprague and Geers and Knowles and Gear metrics are what is termed ‘‘non-symmetric,’’ i.e. produce different values when the measurement and calculated responses are interchanged; Russell [2, 3] has proposed a symmetric metric. For the intended application of the Knowles and Gear metric and Sprague and Geers metrics, i.e. comparing

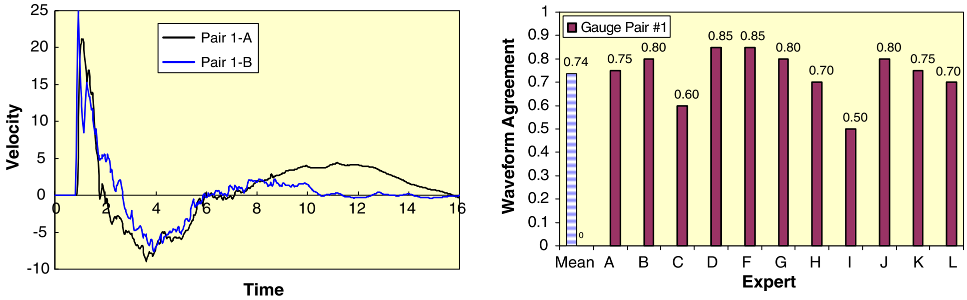

Fig. 4 Waveform Pairing #1 and SME evaluation

measurements to computed responses, this lack of symmetry is not a concern, as it is traditional to take the measurement as the basis. However, for the present application where two measured wave forms are to be compared, the choice of the basis wave form affects the resulting metric value. To place these metric evaluations on a more equal footing with the above SME opinions, the metrics are evaluated both ways, i.e. each wave form in the pair is used as the basis, and the two results averaged. Details of the metric evaluations are provided in a subsequent section.

3.3 Comparison of SME and validation metrics

The metrics are designed to express differences between wave forms with a zero metric value indicating perfect agreement between the wave forms and values of unity, or larger,[3] indicating a difference of 100%, or greater. However, the SME question, cited above, was asked such that good agreement corresponded to unity and poor agreement to a zero value. This phrasing of the SME question was intentional as it was felt to be easier to solicit a quantitative assessment of agreement rather than disagreement (differences). In the comparisons with the metrics, the SME opinions on wave form agreement are transformed to opinions on wave form differences by subtracting the SME value from unity.

Figure 9 presents the comparison of the SME and metric evaluations for the five waveform pairings shown previously. The SME evaluations are presented as the mean of the expert opinions with a 'uncertainty bar' indicating one standard deviation about the mean. For all waveforms, the Sprague and Geers metric is within the one standard deviation of the SME evaluation, and the same is true for the Knowles and Gear metric with the exception of the second waveform pairing, which has the largest metric value among the five waveform pairings.

In addition to agreeing fairly well with the range of SME evaluations, the two metrics provide quite similar evaluation results for the five waveforms. The metric-to-metric agreement is particularly good for waveform Pairings #1 and #3, where the SME evaluation mean is about 25% indicating fair agreement between the waveform pairs.

In contrast, there is nearly factor of two difference between the two metric evaluations for waveform Pairings #4 and #5, where the SME evaluation of 13–14% indicates good agreement between the waveforms. The Knowles and Gear metric's close agreement with the SME evaluation for these two cases is probably indicative of the effectiveness of the weighting used in the Knowles and Gear metric, which places an emphasis on the maximum (large) values of the measured waveform, i.e. the long low level 'tails' of these waveform pairings are discounted in the Knowles and Gear metric; see Figs. 7 and Figure 8. Conversely, the Knowles and Gear metric weighting appears to work against a favorable comparison with the SME evaluation for waveform Pairing #2, see Fig. 5, where the waveforms have long, and relatively large magnitude durations, and shorter duration low level tails.

It should also be noted that the wave characteristic weighting used in the SME evaluations are unknown and likely highly variable from pairing to pairing. But the present comparison of SME and metric evaluations indicates they are in good agreement.

3.4 Observations

The results of this illustration indicate that the present validation metrics can mimic Subject Matter Expert opinion, at least within the variation of the SME opinions. The results, while interesting and favorable for the validation metrics presented, are subject to the limitations noted in the text. In particular, the wave characteristic weighting used in the SME evaluations are unknown and likely highly variable from pairing to pairing, and author's inexperience in properly soliciting expert opinion.

4 Summary and conclusions

Validation metrics proposed by Sprague and Geers and Knowles and Gear have been presented, and demonstrated

[3] The magnitude portion of both metrics are normalized, i.e. if the computed wave form is 1.5 times the measured waveform the magnitude portion of the metrics will yield a value of 150%.

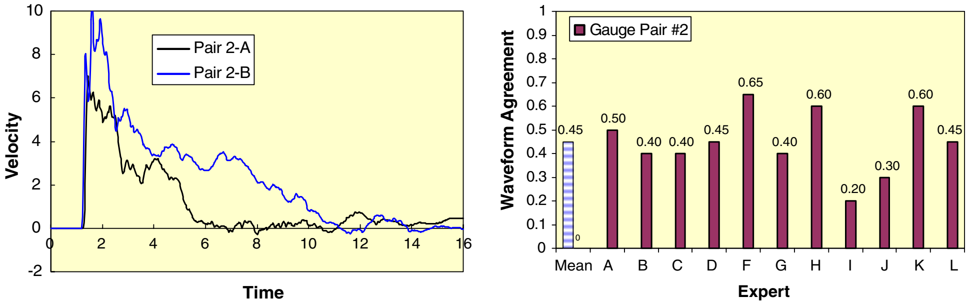

Fig. 5 Waveform Pairing #2 and SME evaluation

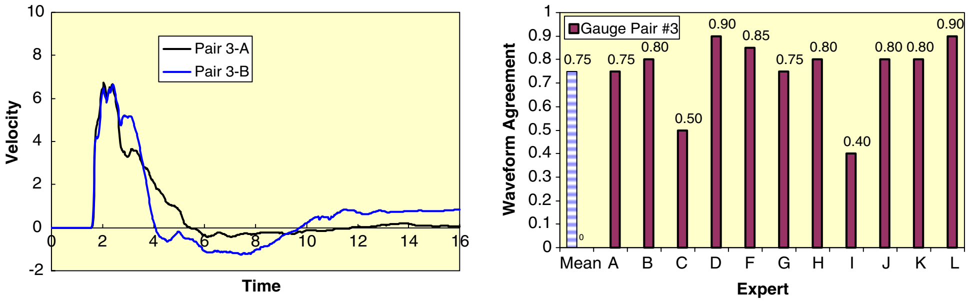

Fig. 6 Waveform Pairing #3 and SME evaluation

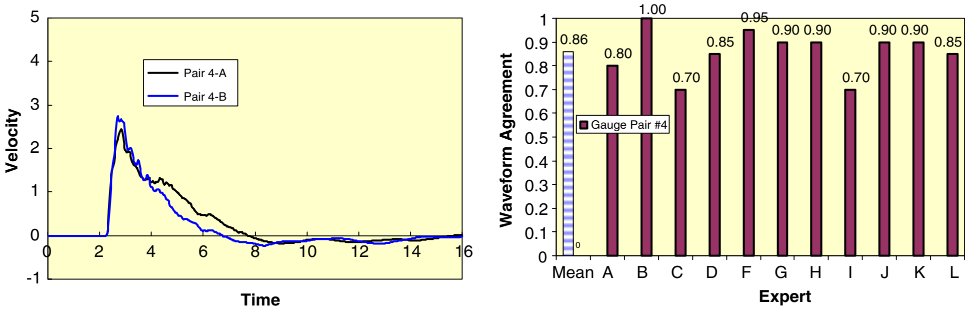

Fig. 7 Waveform Pairing #4 and SME evaluation

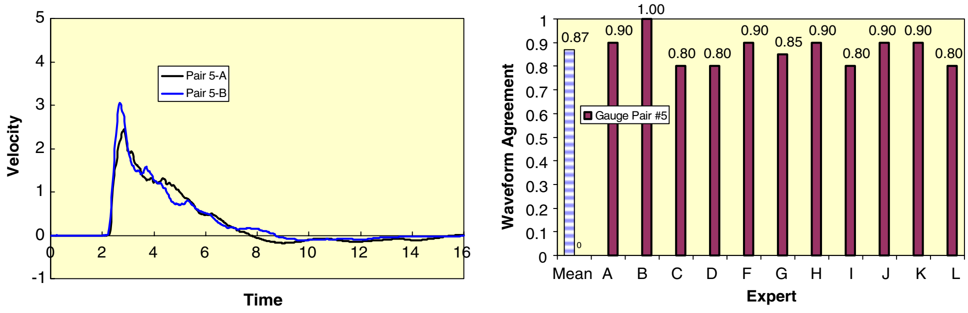

Fig. 8 Waveform Pairing #5 and SME evaluation

by application to experimental velocity wave forms and the results from simulations by three independent organizations. The two proposed metrics have also been demonstrated by example to agree with Subject Matter Expert opinion, which is an important criterion in evaluating proposed validation metrics.

While there is certainly a need for more research on validation metrics, there is an even greater need to encourage the application of existing validation metrics whenever experimental and numerical results are compared. The field of computational mechanics would benefit from the widespread use of metrics for comparisons, and as an augmentation of the traditional assessment for comparing results.

The value of applying validation metrics seems to have a positive affect on all aspects of validation assessments. For those providing the numerical results the quantification of differences and presentation on the same basis with

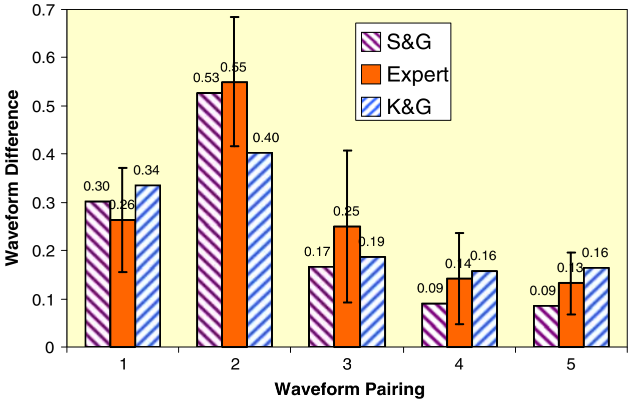

Fig. 9 Comparison of Subject Matter Expert and metric evaluations of five waveform pairings

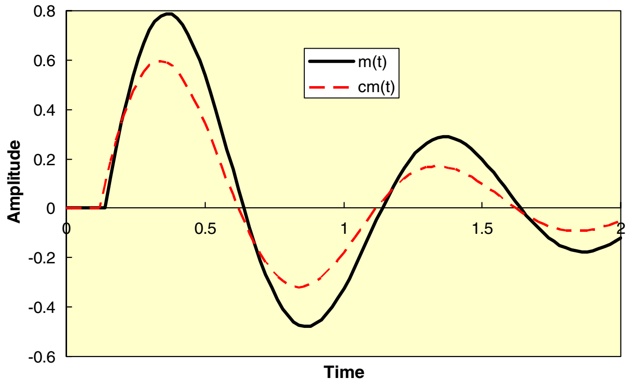

Fig. 10 Idealized response histories for illustrating magnitude error

other numerical results can guide further investigation of differences among simulations techniques. For those providing the experimental results, the comparison of the measurements with self consistent numerical simulations can isolate possible problem areas with instrumentation.

The greatest impact of the application of validation metrics will be among the evaluation groups tasked with the difficult job of judging the appropriateness of simulation techniques to make predictions of critical defense and commercial applications. Metric based results can guide and focus engineering and management judgment, and provide guidance on what and how much additional validation may be necessary.

Acknowledgments The author is grateful to Dr. James Gran of SRI International for providing the digital experimental data used in this work. Professor Tom Geers, Drs. Gene Sevin, David Russell, Cyrus Knowles and William Gear are also most gratefully acknowledged for providing source material that is otherwise not readily available. Special thanks to Dr. Patrick Roache and David Moorcroft for reviewing an early draft and providing suggestions for improvement. My friend Dr. William Oberkampf provided significant contributions to improving the clarity of the final draft. The support of the US Army Engineering Research and Development Center, and the Defense Threat Reduction Agency, under contract DACA42-03-P-0308, made this work possible.

Appendix: Analytical wave form comparisons

The three cases presented in this Appendix are offered as further illustrations of the two considered validation metrics, as comparison cases for verification of local implementations of the metrics, and for comparison with possible future validation metrics.

Magnitude and decay rate differences

To illustrate magnitude and decay errors, Geers offers the following analytical function

$$c_m(t) = 0.8\mathrm{e}^{-(t-\tau_m)/0.8} \sin 2\pi(t-\tau_m) \tag{18}$$

where the parameters τ_m is used to adjust the time of arrival of the wave form; the functions are zero before the time-of-arrival. Figure 10 shows the wave forms proposed by Geers [1] and given by (15) and (18). Unfortunately, Geers did not specify the values of the time of arrival parameters, nor the integration duration used for the evaluation of the metrics. These parameters were estimated from the unlabeled axes plots presented by Geers as Figs. 1, 2 and 3 in his 1984 publication. The lack of a duration of application is not critical as the exponential decay of the wave forms minimizes late time contributions to the time integrals used by the Geers metrics. The times of arrival used in the present example are $\tau = 0.14$ and $\tau_m = 0.12$ with a time integration interval of $0 \leq t \leq 2$ and time increments of $\Delta t = 0.02$.

Table 4 presents a summary of the metric components for the two considered metrics, applied to this magnitude and decay rate comparison. The absolute value of the magnitude components of the Sprague and Geers (–28.4%) and Knowles and Geer (27.3%) metrics are nearly identical. However, this example illustrates an important difference between the two magnitude metrics. The Knowles and Geer magnitude metric does not report the sign of the magnitude difference, while a positive Sprague and Geers magnitude metric indicates the calculated response is greater in magnitude than the measurement, and the converse for a negative magnitude metric. This limitation of

Table 4 Validation metric components for Geers magnitude and decay illustration

	S&G	K&G
Magnitude	–28%	27%
Phase or TOA	5%	13%
Combined	29%	26%

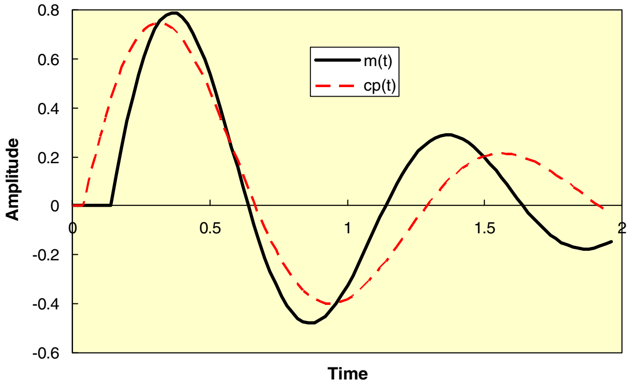

Fig. 11 Idealized response histories for illustrating phase error

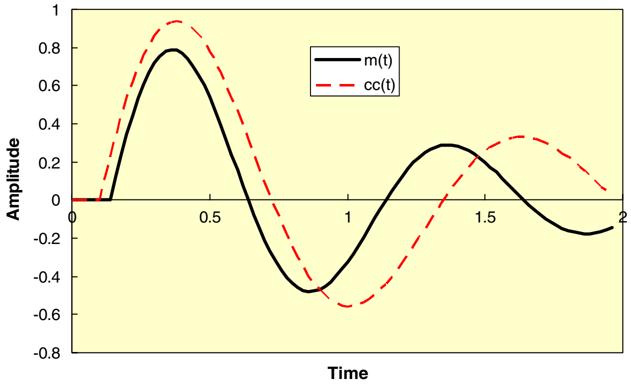

Fig. 12 Idealized response histories for illustrating combined errors in magnitude and phase

Table 5 Validation metric components for Geers phase difference illustration

	S&G	K&G
Magnitude	–0.7%	43%
Phase or TOA	16%	61%
Combined	16%	47%

Table 6 Validation metric components for Geers combined magnitude and phase difference illustration

	S&G	K&G
Magnitude	30%	53%
Phase or TOA	20%	22%
Combined	36%	50%

the Knowles and Geer magnitude metric appears minor when making a single comparison, it becomes more serious when a large suite of comparisons are made.

Phase error

To illustrate phase errors, Geers [1] offers the following analytical function:

$$c_p(t) = \mathrm{e}^{-\left(t-\tau_p\right)} \sin 1.6\pi\left(t - \tau_p\right) \tag{19}$$

Figure 11 shows the wave forms proposed by Geers and given by (15) and (19). The times-of-arrival used for the application of the metrics were $\tau = 0.14$ and $\tau_p = 0.04$ with a time integration interval of $0 \leq t \leq 1.96$ and time increments of $\Delta t = 0.02$.

Table 5 presents a summary of the components of the two metrics, applied to this phase error comparison. For this case the Sprague and Geers magnitude metric is small at –0.7% and the phase metric is 16.2%. However, the magnitude metric for the Knowles and Gear metric is quite large at 43.2%. This large discrepancy in the magnitude metric values illustrates another difference between the two metrics. The Knowles and Gear metric does not address errors due to phase differences, rather such differences are included in the magnitude metric. Consider for example a sine and cosine wave forms with identical times-of-arrival. The Knowles and Gear magnitude metric would be about 100% while the Sprague and Geers magnitude would be nearly zero.

The combining of magnitude and phase differences in the Knowles and Gear magnitude metric can be misleading. For the trivial case of the sine and cosine wave if these wave forms are interpreted as velocity or pressure wave forms, then their time integration is displacement and impulse, respectively. Imagine a structure responding to these two wave forms, although both wave forms produce the same integrated effect, i.e. displacement or impulse, the Knowles and Gear magnitude metric indicates there is a large difference, which would likely not be the case for the structure's response.

Combined magnitude, decay rate, and phase errors

To illustrate combined magnitude, decay, and phase errors, Geers [1] offers the following analytical function:

$$c_c(t) = 1.2\mathrm{e}^{-(t-\tau_c)/1.2} \sin 1.6\pi(t - \tau_c) \tag{20}$$

Figure 12 shows the wave forms proposed by Geers and given by (15) and (20). The times-of-arrival used for the application of the metrics were $\tau = 0.14$ and $\tau_c = 0.1$ with a time integration interval of $0 \leq t \leq 1.96$ and time increments of $\Delta t = 0.02$.

Table 6 presents a summary of metric components for the two metrics as applied to this magnitude, decay rate and phase error comparison. The Sprague and Geers magnitude metric is 30% with a 20% phase metric. The Knowles and Gear magnitude metric is 53% with a TOA metric of 22%. As noted above for the phase error case, the Knowles and

Gear magnitude metric combines both magnitude and phase differences.

Validation metric numerical implementation

The validation metrics presented require the measurement and simulations results to be sampled at the same discrete times. This can most easily be addressed when the parameters of the validation exercise are specified, i.e. the experimental group provides the sampling interval for the measurements and the numerical simulation groups report their results at the specified sampling interval. When the sampling intervals vary, as occurred in the present suite of experiment to simulation comparisons, the simulation results can be interpolated to provide results with the same sampling rate used for the measurements.

Sprague and Geers metric

The time integrals used in the Sprague and Geers metrics are approximated by summations using simple trapezoidal integration, i.e.

$$\int_a^b f(t)\mathrm{d}t \approx \frac{b-a}{2N}\sum_{i=1}^{N}\left[f(t_i)+f(t_{i+1})\right] \tag{21}$$

where N is the number of trapezoidal intervals such that $\Delta t = (b - a)/N$ and $f(t_i)$ is the integrand evaluated at the indicated end of the interval.

All of the metric terms proposed by Sprague and Geers, i.e. (2) and (3), use ratios of the time integrals, so the coefficients preceding the summation in the trapezoidal integration cancel, i.e. from (2)

$$\frac{\vartheta_{cc}}{\vartheta_{mm}} \approx \frac{\sum_{i=1}^{N}\left[c(t_i)+c(t_{i+1})\right]^2}{\sum_{i=1}^{N}\left[m(t_i)+m(t_{i+1})\right]^2} \tag{22}$$

The use of uniform time sampling, or interpolation to uniform time sampling, greatly simplifies the metric component evaluations.

Knowles and Gear metric

The Knowles and Geer magnitude metric, (5), is expressed as a ratio of two series, and as in the Sprague and Geers metric, if uniform time sampling is used the magnitude metric evaluation is greatly simplified.

$$M_{\mathrm{KG}} = \sqrt{\frac{\sum_{i=1}^{N}\left(\frac{|m_i|}{m_{\max}}\right)^p(\tilde{c}_i - m_i)^2}{\sum_{i=1}^{N}\left(\frac{|m_i|}{m_{\max}}\right)^p(m_i)^2}} \tag{23}$$

Details of the Subject Matter Expert opinions and metric evaluations

SME responses

Table 7 presents the zero-to-one responses of the 11 SME, indicated by the letters A through L, on the agreement of the five waveform pairings; recall zero indicates poor agreement and one good agreement. The mean and standard deviation for the 11 responses are provided in the last two columns of the table. The generally low value of the standard deviations is perhaps indicative of the experts being selected from a group that has similar backgrounds/experiences with the selected waveforms, and has been working together for several years.

Unsolicited SME comments

This section includes most of the unsolicited comments and qualification statements provided by the SME's. The common thread is that since the waveforms to be compared are identified as *velocity* histories, some of the SME's also used the corresponding (estimated) displacement to assess the waveforms. Also, the SME's are aware that the metrics make no such assessment of displacement and that the intent of the SME questioning was to compare the SME evaluation with the K&G and S&G metrics. In retrospect, it would have been better not to identify the waveforms as velocity histories, as other types of wave forms, e.g. displacement and stress, may not have a corresponding integral aspect to their expert assessment.

Table 7 Summary of Subject Matter Expert waveform pairings evaluations

Pairing	A	B	C	D	F	G	H	I	J	K	L	Mean	STD
1	0.75	0.80	0.60	0.85	0.85	0.80	0.70	0.50	0.80	0.75	0.70	0.74	0.11
2	0.50	0.40	0.40	0.45	0.65	0.40	0.60	0.20	0.30	0.60	0.45	0.45	0.13
3	0.75	0.80	0.50	0.90	0.85	0.75	0.80	0.40	0.80	0.80	0.90	0.75	0.16
4	0.80	1.00	0.70	0.85	0.95	0.90	0.90	0.70	0.90	0.90	0.85	0.86	0.09
5	0.90	1.00	0.80	0.80	0.90	0.85	0.90	0.80	0.90	0.90	0.80	0.87	0.06

Table 8 Summary of response from SME B and average of responses

SME B	Pre-test	Post-test	Average
1	0.90	0.70	0.80
2	0.50	0.30	0.40
3	0.90	0.70	0.80
4	1.00	1.00	1.00
5	1.00	1.00	1.00

Table 9 Knowles and Gear metric evaluations of time-of-arrival (TOA), magnitude (MAG), and weighted metric (WM) for five waveform pairings

Pairing	TOA		MAG		WM		Ave. WM
	A–B	B–A	A–B	B–A	A–B	B–A	
1	0.07	0.08	0.40	0.38	0.35	0.33	0.34
2	0.05	0.05	0.50	0.45	0.42	0.38	0.40
3	0.01	0.01	0.22	0.22	0.19	0.18	0.19
4	0.01	0.01	0.21	0.17	0.17	0.14	0.16
5	0.01	0.01	0.20	0.19	0.17	0.16	0.16

Table 10 Sprague and Geers metric evaluations of Phase, magnitude (MAG), and comprehensive metric (COMP) for five waveform pairings

Pairing	Phase		MAG		COMP		Ave. COMP
	A–B	B–A	A–B	B–A	A–B	B–A	
1	0.21	0.21	–0.19	0.24	0.29	0.32	0.30
2	0.16	0.16	0.62	–0.38	0.64	0.41	0.53
3	0.13	0.13	0.11	–0.10	0.17	0.16	0.17
4	0.09	0.09	–0.01	0.01	0.09	0.09	0.09
5	0.07	0.07	0.05	–0.05	0.09	0.09	0.09

The SME comments follow: general criteria used to judge the curves:

- Peak amplitude
- Displacement (area under curve), focusing especially on the positive phase
- Overall waveform shape
- Rise time and pulse width
- Arrival time

The Knowles and Gear metric puts a grater weight on what happens near the peak. I do too, but maybe not as much. For example, sometimes you have a very sharp and narrow spike at the peak, so the blind numerical peak value is not really representative. I tend to discount this. I put more weight on the overall shape of the main pulse, and maybe the period right after it (the main tail).

I also tend to put a fair amount of weight on displacements. Since these were not provided, I had to eyeball them. I do not know how much weight others put on displacements, but as far as I know none of the standard quantitative metrics put ANY weight on them, at least directly. (Displacements obviously affect the magnitude portions of the metrics, but indirectly.) I am not sure what you do about this.

Subjective evaluation of SME B's response

Subject Matter Expert 'B' provide a response that consisted of two numbers for each waveform pairing, see Table 8, and the qualification statement:

> My reply is based on the following assumptions. One of the waveforms is an experimental record and the other is a pre- or post-test prediction. My assessment also assumes that the gage records have been peer reviewed. From these records it is obvious that gage 1 is inconsistent (displacements) with gages 2–5 and therefore is less creditable.

When asked, SME B was unwilling to change the provided response to conform with the responses obtained from the other SME's. The author decided to average the two numbers provided by SME B and include the average as the SME's response. The overall low standard deviations of the SME responses perhaps justify this subjective decision.

Metric evaluations

This section presents the details of the time-of-arrival (phase), magnitude, and combined metric evaluations for the Knowles and Gear (Table 9) and Sprague and Geers (Table 10) metrics. The average of the combined metric, for the two non-symmetric evaluations, is used in the comparisons with the SME evaluations.

References

1. Geers TL (1984) An objective error measure for the comparison of calculated and measured transient response histories. Shock Vib Bull 54:99–107
2. Russell DM (1997a) Error measures for comparing transient data: part I: development of a comprehensive error measure. In: Proceedings of the 68th shock and vibration symposium, pp 175–184
3. Russell DM (1997b) Error measures for comparing transient data: part II: error measures case study. In: Proceedings of the 68th shock and vibration symposium, pp 185–198
4. Sprague MA, Geers TL (2003) Spectral elements and field separation for an acoustic fluid subject to cavitation. J Comput Phys 184:149–162

5. Oberkampf WL, Trucano TG (2002) Verification and validation in computational fluid dynamics. Prog Aerosp Sci 8:209–272
6. Oberkampf WL, MF Barone (2004) Measures of agreement between computation and experiment: validation metrics. AIAA 34th Fluid Dynamics Conference, Portland OR, June 2004
7. Knowles CP, Gear CW (2004) Revised validation metric. Unpublished manuscript, 16 June 2004 (revised July 2004)
8. Groethe MA, Gran JK (2000) Ground-shock enhancement using multiple explosive charges. 16th International Symposium on Military Aspects of Blast and Shock (MABS), Keble College, Oxford, England, UK, pp 10–15 September 2000
9. Sprague MA, Geers TL (2004) A spectral-element method for modeling cavitation in transient fluid–structure interaction. Int J Numer Methods Eng 60(15):2467–2499