A method for assessing the quality of datasets for use in model validation

School of Engineering

University of Liverpool

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor of Philosophy by

Chloe McElvaney

March 2023

Abstract

Computational models are used widely in industry to forecast behaviour and make predictions of the performance of structural components. The outputs from computational models can be used to inform decisions of high socioeconomic consequence. For this reason, it is recommended that models are validated against corresponding measurement data to ensure that the model is an accurate representation of reality, relative to its intended use. A key challenge associated with existing validation frameworks is the assumption that there will be a richness of measurement data available for comparison with the corresponding predicted data. For many industries, acquiring new measurement data can be time consuming, expensive and not always viable due to physical and ethical constraints. For these circumstances, new methods of establishing confidence in models are needed.

For systems where the measurement data is lacking richness, the approach explored in this thesis is to establish the quality of the measurement data and incorporate a quality measure into existing validation methods. The quality of the measurement data is evaluated using a new methodology which has been developed using ideas incorporated in the analytical hierarchy process and the rational decision-making process.

The quality of the measurement data is established by a panel of participants who score the data according to how well it possesses a set of attributes which have been taken from a good measurement practice guide. The scores assigned to the dataset, along with weightings which indicate the importance of the attributes, are used to calculate a measure of quality QF, termed the quality factor. The quality factor allows an assessment of whether the quality of the measurement data is sufficient for its intended use. When incorporated into a validation metric, it provides a measure of the extent to which predictions are representative of measurement data for their intended use, taking account of the quality of the measurement data. In this thesis, the quality factor has been incorporated into an existing validation metric which calculates the probability that a set of predictions belongs to the same population as the measurements, for a given uncertainty in the measurement data and intended use.

The new methodology has been successfully demonstrated in three case studies in the discipline of mechanics of solids. The first case study uses a series of sparse strain gauge data acquired from a loaded specimen. The second and third case studies use full-field displacement data of a bonnet liner subject to a projectile impact, and a plate subject to thermoacoustic loading. This marks the first use of the analytical hierarchy process and rational decision-making process for validation purposes, and when the quality factor is incorporated into a validation metric, this results in the first validation metric which incorporates the quality of the measurement data into the validation outcome.

Acknowledgments

First and foremost, I would like to express my profound gratitude to my supervisors Professor Eann Patterson and Dr Ksenija Dvurecenska for their support and guidance during my PhD study. I fully appreciate the patience they have shown me, particularly with my writing. I would also like to thank my industrial supervisors Samantha Wilkinson and Dr Allan Harte for sharing their valuable industrial expertise.

I would also like to thank all those who generously gave up their time to participate in my case studies. Their participation allowed me to test and develop my research ideas.

The project was funded by the UK Engineering and Physical Sciences Research Council in the form of a PhD studentship from the GREEEN CDT, and I am very grateful for the financial support provided.

I would like to express my deep gratitude to my partner and mum, both of whom have provided me with endless support and encouragement throughout the ups and downs of my project. A special thank you to my partner for combing through every page of this thesis and dealing with my stress!

Finally, I would like to dedicate this thesis to my beloved cat Sox, who sadly passed during the early stages of writing. 🐾

Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Introduction

Computational models are developed to study and predict the behaviour of complex systems across a wide range of domains such as: public policy, transport, finance, business and manufacturing [1]. Models are developed to provide predictions of a complex system, explore future ‘what if’ scenarios, understand theory or design, provide visualisation and to aid new insights. The outputs from models can be used to aid decisions which may have a high socioeconomic consequence. Therefore, it is important for modellers and decision-makers to have confidence that their model is working correctly and a suitable representation of what is being modelled. Performing verification answers the question of whether the model has been built correctly, and validation answers the question of whether the model is suitable.

Generally, validation is performed by comparing prediction data from a model to corresponding measurement data acquired from a physical experiment. The data used to validate the predictions from the model must be different from the data used to build the model. A standard published by the American Society for Mechanical Engineers [2] defines validation as the *‘process of determining the degree to which a model is an accurate representation of corresponding physical experiments from the perspective of the intended uses of the model’*. This definition has been derived from a Department of Defence (DoD) instruction [3] and the American Institute of Aeronautics and Astronautics (AIAA) guide for the verification and validation of computational fluid dynamics simulations. In the DoD instruction and AIAA guide [4], validation was defined as ‘the process of determining the degree to which a model is an accurate representation of the real-world from the perspective of the intended uses of the model’.

There are existing frameworks such as the CEN guide [5] for the validation of computational solid mechanics models, which guide the user through the necessary steps of validation. The steps required are dependent upon the measurement data available and the intended use of the model. The work described in the CEN guide builds on the research outputs of two completed projects: Standardisation Project for Optical Techniques of Strain Measurement (SPOTS) [6] and Advanced Dynamic Validations using Integrated Simulation and Experimentation (ADVISE) [7]. The SPOTS project (2003-2005) resulted in the development of a calibration methodology for optical systems which measure strain fields on a planar surface [8] [9] – this provided a route for acquiring high quality validation data from experiments. The ADVISE project (2008-2011) developed this work further by developing a methodology for comparing large datasets and including dynamic and out-of-plane loading in the calibration methodology [5]. The outputs from the SPOTS and ADVISE projects led to the development of the Validation of Numerical Engineering Simulations: Standardisation Actions (VANESSA) project. The goal of the VANESSA project (2013-2014) was to develop a standards framework which incorporated the validation methodology and associated calibration procedures. The CEN Workshop Agreement on the Validation of Computational Solid Mechanics Models [5] (referred to as the CEN guide in this thesis) was developed in the VANESSA project.

The CEN guide recommends that full-field measurement data, such as strain and displacement maps, are used to validate computational solid mechanics models. Such full-field data can be obtained from optical techniques such as digital image correlation (DIC). However, acquiring full-field measurement data can be very challenging for many industries. For example, in the nuclear industry, data acquisition is limited and can be restricted to specific regions of interest. Specifically, in a reactor environment, it is difficult to acquire measurements due to the hostile environmental conditions present in and around the reactor. Additionally, performing new experiments is time-consuming and expensive. For these circumstances, existing data from sparse and historical datasets should be utilised if their quality is suitable for the intended purpose. The research in this thesis aims to address this challenge through the development of a new methodology which determines and incorporates data quality into validation.

The following three case studies were selected to explore the efficacy of the new methodology set out in this thesis:

1. Tensile plate with a hole
2. Impact on bonnet liner [10]
3. Plate subject to thermoacoustic loading [11]

The data from the first case study was acquired for use as a sparse case study, and the data from the second and third case studies is historical and previously published.

Validation assesses if a model is an accurate representation of the reality it is modelling. If it is found to be an accurate representation, this will provide the modellers with confidence in the model they have built, and the decision-maker with confidence in the predictions. However, the output from the validation exercise does not account for potential quality issues with the measurement data which has been used to determine if the model is valid. The work described in subsequent Chapters aims to assess this gap.

1.2 Aim and objectives

The aim of the research will be to develop a reliable and transferable quantitative validation technique, which will be widely accepted in academic and industrial sectors, and apply it to mechanistic models with sparse data. This will be achieved through the following objectives:

a) To quantitatively assess the suitability of measurement data for validation.

b) To consider the effect of data sparsity on validation outcomes.

c) To validate models with sparse data by extending an existing probabilistic validation metric.

1.3 Thesis outline

Following this introduction, prior work which has contributed towards the ideas and motivation surrounding the work in this thesis is presented. The thesis has been divided into eleven Chapters. Chapter 2 contains a literature review of validation methodologies, a demonstrative example of the validation of computational solids mechanics models and an overview of validation metrics and their desired features. This Chapter also contains a literature review of the analytical hierarchy process and the rational decision-making process as these topics have been embedded within the new methodology presented. Preliminary work investigating the use of historical data for validation is outlined in Chapter 3 with a supporting case study using nuclear graphite data. The new data quality methodology is outlined in Chapter 4. The quality output from the methodology, termed the quality factor, is discussed in Chapter 5. The three case studies used to demonstrate the data quality methodology – tensile plate, bonnet liner and thermoacoustic plate, are introduced, assessed and discussed in Chapters 6, 7 and 8 respectively. Chapter 9 contains a discussion of the methodology and its key limitations. Conclusions and proposed future work are then outlined in Chapters 10 and 11 respectively.

Chapter 2: Literature review

The literature review presented provides a detailed review of the development of validation methodologies beginning with early discussions of validation that started in the field of economics. This Chapter also includes a review of the validation of computational solid mechanics models and the application of orthogonal decomposition for comparing measured and predicted data.

In this thesis, the measure of quality from the new data quality methodology, termed the quality factor, has been incorporated into an existing probabilistic validation metric. Therefore, this literature review also contains a review of validation metrics and specifically, the probabilistic validation metric. Recent work which is exploring establishing model credibility instead of pursuing robust validation, particularly when real-world data is not available, is also discussed. Finally, the analytical hierarchy process and the rational decision-making process are reviewed as ideas from these processes have been incorporated into the new methodology outlined in Chapter 4.

2.1 Introduction to validation

Prior to the 1970s, little literature existed on methodologies for performing Verification and Validation (V&V). An advancement was made in the 1950s when the development of digital computers led to model validation becoming a concerning topic amongst those in the simulation community [12]. Two decades later, literature surrounding V&V and its application emerged in economics. One of the main challenges that was observed was the misuse of the terms ‘verification’ and ‘validation’; these terms were used interchangeably and the distinction between them was not clear. Fishman and Kiviat [13] were among the first to define clear definitions of verification and validation. In their 1968 paper which discussed simulation statistics in economic science, they defined verification as the process which 'determines whether a model with a particular mathematical structure and data base actually behaves as an experimenter assumes it does'; validation was defined as a process which 'tests whether a simulation model reasonable approximates a real system'. Although this paper did not include a step-by-step methodology for performing validation, providing a definition of validation was a significant advancement for the field.

Naylor and Finger [14] also contributed significantly to the field in their 1967 paper. In this paper, they analysed three methodologies used in economics validation. The three methodologies discussed, each of which philosophically founded, were rationalism, empiricism and positive economics. Naylor and Finger proposed a new multistage validation methodology which incorporated the three discussed methodologies, as it was claimed that a union of the three methods was required to truly distinguish a model that is 'true' from one that is not. The extent of the agreement between simulated data and observed data was also investigated in this paper through goodness of fit tests such as analysis of variance, chi-square test and spectral analysis. While attempting to describe the extent to which the simulation data matches the observed data is significant for decision-makers, their paper did receive much criticism. Schrank and Holt [15] noted in their critique that an explanation of the application of the goodness of fit tests was not provided. They also speculated that Popper's criterion [16] could be a possible basis for model validation. Popper's criterion states that for a theory to be considered scientific, it must be tested and conceivably proven to be false - i.e., a model would be considered valid by checking that the model is capable of producing correct results for all possible experimental conditions. In practice, this is difficult to ascertain due to the time and expense associated with performing experiments.

In the 1990s, it was suggested by Kleindorfer [17] [18] that the multistage validation concept proposed by Naylor and Finger [14] grounded the objectivism approach in philosophical science. Kleindorfer stated that an objectivist approach looks for a gold standard validation method which is applicable for any model. Whereas, a relativist approach asserts that the validity of a model is subjective and thus there is no true answer.

A significant issue that remained was the absence of a clear validation methodology which included a set of actions, and a criterion for choosing between the actions. In 1971, this gap was addressed by Van Horn [19] who provided a methodology for performing validation which was based on the Naylor and Finger's multistage validation approach. Although Van Horn's paper was directed towards management science applications, the methodology was applicable in the field of engineering. The proposed methodology included three key steps: 1. The construction of the model, 2. The testing of empirical assumptions and 3. Comparing input-output outcomes.

Generally, validation methodologies can be divided into two categories: subjective techniques and objective techniques. Sargent [20] stated subjective techniques could be used for operational validity, but the applicability of the technique would be dependent upon the observability of the system. The observability of the system relates to how much data is available; an observable system is one for which a sufficient amount of data can be collected, a partially observable system is one for which only a limited amount of data can be collected, and the system is classified as unobservable if no data can be collected.

If observational data is not available for comparison with a corresponding model, historical data can be used to establish operational validity [21]. As highlighted by the American Society for Mechanical Engineers (ASME) guide for V&V in computational solid mechanics [22], many issues arise with the use of historical data; the effective use of historical data relies upon assumptions and uncertainties within the data being well-documented. Additionally, a key issue that arises is lack of completeness – i.e. the historical data may lack regions of interest which are important for the validation exercise. Historical data collected from literature surveys can also be used when the available observational data is sparse. Norman and Blattnig [23] worked on validating NASA models for space radiation applications using sparse historical experimental datasets. In their report, it was emphasised that the historical data they were using was collected for a different intended use and purpose. This posed the question of whether the datasets would be suitable for validating a model which has a different intended use. This is a key challenge that occurs with the use of historical data for validation purposes, and often this leads to the requirement for performing new validation experiments to compensate for data gaps.

Balci outlined a number of subjective techniques in his discussion of the acceptability and credibility of simulation results [24]. These included the Turing test, face validation and graphical comparisons. The Turing test was proposed by mathematician Alan Turing, and was designed to answer the question of whether machines can think [25]. This test can be utilised for validation purposes to establish model confidence through the testing of expert knowledge. This is conducted by presenting experts with data from the experiment and data from the model, confidence in the model is then enhanced if they are unable to distinguish between the two datasets. Face validation can be used in the preliminary stages of validation – this is where experts are asked if the data from the model is in line with their expectations. This helps determine if the model is conveying the correct logistics and concepts. Graphical comparisons can use histograms, box plots and behaviour graphs to compare variables as a function of time for both the model and the experiment [20]. However, as pointed out by Oberkampf and Barone [26], at that time, graphical comparisons were qualitative and did not incorporate or consider uncertainties. Nevertheless, preliminary approaches such as face validation and graphical comparisons were able to highlight model faults early.

Objective techniques, such as statistic tests, are also used to compare experimental data with corresponding model data. Objective comparisons are preferable for achieving robust validation, however as pointed out by Sargent [20], it is not always possible to use statistical tests to compare the outputs from the experiment and the model. A statistical test will often require that a sufficient number of observations are available, but this number can differ according to the system under study.

Many statistical tests have been proposed for model validation purposes in the literature, including: confidence intervals and Hotelling's T^2 test [27]. Confidence intervals provide objective comparisons by calculating differences between distributions and measures such as mean and variance. A large number of available observations will result in a narrow confidence interval, which in turn results in a more precise estimate. The width of the confidence interval is also determined by the quality of the observations.

An outline of Hotelling's two-sample T^2 test is provided by Balci and Sargent [28] in their paper which focuses on the validation of multivariate response models. The procedure tests if the model is valid, or invalid, for an acceptable accuracy range, under a specific set of experimental conditions. In practice, a Boolean result stating if a model is valid or invalid does not provide the decision-maker with insights into the merits and drawbacks of the models. Furthermore, it does not provide the decision-maker with the extent to which their model is valid, or invalid.

As mentioned at the start of the literature review, Naylor and Finger [14] were the first to provide a definition of validation. Sargent [29] extended this definition by adding that validation should be dependent upon the model’s intended use. An issue that remained following these definitions, was the lack of actions for conducting model validation.

The first guide for V&V was published by the American Institute of Aeronautics and Astronautics (AIAA) in 1998 for use in computational fluid dynamics [30]. This facilitated the development of the ASME guide in 2006 which was developed for applications in computational solid mechanics [22]. The ASME guide provides a conceptual framework for validating computational solid mechanics models. In this guide, verification is described as ‘the process of determining that a computational model accurately represents the underlying mathematical model and its solution’. Validation is defined as ‘the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model’.

Figure 1 is a flowchart from the ASME guide which outlines the steps that need to be performed for V&V. The process begins with a definition of the model’s intended use and then the flowchart divides into two streams: one for the model, which results in simulation outcomes, and one for the experiment, which results in experimental outcomes. The experiments are designed specifically with the aim of providing observational evidence to assess against the outcomes of the model. Both of these streams include quantification of associated uncertainties. Following the collection of simulation data and experimental data, a quantitative comparison is made. A validation metric is applied to assess the comparison between the two datasets. If an acceptable agreement is found between the model and the experiment, the model will be declared valid for its intended use. If not, the validation process will be repeated and this may involve refining the model or performing additional experiments.

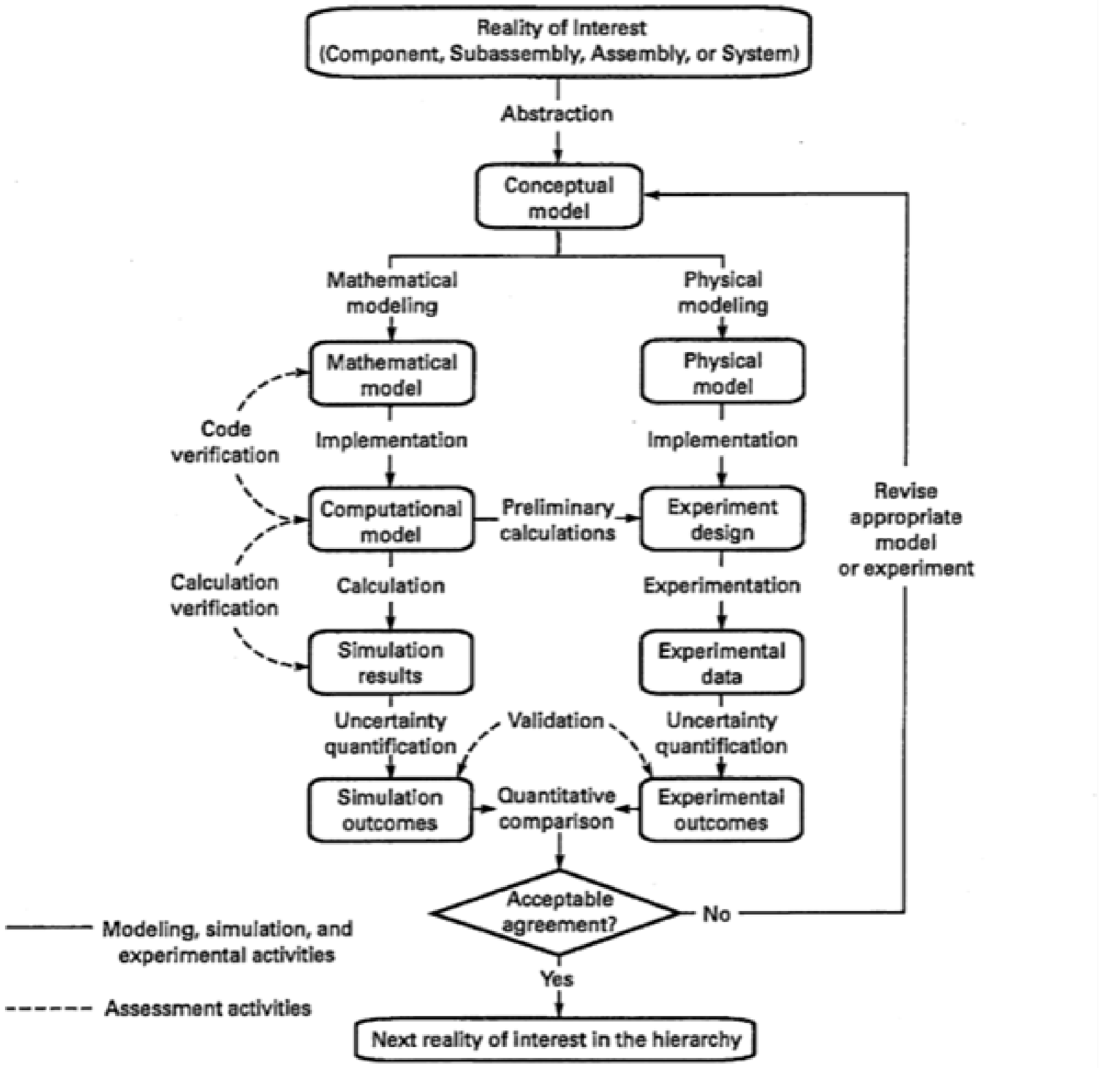


Figure 1: Verification and validation flowchart taken from the ASME guide *[22]*.

One of the drawbacks of the ASME guide is that the accuracy requirements which are used to assess the model are not included. The accuracy requirements are necessary for determining if the model should be accepted or rejected by the decision-maker [22]. An updated ASME standard [2] now defines validation as the 'process of determining the degree to which a model is an accurate representation of corresponding physical experiments from the perspective of the intended uses of the model'.

2.2 Validation of computational solid mechanics models

As described by the 2014 CEN guide [5], a computational solid mechanics model examines how an object responds when it is subject to loading – this is important for foreseeing and preventing potential structure failure. Such models are often based on finite element analysis (FEA) models. There are three main steps involved in FEA; the first is pre-processing which involves creating the geometric mesh, modelling the material and defining any loading and boundary conditions. Simple elements are used to create the mesh and the computational cost increases with the complexity of the finite element used to simulate the physical form under study. This step is followed by solution, where convergence and error analysis occur. FEA models are numerical approximations to the exact solution and the accuracy of this approximation is improved by increasing the number of elements used in the model. The final step is post-processing of results and this comprises validation of the FEA model via a quantitative comparison with available experimental data.

As acknowledged by Sebastian et al [31], strain is a component of interest when validating computational solid mechanics models as it is a measurable parameter which directly relates to the component's failure. Stress can be determined from strain using a Young's Modulus relationship provided linear elasticity applies.

Prior to the availability of full-field data, validation was conducted by comparing points corresponding to areas of maximum stress/strain, which were located using a computational model [32]. For a typical validation experiment, resistance strain gauges were located at these corresponding locations on a loaded physical prototype [33]. The gauges contain thin strips of foil which deform due to strain, this results in a change in their resistance which is detected by instrumentation. The strain gauges are calibrated so that the resistance measurements can be converted into strain values, and from these values stress can be evaluated.

The strain profiles obtained from the experiment are then compared to those obtained from FEA models to determine if there is a good agreement between the measurements and the predictions. The stress distribution obtained from the model can also be quantitatively compared to the stress distribution obtained using techniques such as photo-elasticity and digital image correlation. In this context, qualitative validation is performed by comparing images by eye and making a judgement of how well they compare, without applying metrics to describe the level of agreement between them.

The above approach relies on the ability of the computational model to correctly identify areas of maximum stress and it is limited by the number of strain gauges. Moreover, robust validation cannot be achieved by the comparison of a few data points. As acknowledged in the CEN guide [5], another disadvantage of this approach is that model validation is not performed in regions of apparent predicted low stress where component failure may still occur.

Full-field data is preferred for validation as it provides strain and displacement data over the entire surface and contains a large number of data points comparable to the number of nodes present in FEA models, therefore allowing for a better quantitative comparison. As highlighted by Sebastian et al [31], obtaining full-field data using optical techniques such as digital image correlation, digital speckle pattern interferometry and thermoelastic stress analysis is relatively easy and inexpensive. To validate a computational solid mechanics model, data from a model is compared to full-field data from a physical experiment. As described by Hack and Patterson [34], the data generated from the model and experiment are expressed in their own respective reference frame. However, in order to validate the model, each data set must be converted to a common reference frame to allow for quantitative data comparison.

For quantitative comparison, data reduction is necessary as fields of data contains a large number of data-points which means that a point-by-point comparison is very time consuming

[32]. The compression or reduction of data can be performed using a process called orthogonal decomposition. In this process, fields of data such as strain and displacement maps are treated as images and are then decomposed using orthogonal polynomials such as Zernike, Chebyshev and Krawtchouk. For example, as described in the CEN guide [5], an image of strain or displacement *I(i,j)* can be decomposed using a series expansion of Tchebichef polynomials *t(i,j)*:

$$I(i,j) = \sum_{l=0}^{n} s_l \, t_l(i,j) \tag{1}$$

The coefficients of the polynomials are represented by s_l:

$$s_l = \sum_{i,j}^{N} I(i,j) \, t_l(i,j) \tag{2}$$

In equation (1), *n* represents the number of coefficients and in equation (2), *N* represents the number of data points.

The coefficients s_l are collated into a feature vector which describes the original image. To ensure that the feature vector is an appropriate representation of the original image, a reconstruction of the original image is performed. For a displacement or strain image *I(i,j)*, the reconstruction of the image *Î(i,j)* is found by reconstructing the data field from the feature vector. The average squared residual u^2_{deco} is then evaluated as:

$$u_{deco}^2 = \frac{1}{N} \sum_{i,j}^{N} (\hat{I}(i,j) - I(i,j))^2 \tag{3}$$

For the reconstruction to be acceptable, u_{deco} must not be greater than the measurement uncertainty obtained from the instrument. Additionally, there should not be a location which shows a clustering of residuals greater than three times u_{deco}. A cluster refers to a group of adjacent pixels which make up 0.3% or more of the total number of pixels in the region of interest [5].

Zernike polynomials are based on a polar co-ordinate system and are effective as they are invariant to rotation, scale and translation. This allows for direct data comparisons regardless of coordinate system, pitch, orientation and sampling grid differences [32]. Chebyshev and Krawtchouk polynomials are based on a cartesian co-ordinate system and as highlighted by Sebastian et al [31], they are an order of magnitude faster to implement for strain fields that have been acquired optically. Thus, reducing time and cost.

Orthogonal decomposition is an effective method as it reduces the dimensionality of the field data from both the model and experiment whilst maintaining information about the data in the corresponding feature vectors. The resultant feature vectors contain typically only 20-100 terms, allowing for a more efficient statistical comparison [5]. As acknowledged by Pakti et al [33], orthogonal decomposition has the advantage of not overlooking important regions of interest. However, Zernike and Chebyshev polynomials used in the orthogonal decomposition process are not proficient at providing an accurate description of strain and displacement fields when an artefact, such as a cut-out, hole or discontinuity related to the geometry, is present in the image [31]. Lampeas et al [32] did provide an alternative approach to decomposing geometries with discontinuities by highlighting that Zernike, Chebyshev and Krawtchouk polynomials are all applicable to continuous rectangular planar or circular regions. Therefore, you can split a geometry into a number of rectangular planar, and or circular regions and then perform decomposition separately on each. A new decomposition algorithm developed by Christian et al [35] addressed the challenge associated with decomposing irregularly shaped stress and deformation datasets. The algorithm enables decomposition to be applied to datasets which contain large holes or contain regions of missing data and it has been implemented into a programme called THEON [36] – a software package which decomposes and compares 2D spatial data regardless of geometry using Chebyshev polynomials. The algorithm for decomposition is based on QR factorization, which states that a n x m matrix A can assume the form A = QR, where Q is an orthogonal matrix and R is an upper triangular matrix [37].

To test if the model is a good representation of reality, the coefficients (also referred to as shape descriptors) of the feature vector representing the experiment and the coefficients of the feature vector representing the model are compared using a simple linear correlation [31], as shown in Figure 2. The model is deemed valid if the coordinate pairs from the two feature vectors lie within an area defined by $S_m = S_e \pm 2u(S_e)$; where S_m is the model feature vector, S_e is the experimental feature vector and $u(S_e)$ is the uncertainty in the experimental feature vector.

In Figure 2, the model on the left-hand side is considered valid and therefore an acceptable representation of reality as all the points fall within the uncertainty band. However, the model on the right-hand side is considered invalid, and therefore unacceptable, because several points lie outside of this region.

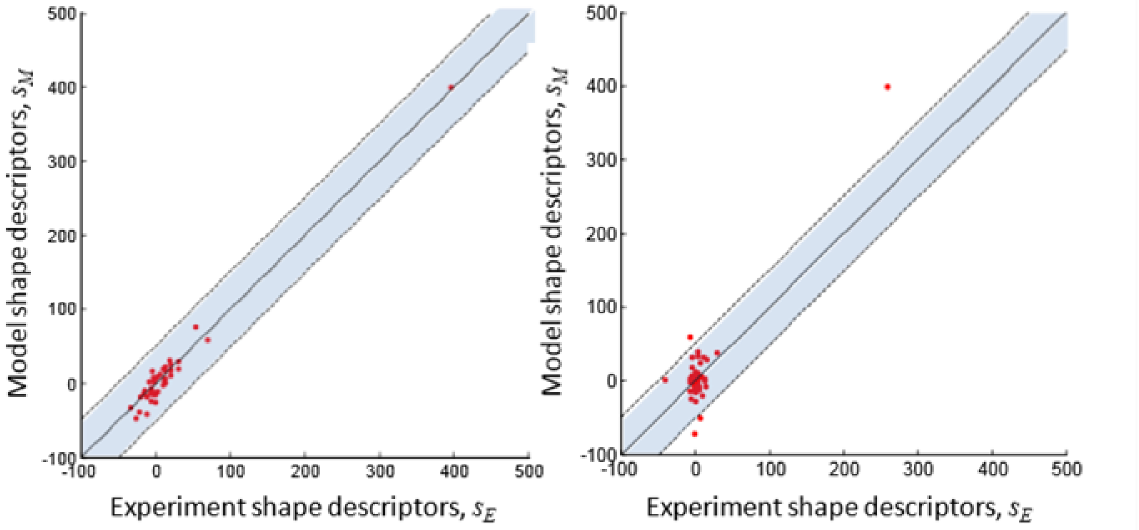


Figure 2: Graphical comparison of components of the experimental and model feature vector resulting from orthogonal decomposition. The model is considered to be valid if the red points fall within the dashed lines which have been defined by the minimum measurement uncertainty. Figure taken from ref *[5]*.

This approach to validation provides an efficient step-by-step methodology for quantitatively comparing full-field data acquired from a model and an experiment, whilst incorporating uncertainty in the experimental data. A flaw in this approach is that the measurement uncertainties are not transformed into the low-dimensional or feature vector domain. Alexiadis et al [38] addressed this flaw by developing a methodology which produces a distribution, using an approximate Bayesian computation, which represents the measurement uncertainty in the feature vector domain.

A further flaw of this approach is that the output is Boolean – i.e., the model is declared to be either valid, or invalid. In the next section, a validation metric which describes the extent to which the model is valid is discussed.

2.3 Validation metrics

A validation metric is described as a mathematical measure of the difference between two outcomes. In validation, a validation metric is applied to assess the comparison between the data acquired from the model, and the data acquired from the experiment. The choice of validation metric is dependent upon the data available and the outcome desired by the decision-maker. Oberkampf and Barone [26] stated that a validation metric can be described as a mathematical procedure that operates on a System Response Quantity (SRQ) of interest. An SRQ can be any physically measurable or inferable quantity that is measured in an experiment, or predicted by a model. It can be single-valued, like the point of maximum stress on a specimen, or it can take the form of a probability distribution [26].

For example, Ferson et al [39] investigated the predictive capability of probabilistic SRQs in their paper which focused on the thermal challenge problem. It was highlighted in this paper that a validation metric measures the mismatch between data from a model, and data from an experiment; where a low value will correspond to a good agreement, and a high value implies poor agreement between the two datasets.

In the literature, there is much guidance surrounding the recommended and desired features of a validation metric [40]. In general, it is stated that a validation metric should be quantitative and should incorporate the uncertainties related to the SRQ of interest. Additionally, it should account for errors that arise due to the postprocessing of experimental data. Ferson also added that another desirable feature of a validation metric is for it to be able to separate validation, from predictive capability; predictive capability refers to how much we can trust the predictions from the model, and validation refers to how good the model is. However, as acknowledged by Liu [41] , the desired features of the metric are subject to its intended use.

Dvurecenska et al [42] developed a novel validation metric, VM, for use on full-field data such as strain and displacement maps. It was recognised that the outcome of the CEN process [5] is a Boolean statement which declares if the model is valid or invalid for its intended use. This does not provide the decision-maker with an insight into how well the model's predictions compare to reality. In addition, at this time, a validation metric capable of handling the fields of data recommended by the CEN guide did not exist. Dvurecenska et al [42] addressed this gap by developing a probabilistic validation metric for use on fields of data, which incorporates uncertainty quantification.

Dvurecenska's [42] validation metric uses the measured and predicted feature vectors obtained from orthogonal decomposition as inputs. In the first step, the normalised relative error is computed using equation (4):

$$e_k = \left| \frac{S_{Pk} - S_{Mk}}{max_{m \in S_M} |S_{Mm}|} \right| \tag{4}$$

where S_{Pk} and S_{Mk} (referred to as S_M and S_E respectively in Figure 2) are the kth components from the predicted and measured feature vectors and the denominator is the magnitude of the largest absolute value from the measured feature vector.

The weight of each of the errors is defined in the second step as:

$$w_k = \frac{e_k}{\sum_{k=1}^{n} e_k} \times 100 \tag{5}$$

Where n is the number of coefficients in the feature vector. The error threshold, e_{th}, is calculated using the total uncertainty, u_{exp}, associated with the experimental data:

$$e_{th} = \frac{2u_{exp}}{max_{m \in S_M} |S_{Mm}|} \times 100 \tag{6}$$

In equation (6), the total uncertainty associated with the experimental data u_{exp} is calculated by combining the minimum measurement uncertainty u_{cal} and the average reconstruction residual u_{deco} as shown:

$$u_{exp} = \sqrt{u_{cal}^2 + u_{deco}^2} \tag{7}$$

The minimum measurement uncertainty u_{cal} can be obtained by performing a calibration process [5]. Patterson et al have proposed a reference material for the calibration of optical systems for full-field strain measurements [8]. The average reconstruction residual u_{deco} (see equation 3) measures the accuracy with which the feature vector represents the original data field [5].

In the final step, *VM* is computed by comparing the weighted errors to the error threshold:

$$VM = \sum_i w_i ||_{e_k < e_{th}} \tag{8}$$

where || is an indicator value which is equal to zero when the weighted error is greater than the error threshold, and equal to one when the weighted error is less than the error threshold.

The resulting statement provided to the decision-maker, using equation (8), is a clear statement regarding the probability of the model's predictions belonging to the same populations as the measurements. It is important to be able to communicate validation outcomes clearly to non-experts and decision-makers. Dvurecenska extended this work further by incorporating the VM outcome into a three-part statement which includes the probability of the model's predictions belonging to the same population as the measurements, for a given uncertainty in the measurement data and a given intended use. Such a statement may be: There is an 82% probability that predictions belong to the same population as measurements, when simulating out-of-plane displacement, given 10% relative uncertainty in the measurement data.

A probability can be interpreted easily by non-experts as they are used extensively in society to provide information about the likelihood of an event occurring. Many people use probability assessments in their daily life to make decisions. For example, people may avoid walking to work and instead opt for public transportation if there is a high probability of precipitation. When making decisions, people will assess the resulting probability of precipitation and make their own judgment of the risk, and these judgements can vary greatly. Gigerenzer et al [43] demonstrated that the statement 'A 30% chance of rain tomorrow' evoked various public interpretation. This ambiguity can be reduced by clearly explaining how the probability should be interpreted. One way to do this is by providing bands which describe the likelihood of rainfall occurring. An example of such a band would be almost no rain occurring if the probability lies between 10% and 30%. When the decision-makers review the outcome of the probabilistic validation metric, it is down to their discretion to assess the value and decide if it is appropriate for the intended use of the model.

Measurement uncertainty is harder to interpret by decision-makers and non-experts [44]. There can also be confusion regarding what information on measurement uncertainty should be communicated to decision-makers [45]. The total measurement uncertainty quoted in Dvurecenska's three-part statement can be overlooked if it is not fully understood by decision-makers, or if they are not able to interpret the significance of the value. It is easier to interpret probabilities as they lie between a range of 0 and 1. Whereas, measurement uncertainties are an inherent property of any quantitative measurement result and they express a lack of knowledge of the true value of the result [45]. This can make it harder for decision-makers to understand the significance of the measurement uncertainty and instead focus solely on the probabilistic outcome provided in the statement.

One of the consequences of the probabilistic validation metric is that when the relative error from the feature vectors is less than the measurement uncertainty, the value of the validation metric is 100%. This implies that there is a 100% probability that the predictions belong to the same population as the measurements – this is illustrated in Figure 3. While the statement does also refer to the uncertainty in the measurements, this part of the statement can be overlooked by decision-makers as they may not fully understand the implications. As described above, it is easier for decision-makers to interpret the probability aspect of the statement and use this to aid decisions. This may mislead the decision-makers into thinking that it is certain that the predictions from their model belong to the same population as the measurements used to validate the model. In reality, this statement needs to be exercised with caution due to the presence of the measurement uncertainty.

In this thesis, a modified version of the probabilistic metric has been developed which includes the measurement data's measure of quality – which has been obtained using the new methodology described in this thesis. The modified metric cannot have a value of unity because the value is moderated by the lack of quality of the data and the measurement uncertainty, which will never be zero-valued. This metric is discussed in Chapter 5. The moderation of an existing validation metric using the quality measure from the new methodology demonstrates the novelty of the work conducted in this thesis and the validation applications of the new methodology. By moderating an existing validation metric using the quality of the measurement data, validation outcomes are improved and the decision-maker will be provided with more information.

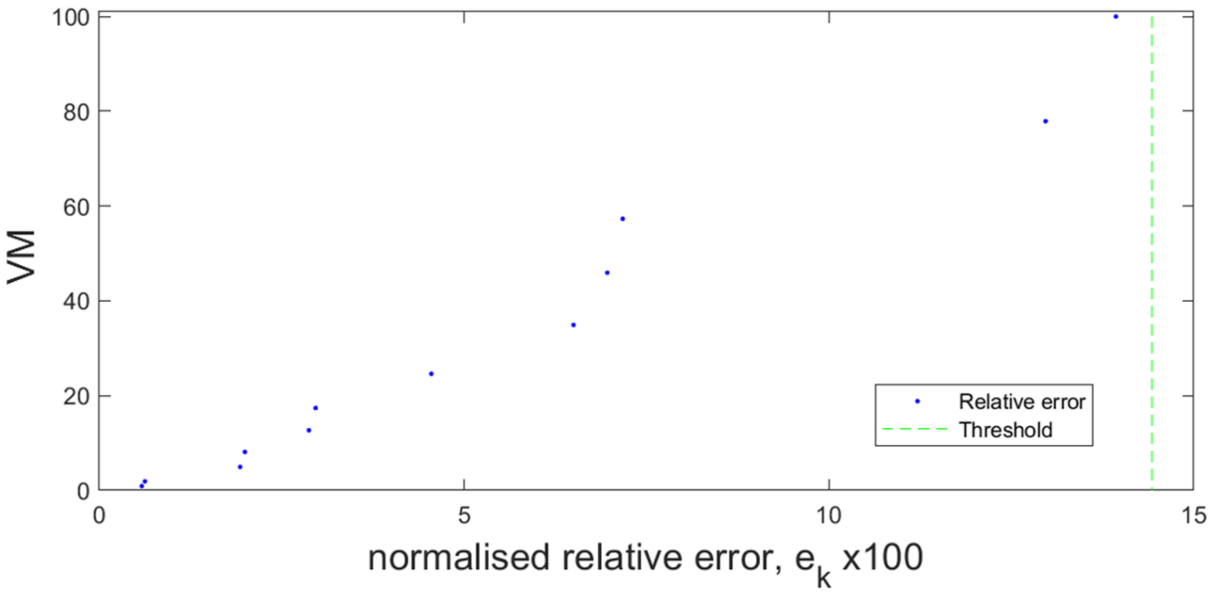


Figure 3: Dvurecenska's *[42]* validation metric which provides the probability that predictions belong to the same population as measurements. In the figure, the probability is calculated by summing the cumulative relative errors (blue dots) which sit below the threshold (green dashed line). As all of the relative errors sit below the threshold, the resulting probability is calculated to be 100%.

2.4 Credibility and validation

It has been proposed by Patterson that engineering models can be divided into two categories: informative and predictive [46]. Informative models are based on retrodiction and their value is seen to be heuristic, and predictive models are used to inform decisions that have socioeconomic consequence and so they require the establishment of model credibility. Credibility is the willingness of people to make decisions based on data from a model [47]. A 2 x 2 matrix (see Figure 4) has been developed to show the level of credibility that can be established by engineering models used for prediction and retrodiction.

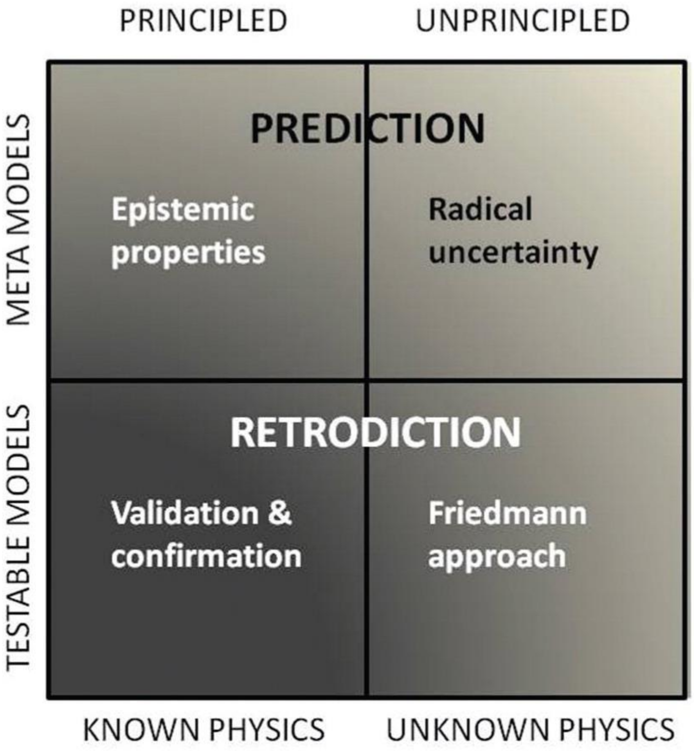


Figure 4: 2x2 matrix developed by Patterson [46] to describe the level of credibility that can be established (indicated by level of greyscale) for meta models and testable models based on known and unknown physics. The approaches to establishing model reliability are included in each region.

The matrix shows a separation between computational models concerned with retrodiction and engineering meta-models which predict future events with no real-world data available. Engineering meta-models are employed in fields of engineering where developed computational models are impossible to test comprehensively due to the lack of real-world data available. The right-hand edge of the matrix represents the boundary between known and unknown physics; unknown physics includes phenomena that has not yet been observed in nature or described using existing scientific laws or principles [46]. The ability to validate and establish credibility is indicated by the density of greyness – i.e., it is easiest to validate and establish credibility for testable models with known physics which sit in the bottom left hand region. In this region, existing validation methodologies from guides such as the CEN [5] and ASME [22] guide can be employed. However, the top right-hand region consists of meta models which are unprincipled and exhibit unknown physics. This region represents radical uncertainty and it is difficult to establish credibility in this region due to lack of information. For models which sit in the bottom right-hand region, engineers can use empirical evidence to support engineering models [46]. Models in the top left region suffer from lack of observational evidence from the real-world. For models within this region, the proposed approach is focused on the epistemic values of the model.

The 2x2 matrix shown in Figure 4 was developed further by Patterson and Whelan [48] to include validation approaches, based on Kleindorfer's various positions in the philosophy of science [17], which could be used to validate computational biology models which have a lack of available measurement data. The new matrix, shown in Figure 5, classifies models according to whether they are principled or unprincipled, testable or untestable, or whether the model is affiliated with known and unknown physics and the availability of real-world data. The philosophical validation approaches included are: Kuhnianism, Empiricism, Bayesianism, Falsificationism, Instrumentalism, Lakatos MSRP (Methodology of Scientific Research Programmes), Hermeneutics and Rationalism [48]. Similar to Figure 4, the right-hand box region represents models which are unprincipled and untestable, thus leading to radical uncertainty. Therefore, this region is relatively void of potential validation approaches and it is recommended that any validation approach should aim to move the models out of this boxed region. For models in this region, it is better to establish credibility in the model rather than perform a validation exercise. Patterson et al [49] developed seven credibility factors which were applied in the field of toxicology. These are: Assumption confirmation, qualitative concordance, quantitative concordance, explanatory power, internal consistency, external consistency and simplicity.

The shift from validation to credibility is appropriate for untestable models which lack or do not contain real-world data to compare against model predictions. Credibility needs to be established through a process of social epistemology [49] which promotes transparency, communication and generates shared knowledge.

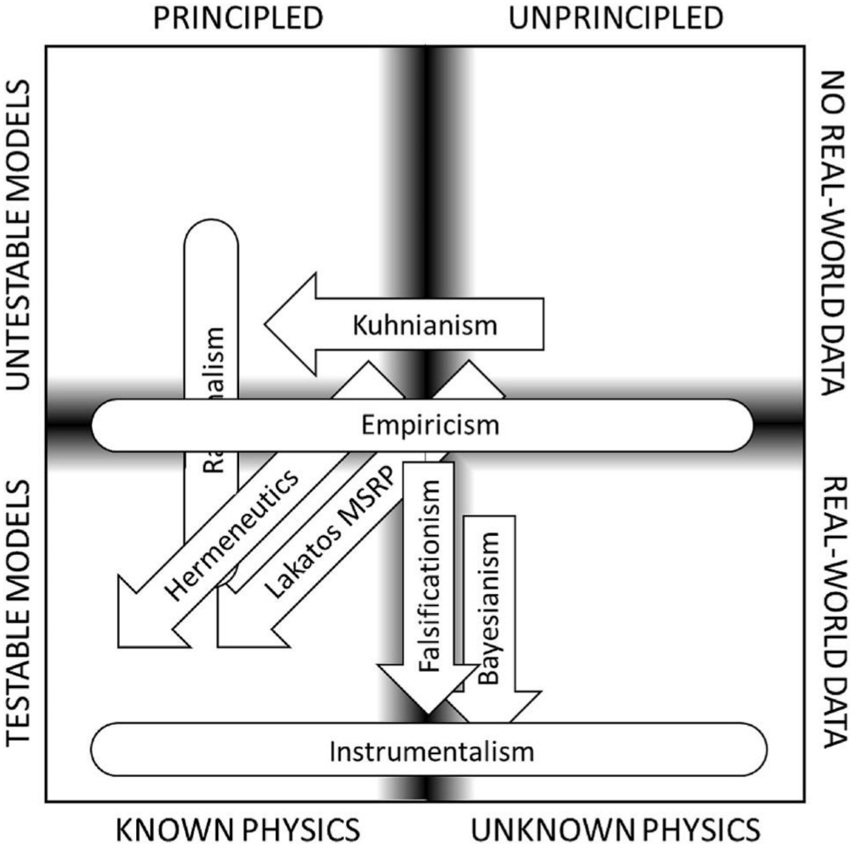


Figure 5: Validation approaches based on positions in philosophical science for models which are testable and untestable, principled and unprincipled, based on known and unknown physics, with and without real-world data. Figure taken from ref *[48]*.

2.5 Analytical hierarchy process

The analytical hierarchy process is a multi-criterion decision-making tool which is used in a wide range of applications which involve complex decision-making [50]. It is one of the most widely used multi-criterion decision-making tools and it is based on the principle that when making decisions, the experience and knowledge of people is at least as valuable as the data used [51].

Vaidya and Kumar [50] wrote a comprehensive literature review about the analytical hierarchy process, which investigated 150 papers from a wide range of areas. In their paper, they found that the process had been applied in areas such as selection, evaluation, benefit-cost analysis, planning and development, forecasting and medicine. The specific applications of the analytical hierarchy process were observed in areas such as: manufacturing, engineering, education, industry and government. Additionally, the analytical hierarchy process is also used to select competing alternatives in a multi-objective environment. For example, NASA's Lyndon B. Johnson Space Center used the analytical hierarchy process in a study aimed at selecting a propulsion system for the Lunar Lander [52].

The analytical hierarchy process is carried out in two stages: the hierarchic design stage and the evaluation stage [51]. In the first stage, the key steps are stating the problem of interest, considering the objectives and desired outcomes, and structuring the problem using a hierarchy [50]. The top of the hierarchy will represent the overall objective of the study and the levels below will represent the criteria of interest. The lowest level will represent the alternatives that are being assessed. An example of a simple hierarchy is shown in Figure 6. In this example, the overall objective is to choose a college. The desired criteria under consideration are: location, ambience, reputation and academics. The bottom level of the hierarchy represents the alternatives – i.e. in this example, the colleges that are being considered in the study. The desired criteria can be grouped and divided across multiple levels of the hierarchy. It is required that the hierarchy is detailed enough to capture the complexity of the problem, but simple enough to ensure that any changes can be easily implemented.

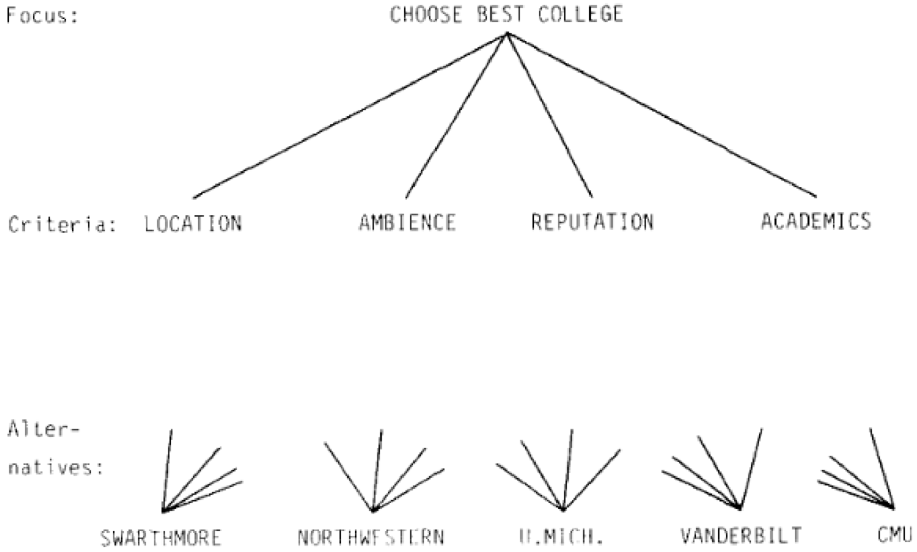


Figure 6: Example of a hierarchical structure used in the analytical hierarchy process. In this example, the goal is to choose the best college, the criteria under consideration are location, ambience, reputation and academics. The four alternatives are shown on the bottom level of the hierarchy. Figure taken from ref *[53]*.

In the evaluation stage of the process, a number of calculations are performed to determine which alternative is most favourable with respect to the criteria. An overview of the methodology is presented below, but more details can be found in references [53] and [54].

First, the weights (or priorities) of the criteria are calculated by performing pairwise comparisons of each of the criterion, with respect to the overall objective of the study [53]. The weights provide an insight into the relative importance of each of the criteria. The criteria are compared using Saaty's 1-9 scale [53]. The 1-9 scale represents the intensity of the importance, or how much more important one criterion is when compared to another. A value of 1 represents equal importance – i.e. the two criteria contribute equally to the overall objective. Assigning a score of 9 indicates that one criterion is extremely important compared to another. A summary of the 1-9 scale is shown in Table 1.

The results of the pairwise comparisons between the criteria form a positive reciprocal matrix where a_{ij} is the comparison between element i and j [54]. Referring to the hierarchy shown in Figure 6 if the location criterion is found to be extremely important when compared to the ambience criterion, a value of 9 will be placed in the corresponding position in the matrix. The reciprocal value, i.e. 1/9, will then be placed in the transpose position.

Table 1: Summary of the 1-9 scale used to assess pairwise comparisons of criteria in the analytical hierarchy process. The number corresponds to the intensity of the importance. An importance of 1 indicates that the two criteria contribute equally.

Intensity of importance	Definition
1	Equal importance
2	Weak
3	Moderate importance
4	Moderate plus
5	Strong importance
6	Strong plus
7	Very strong or demonstrated importance
8	Very, very strong
9	Extreme importance

The resulting positive reciprocal matrix, *A*, is defined as:

$$A = [a_{ij}] = \begin{bmatrix} 1 & a_{12} & \dots & a_{1n} \\ a_{21} & \dots & a_{ij} & \dots \\ \dots & a_{ji} = 1/a_{ij} & \dots & \dots \\ a_{n1} & \dots & \dots & 1 \end{bmatrix}$$

The weights of each of the criteria are found by dividing each element in the matrix by the sum of its column, and then calculating the mean of each row. This gives a vector where each element represents the weight of the specific criterion [54] [53].

In the next step of the evaluation stage, pairwise comparisons are performed on each of the alternatives for each of the criteria – i.e. each of the colleges are assessed with respect to location, then ambience, then reputation, then academics. Using the method discussed above, this will lead to four vectors which include the resulting weights of each of the alternatives, with respect to each of the criterion. These vectors are then collated to produce a matrix which represents the local priority of each of the alternatives. In the final step, this matrix is multiplied by the vector which contains the criteria weights. This produces a global priority vector where each element corresponds to each of the alternatives – i.e. each of the colleges under consideration. The college with the highest global priority is presented to the decision-maker as the most favourable option. An example of the calculation of the global priority vector is outlined in Appendix A.

2.6 Rational decision-making process

The rational decision-making process provides users with a structured step-by-step method for making decisions. It requires individuals to use facts and information to reach a decision [55]. The process has been cited extensively in the literature and is employed across a number of application domains; in the health-care industry, it has been used to recommend how resources should be allocated to the population [56] by determining where they are needed most [57]. It has also been used to recommend what care and treatment should be allocated to patients who have been diagnosed with dementia. Recommendations were determined by dividing the rational decision-making process into three phases: identifying individual needs, exploring options and making a choice. In engineering, the rational decision-making process has been used to determine the best full-field optical technique that should be employed for structural mechanics applications [58].

For the rational-decision making process to work effectively, the problem of interest needs to be clearly understood and well defined with clear objectives. The different alternatives must then be listed and evaluated. Based upon the evaluations, the best alternative can be selected and implemented. The applicability of the rational decision-making process is most effective for problems which have been defined clearly with well understood objectives [59].

The steps involved in the rational decision-making process are illustrated in Figure 7. The first step is identifying and defining the problem of interest. For step 2, the decision-maker establishes the criteria that are going to be relevant for the decision-making exercise. Without criteria, there will be no basis for comparing the different alternatives/options available [60]. These criteria will not be equally important to the problem of interest; thus, they are weighted in step 3. In steps 4 and 5, the available alternatives are listed and evaluated [61]. Following this evaluation, the most desirable option for the problem of interest can be recommended. If this solution is not implemented successfully, the process will be repeated.

A number of situations where the rational decision-making model is not effective have been identified [60]. These include situations where the required information is not correct or unavailable, situations where the problem of interest is changing significantly over the period of the decision-making process and situations where there are no defined set of criteria by which the alternatives can be assessed.

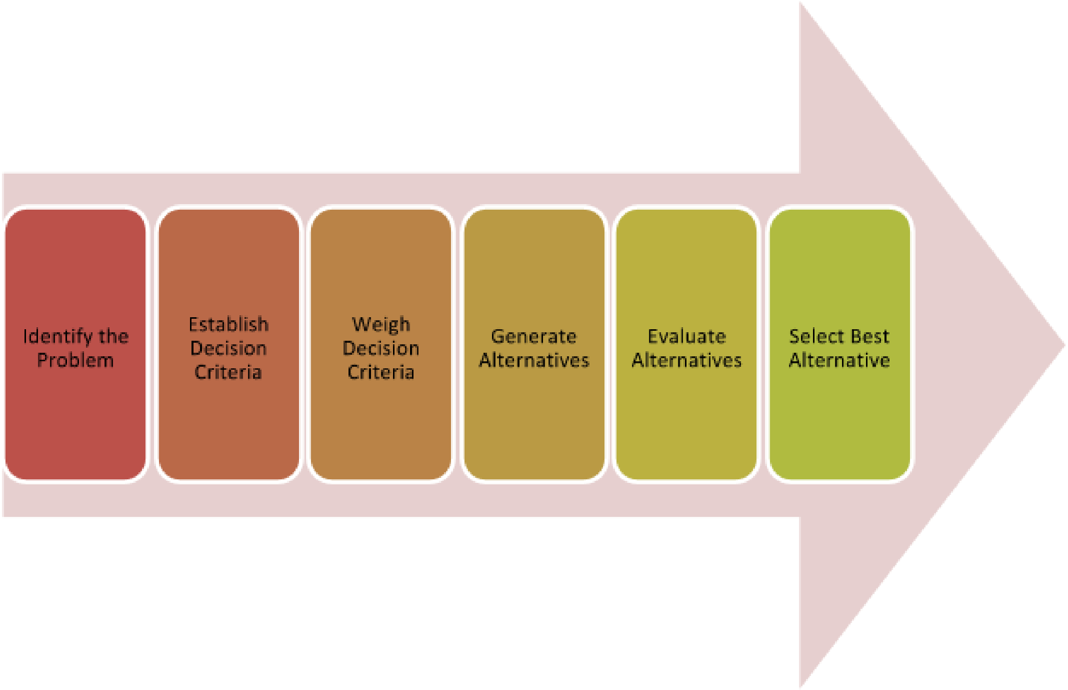


Figure 7: Steps required in the rational decision-making process *[61]*.

2.7 Summary of review

Validation is required to establish confidence in models and to provide decision-makers with evidence to determine the degree to which their model is an accurate representation of corresponding physical experiments, from the perspective of the intended uses of the model. Until Fishman and Kiviat, the term 'validation' was often confused with 'verification' and clear definitions of the two terms did not exist. Van Horn added to the definitions provided by Fishman and Kiviat by stating that validation was dependent upon the intended uses of the model. Existing validation guides and standards now provide users with guidance and the recommended steps for performing model validation.

In general, model validation is performed by comparing measurement data with corresponding prediction data. The two datasets are compared using a validation metric and this outcome will be compared against a specific decision-criteria. Experiments for acquiring measurement data are tailored to the model of interest. In many circumstances, it is difficult to acquire measurement data to support the validation process. For complex systems such as nuclear systems, hypersonic flight and biological systems which exhibit emergent behaviour, the measurement data available can be characterised as sparse. Such data will not be sufficient to support model validation using existing validation guides. For these circumstances, the decision-maker must rely upon sparse validation data or historical validation data to validate their model. For such data to be used for validation purposes, decisions must be made regarding their suitability.

For untestable models which do not have supporting real-world measurement data, it has been suggested that it would be more appropriate to establish model credibility rather than perform a validation approach. The level of credibility that can be established by a model is governed by the amount of information available – i.e. if the model is principled or unprincipled, associated with known or unknown physics. In reality, there is a fuzzy boundary between the models with and without real-world data available. For models within this region, there is a lack of measurement data available and additional measurement data cannot be obtained. Therefore, existing sparse and historical data must be utilised.

Validation experiments are specifically designed to produce high quality measurement data for comparison with predictions, but the quality of the measurement data is not embedded into the validation outcome. Existing literature does not evidence a validation metric which incorporates a quality measure of the data used to validate the model. These gaps have been addressed in the work presented in this thesis.

Chapter 3: Use of historical data for validation

The work described in this section was conducted in the first year of the PhD and it serves as motivation for the ideas used to develop the novel data quality methodology. The work presented highlights the issues associated with the use of historical data in the validation process.

Following the development of the CEN guide which provided a protocol for validating computational models using full-field data with the aid of orthogonal decomposition, an inter-laboratory study was performed to assess the protocol. Following the inter-laboratory study, three areas of further investigation were identified [62]:

a. Measuring the quality of predictions
b. Matching of regions of interest from the prediction and measurement fields
c. The importance of designing experiments for the specific purpose of performing a validation of a model

These areas of investigation were explored by The Matrix Optimization for Testing by Interaction of Virtual and Test Environments (MOTIVATE) project [63] [64], which was funded under the European Commission's Horizon 2020 program. This project was a collaboration between Airbus Operations, the University of Liverpool, Empa, Dantec Dynamics GmbH and Athena Research and Innovation Center.

The first area of investigation, measuring the quality of predictions, was addressed by Dvurecenska et al in the development of their probabilistic metric [42]. The second area of investigation, matching regions of interest, was addressed by Christian et al [35] in the development of their decomposition algorithm which allows fields of data with irregular shapes to be decomposed and compared [62]. These pieces of work have been discussed in Chapters 2.2 and 2.3 respectively. The third area of investigation relates to difficulties that arise due to lack of information about measurement data from experiments. This area of investigation was addressed by extending the flowcharts found in the 2006 ASME guide [22], the recently revised 2020 ASME standard [2], and CEN guide [5] to include consideration of historical data.

Figure 8 shows the MOTIVATE validation flowchart. This flowchart has been divided into two separate branches: one for modelling and one for physical testing. Once the steps within these branches have been completed, a quantitative comparison is performed using a validation metric and this will feed into the decision-maker's review [64]. In the physical branch of the flowchart, there is a sub-flowchart (see Figure 9) for evaluating the suitability of historical measurement data. Likewise, in the modelling branch of the flowchart, a sub-flowchart which evaluates the suitability of historical simulation data (see Figure 10) has also been incorporated.

To test the efficacy of the historical measurement flowchart, a case study was conducted in the first year of the PhD using nuclear graphite data. This case study is discussed in the next section.

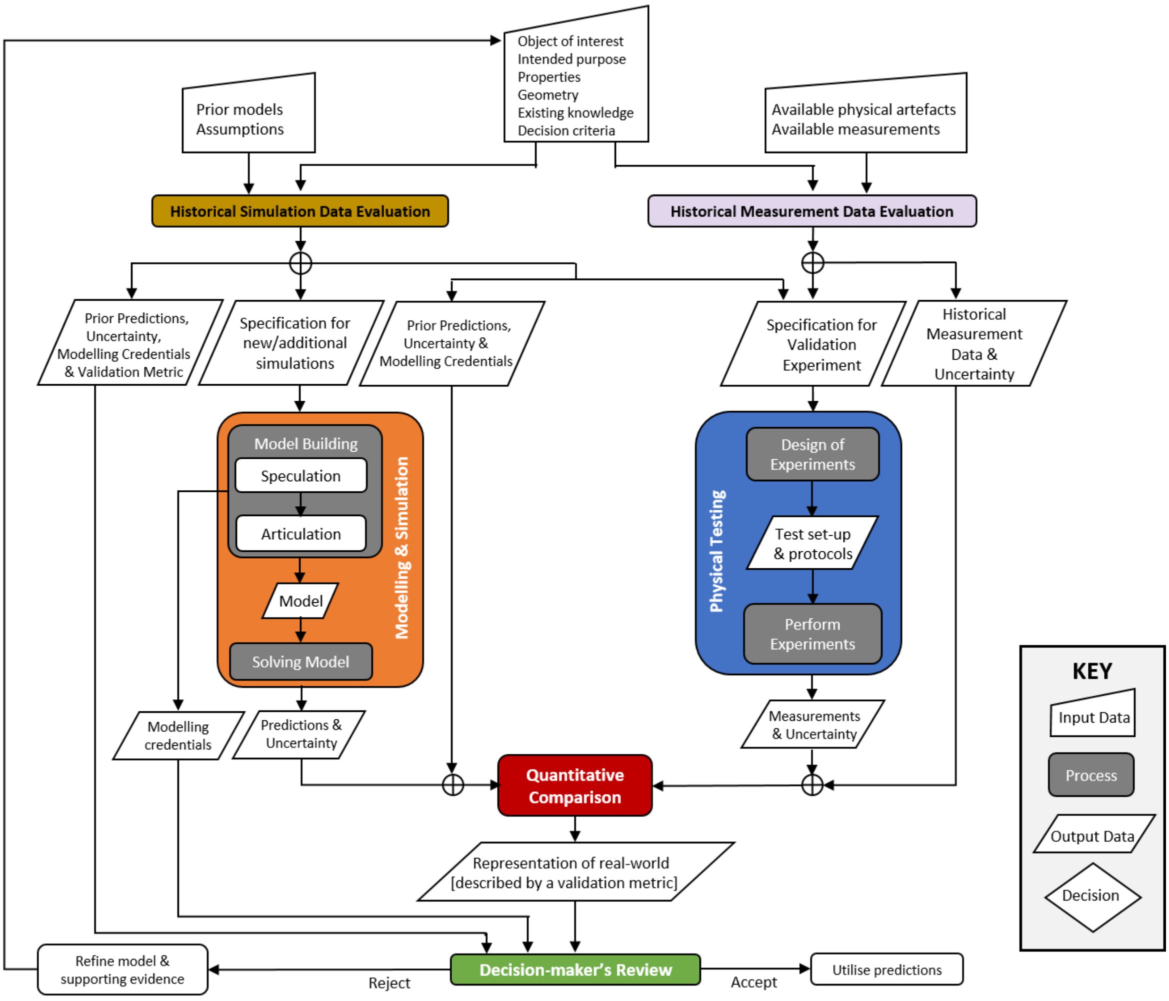

Object of interest
Intended purpose
Properties
Geometry
Existing knowledge
Decision criteria
Prior models
Assumptions
Available physical artefacts
Available measurements
Historical Simulation Data Evaluation
Historical Measurement Data Evaluation
Prior Predictions, Uncertainty, Modelling Credentials & Validation Metric
Specification for new/additional simulations
Prior Predictions, Uncertainty & Modelling Credentials
Specification for Validation Experiment
Historical Measurement Data & Uncertainty
Model Building
Speculation
Articulation
Model
Solving Model
Modelling & Simulation
Physical Testing
Design of Experiments
Test set-up & protocols
Perform Experiments
Measurements & Uncertainty
Modelling credentials
Predictions & Uncertainty
Quantitative Comparison
Representation of real-world [described by a validation metric]
Refine model & supporting evidence
Reject
Decision-maker's Review
Accept
Utilise predictions
KEY
Input Data
Process
Output Data
Decision

Figure 8: Validation flowchart developed in the MOTIVATE project *[63]*. The flowchart includes two separate branches for physical testing and modelling, both of which incorporate a stage for evaluating the suitability of historical data for use in validation.

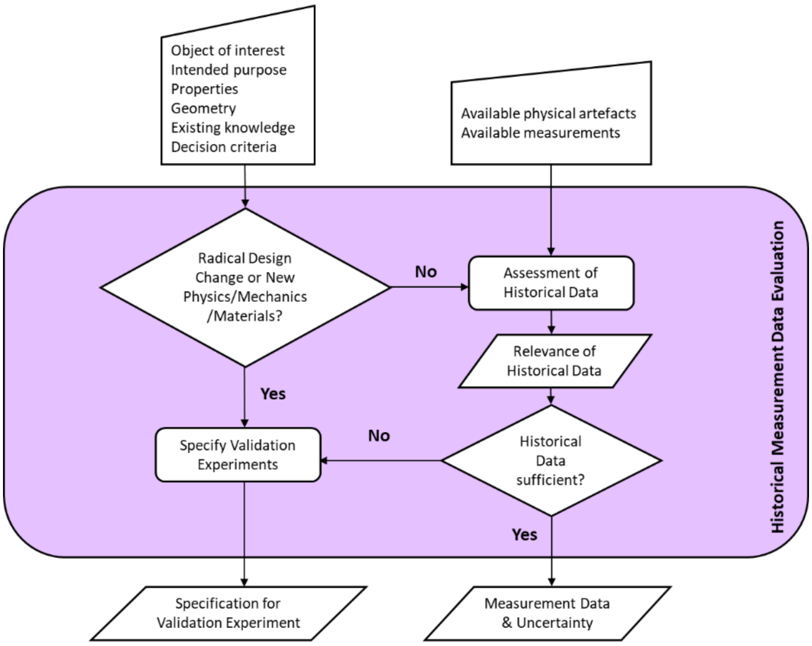

Object of interest
Intended purpose
Properties
Geometry
Existing knowledge
Decision criteria
Available physical artefacts
Available measurements
Radical Design Change or New Physics/Mechanics /Materials?
No
Assessment of Historical Data
Relevance of Historical Data
Yes
No
Specify Validation Experiments
Historical Data sufficient?
Yes
Historical Measurement Data Evaluation
Specification for Validation Experiment
Measurement Data & Uncertainty

Figure 9: MOTIVATE [63] sub-flowchart for evaluating the suitability of historical measurement data for validation.

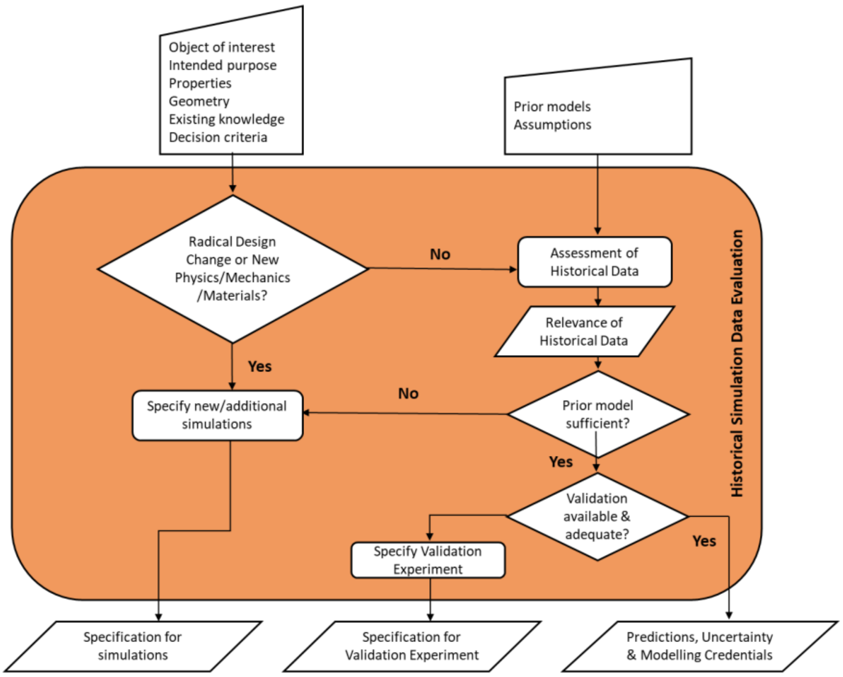

Object of interest
Intended purpose
Properties
Geometry
Existing knowledge
Decision criteria
Prior models
Assumptions
Radical Design Change or New Physics/Mechanics /Materials?
No
Assessment of Historical Data
Relevance of Historical Data
Yes
No
Specify new/additional simulations
Prior model sufficient?
Yes
Validation available & adequate?
Yes
Specify Validation Experiment
Historical Simulation Data Evaluation
Specification for simulations
Specification for Validation Experiment
Predictions, Uncertainty & Modelling Credentials

Figure 10: MOTIVATE *[63]* sub-flowchart for evaluating the suitability of historical simulation data for validation.

3.1 Graphite historical data case study

In this case study, the historical measurement data flowchart shown in Figure 9 was used to determine if historical data would be suitable to validate a 2D polycrystalline graphite model [65] [66]. The polycrystalline model was developed to predict the long-term behaviour of graphite in a reactor environment. Graphite is used in reactors such as Advanced Gas Cooled Reactors (AGRs) to provide structural integrity in the core and to moderate neutrons to thermal energies to optimise the fission reactions taking place. In a reactor environment, graphite is subject to two degradation mechanisms which impact key macroscopic properties: fast neutron irradiation and radiolytic oxidation. Fast neutron irradiation results in differential dimensional change across the graphite brick [67]. This results in the build-up of internal stresses within the brick which can lead to the development of cracking. Cracking is a significant safety concern because it can cause channel deformation which may impede the entrance and exit of fuel rod (a long tube which is filled with a string of fuel pellets) from the channel.

Radiolytic oxidation is a consequence of the CO_2 coolant which is used in nuclear reactors such as Advanced Gas Cooled Reactors (AGRs). In a reactor environment, the CO_2 present will react with gamma radiation to produce an oxidising species. The oxidising species produced then react with the adjacent graphite surface, which leads to weight loss of the surface. This is summarised in the reaction below:

$$CO_2 + \gamma \rightarrow Ox \tag{9}$$

$$Ox + Graphite \rightarrow \text{Oxidation products} \tag{10}$$

The weight loss from the graphite components results in the development of internal porosity which leads to a reduction in strength [68]. Thus, it serves as a key contributing factor to the limiting lifetime of existing reactors.

Extending the lifetime of the existing fleet of AGRs requires the development of robust safety cases that contain reliable evidence for the future behaviour of graphite in the reactor environment. This provided the motivation for the development of the 2D polycrystalline model which is used in this case study exemplar.

The historical data tested in this case study was German ATR-2E data [69] which was extracted from the IAEA Coordinated Research Project (CRP) database. This dataset was selected as it has previously been compared to the outputs from the 2D polycrystalline model in literature [65]. Furthermore, this dataset was chosen because it did not contain data gaps and it included well-defined details of sample characteristics, elastic properties and irradiation conditions. The ATR-2E samples were irradiated in the High Flux Reactor (HFR) in the Netherlands. This is a water-cooled and water-moderated multi-purpose Material Test Reactor (MTR) which is often used for irradiation programmes, fundamental research and isotope production [70].

To test the suitability of the ATR-2E historical measurement data, the flowchart represented by Figure 9 was applied to it. To navigate through the steps of the flowchart, the specific definitions and instructions provided in Table 2 were used.

Table 2: Specific user instructions for the stages incorporated in the historical measurement data flowchart.

Item in the flowchart	Description/steps to take
Object of interest	Define the structure and application context
Intended purpose	Identify the use of the measurement data
Properties	i.e. material and mechanical properties
Geometry	Identify the detailed geometry
Existing knowledge	Do we understand the behaviour well?
Decision criteria	This will be used to assess the outcome of the validation procedure
Radical design change of new physics	Is the design different to previous designs? Is the behaviour likely to be in a different physics domain?
Assessment of historical data	Is the documentation complete with uncertainty analysis? Are boundary conditions known?
Specification for validation experiment	Define load cases, measurement parameters, target uncertainty and data format

The first stage (beginning with object of interest) of the historical measurement data flowchart was completed with ease due to the abundance of existing knowledge available regarding the data and the grade of the irradiated graphite. Reactor conditions such as max fluences, irradiation temperatures and applied stresses were also well defined. The geometry of the ATR-2E graphite and its associated property changes due to fast neutron irradiation were also clearly documented [69].

Referring to the second stage of the flowchart, which focuses on radical design change or new physics/mechanics/materials, a radical design change was not documented. The ATR-2E graphite grade was incorporated in many German irradiation programmes and it was available on a large scale. As no radical design change or new physics were documented, this guided the flowchart to the ‘assessment of historical data’ stage.

For the assessment of historical data stage, it was concluded that the sample was well defined and documented to a high standard. However, one challenge that arose was the lack of documentation of uncertainties associated with the data. This made it difficult to ascertain if the historical data was of sufficient quality for its intended use. It was concluded that without knowledge of uncertainties, it is hard to pass a decision regarding the suitability of the data. Therefore, the recommendation was to conduct further validation experiments to obtain the necessary uncertainty information.

When data is sparse and limited, the approach explored in this thesis is to assess the quality of the dataset and incorporate this into the validation outcome. The new methodology described in this thesis provides the decision-maker with a statement of the dataset's quality alongside a recommendation of its suitability for the validation exercise. This methodology is described in the next Chapter.

Chapter 4: Development of data quality methodology

The methodology presented in this section has been developed to assess the quality of measurement data for validation. The quality of the measurement data is assessed to determine if the data is fit for its intended purpose. A key advantage of the methodology is that the final quality measure QF, termed the quality factor, can be incorporated directly into validation metrics. Thus, providing decision makers with more information about the validation data used to determine how representative their model is of reality.

The pilot data quality methodology, which was implemented in the first case study is described in Section 4.1. Following the results and feedback from the first case study, several improvements were made to the methodology. These improvements and justifications are described in Section 4.2 alongside an overview of the final methodology version – which was implemented in the second and third case studies described in this thesis.

4.1 Pilot data quality methodology

The methodology presented here has been developed using concepts from the analytical hierarchy process and the rational decision-making process. The analytical hierarchy process has been used extensively in applications which involve complex decision-making. The process compares multiple alternatives using a set of criteria and provides a recommendation for the most favourable option. The rational decision-making process provides an evidence-based intuitive process for making decisions.

In the new method, the quality of a measured dataset is evaluated by assessing how well it possesses eleven attributes which have been taken from a National Physical Laboratory Good Measurement Practice Guide [71]. The National Physical Laboratory have published many good practice guides which are designed to improve measurement understanding. For this methodology, a good practice guide which provides a beginner’s guide to measurement was chosen as it provided a more generalised overview.

The first six attributes used in the methodology are described in the guide as the six guiding principles that should be followed in order to obtain a good measurement. These have been referred to as the fundamental attributes. While, the last five attributes are described in the guide as the additional factors which can also affect the measurement result. These refer to the desirable attributes. When combined, both sets of attributes provide a thorough basis for determining data quality and highlighting areas for improvement. These attributes have been listed in Table 3.

To describe how well the dataset possesses the attribute, a score between 1 and 5 is assigned – a score of 1 indicates that the dataset does not possess the attribute and a score of 5 indicates that it fully possesses the attribute – these scores are illustrated in Table 4. For six attributes (regular review, demonstrable consistency, instruments, the object to be measured, sampling and environmental factors), a not applicable (N/A) option is also available in level 5 of Table 4. This option is provided as the technical panel members and subject matter experts may feel that the 1-5 scoring system is not appropriate for their specific case study . A flowchart highlighting the steps of the methodology is presented in Figure 11.

Table 3: The list of eleven attributes used to assess the quality of the dataset, taken from a Good Measurement Practice Guide of the National Physical Laboratory *[71]*. The first six attributes (termed fundamental), are the six guiding principles that should be followed in order to obtain a good result. The last five attributes (termed desirable) are the additional factors which can affect measurements.

Attribute	Description
1. The right measurements	A measurement is made for a reason and this needs to be clearly defined and understood.
2. The right tools	Measurements should be made using equipment and methods that have been demonstrated to be fit for purpose.
3. The right people	Measurements must be carried out by appropriate operators.
4. Regular review	Measuring instruments are often damaged, so regular checks should be carried out.
5. Demonstrable consistency	Measurements made in one location should be consistent with those made elsewhere.
6. The right procedures	Measurements should be carried out in accordance with written procedures.
7. Instruments	While calibrations and preliminary checks can confirm that measuring instruments are behaving as they should prior to measurement, a number of factors can impair their performance during the measurement. For example, electrical measuring instruments can be affected by electrical noise and acoustic measurements should be made in anechoic conditions.
8. The object to be measured	The object to be measured can change over time. In such cases, a consideration should be made when planning measurements.
9. Sampling	Measurements must be made in sufficient numbers and must be representative.
10. Operator skill	Is the operator sufficiently qualified to acquire the measurement?
11. Environmental factors	The environment can affect the results of the measurement and for some experiments, environmental factors need to be controlled, or corrections may need to be applied.

Table 4: Guidance for assigning scores to describe how well the dataset possesses the eleven attributes taken from a National Physical Laboratory good practice guide *[71]*.

	Level 1	**Level 2**	**Level 3**	**Level 4**	**Level 5**
Attributes	**Very dissatisfied**	**Dissatisfied**	**OK**	**Satisfied**	**Very satisfied**
1. The right measurements	The measurement is not related to the problem				The measurement is fully appropriate for the problem being studied
2. The right tools	The instruments have not been calibrated				The instruments have been fully calibrated
3. The right people	No training and instruction provided				The 'operator' has received the correct training and instructions
4. Regular review	No instrument check prior to use				N/A or Instruments are regularly checked and assessed
5. Demonstrable consistency	Measurement is only valid in one place				N/A or measurement has been replicated in different environments
6. The right procedures	Measurement not carried out in accordance with written procedure				Measurement carried out in accordance with documents provided by manufacturer
7. Instruments	No measures have been put in place to minimise electrical noise disruptions				N/A or proper earthing of equipment ensured
8. The object to be measured	The object to be measured varies and no corrections have been applied				N/A or environmental conditions have been controlled, corrections and averages applied or dynamic measurements taken
9. Sampling	Measurement technique has not been well-designed and is not representative				N/A or measurement technique is representative of important variations and there is knowledge of any expected changes
10. Operator skill	The operator is not skilled				The operator has correctly set up the measuring equipment and prepared the object to be measured
11. Environmental factors	Environmental factors not taken into account				N/A or corrections applied to take account of any environmental factors

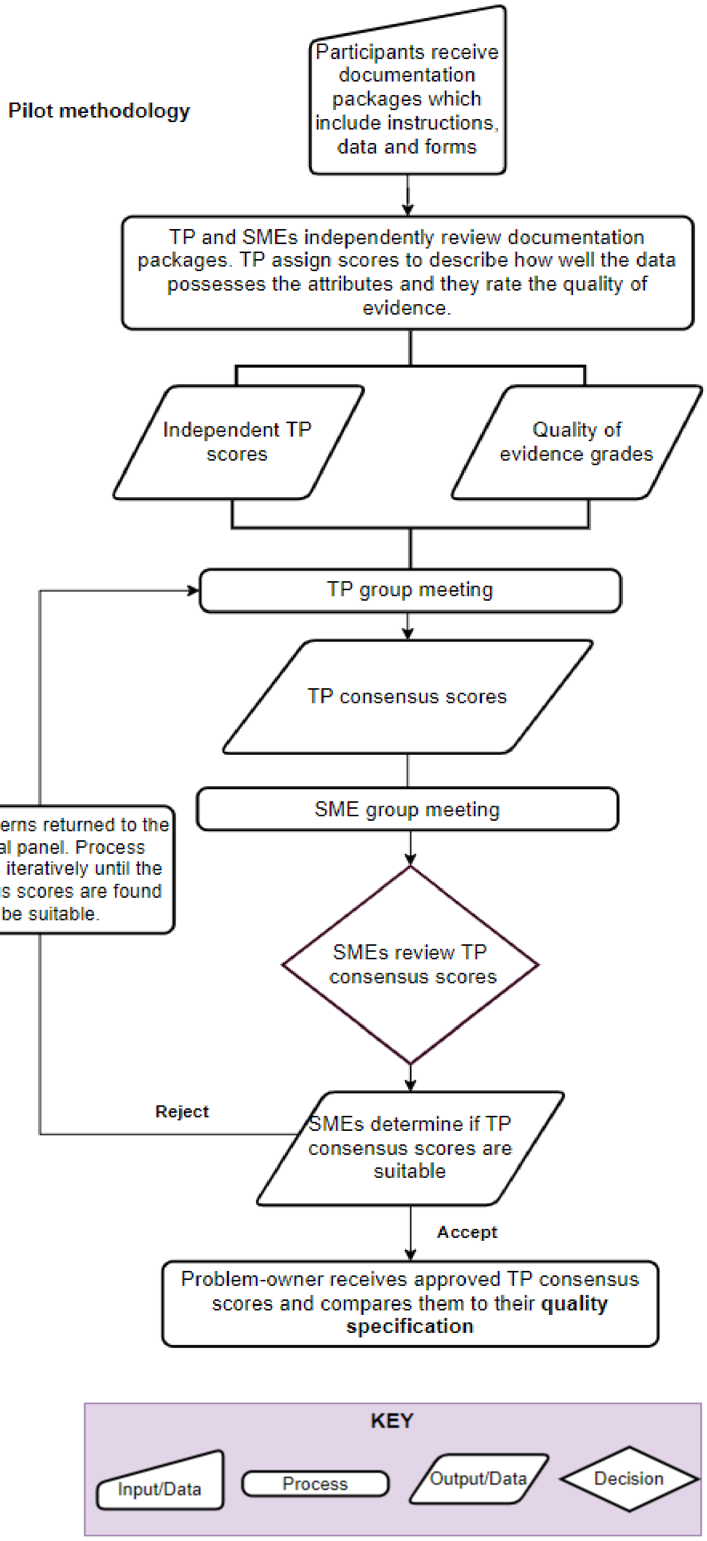

Pilot methodology
Participants receive documentation packages which include instructions, data and forms
TP and SMEs independently review documentation packages. TP assign scores to describe how well the data possesses the attributes and they rate the quality of evidence.
Independent TP scores
Quality of evidence grades
TP group meeting
TP consensus scores
SME group meeting
SME concerns returned to the technical panel. Process continues iteratively until the consensus scores are found to be suitable.
SMEs review TP consensus scores
Reject
SMEs determine if TP consensus scores are suitable
Accept
Problem-owner receives approved TP consensus scores and compares them to their quality specification
KEY
Input/Data
Process
Output/Data
Decision

Figure 11: Flowchart outlining key steps of pilot data quality assessment method and the participants involved in the methodology. TP denotes technical panel and SME denotes subject matter expert.

The methodology relies on an expert panel of participants who are split into two groups: a technical panel and subject matter experts. Previous work using the Delphi method has reported that a reliable outcome can be obtained if the panel consists of a minimum of 3 to 9 members [72]. The technical panel members and subject matter experts must have a deep knowledge of the field and relevant associated professional experience and qualifications – a higher level of expertise is required from subject matter experts. A subject matter expert is referred to as an individual who is recognised as having a special skill or a specialised knowledge of a process in a particular field [73]. As such, a subject matter expert must have demonstrable education, training, evidence continuous learning and experience.

The subject matter experts are responsible for reviewing the consensus scores provided by the technical panel to confirm that they are suitable, i.e., if the scores presented by the technical panel, and associated justifications, are appropriate for the validation exercise. When reviewing the consensus scores provided by the technical panel, the subject matter experts can choose to proceed with the technical panel consensus, or they can make changes to the scores, with justifications. As a result, it is important that the subject matter experts are impartial, unbiased and so have not been involved in the collection of the measured data. Evidence is provided to the technical panel members and subject matter experts by the problem-owner to aid the scores assigned.

Throughout this thesis, the problem-owner refers to the individual who assigns the problem. The panel of participants consist of the technical panel members and the subject matter experts.

In the first step of the methodology (see Figure 11), the problem-owner provides the members of the technical panel, subject matter experts and facilitator with a documentation package which includes:

- Specific role instructions

- Description of the problem
- Relevant data and documentation

The forms which were sent to the participants for the three case studies can be found in Appendices B-D.

Following this, each member of the technical panel and each subject matter expert reviews all of these resources individually and independently. The technical panel members are then asked to assign scores between 1-5 which describe how well the dataset possesses the attributes shown in Table 3. They are also asked to grade the quality of the evidence they received to aid their score assignments. The evidence is graded for quality as it provides decision-makers with confidence when making recommendations based off the evidence.

The evidence was graded according to The Grading of Recommendation Assessment, Development and Evaluation (GRADE) approach [74]. In this approach, the quality of evidence is divided into four classifications: 1. Very low, 2. Low, 3. Moderate and 4. High [75]. The GRADE approach was used as it provides a level of confidence in the evidence used by the technical panel to aid attribute scoring. The GRADE method has been employed by organisations such as the World Health Organization (WHO) to provide a systematic approach for making clinical practice recommendations [76]. This method was chosen to assess evidence quality as it has been endorsed by over 100 organisations worldwide.

In the next stage of the methodology, the technical panel meet in a group. The aim of this meeting is for the technical panel to discuss the individual scores they have assigned to the dataset, and reach a consensus through group discussion. Once consensus has been reached, the consensus scores are sent to the subject matter experts for review. The subject matter experts will then meet to discuss the technical panel consensus scores and determine if they are suitable. If the consensus scores are found to be unsuitable, the subject matter experts will pass their concerns and justifications to the technical panel members. A second group meeting will then take place between members of the technical panel.

Once the subject matter experts have approved the technical panel consensus scores, they are sent to the problem-owner. The problem-owner will then compare the scores to their own quality specification to determine if the data is fit for purpose. The quality specification consists of the scores required of the dataset, with respect to its intended use. The resulting quality specification was used to calculate the preliminary quality factor equation described in Chapter 5.1. However, this was later replaced with an updated quality factor described in Chapter 5.2, which does not rely upon a problem-owner prescribed quality specification.

The methodology described above was implemented in the first case study which served as a development tool. Following the results from this case study, several improvements were made to the methodology. These changes are described in the next section.

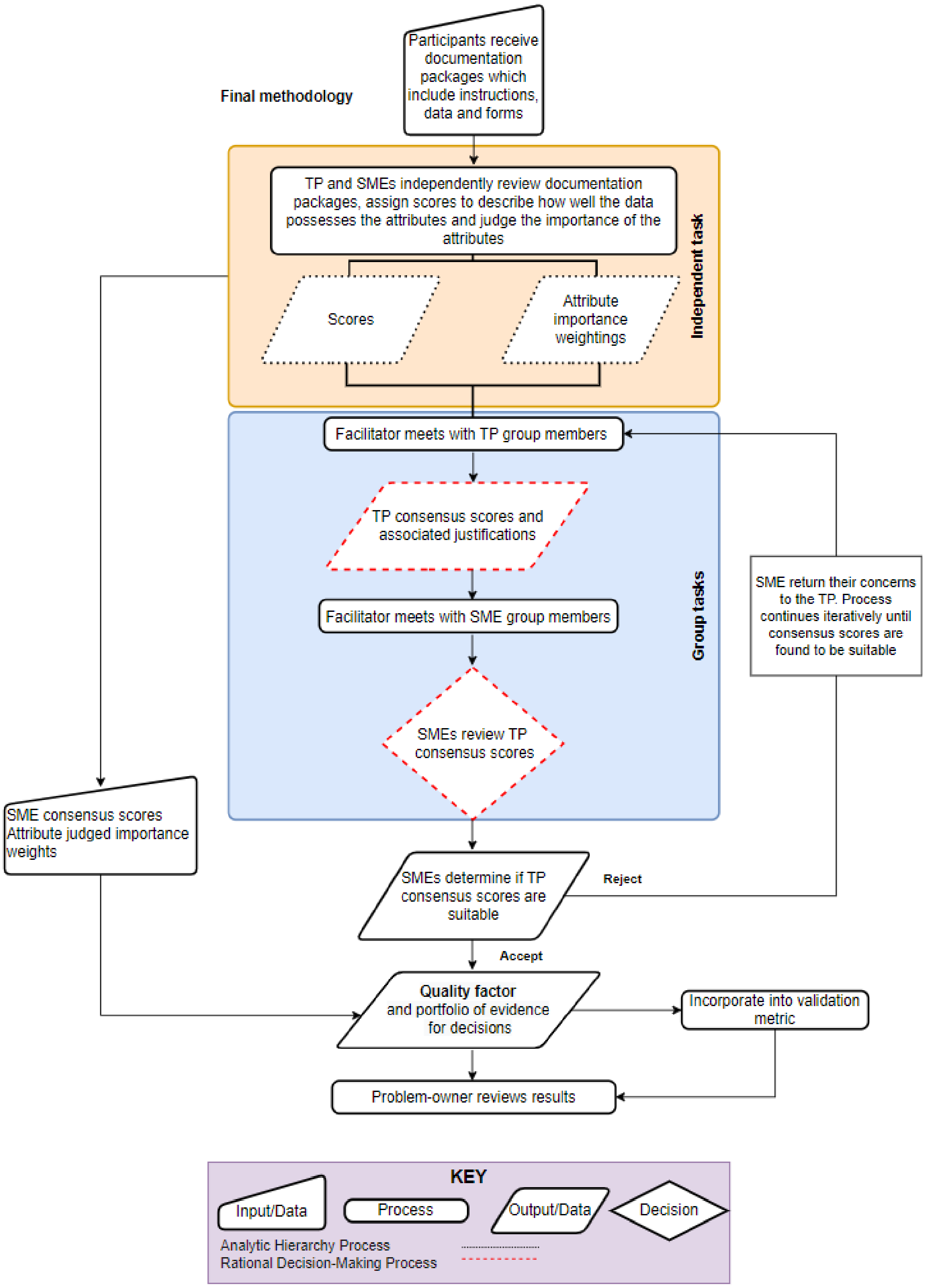
4.2 Final data quality methodology

Following the results from the first case study, four key changes were implemented to the methodology:

1. The technical panel members and subject matter experts were asked to judge the importance of the attributes, relative to the case study.
2. The subject matter experts were asked to assign scores to describe how well the dataset possesses the attributes.
3. A facilitator was introduced to guide and mediate group discussion.
4. The grading of evidence stage was removed from the methodology.

The final methodology is shown in Figure 12. It was reported in the first case study that the group discussion between the technical members was convoluted and unstructured. Therefore, a facilitator was introduced into the final methodology. The primary role of the facilitator is to manage group discussion and mitigate negative group dynamics to ensure that the goal of the discussion is achieved [77]. Pierce et al have published six competencies that facilitators should be able to demonstrate, these are: engaging in professional growth, creating collaborative partnerships, creating an environment of participation, utilizing multisensory approaches, demonstrating integrity and guiding the group to consensus and desired outcomes [78].

Referring to Figure 12, as before, the participants receive documentation packages which they review independently. In the final version of the methodology, the technical panel members and the subject matter experts are asked to assign scores to describe the possession of the attributes. They have also all been asked to indicate the importance of the attributes.



Final methodology
Participants receive documentation packages which include instructions, data and forms
TP and SMEs independently review documentation packages, assign scores to describe how well the data possesses the attributes and judge the importance of the attributes
Scores
Attribute importance weightings
Independent task
Facilitator meets with TP group members
TP consensus scores and associated justifications
Facilitator meets with SME group members
SMEs review TP consensus scores
Group tasks
SME return their concerns to the TP. Process continues iteratively until consensus scores are found to be suitable
SME consensus scores Attribute judged importance weights
SMEs determine if TP consensus scores are suitable
Reject
Accept
Quality factor and portfolio of evidence for decisions
Incorporate into validation metric
Problem-owner reviews results
KEY
Input/Data
Process
Output/Data
Decision
Analytic Hierarchy Process
Rational Decision-Making Process

Figure 12: Flowchart for final methodology. TP denotes technical panel and SME denotes subject matter expert. The black dashed line relates to steps which incorporate the analytical hierarchy process and the red dashed line corresponds to steps which incorporate the rational decision-making process.

The participants indicate the importance of each attribute on a scale from one to five: 1. Not important, 2. Preferred, 3. Important, 4. Highly desirable and 5. Essential. The technical panel members and subject matter experts are asked to judge the importance of the attributes relative to the problem of interest.

Following the results from the first case study, the technical panel members were no longer asked to grade the quality of evidence. This stage was removed from the methodology as it was observed that the scores assigned by the technical panel, to describe the possession of the attributes, were not affected by very low grades of evidence. Furthermore, the resulting grades of evidence were not incorporated into the final output of the methodology. Therefore, a justification for including this step could not be made.

In the next stage of the methodology, the technical panel meet with the facilitator. The aim of this meeting is for the technical panel to discuss the individual scores they have assigned to the dataset, and reach a consensus through group discussion. The facilitator is responsible for recording key decisions, guiding discussion and ensuring that the environment is inclusive for all members. To achieve a consensus, the Nominal Group Technique [79] is followed as it allows participants to assess the data independently before reaching a consensus in a group environment, which might require several iterations of independent assessment and group discussion until a consensus is reached. A second group meeting is then held between the facilitator and the subject matter experts. In this meeting, the facilitator presents the consensus scores assigned by the technical panel for the subject matter experts to discuss and review. They may choose to leave the scores unchanged, or they will modify the consensus scores following their discussion.

The final scores approved by the subject matter experts which describe the possession of the attributes, and the judged importance attribute weights, are used to calculate a quantitative measure of the quality of the data – the quality factor, QF. The definition of the quality factor and its interpretation is discussed in Chapter 5.

The quality factor is calculated by the problem-owner and is provided to the decision-maker, along with a portfolio of evidence for the decisions made during the group discussions. The portfolio of evidence contains the completed forms that were sent to participants. In these forms, the participants are asked to justify the decisions they have made and the facilitator is asked to comment on the interactions within the group.

The value of the quality factor required by the decision-maker will be dependent upon the intended use of the data and its associated socioeconomic impact. In the pilot methodology, the quality factor is calculated using a quality specification which is provided by the problem-owner. This stage of the methodology was not included in the final methodology as it is difficult for a problem-owner to determine what scores are needed by a dataset in this circumstance.

The flowchart shown in Figure 12 also includes a key which highlights where the ideas from the analytical hierarchy process and rational decision-making process have been incorporated. The scoring and weighting of attributes relates to the analytical hierarchy process and the evidenced-based process for making decisions and reaching a consensus is derived from the rational decision-making process. Both processes compare alternatives using a set of criteria – referred to as attributes in this methodology.

The next Chapter outlines the method for calculating the quality factor and guidance to decision-makers on its interpretation. An example of how the quality factor can be incorporated into a validation metric is also included within that Chapter.

Chapter 5: The quality factor

The final output from the data quality methodology is a measure of quality QF, termed the quality factor. Chapter 5.1 presents the preliminary equation used for calculating the quality factor based upon a required quality specification. This equation has been updated in Chapter 5.2 which presents the final equation for calculating the quality factor, using 'perfect' data instead of a quality specification. In, Chapter 5.3 a matrix for assessing the quality factor against the socioeconomic consequence of the problem of interest is described. Finally, Chapter 5.4 outlines how the quality factor has been incorporated into an existing probabilistic validation metric.

5.1 Quality factor – Preliminary equation

Previously, the definition of the quality factor was based upon the idea of a safety factor. In engineering, a safety factor refers to how much stronger a system is than it needs to be for a given intended load. The value of the safety factor is calculated by dividing the maximum stress of the structure by the working or design stress [80]. The higher the value of the safety factor, the safer the structure will be. To ensure the design is safe, the value of the safety factor must be greater than 1. The value required will be dependent upon the consequence of the structure failing – in some circumstances a higher safety factor will be required by design or by law [80].

As mentioned in Chapter 4, in the pilot version of the data quality methodology, the problem-owner was asked to provide a quality specification – i.e. a set of scores to describe how well the dataset possesses the attributes. This quality specification is used to calculate the preliminary quality factor, QF_{pre}, described by equation (11). In equation (11), the quality factor is calculated by summing the product of attribute weightings, $w_{i,}$ and consensus scores,

s_i, for the real data, i.e. the data being assessed by the participants. This is then compared to the threshold data, i.e. the data required according to the quality specification.

$$QF_{pre} = \frac{\sum_i (w_i \times s_i)_{real}}{\sum_i (w_i \times s_i)_{th}} \tag{11}$$

As attribute weightings were introduced after the pilot methodology, assumptions about attribute weightings were made for the first case study. These are discussed in Chapter 6.

Referring to equation (11), it was stated that a quality factor value ≥ 1 indicated that the dataset was suitable for its intended use and a value < 1 indicated that the dataset was not of sufficient quality, and therefore was unsuitable for its intended use.

The advantage of this equation was that it resulted in a value which could be interpreted clearly – i.e. if the value is above 1, the dataset is suitable. This is analogous to the use of safety factors in engineering. However, further bias is introduced when the quality specification is set by the problem-owner. This bias is removed in the updated quality factor equation, equation (13), which compares the data against an ideal dataset.

5.2 Quality factor - Final

The updated and final quality factor value is calculated using two inputs: the consensus scores from the subject matter experts for the possession of each attribute, s_i by the dataset and the attribute importance or weightings, w_i. Each of the technical panel members and subject matter experts are asked to indicate the importance of the attributes using a 1-5 scale where 1 indicates that the attribute is not important and 5 indicates that it is essential. The importance of each attribute w_i is then obtained by calculating the mean. The rating, *R*, of the dataset is defined as the sum of the product of the consensus of the scores from the subject matter experts, s_i, and w_i, the attribute weightings:

$$R = \sum_i (w_i \times s_i) \tag{12}$$

The rating obtained for the real dataset is then normalised by a rating for a 'perfect' dataset. The real dataset refers to the dataset which has been assessed by the technical panel and subject matter experts. For the 'perfect' dataset, the same attribute weightings used for the real dataset are used but the scores are set to a maximum value of five for all attributes. The same attribute weightings are used for the real dataset and the perfect dataset because the importance of the attributes is dependent of the problem of interest. Thus, the final QF is described as:

$$QF = \frac{R_{real\,data}}{R_{perfect\,data}} = \frac{\sum_i (w_i \times s_i)_{real}}{\sum_i (w_i \times s_i)_{perfect}} = \frac{\sum_i (w_i \times s_i)_{real}}{\sum_i (w_i \times 5)_{perfect}} \tag{13}$$

This gives the following range of QF values: $0.2 \leq \mathrm{QF} \leq 1$, where a value of one indicates that the panel of participants have assigned perfect scores to the dataset.

5.3 Determining suitability of data quality

To determine if data with a given quality factor value is suitable for its intended use, or the problem of interest, the range of quality factor values have been divided into four equal and distinct populations which represent: Very poor, Poor, Good and Excellent. To determine the ranges of values within these descriptors, a statistical analysis was performed to simulate 10 000 expected observations of QF for both a random and biased trial. The results from this analysis are shown in Figure 13.

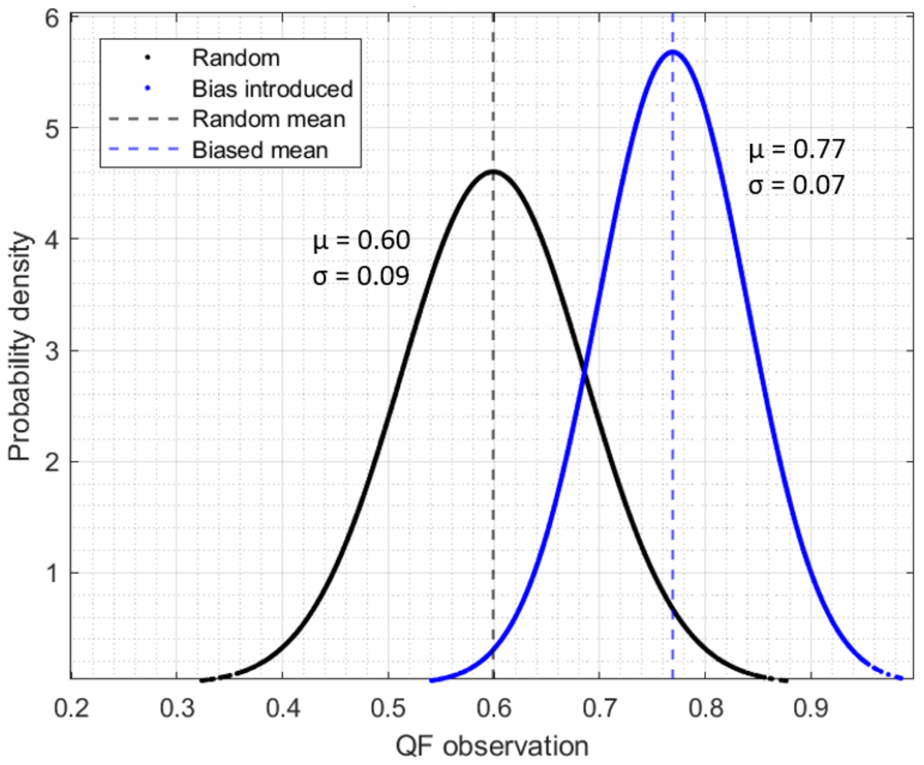


Figure 13: Expected values of the quality factor for a random trial (black curve) and a biased trial (blue curve), with 10 000 iterations. For the biased trial, scoring patterns from the second and third case study were incorporated. The respective mean and standard deviation for both trials are displayed.

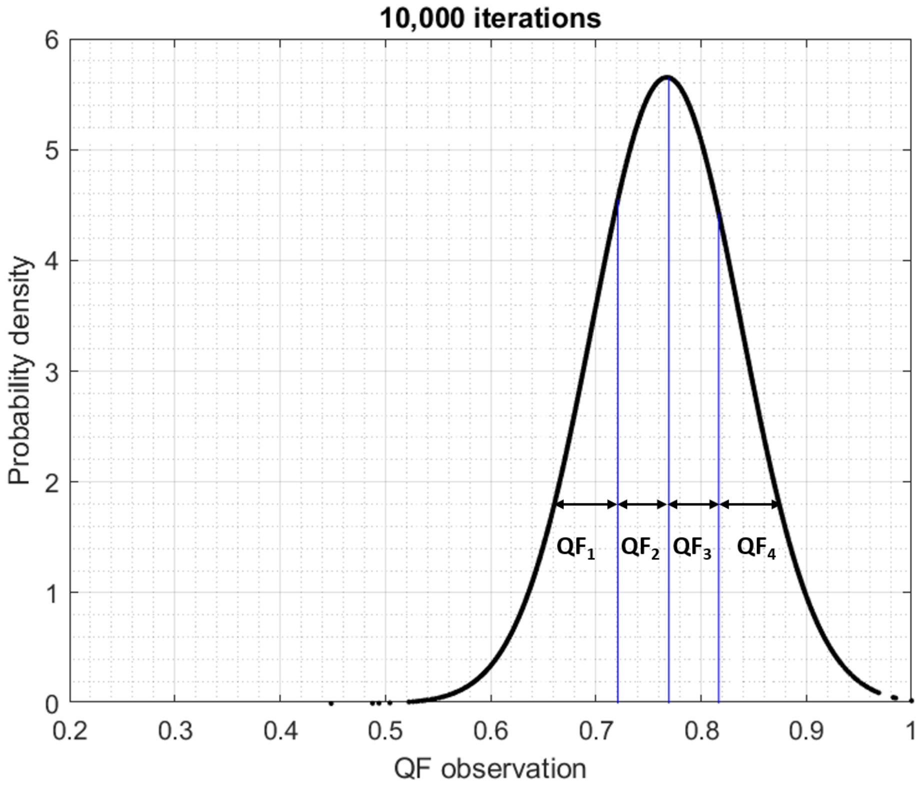


Figure 14: Four quantiles of QF values calculated using the biased normal distribution shown in Figure 13.

For the random trial, a random number generator was used to simulate the subject matter expert consensus scores and the attribute importance weightings. The resulting QF values from the random trial are shown by the black curve in Figure 13. In reality, when individuals are presented with questionnaires which have 5 point-scale outputs, biases are introduced [81]. For example, there is a tendency for individuals to choose the middle answer. To incorporate the effect of bias into the statistical analysis, previous scoring patterns were studied from case study #2 – impact on bonnet liner and case study #3 – thermoacoustic plate. Using the data from these case studies, the probability of the participants assigning each of the 1-5 scores to describe the possession of the attributes was calculated. The calculated probabilities were incorporated into a random number generator with applied bias to produce the range of QF values represented by the blue curve in Figure 13.

For normally distributed data, 68% of the data lies within 1 standard deviation of the mean; 95% of the data lies within 2 standard deviations of the mean and 99.7% of the data lies within 3 standard deviations of the mean [82] – this is referred to as the 68-95-99.7 rule in statistics. Using the random and biased mean and standard deviation values listed in Figure 13, for the random trial, 68% of the values fall within the following $0.51 < QF < 0.69$ range. For the biased trial, 68% of the values fall within the range $0.70 < QF < 0.84$.

To determine the four quality factor ranges, the biased normal distribution shown in Figure 13 was divided into four quantiles, as shown in Figure 14. Four quantiles were chosen as they divide the area underneath the curve into four equal parts. From this figure, it can be seen that very poor quality has been assigned to the range $0.2 \leq QF < 0.72$, poor quality to the range $0.72 \leq QF < 0.77$, good quality to the range $0.77 \leq QF < 0.82$ and excellent quality corresponds to the range $QF \geq 0.82$.

The resulting quality factor ranges were then incorporated into the matrix shown in Table 5. This matrix has been developed using the principles of risk assessment matrices. Risk assessment matrices are used by safety professionals to assess and evaluate the risk of an event occurring. To assess the risk, the severity or consequence of the event is assessed against the probability, i.e., the likelihood of the event occurring. The severity is often ranked on a four-point scale where a value of 4 represents a catastrophic consequence and a value of 1 represents negligible consequences occurring as a result of the event. Likewise, the probability is expressed on a five-point scale where a value of 5 indicates that the event is frequent and a value of 1 indicates that the event is improbable [83].

The resulting risk assessment values are then found by multiplying the values assigned to the severity and the probability. A high value represents a greater probability of harm occurring and a greater severity of that harm, should it occur; while a low value represents the extreme opposite situation.

As shown in Table 5, the socioeconomic consequences have also been divided into four categories: High, Medium, Low and Very Low. There are four recommendations provided to the problem-owner which depend upon the quality of the data and the socioeconomic consequence; these are: sufficient quality, some quality issues, serious quality issues and insufficient quality. For a dataset which has an excellent quality factor and a very low socioeconomic consequence for the problem of interest, it can be inferred from the assessment matrix that the dataset is of sufficient quality. Whereas, if the data is very poor and the socioeconomic consequence of the problem of interest is high, then it can be inferred that the dataset is of insufficient quality.

If the dataset is found to be of sufficient quality, it can be employed by the decision-maker without concerns. If there are some quality issues associated with the dataset, the decision-maker must proceed with caution and consider the likely impact of the issues on the decision-making. For data with serious quality issues, the measurement systems must be reviewed and improved. Finally, for data which is of insufficient quality, the measurement systems must be reviewed substantially and improved, or replaced, before re-acquiring data.

Table 5: A matrix, based on a risk assessment matrix, for interpreting the implications of the quality factor, QF whose range from 0.2 to 1.0 is divided into four periods against the level of socioeconomic consequences associated with the decision which the assessed dataset is intended to inform; the product of the QF descriptor and consequence level defines the recommendations which are provided to the decision-maker.

	Socioeconomic consequence of problem of interest			
Quality Factor descriptor	High - **4**	Medium - **3**	Low - **2**	Very Low - **1**
Very poor – **4** *$0.2 \leq QF < 0.72$*	Insufficient quality (16)	Insufficient quality (12)	Serious quality issues (8)	Some quality issues (4)
Poor – **3** *$0.72 \leq QF < 0.77$*	Insufficient quality (12)	Serious quality issues (9)	Serious quality issues (6)	Some quality issues (3)
Good – **2** *$0.77 \leq QF < 0.82$*	Serious quality issues (8)	Serious quality issues (6)	Some quality issues (4)	Sufficient quality (2)
Excellent – **1** *$QF \geq 0.82$*	Some quality issues (4)	Some quality issues (3)	Sufficient quality (2)	Sufficient quality (1)

Description of recommendation:	
Sufficient quality	Dataset can be employed without concerns
Some quality issues	Some issues with the quality of the dataset, proceed with caution and consider the likely impact of the issues on decision-making
Serious quality issues	Measurement systems and procedures must be reviewed and improved before re-acquiring data.
Insufficient quality	Measurement systems and procedures must be reviewed substantially and improved or replaced before re-acquiring data.

5.4 Incorporation of the quality factor into a validation metric

The quality factor has been incorporated into a validation metric which quantifies the comparison between predictions from a model and measurements from an experiment.

As discussed in Chapter 2.3, Dvurecenska et al [42] developed a validation metric, VM, which calculates the probability that a set of predictions belong to the same population as the measurements. This metric was selected for use in this thesis as, to the best of the author's knowledge, it is the only validation metric capable of handling full-field data.

To incorporate the quality factor into the probabilistic validation metric, a modified metric has been developed by combining the measurement error, u obtained from a calibration of the measurement instrumentation and the reconstruction error due to the decomposition process, and error in the data acquisition represented by (1 – QF) with the probability that the predictions do not belong to the same population as the measurements, i.e., (1 – VM). These three terms are each an estimate of the discrepancy between the measured and predicted data and can be combined as a sum of squares of residuals, which following Dvurecenska et al [42] could be interpreted as probability that the measurements and prediction do not belong to the same population. To ensure that the result is consistent with Dvurecenska et al and therefore more intuitive for users, it is more useful to express the result such that the modified validation metric, VM_{mod} can be interpreted as the probability that the predications and measurements belong to the same population, i.e.,

$$VM_{mod} = 1 - \sqrt{(1 - VM)^2 + \, u^2 + \, (1 - QF)^2} \tag{14}$$

When using this formula, the value of u should be normalised to ensure that the value is dimensionless and consistent with the other terms in the formula.

As mentioned in Chapter 2.3, when the relative errors from the feature vectors are less than the measurement uncertainty, the corresponding VM outcome is 100%. Although this result is quoted alongside a statement which includes the model's intended use and the uncertainty in the measurement data, it can still be misleading for decision-makers. Therefore, an alternative approach, described above, looks at determining the probability that the predictions do not belong to the same population as the measurements. Referring to equation (14), the total measurement uncertainty will never be zero. Therefore, even if the quality of the data is considered to be perfect, $VM_{mod} \neq 100\%$.

The case studies used to demonstrate the new data quality methodology are presented in the next three Chapters. The modified validation metric result shown in equation (14) has been applied to the latter two case studies as these case studies use full-field data.

Chapter 6: Case study #1 – Tensile plate

Case study aim: To serve as a development tool for the methodology

Methodology: Pilot methodology – described in Chapter 4.1

Case study context: To determine if data is suitable for undergraduate laboratory data

6.1 Introduction

For the first data quality case study, data from a loaded tensile plate with a hole was used. The aim of this case study was to test the steps of the methodology and highlight any potential areas for improvement. The case study involved five technical panel members and two subject matter experts. The technical panel was comprised of five PhD students, all of whom were familiar with the dataset and its acquisition. The subject matter experts were a postdoctoral researcher and a lecturer. Table 6 describes the expertise and relevant years of experience of the technical panel members and subject matter experts.

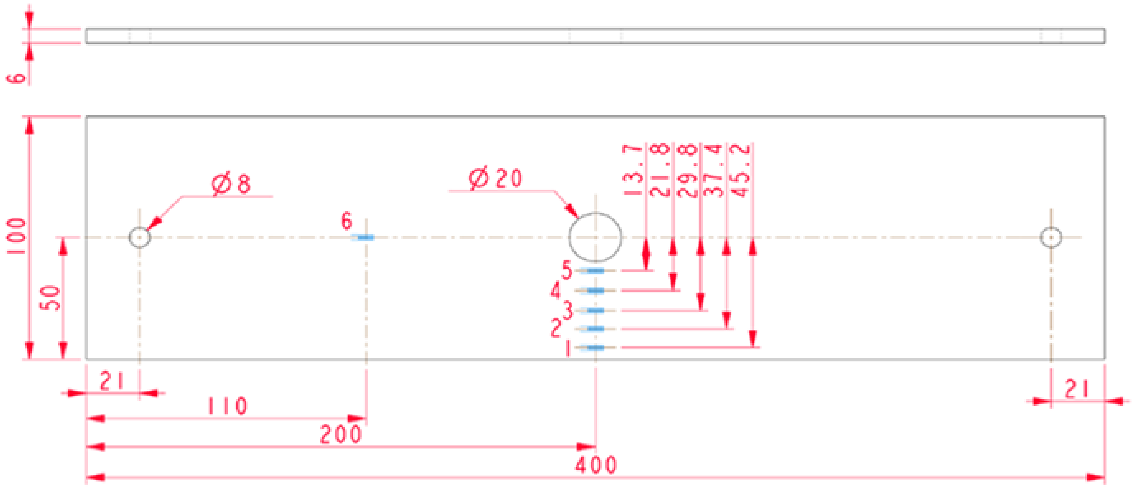
Table 6: Expertise and relevant years of experience for participants in case study #1 – tensile plate with hole. In total, there were five technical panel members and two subject matter experts.

Role	Expertise	Years of experience relevant to case study
Technical Panel	Model validation and credibility	1
Technical Panel	Strain measurement techniques	1
Technical Panel	Strain measurement techniques	1
Technical Panel	Strain measurement techniques	4
Technical Panel	Strain measurement techniques and digital image correlation	5+
Subject Matter Expert	Validation methods and use of full-field measurements	5+
Subject Matter Expert	Digital image correlation	5+

A Hounsfield tensometer was used to apply a uniaxial load to an aluminium specimen with a 20 mm diameter hole at its centre. A force of 8 kN was applied to the specimen and this was measured using a 1000 kgf load cell; where a 1kgf = 9.81 N.

The aluminium specimen had six resistance strain gauges bonded to its surface – five of these strain gauges were located near the hole at the centre and the sixth was located away from the hole, i.e. at a far-field location. Strain gauges measure strain – the deformation of a specimen as a result of an applied stress. A strain gauge is a sensor which measures electrical resistance and this resistance changes due to the strain experienced by the specimen [84].

The dimensions of the plate and locations of the strain gauges are shown in Figure 15.



E = 70.0 GPa

ν = 0.33 [–]

All dimensions are in mm

Figure 15: Diagram of aluminium plate with a hole in the centre used for the first data quality case study. The dimensions of the plate and the locations of the strain gauges are illustrated.

Appendix B contains the documentation which was sent to the case study participants. The documentation package included details of the data quality methodology, the strain data and details of its acquisition and evidence to aid the scores provided by the participants to describe how well the dataset possesses the attributes. In reality, documentation about data often contains gaps and ambiguities; hence, the evidence to support the attributes was made to be deliberately vague. For some attributes, the evidence was missing to investigate the impact of missing information.

For this case study, the pilot methodology described in Chapter 4.1 was implemented.

6.2 Results

The evidence provided to the technical panel for each of the attributes is displayed in Table 7. The individual scores and the consensus scores for the possession of the attributes allocated by the technical panel are shown in Figure 16. The methodology implemented in this case study asked the participants to grade the evidence they received to support the scores they assigned. The grades were split into four classifications: 1. Very low, 2. Low, 3. Moderate and 4. High. Figure 17. compares the consensus scores assigned by the technical panel, to the corresponding grade assigned to the evidence. The justifications for the consensus scores assigned by the technical panel are displayed in Table 8. The subject matter experts for this case study agreed with the consensus scores set by the technical panel, therefore no changes were made.

The raw data for this case study has been summarised in Table E1, in the Appendix section of this thesis.

Table 7: Evidence provided to participants to aid scoring how well the dataset possesses the attributes for case study #1 – tensile plate with a hole.

Attribute	Evidence
1. The right measurements	Judgment required
2. The right tools	The instruments were calibrated fully prior to use.
3. The right people	Formal training has been received
4. Regular review	The instruments were checked twice before use.
5. Demonstrable consistency	The same experiment has been performed in a different laboratory using a different set of strain gauges.
6. The right procedure	The methodology was carried out in accordance with the experiment's written procedure.
7. Instruments	Judgement required.
8. The object to be measured	No significant conditions detected in the lab.
9. Sampling	Measurement technique is representative, and no expected changes were expected.
10. Operator skill	The measuring equipment was set-up by the lab technician.
11. Environmental factors	Judgement required.

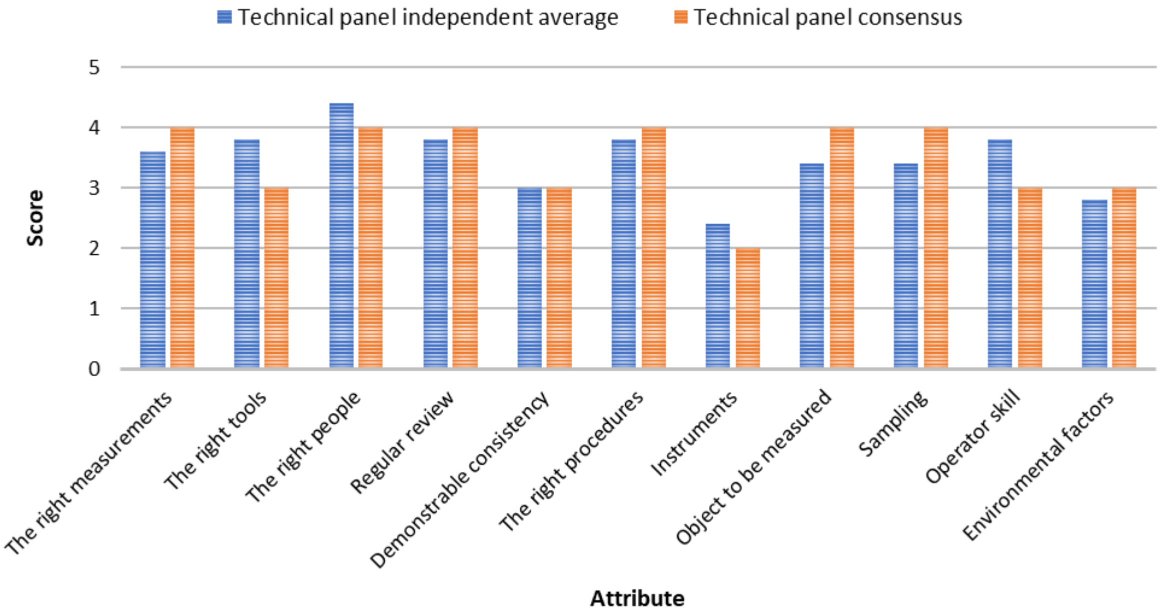


Figure 16: Average of individual scores assigned by the technical panel, to describe how well the dataset possesses the attributes, compared to the consensus scores obtained during the group meeting for case study #1 – tensile plate with a hole.

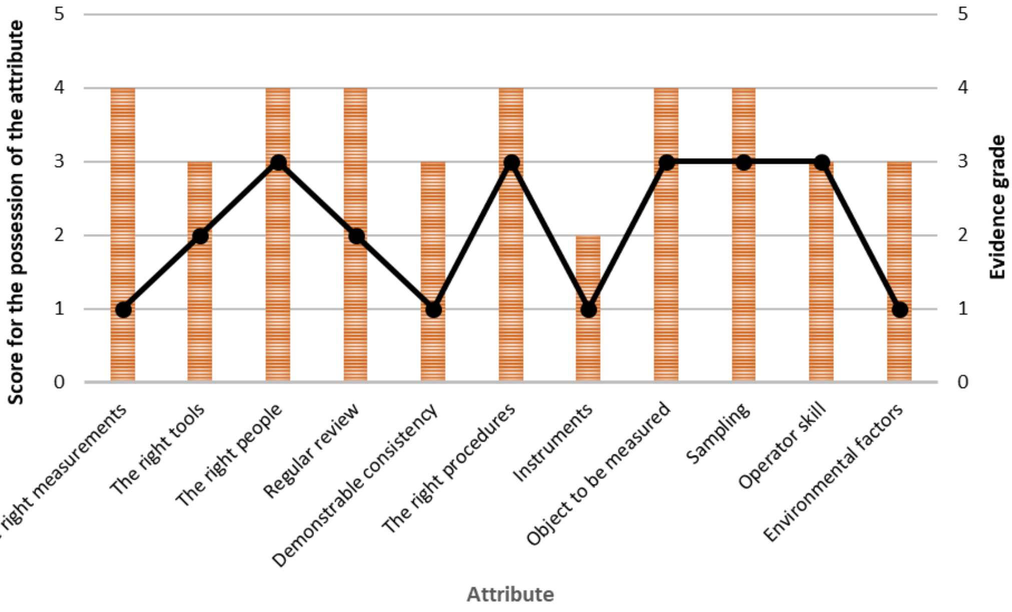


Figure 17: Consensus technical panel scores to describe the possession of the attributes, compared to the corresponding evidence grade assigned by the technical panel. For case study #1 – tensile plate with a hole.

Table 8: Technical panel consensus scores and associated justification for scores assigned to describe the possession of the attributes, for case study #1 – tensile plate with hole.

Attribute	Consensus score (1-5)	Justification stated
1. The right measurements	4	Measurement chosen is standard for analysis.
2. The right tools	3	Evidence provided confirmed the full calibration of the instrumentation.
3. The right people	4	The operator has the theoretical and experimental background required.
4. Regular review	4	Evidence states that the instrument has been checked twice.
5. Demonstrable consistency	3	Experiment has been repeated in a different environment but with a different set of strain gauges.
6. The right procedure	4	Procedure appears to be accurate.
7. Instruments	2	Measurements can be affected by electrical noise and no evidence has been provided to mitigate.
8. The object to be measured	4	Evidence provided confirms the absence of environmental conditions.
9. Sampling	4	No expected changes.
10. Operator skill	3	Operator appropriately skilled.
11. Environmental factors	3	Experiment not likely to be affected by environmental factors.

6.3 Discussion

In this case study, a sparse series of strain gauge data was acquired from a loaded aluminium plate with a hole at its centre.

The individual scores and consensus scores for the possession of the attributes by the dataset assigned by the technical panel are illustrated in Figure 16. It can be seen from this figure that there is close agreement between the initial scores given by the technical panel independently, and the resulting consensus which was achieved through the group discussion. Referring to Table 7, the problem-owner did not provide evidence for three of the attributes: the right measurements, instruments and environmental factors. For these attributes, the technical panel were required to apply their own knowledge and personal judgement when assigning the scores to describe how well the dataset possesses the attributes. Despite missing information, the only score which appears to have been hindered is associated with the attribute seven, the right instrument. For this attribute, three out of five of the technical panel members assigned the dataset a score of 2 with the justification (see Table 8) that insufficient evidence was provided. For the final attribute, environmental factors, a higher average score and consensus score was assigned to describe the possession of the attribute, despite missing information. As shown in Table 8, the justification for the score assigned was that the experiment would not be affected by environmental factors.

In the pilot methodology, which was employed in this case study, the technical panel were asked to grade the evidence they received. The grades were split into four classifications: 1. Very low, 2. Low, 3. Moderate and 4. High. The consensus scores assigned by the technical panel are compared to corresponding evidence grade ratings assigned in Figure 17. Analysing these figures in conjunction with Table 7, which summarises the evidence provided to the technical panel, highlights that for some attributes, the associated consensus scores were not influenced by the grade of evidence assigned. For example, for the first attribute – the right measurements, a consensus score of 4 out of 5 was assigned to describe how well the dataset possessed the attribute despite lack of evidence (and consequently a 1. Very low evidence grade rating) for this specific attribute. Similarly, for the final attribute – environmental factors, a consensus score of 3 out of 5 was assigned alongside a very low evidence grade rating. For this attribute, no evidence was provided to aid the score and judgment was required. Referring to Table 8, which outlines the justifications for the consensus scores, the technical panel have chosen to assign reasonable consensus scores despite the provision of evidence because of their own expertise. For the first attribute, the right measurements, the technical panel are aware that the measurement chosen is standard for the analysis. Likewise, for the final attribute, environmental factors, the technical panel have stated that they are aware that the experiment will not be affected by environmental factors. Thus, the lack of evidence has not impacted their consensus score.

In contrast, for the seventh attribute shown in Figure 17 – instruments, no evidence was provided to the technical panel. The evidence for this attribute was rated 1. very low and the consensus score provided to the attribute was 2 out of 5. In Table 8, which summarises the technical panel consensus scores and associated justifications, it is stated that the instruments attribute was scored low due to the absence of evidence.

The results from Figure 17 indicate there is no clear correlation between the grade of evidence and the final technical panel consensus score. This lack of correlation may be due to a number of reasons. For example, the technical panel members and subject matter experts were familiar with the dataset and its acquisition. Therefore, they may require a lower quality of evidence for their score assignments. Additionally, the aim of the exercise was to determine if the data was of sufficient quality for an undergraduate laboratory. As the socioeconomic consequence of misevaluating the data is low in this case, a lower quality of evidence is likely sufficient. When evidence is missing, the participants put their trust in the reputation of the operator. However, as highlighted by Origgi [85], relying on reputation challenges epistemic responsibility.

This stage of the pilot methodology was not incorporated into the final methodology for two reasons: 1. As discussed above, the quality of the evidence was not having a notable impact on the resulting scores to describe how well the data possessed the attributes, and 2. The quality of the evidence was not used to moderate the final quality factor assigned to the dataset.

A quality factor of 0.69 was obtained for this case study using equation (13). The quality factor calculation requires consensus scores and weightings which indicate the importance of the attributes. In this case study, the technical panel and subject matter experts were not asked to indicate the importance of the attributes as this step was introduced into the methodology following the completion of this case study. Therefore, when calculating the rating of the real data using equation (12), the inputs were the subject matter expert approved technical panel consensus scores shown in Figure 16, and for the attribute weightings, it was assumed that all of the attributes were 'essential', i.e., a weighting of 5.

A quality factor value of 0.69 corresponds to a very poor-quality descriptor as shown in Table 5. Prior to the availability of full-field optical techniques, the quality of this data would likely be considered to be higher. The socioeconomic consequence of the problem of interest can be considered to be low for this case study. The recommendation to the decision-maker would therefore be that there are some quality issues with the data and so decisions should be made with caution.

Following the feedback from this case study, several improvements were made to the methodology including requesting justifications for scores and the introduction of a facilitator to mediate and record discussions. These changes were made as they provide the problem-owner with an insight into why decisions were made. For example, if the data receives a low score for a specific attribute and a justification is provided, it highlights areas of improvement within the dataset for the benefit of the problem-owner. Introducing a facilitator provides structure to the group discussions and mitigates negative dynamics that can arise in these environments.

The aim of the first case study was to test if the methodology could produce a tangible result for a simple case study, which was demonstrated successfully.

Chapter 7: Case study #2 – Bonnet liner impact

Case study aim: To demonstrate the methodology using an industrial case study

Methodology: Final methodology described in Chapter 4.2

Case study context: To determine if dataset is suitable for validation of a FE model

7.1 Introduction

The primary aim of the second data quality case study was to demonstrate that the data quality methodology can be implemented in an industrial case study. The data evaluated in this case study was DIC data from a composite bonnet liner which had been subjected to a high velocity (70 m/s), low energy impact (< 300 J) [10]. The dimensions of the bonnet liner specimen are shown in Figure 18 alongside a photograph of the 50 mm projectile impact prior to impact. A single-stage helium-driven gas gun was used to fire the projectile at the bonnet liner target area.

Using high-speed stereoscopic DIC, displacement maps were generated for the approximately 1 m^2 area of the bonnet liner at 0.2 ms increments for a total of 0.1 s, resulting in 500 datasets. The experimental dataset was collected so that it could be compared to predictions acquired using finite element analysis. The original study was performed as part of the ADVISE project [7] to demonstrate the validation process and this was later incorporated into the CEN guide [5].

The bonnet liner was modelled using the finite element code Ansys- LS-Dyna [10]. Results from this study indicated that the finite element model was valid in the early stages of the impact but not in the later stages of the event. The predicted and measured fields of out-of-place displacements for three different time intervals: 40, 50 and 60 ms after impact, are shown in Figure 19. To allow for an efficient comparison between the predicted and measured data,

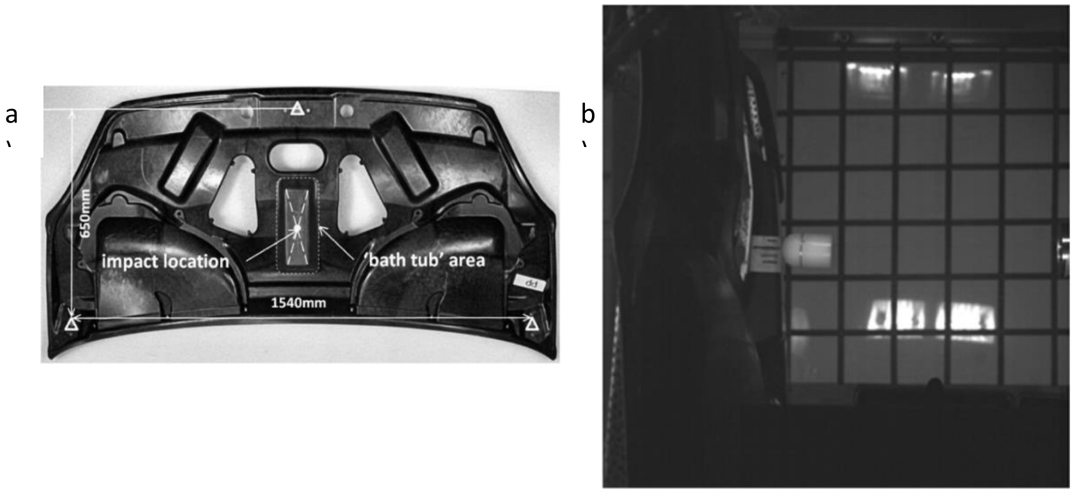


Figure 18: a) Car bonnet liner specimen showing the dimensions and impact location, and b) photograph of the projectile impact *[10]*.

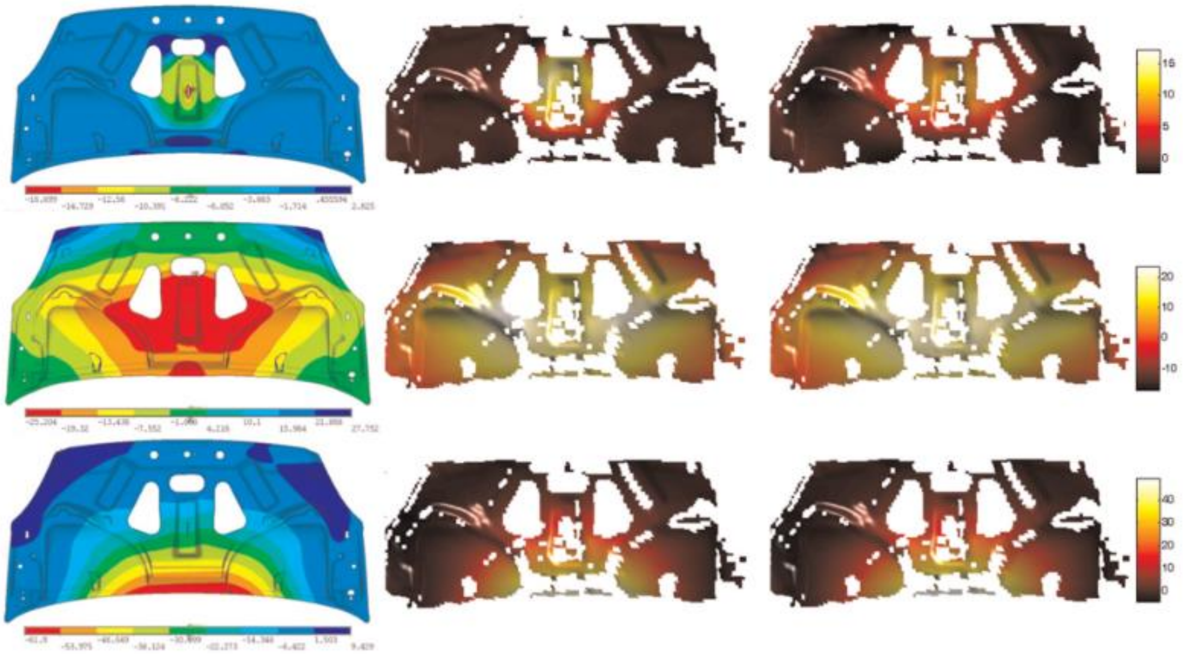


Figure 19: Out-of-plane displacement fields for the car bonnet liner shown at 40, 50 and 60 ms (from top to bottom) after impact. Predictions from the finite element are shown on the left, measurement data from DIC is shown in the centre, and the data on the right shows a reconstruction of the DIC data using 20 AGMDs *[10]*.

both fields were decomposed using adaptive geometric moment descriptors (AGMD). A reconstruction of the DIC data is also shown in Figure 19, using 20 AGMDs [10].

The participants for this case study consisted of one facilitator, four technical panel members and two subject matter experts. To aid scoring, the participants were provided with a journal article [10] which describes the acquisition of the case study data. The expertise of the participants ranged from PhD students to academics and industrialists. The years of relevant experience associated with the facilitator, technical panel members and subject matter experts is summarised in Table 9. The forms sent to the participants can be found in Appendix C of this thesis.

Table 9: Expertise and relevant years of experience for participants in case study #2 – impact on bonnet liner. In total, there were four technical panel members and two subject matter experts. A facilitator was also introduced in this case study to support the technical panel members and subject matter experts.

Role	Expertise	Years of experience relevant to case study
Facilitator	Strain measurement techniques, Digital image correlation, model validation	5+
Technical Panel #1	Digital image correlation and damage characterization	2
Technical Panel #2	Structural assessments of aerospace composites	4
Technical Panel #3	Material science and optical engineering	5+
Technical Panel #4	Digital image correlation, optical engineering	10+
Subject Matter Expert #1	Digital image correlation	8
Subject Matter Expert #2	Digital image correlation, mechanical characterization	10+

7.2 Results

The weights assigned by the technical panel and subject matter experts to indicate the importance of the attributes are shown in Table 10. The importance of the attributes is judged according to five options: 1. Not important, 2. Preferred, 3. Important, 4. Highly desirable and 5. Essential.

Figure 20 shows the independent scores assigned by the technical panel and subject matter experts for the possession of the attributes, prior to reaching a consensus via group discussion. These independent scores are then compared against the respective consensus scores in Figure 21 (also represented as a radar chart in Figure 22). The error bars represent +/- one standard deviation.

Figure 23 illustrates an application of the data quality methodology in a validation process. This figure shows the probability that the predictions from a model of the displacements of the bonnet liner following an impact are from the same population as the measurements acquired using digital image correlation. The black solid line indicates the results from Dvurecenska et al [42] and the dashed line those from the modified validation metric incorporating the quality factor and measurement uncertainty as described in equation (14) in Chapter 5.4.

The raw data for this case study has been summarised in Table E2, in the Appendix section of this thesis.

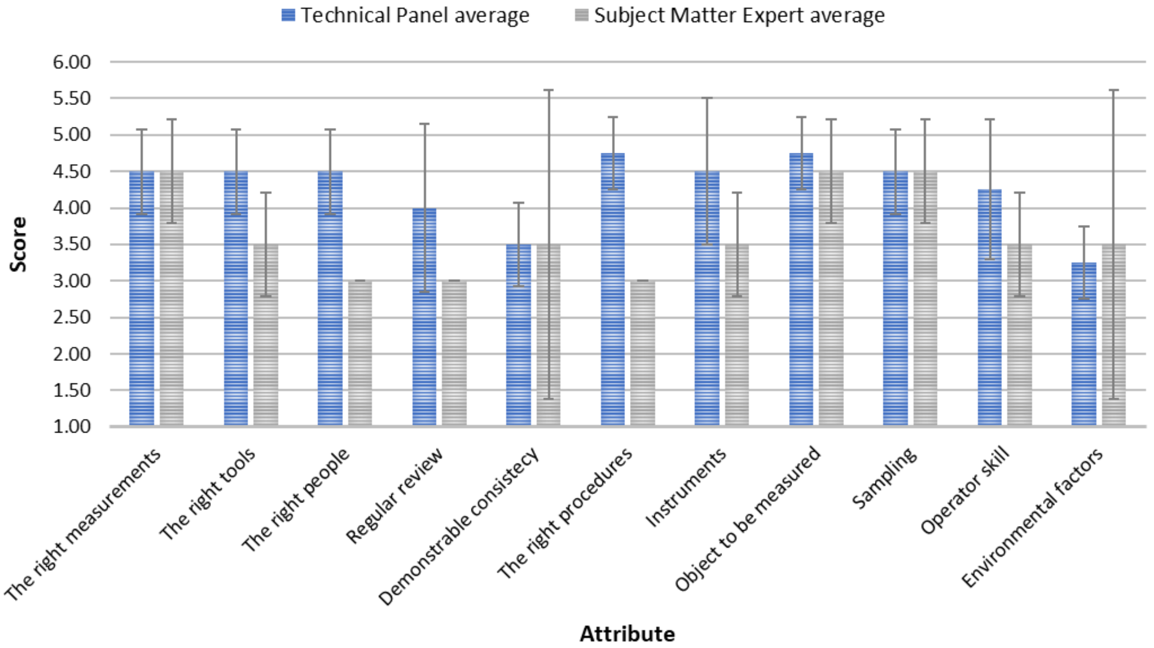


Figure 20: Independent scores assigned by the technical panel and subject matter experts to describe how well the dataset possesses the attributes, prior to reaching a consensus via group discussion for case study #2 – impact on bonnet liner. The error bars represent +/- 1 std.

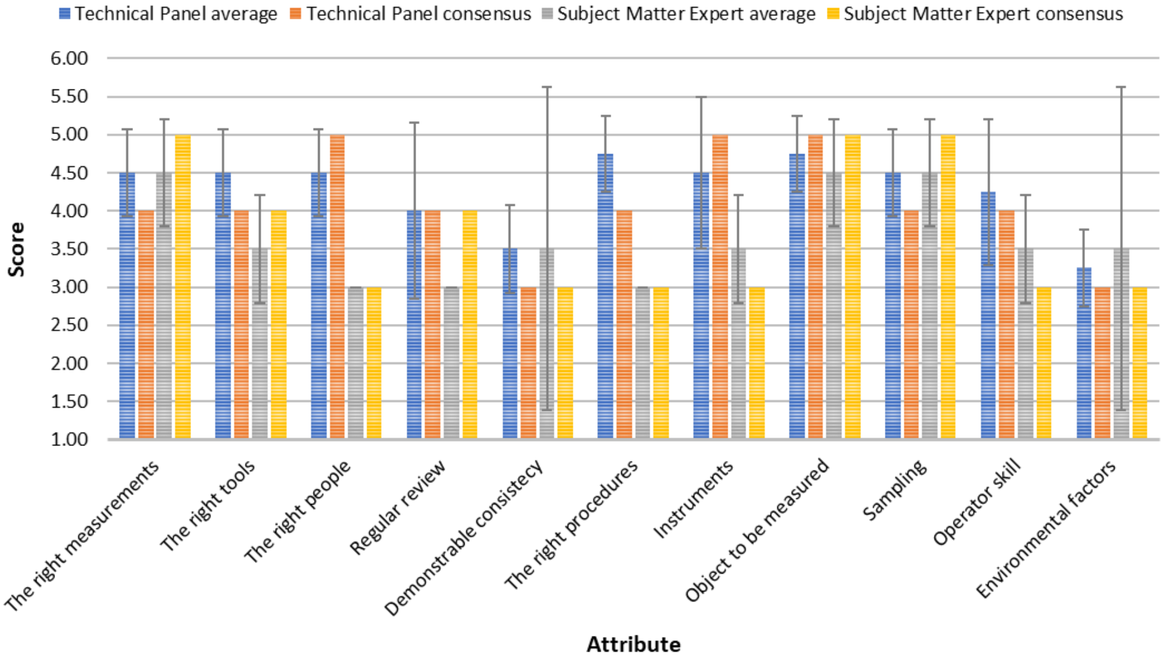


Figure 21: A comparison of scores assigned to the attributes by the technical panel and subject matter experts, to describe how well the dataset possesses the attributes, pre and post group discussion for case study #2 – impact on bonnet liner.

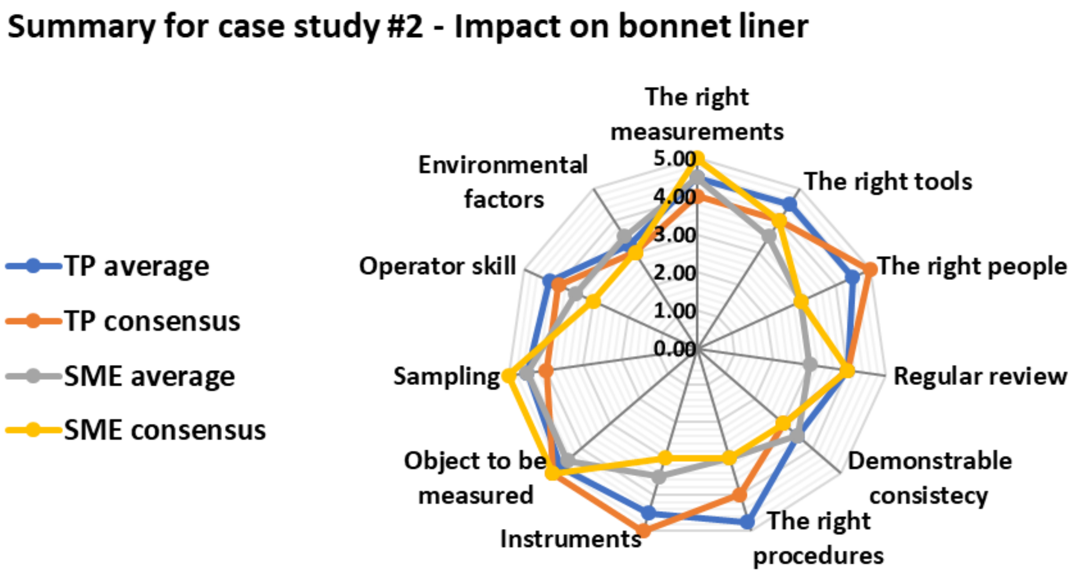


Figure 22: Radar chart representation of Figure 21. The blue line and grey lines represent the technical panel and subject matter expert averages resulting from individual scoring. The orange and yellow lines represent the technical panel and subject matter expert consensus scores for case #2 – impact on bonnet liner.

Table 10: Importance weights for the attributes assigned by the technical panel and subject matter experts for case study #2 – impact on bonnet liner. TP denotes technical panel and SME denotes subject matter expert.

Attribute	TP	TP	TP	TP	SME	SME
1. The right measurements	4	4	5	5	5	5
2. The right tools	5	4	5	5	5	5
3. The right people	5	5	4	4	4	4
4. Regular review	3	4	3	2	4	4
5. Demonstrable consistency	5	1	4	4	3	5
6. The right procedures	4	5	5	4	3	3
7. Instruments	3	4	4	4	5	4
8. Object to be measured	4	5	5	3	2	3
9. Sampling	4	5	5	2	2	5
10. Operator skill	4	5	4	4	4	3
11. Environmental factors	3	3	4	3	1	2

1. Not important	*2. Preferred*	*3. Important*	*4. Highly desirable*	*5. Essential*

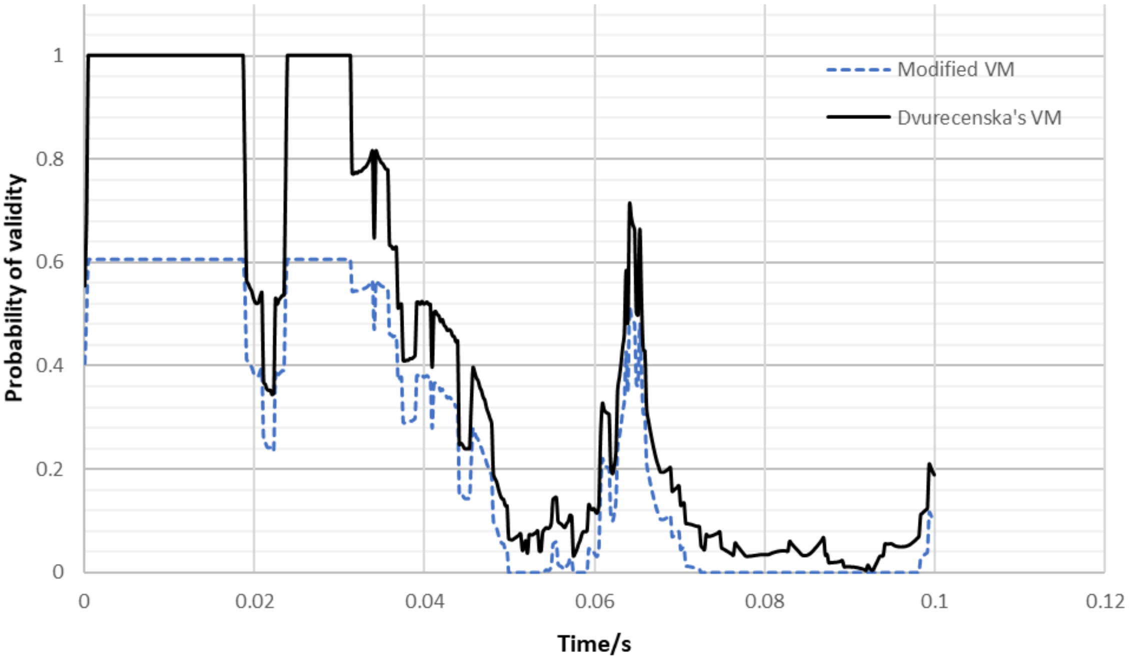


Figure 23: Validation metric as a function of time (measurement data was acquired for 0.1s) following impact on car bonnet liner based on predicted and measured data from Burguete et al *[10]*; values for both the unmodified validation metric, VM from Dvurecenska et al *[42]* (black solid line) and the modified validation metric, VM_{mod} defined in equation (14) – denoted by the blue dashed line.

Table 11: Technical panel consensus scores and associated justification for case study #2 – impact on bonnet liner.

Attribute	Consensus score (1-5)	Justification stated
1. The right measurements	4	DIC system and displacement measurements are appropriate, based on the experience of the panel. Measurement uncertainty pre-statement specific to the impact is missing from the paper.
2. The right tools	4	Missing some details about the calibration, however a reference to a paper with the details of the calibration procedure is included.
3. The right people	5	The members of the panel have a good knowledge of the people who performed the experiment. Based on personal experience, believe that people involved received a good training.
4. Regular review	4	In the paper it is stated that equipment was recalibrated between the runs. From personal experience the panel are satisfied that there is no need for a regular specialist review.
5. Demonstrable consistency	3	Not enough repeats of experiments reported in the paper; only a single dataset is reported, but it covers a large area.
6. The right procedure	4	The information provided in the paper leads the panel to believe that a good practice was implemented, however a specific procedure or methodology was not referenced.
7. Instruments	5	All members agree that appropriate precautions have been considered and implemented.
8. The object to be measured	5	With respect to the specimen, various relevant aspects were considered, e.g. fixings, frame. Speckle pattern is a slight concern, but is agreed to be outside the scope for this attribute.
9. Sampling	4	Graphs/figures are well resolved, however only one timeseries were discussed in the paper.
10. Operator skill	4	The operators involved in this work have considered various factors with respect to the set up and sample preparation. No other evidence is available.
11. Environmental factors	3	Members believe that details are missing to satisfy this attribute, but some confidence is given that the measurements were taken in a controlled laboratory environment. Based on personal experience of following good practice, think that the environmental factors would have been considered.

7.3 Discussion

In this case study, digital image correlation data of a car bonnet liner subject to a high velocity, low energy projectile impact was assessed.

The results from the second case study provide more insight into the impact of group discussions. Figure 20 displays the average scores assigned by the technical panel and subject matter experts prior to achieving a consensus via group discussions. Figure 21 compares these individual assignments with the consensus scores reached by the technical panel and the subject matter experts.

Referring to Figure 21, for seven out of eleven attributes, the consensus of the technical panel (orange bar) is lower than the technical panel average of the individual assignments (blue bar). The consensus scores are assigned in a group environment which contains more collective knowledge and expertise. A group of members will be able to find more weaknesses and more areas of criticism for a particular attribute, thus leading to a lower consensus score when compared to individual assignments.

For one attribute – regular review, the consensus score matches the average score assigned prior to the group meeting. This attribute also had the highest associated standard deviation – with two technical panel members (TP #1 and TP #2 in Table 9) assigning a score of 3 out of 5, and two technical panel members (TP #3 and TP #4 in Table 9) assigning a score of 4 out of 5. These scores can be found in table E2 in the appendix. Referring to Table 9, which outlines the expertise of the technical panel members, it can be seen that the members who scored the lower scores also have fewer years of experience relevant to the case study. Whereas, the higher initial scores were assigned by members with more experience. The consensus score of 4 out of 5, was stated by the facilitator to be a compromise between the two scores.

It was reported by the facilitator that there was no obvious negative group dynamic, i.e. the discussion was not dominated by any one member and each participant had the opportunity to share their opinion.

It is clear from Figure 21 that the subject matter experts have been influenced by the scores assigned by the technical panel. For example, for the fourth attribute, regular review, the subject matter experts both gave the dataset an initial score of 3 out of 5. During their meeting, they then increased this score to 4 out of 5 which matches the consensus reached by the technical panel. However, for the third attribute, the right people, they have remained consistent with their individual assignments and have not been obviously swayed by the consensus reached by the technical panel. The ninth attribute, sampling, also led to interesting results; the technical panel and subject matter experts agreed on individual scores but following the group meetings, the technical panel decided to reduced their scores whereas the consensus of the subject matter experts was to increase their score.

For nine out of eleven attributes, the subject matter expert consensus either matched the technical panel consensus, or was lower. It was reported by the facilitator that the subject matter experts felt that there was not sufficient information and evidence in the journal paper to support their scoring. It was also noted that some attributes require a subjective judgement to be made, i.e. based on personal experience. Due to insufficient evidence, it is likely that the subject matter experts decided to proceed with caution when assigning scores to describe the possession of the attributes by the dataset. Whereas, the technical panel members did not express difficulty to the facilitator regarding scoring the dataset using the evidence provided in the journal paper. Furthermore, referring to Table 11, which states the justifications provided for the technical panel consensus scores, it can be seen that for attribute three – the right people – the technical panel members have assigned a maximum consensus score of 5 out of 5 because of the reputational credibility of those who acquired the data. This is supported by the justification provided: ‘The members of the panel have a good knowledge of the people who performed the experiment’. However, for the same attribute, the subject matter expert consensus is 3 out of 5, which is in line with the average of their individual assignments.

A quality factor of 0.75 was obtained using the SME consensus scores (yellow bar in Figure 21) and the mean of the attribute importance scores (see Table 10). According to Table 5, this corresponds to a poor-quality descriptor. Although rich full-field data was used in the case study, the dataset did contain regions of missing data. Furthermore, the data was acquired in an industrial lab by an international group over a period of 3-4 days. For these reasons, the resulting quality factor appears to be reasonable.

The facilitator reported that some attributes were misinterpreted and suggested that further clarification would be useful. This could reduce ambiguity and would likely lead to greater consistency across the scores. Following this feedback, more guidance regarding how to assign scores was provided to participants for future implementation.

Figure 23 illustrates an application of the data quality methodology in a validation process. This figure shows the probability that the predictions from a model of the displacements of a bonnet liner following an impact are from the same population as measurements acquired using digital image correlation from the second case study. The black solid line indicates the results from Dvurecenska et al [42] and the dashed line those from the modified validation metric incorporating the quality factor and measurement uncertainty as described in equation (14). If the predictions from the model are unreliable and yield a low initial result from the unmodified validation metric, e.g., $t > 0.07$ s, then the modified VM equation only has a small effect; however, when the result from the unmodified validation metric is close to unity then the modified version has a greater impact on the value. The modified validation metric cannot have a value of unity, even when there is statistical congruence of the predictions and measurements because its value is moderated by the quality factor, QF and measurement uncertainty, u which will never be zero-valued. Thus, the apparent certainty (VM = 1) that the predictions and measurements belong to the same population, immediately following the impact in Figure 23, is removed when using the modified validation metric, VM_{mod} defined in equation (14).

This represents an improvement on the work of Dvurecenska et al [42] in which it was possible for the validation metric to have a value of unity implying that it is certain the predictions belong to the same population as the measurements given the measurement uncertainty; while this is technically correct, it is potentially misleading for decision-makers who do not appreciate the implications of the measurement uncertainty. The statement provided to decision-makers includes the outcome of the validation metric, the measurement uncertainty and the intended use of the model. However, decision-makers can overlook the measurement uncertainty as it can be difficult for non-experts to interpret; whereas, the probability that results from the value of the validation metric is accessible to non-expert target audiences. A further significant improvement is that the modified validation metric incorporates a value to describe the quality of the measurement data. To the author's knowledge, the literature does not reference any existing validation metrics which incorporate a quality of measurement data.

For the first 0.02 seconds following the impact, the value of the modified validation metric, VM_{mod} is just above 0.6. This is a 0.4 reduction from the unity value produced by Dvurecenska's VM over this time period. After 0.02 seconds, the value of the modified validation metric reduces substantially into the range 0.2 to 0.4, which is unlikely to be considered adequate for supporting any decisions with socioeconomic consequences. This is consistent with the results of Dvurecenska et al [42] and Burguete et al [10] who highlighted that a crack initiation on this timescale in the experiment which the model did not have the capability to simulate. Burguete et al highlighted that the model is valid up to 0.016s after impact but not in the later stages of the event. In addition, Burguete et al explored the percentage difference between the results from simulation and the results from the experiment using the Euclidean distance between the feature vectors. This work shows that the normalised differences are smaller than 2.5% in the first half of the event, but oscillate up to 10% in the second half of the impact event. It was suggested that introducing damping into the FE model could help reduce these errors.

Chapter 8: Case study #3 – Thermoacoustic plate

Case study aim: To investigate the quality of PhD data

Methodology: Final methodology described in Chapter 4.2

Case study context: To determine if dataset is suitable for validation of a FE model

8.1 Introduction

For the final data quality case study, full-field out-of-plane displacement data acquired from a Hastelloy-X (nickel base alloy) plate subject to thermo-acoustic loading was used [11]. The dimensions of the Hastelloy-X plate specimen and the speckle pattern which was applied for DIC is shown in Figure 24 a).

Heating to the plate was provided using Halogen quartz lamps with a power output of 1 kW and a colour temperature of 3210 K. Transverse heating across the plate was achieved using four vertically-oriented lamps and longitudinal heating was achieved using two horizontally-oriented lamps. Transverse heating was applied at the middle of the plate and longitudinal heating was provided on the right edge. The transverse and longitudinal heating regimes are illustrated in Figure 24 b). For reference, a uniform room temperature distribution (25 degrees Celsius) was also studied.

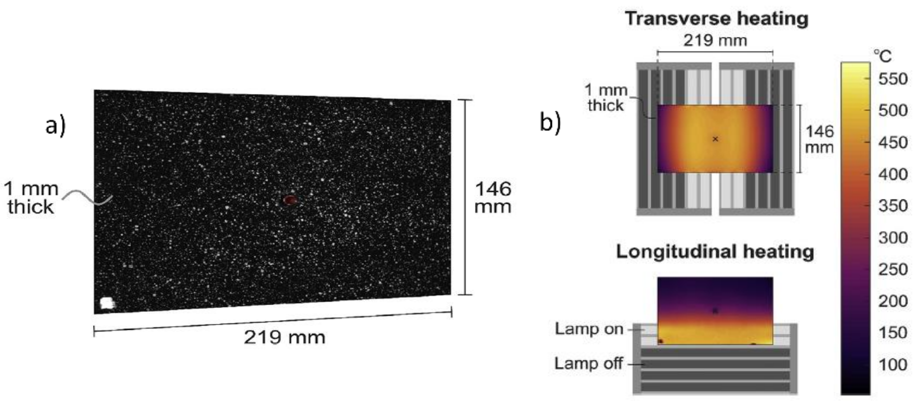


Figure 24: a) Hastelloy-X plate specimen dimensions and speckle pattern used for DIC and b) Configuration of lamps for transverse and longitudinal heating. Figure taken from ref *[11]*.

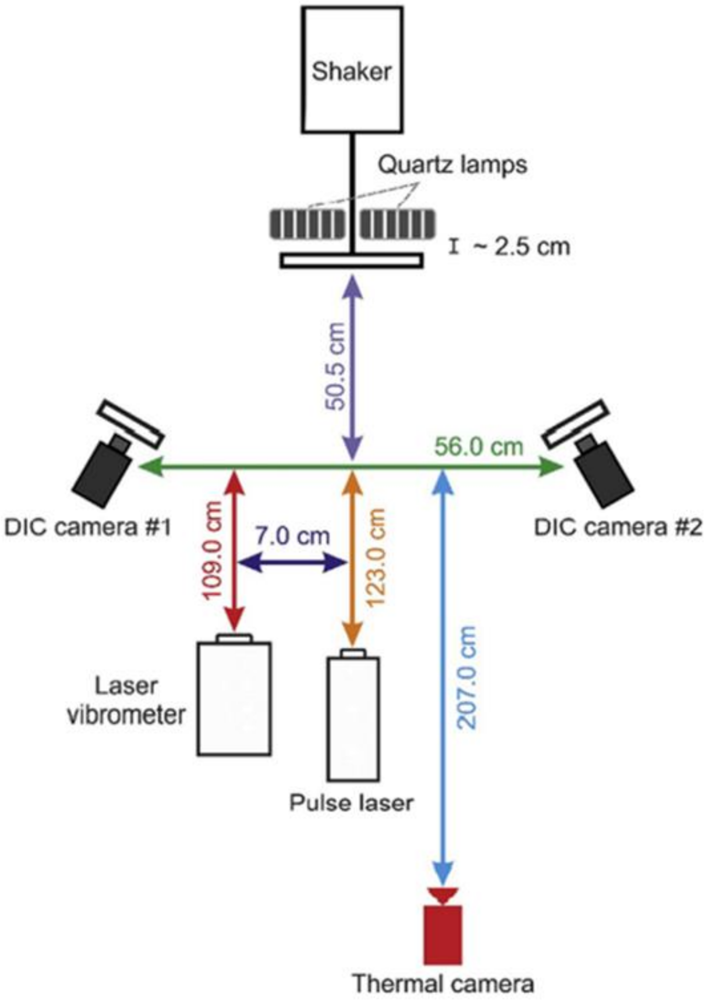


Figure 25: Experimental set-up for Hastelloy-X plate subject to thermal and mechanical loading. The thermal loading is provided by the quartz lamps and the mechanical loading is provided by a shaker located behind the specimen *[11]*.

Mechanical loading was provided using a commercially available V100 DataPhysics shaker and a 1 kW power amplifier system. Two types of loading were applied: broadband and single-frequency sinusoidal loading. With broadband loading, a signal between 0 and 800 Hz is used to determine the plate's resonant frequencies. Once these had been determined, a function generator which controls the V100 DataPhysics shaker was used to create a sine waveform which excited the plate at the measured resonant frequencies. The resonant frequencies for first 11 modes were determined for each heating regime with six independent repeats.

The experimental setup is presented in Figure 25. Stereoscopic DIC (Q-400 system, Dantec Dynamics) was used to acquire out-of-plane displacement data using two 1624 x 1234-pixel CCD cameras. Image capture was performed using Istra 4D software and image correlation was performed with facets of 25 x 25 pixels with a centre-to-centre spacing of 21 pixels. Temperature maps were recorded using a thermal camera. A pulse laser (4 ns pulse of green light, 532 nm) provided a stroboscopic illumination of the plate. The two DIC cameras were fitted with optical filters with a centre wavelength of 532 nm and 4 nm bandwidth to ensure that any light outside of the wavelength of the laser was blocked.

The measured and predicted out-of-plane displacement data for each of the temperature regimes are shown in Figure 26 for the first eleven resonant frequencies.

The participants for this case study consisted of three technical panel members, two subject matter experts and one facilitator. The expertise and years of relevant experience for the facilitator, technical panel and subject matter experts is summarised in Table 12. The forms sent to the participants can be found in Appendix D of this thesis.

Table 12: Expertise and relevant years of experience for participants in case study #3 – thermoacoustic plate. In total, there were three technical panel members, two subject matter experts and one facilitator.

Role	Expertise	Years of experience relevant to case study
Facilitator	Strain measurement techniques, digital image correlation, thermoelastic stress analysis	5+
Technical Panel	Digital image correlation	5+
Technical Panel	Digital image correlation	5+
Technical Panel	Digital image correlation, experimental mechanics	10+
Subject Matter Expert	Digital image correlation, thermoelastic stress analysis	10+
Subject Matter Expert	Full-field optical stress analysis methods	10+

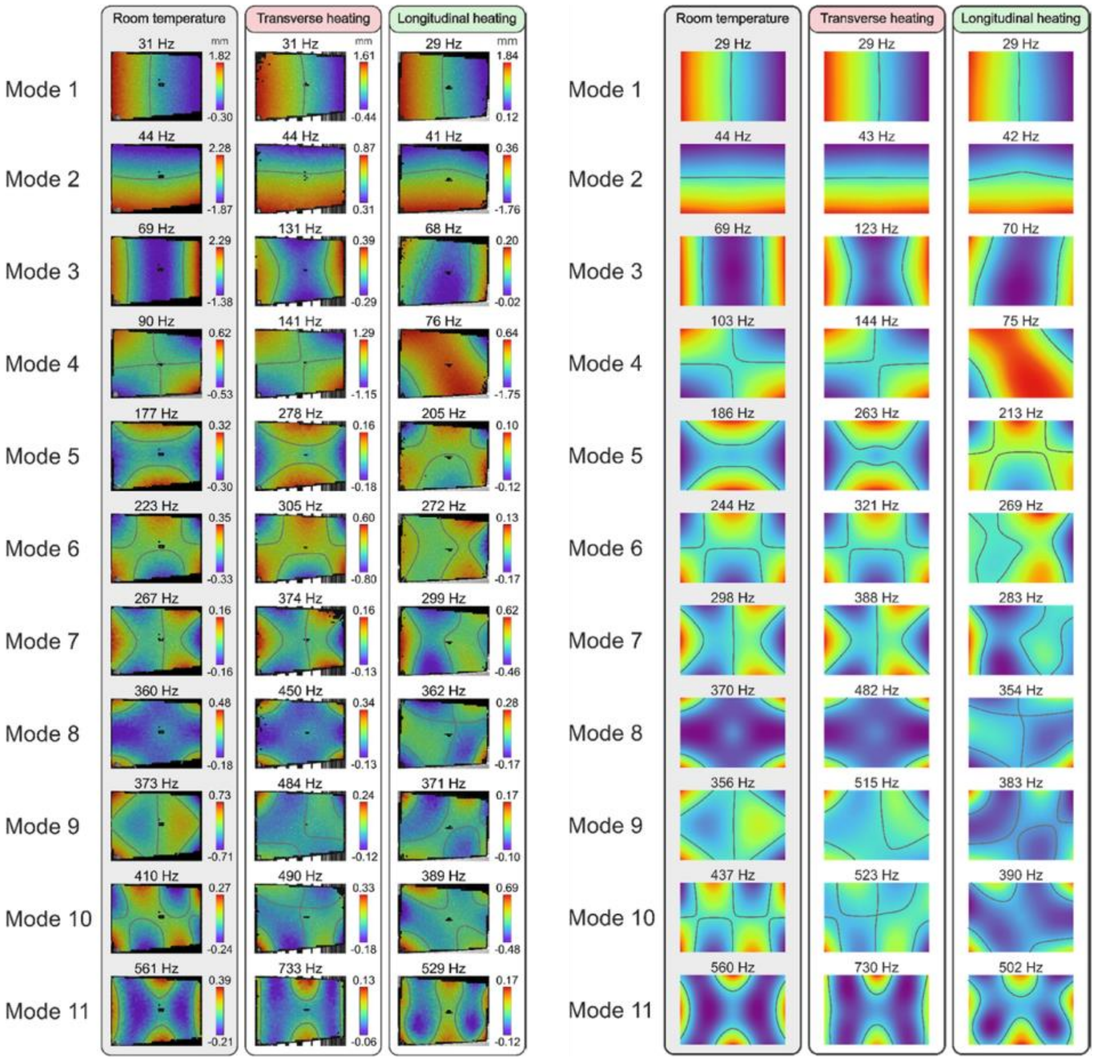


Figure 26: Case study data. Measurement data (left) and prediction data (right) for the first 11 resonant frequencies of a Hastelloy-X plate subject to thermo-acoustic loading. Three temperature regimes are shown: room temperature, transverse heating and longitudinal heating. Figure taken from ref *[11]*.

8.2 Results

The independent scores assigned by the technical panel and subject matter experts for the possession of the attributes, prior to group discussions, are shown in Figure 27. These scores are compared against the respective technical panel consensus scores and subject matter expert scores in Figure 28 (also shown as a radar chart in Figure 29). The error bars represent +/- one standard deviation. The importance weights for the attributes, assigned by the technical panel and subject matter experts, are illustrated in Table 13.

To investigate the probability of the model's validity, the room temperature modal shapes shown in Figure 26 were decomposed using a program called Euclid [86] which decomposes data using Chebyshev polynomials. Fifty coefficients (consistent with Silva et al [11]) were used to describe the measurement and prediction data and the total measurement uncertainties for each modal shape are shown in Table 15. The modal shape images were inputted into Euclid as hdf5 files and missing data, i.e. due to the attachment of the nut, was interpolated using Euclid's nearest neighbour interpolant function. Dvurecenska's validation metric was then employed to the resultant measurement and prediction feature vectors. Using the quality factor from the case study, the modified validation outcomes using equation (14) were calculated. These are listed in Table 14.

The consensus scores allocated by the technical panel and the subject matter experts, along with associated justifications, are listed in Table 15 and Table 16. The raw data for this case study has been summarised in Table E3, in the Appendix section of this thesis.

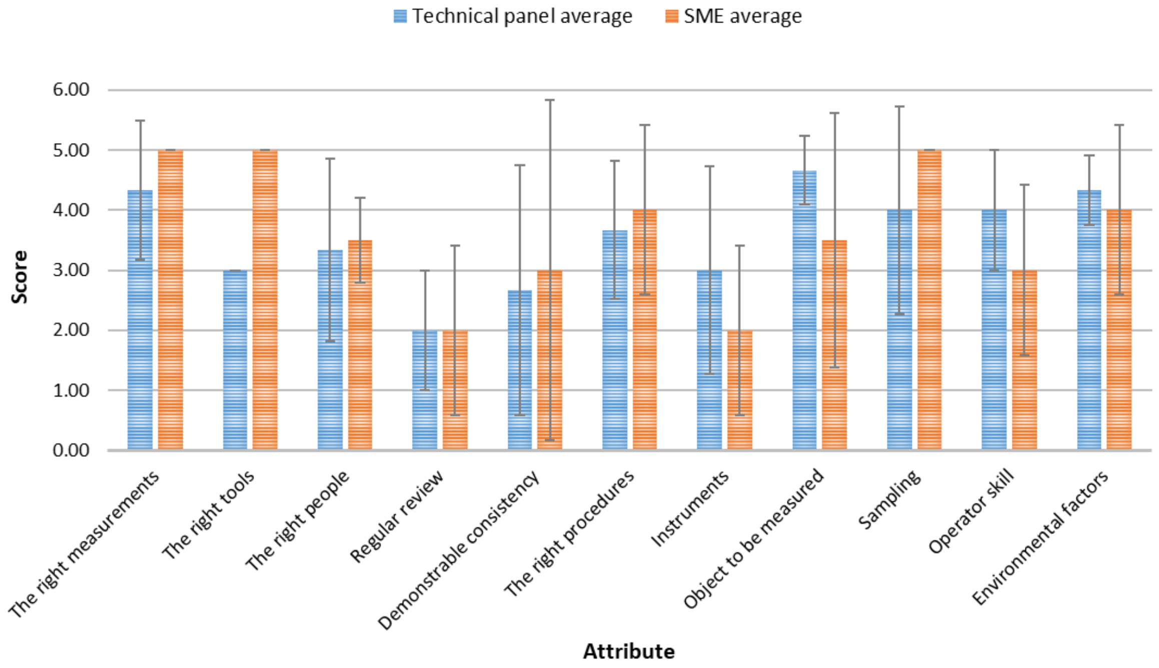


Figure 27: Independent scores assigned by the technical panel and subject matter experts to describe how well the dataset possesses the attributes, prior to reaching a consensus via group discussion for case study #3 – thermoacoustic plate. SME denotes subject matter expert.

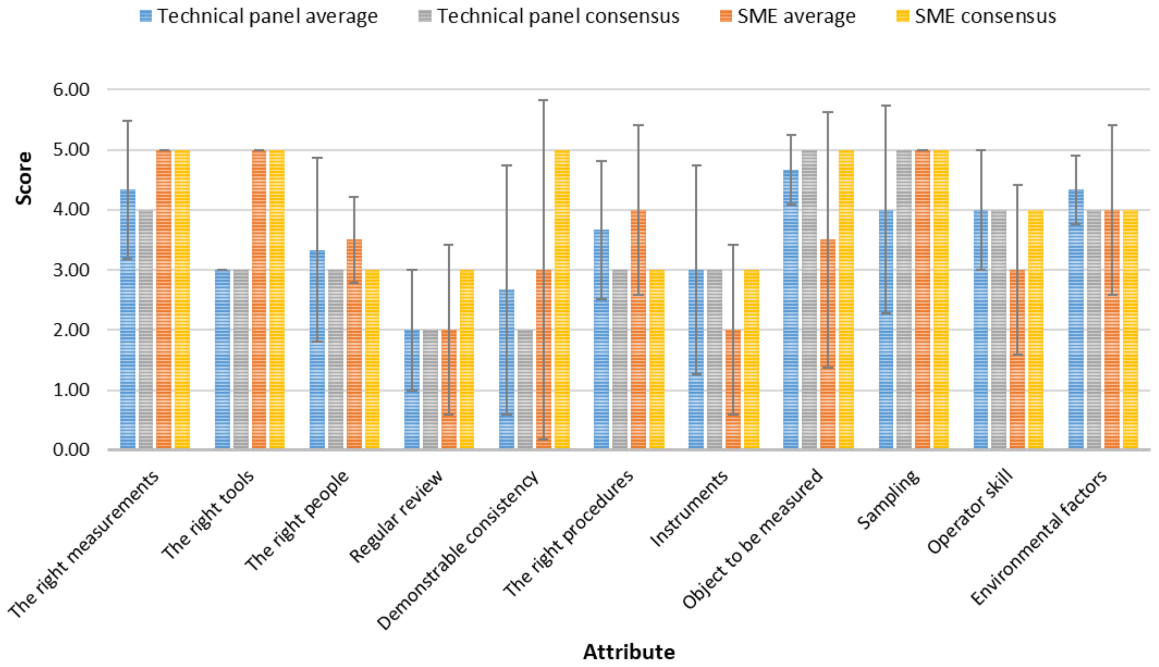


Figure 28: A comparison of scores assigned to the attributes by the technical panel and subject matter experts, to describe how well the dataset possesses the attributes, pre and post group discussion for case study #3 – thermoacoustic plate. SME denotes subject matter expert.

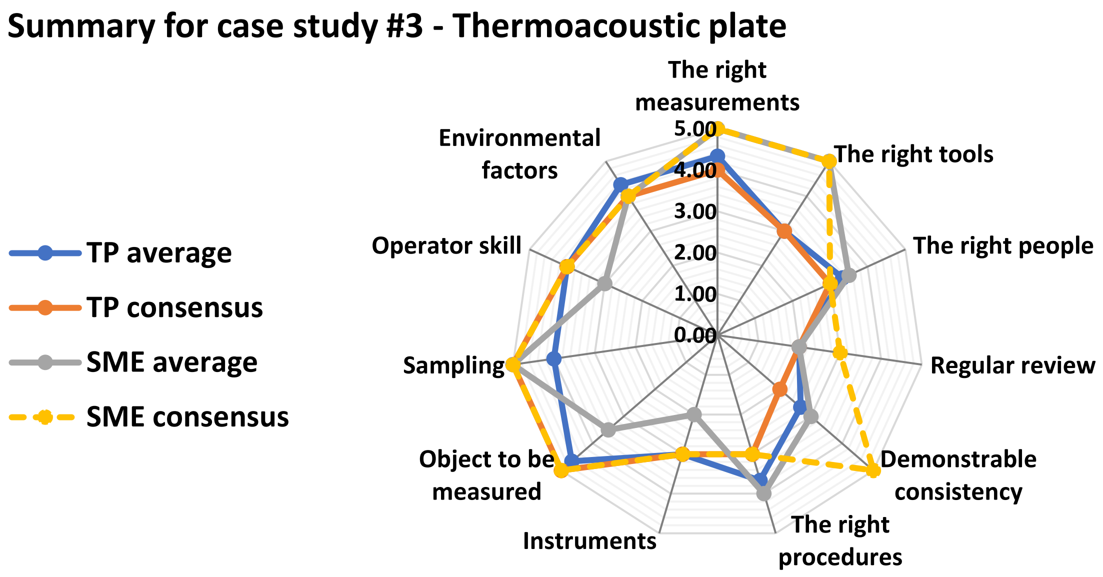


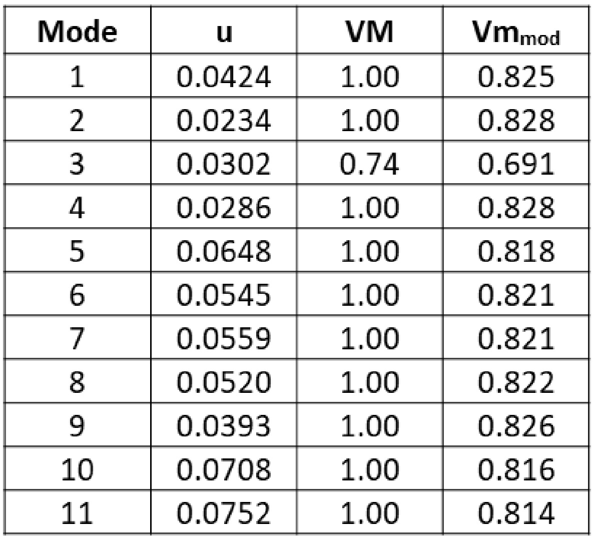
Figure 29: Radar chart representation of Figure 27. The blue line and grey lines represent the technical panel and subject matter expert averages resulting from individual scoring. The orange and dashed yellow lines represent the technical panel and subject matter expert consensus scores for case study #3 – thermoacoustic plate.

Table 13: Importance weights for the attributes assigned by the technical panel and subject matter experts for case study #3 – thermoacoustic plate. TP denotes technical panel and SME denotes subject matter expert.

Attributes	TP	TP	TP	SME	SME
1. The right measurements	5	5	3	5	5
2. The right tools	5	5	4	5	3
3. The right people	5	3	3	5	1
4. Regular review	3	2	2	3	1
5. Demonstrable consistency	5	3	3	1	2
6. The right procedures	4	5	4	4	5
7. Instruments	4	5	3	2	1
8. Object to be measured	5	3	2	2	1
9. Sampling	5	3	4	5	5
10. Operator skill	5	3	4	5	1
11. Environmental factors	3	5	5	2	3

1. Not important	*2. Preferred*	*3. Important*	*4. Highly desirable*	*5. Essential*

Table 14: Validation outcomes for 11 modal shapes shown in Figure 26 for room temperature heating. The total measurement uncertainty is represented by *u*, *VM* is Dvurecenska's *[42]* validation metric and $\mathrm{VM}_{\mathrm{mod}}$ is the modified validation metric described by equation (14).



Mode	u	VM	$\mathrm{Vm}_{\mathrm{mod}}$
1	0.0424	1.00	0.825
2	0.0234	1.00	0.828
3	0.0302	0.74	0.691
4	0.0286	1.00	0.828
5	0.0648	1.00	0.818
6	0.0545	1.00	0.821
7	0.0559	1.00	0.821
8	0.0520	1.00	0.822
9	0.0393	1.00	0.826
10	0.0708	1.00	0.816
11	0.0752	1.00	0.814

Table 15: Technical panel consensus scores and associated justification for case study #3 – thermoacoustic plate.

Attribute	Consensus score (1-5)	Justification stated
1. The right measurements	4	Some concern the conditions didn't match service, but on the whole measurements were appropriate
2. The right tools	3	Some concern about lack of calibration. Test was very unique so couldn't be high but results seemed good
3. The right people	3	Training is not covered in the paper, can only assume that given it is university research training was sufficient.
4. Regular review	2	Lack of detail about checks, universities don't have as many requirements about equipment as industry might.
5. Demonstrable consistency	2	Felt the number of repetitions helped here, although it was argued that falls more under "Sampling". Essentially only one set of equipment used, one environment so consistency not considered. Median score
6. The right procedure	3	Not enough information provided so went in the middle as high scores would warrant a standard being used
7. Instruments	3	TP Members torn between N/A and lower end of scale. Attribute very focused on earthing which wasn't discussed in the paper.
8. The object to be measured	5	Object and its preparation very well described
9. Sampling	5	Six repeats showing data is well sampled. Some initial controversy, one TP hadn't seen mention of number of repeats so initially gave low score. Score of 5 was unanimous after discussion.
10. Operator skill	4	Felt this was more focused on TPs judgement of the skill. Based on description given the experiment seemed to be sensibly performed demonstrating skill.
11. Environmental factors	4	Some control of environmental factors. The comparison to room conditions was liked.

Table 16: Subject matter expert consensus scores and associated justification for case study #3 – thermoacoustic plate.

Attribute	Consensus score (1-5)	Justification stated
1.The right measurements	5	TP overruled as SME felt it was unfair to mark down test due to conditions given how difficult it is to simulate hypersonic flight conditions.
2. The right tools	5	SME member identified disagreement between the two rubrics. They strongly felt that right tools had been used so overruled the TP.
3. The right people	3	Insufficient evidence was provided by the paper, implicit that right people were involved so deemed acceptable.
4. Regular review	3	Approach to reviewing equipment condition acceptable.
5. Demonstrable consistency	5	TP overruled as the paper makes no claims to the consistency of the data and thus it is N/A.
6. The right procedures	3	Unclear if a written procedure was used. SME agreed with logic of the TP
7. Instruments	3	Not clear this attribute description is appropriate for modern structural testing.
8. The object to be measured	5	Shape changes and emissivity changes were taken account of by measurement approach.
9. Sampling	5	Sufficient repetition used, good scientific method used.
10. Operator skill	4	No strong opinion, SME were swayed by the TP justification of their score
11. Environmental factors	4	Heat flux around sensors not considered. Lab condition were adequately controlled.

8.3 Discussion

For this case study, digital image correlation data of a Hastelloy-X plate subject to thermoacoustic loading was investigated.

Figure 27 illustrates the average scores assigned by the technical panel members and subject matter experts prior to group discussions. Figure 28 compares the individual assignments shown in Figure 27 to the consensus scores reached by the technical panel members and subject matter experts in the group meetings. For each of the attributes, the technical panel consensus lies within the standard deviation of the technical panel individual average. For two out of eleven of the attributes (the right measurements and the right tools) the subject matter expert consensus is the same as the subject matter expert average assigned prior to group discussion – i.e. they have chosen to overrule the technical panel consensus score and stick with their independent assignments. For the first attribute, the right measurements, the subject matter experts decided to increase the technical panel consensus due to the difficulty associated with simulating hypersonic flight conditions (see Table 16 for SME justifications). Likewise, for the second attribute, the right tools, the subject matter experts felt that the right tools had been used in the experiment. Therefore, they overruled the technical panel's score of 3 out of 5 for that specific attribute and increased it to 5 out of 5.

For the fifth attribute, demonstrable consistency, Figure 28 illustrates that the subject matter experts decide to overrule both the technical panel consensus and their independent average score assignment. As stated in Table 16, they decided to overrule the technical panel because the attribute was not applicable to the problem of interest.

For five out of eleven of the attributes (the right people, the right procedures, instruments, the object to be measured and operator skill), the subject matter experts overrule their individual score assignments to match the consensus set by the technical panel. The subject matter experts stated that they were swayed to stick with the technical panel consensus score assigned to the operator skill attribute because of the justification by the technical panel.

For the fourth attribute, regular review, there was initial agreement between the technical panel consensus, technical panel individual scores and the subject matter expert scores. Despite this, the subject matter experts decided to increase their assignment. As shown in Table 16, the justification presented for this was that the reviewing approach was considered acceptable. Referring to Table E3 in the Appendix, which contains the raw data for this case study, the first subject matter expert initially assigned a score of 3 out of 5, and the second subject matter expert initially assigned a score of 1 out of 5 to describe the possession of this attribute. Therefore, it is likely that the first subject matter expert has presented the second subject matter expert with sufficient justifications to support an overall consensus score of 3 out of 5. The facilitator also highlighted that both subject matter experts were influenced by their prior knowledge of the test setup and this allowed them to overrule technical panel consensus scores.

The attribute importance weightings assigned by the three technical panel members and two subject matter experts are shown in Table 13. The weightings options are: 1. Not important, 2. Preferred, 3. Important, 4. Highly desirable and 5. Essential. It can be seen that the average score assigned by the subject matter experts is lower than the average assigned by the technical panel – i.e. the technical panel consider the attributes to be more important for the problem of interest. To expand on this, the lowest importance weighting assigned by the technical panel members was 2 – Preferred. Whereas, both subject matter experts assigned scores of 1 – Not important – to describe the importance of some of the attributes. As stated previously, it was reported by the facilitator that both subject matter experts had a prior knowledge of the test set up which influenced their importance scores. This will allow them to better assess which attributes are important to the case study.

Also, there are key differences that can be observed between the weights assigned by the two subject matter experts. For the first subject matter expert, the most common weight assigned is 5 (essential), whereas for the second subject matter expert, the most common weight assigned is 1 (not important). The discrepancies observed will largely be due to the background and expertise of the participants and their interpretation of the attributes and the problem of interest. Additionally, the facilitator noted that the first subject matter expert noted similarity between the importance weightings and the scores to describe the possession of the attributes. The subject matter expert expressed that this was linked to the good quality of the paper provided.

The room temperature modal shapes obtained from digital image correlation and finite element analysis (see Figure 26) are shown in Table 14 using Dvurecenska's validation metric, VM [42] and the modified validation metric VM_{mod} described by equation (14). For all of the modal shapes except for modal shape 3, a VM value of 100% was obtained; i.e. the relative errors between the measurement and prediction feature vectors were less than the total measurement uncertainty resulting from the calibration and decomposition. When modified using the quality factor, all of the 100% values fall to around 82%. These values are similar because, referring to equation (14), the only variable present is the reconstruction error component of the total measurement uncertainty u. Therefore, similar modified validation metric outcomes can be expected for modal shapes which exhibit similar reconstruction errors from the decomposition process. For modal shape 3, the outcome of Dvurecenska's VM was 74.3% and this reduced to 69% using the modified validation metric. This VM outcome was lower (i.e. ≠ 100%) for this specific modal shape because the displacement data contained more regions of missing data, thus contributing a higher relative error.

The quality factor for this case study was calculated using the SME consensus scores (yellow bar in Figure 28) and the mean of the attribute importance scores (see Table 13). The quality factor for this case study was calculated to be 0.83 – this corresponds to an excellent quality factor descriptor. This quality factor is reasonable because the case study data was rich full-field data which was collected over the period of a PhD project. The experimental rig used to acquire the data was also refined in a previous PhD project. Therefore, it is expected that the data would be of a high quality.

Chapter 9: Discussion

The data quality methodology has been demonstrated using three different case studies: 1. strain gauge data from a loaded aluminium plate with a hole in the centre, 2. digital image correlation data of a car bonnet liner subject to a projectile impact, and 3. digital image correlation data of a plate subject to thermoacoustic loading. The output of the methodology is a quantitative measure – the quality factor, QF, which describes the fitness of purpose of the dataset. The quality factor is provided to the decision-maker alongside a portfolio of evidence to support the decision making.

The methodology demonstrates the first successful application of the analytical hierarchy process and rational decision-making process for validation applications. The methodology's applicability to validation has been demonstrated by incorporating the resulting quality factor into an existing validation metric. When incorporated into an existing probabilistic validation metric, this marks the first validation metric which incorporates the quality of the measurement's data into validation outcomes. Additionally, the new modified validation metric addresses the limitations associated with the existing probabilistic validation metric.

The versatility of the methodology is also demonstrated by the data used in the case studies; the first used a sparse series of point measurements obtained from six strain gauges, whereas the second and third case study use fields of displacement data from measurements using digital image correlation.

As mentioned at the start of the thesis, there are various challenges associated with acquiring rich measurement data for use in validation. In circumstances where the quantity of the data cannot be increased, the quality of the data can be assessed and incorporated into the validation outcome instead. Thus, allowing sparse data to be used to validate models. Previous validation frameworks relied upon a richness of measurement data, often obtained from full-field techniques. Therefore, the new methodology addresses the gap of how to incorporate sparse data into validation outcomes.

9.1 Case study quality factors

The resulting quality factors and associated quality factor descriptors for each of the case studies are summarised in Table 17. The first case study, tensile plate with a hole, received a quality factor of 0.69 – the lowest of the three case studies. This is to be expected as the displacement of the plate can only be inferred at the six strain gauge locations; thus, the quality of the data will be lower as a result. The density of the strain gauges is adequate in the regions away from the hole at the centre, where the strain gradient is small, but insufficient close to the hole. In addition, the peak strain occurs at the edge of the hole where there is no strain gauge located.

Table 17: Final quality factor values and associated descriptors for the three case studies demonstrated: tensile plate with a hole, impact on bonnet liner and thermoacoustic plate.

Case study	Data type	Quality factor	Corresponding quality factor descriptor
Strain gauges on a loaded Al specimen	Sparse strain data	0.69	Very poor
Car bonnet liner subject to projectile impact	Historical DIC	0.75	Poor
Hastelloy-X plate subject to thermoacoustic loading	Historical DIC	0.83	Excellent

For the latter two case studies, the impact on the bonnet liner and the thermoacoustic plate, historical digital image correlation data was used. The rich full-field measurements provided full-field displacement information over the surface of the specimens. The measurement data used in these case studies has previously been published in research papers which document the experiment and procedures well. For these reasons, the higher quality factors that are observed in these case studies is expected.

For the second case study, the impact tests were taken over a period of 3-4 days by a team who came together specifically for this purpose. Although the measurement data in this case study is rich, it does contain regions of missing data. It was expected that the highest quality factor would be obtained for the third case study because the measurement data was collected over the period of a PhD. In addition, the experimental rig was refined in a previous PhD project. Therefore, there is more confidence that the measurements have been acquired to a high standard.

9.2 Data quality methodology

The methodology presented in this thesis allows the quality of a dataset to be assessed by measuring how well it possesses a series of good measurement practice attributes. The resulting quality measure QF, the quality factor, has been incorporated into an existing validation metric to allow for validation outcomes to consider the quality of the dataset. By incorporating the quality factor into validation, it allows the probability of predictions belonging to the same population as measurements to be moderated by the quality of the measurement data. This is significant as existing validation methodologies quantify the quality of predictions without incorporating the quality of the measurement data. The modified validation metric is the first validation metric to incorporate a measure of data quality, thus providing the decision-maker with more information and in cases where the measurement dataset's quality is sufficient, more confidence in the model.

The required quality of the data depends heavily upon its intended use and the consequences associated with its use. Therefore, a matrix has been developed which allows the decision-maker to determine if they have a suitable quality factor for a given socioeconomic consequence. Depending upon the quality factor and socioeconomic consequence, recommendations are provided to the decision-maker. There are four recommendations provided to the problem-owner: sufficient quality, some quality issues, serious quality issues and insufficient quality. The matrix allows the decision-maker to decide if the data they have is fit for a particular purpose. It may be that the data is of insufficient quality for one use, but sufficient for a different application.

The methodology has been successfully demonstrated using three case studies and it marks the first application of the analytical hierarchy process in validation. The modified validation metric which incorporates the quality factor represents the first validation metric which includes a measure of data quality. A major advantage of the methodology is that it can be applied to sparse and historical data which suffers from lack of documentation. As mentioned previously, the decision-maker receives a quality factor which describes the quality of the dataset and a portfolio of evidence which includes the decisions made. The portfolio of evidence includes attribute scores and associated justifications. In some circumstances, the justification for a low attribute score will be lack of information about the historical data. In such cases, the decision-maker may decide to acquire the relevant information to support a higher scoring of the attribute. The final value of the quality factor will not reflect lack of documentation but the portfolio of evidence provided will highlight instances where lack of documentation has impacted results.

The methodology relies on the judgment and knowledge of a panel of participants. This can be viewed as a limitation with respect to biases that are introduced. These biases are mitigated by ensuring that the panel of participants are diverse and by introducing a facilitator to mediate discussion. The panel of participants use data and their own personal expertise to allocate scores to describe how well the data possesses the attributes. The advantage of this relativist approach is that participants hold unique insights and expertise which provide informed judgments. Whereas, a non-expert making judgments based solely on data may lead to less informed judgments.

The resulting quality factor is calculated using the scores assigned by the participants, to describe how well the dataset possesses the attributes, and the weights which indicate how important the attributes are. While this approach is subjective, it does provide key insights into what academics and industrialists consider to be important and of high quality in regards to data acquisition. Key insights include identifying which attributes are important for a given case study, highlighting limitations of an experiment or the acquisition of a dataset. A drawback of subjective methodologies is that they are generally affected by various types of biases. To reduce bias in the data quality methodology, a diverse panel of participants are used with a facilitator to manage group discussions and mitigate negative group dynamics.

The quality factor has a resulting range of values which span from 0.2 to 1; where 1 indicates that the dataset has received perfect scores by the subject matter experts to describe how well the attributes are possessed. An advantage of the quality factor having a maximum value of 1 is that it can be easily interpreted by non-experts.

In the next section, a number of improvements which can be incorporated into the methodology have been proposed.

9.3 Improvements to data quality methodology

Measuring expertise:

The quality factor measures the quality of the data with respect to how well the data possesses a series of good practice guide attributes. The scores assigned by the technical panel and the subject matter experts are, as indicated by the case study facilitators, dependent upon their own expertise, judgment and personal knowledge of the data, the field and the problem of interest. Therefore, one improvement for the methodology would be to incorporate a measure of the expertise of the participants into the resulting quality factor. Methods for quantitatively assessing expertise are limited in literature. It is recognised in literature that to identify an expert, there needs to be an assessment against a gold standard. However, such a standard does not exist [87]. The difficulty to define and measure expertise is summarised by Martini who states that 'expertise is a social concept, and measuring expertise is more like measuring a country's wealth, or an individual's happiness: a measuring process that must be constantly updated and corrected' [88]. Expertise cannot be detected directly, but factors such as social accreditation, experience and competence can be examined. The most recent efforts towards identifying and measuring expertise in organisations are outlined by Grenier and Germain [89]. They outline the importance of organisations having methods of defining and understanding expertise. Once expertise is understood, organisations can then develop methods for identifying and measuring expertise in employees, which may include incorporating machine learning approaches using deep neural networks.

Expert panel of participants

Another improvement that could be made to the methodology is repeating the case studies using several independent panels of participants. For the work presented in this thesis, this was not possible to achieve, however in an industry environment this would be more applicable.

This would allow for the resulting quality factor to be formed from several sets of participants, thus allowing for greater diversity in expertise and knowledge and a reduction in subjective bias.

DIC-specific attributes

The attributes used in the methodology have been taken from a National Physical Laboratory good practice guide [71]. These attributes are generalised and focus on the quality of the procedure used to acquire the data. For case studies which use digital image correlation data, a set of DIC-specific attributes could be incorporated to better tailor the methodology. Such attributes can be divided into three categories: sample preparation, software and hardware.

For sampling preparation, the key attribute of interest is the speckle pattern [90]. The quality of the speckle pattern can be assessed using a set of requirements [91]. These include ensuring that the speckle pattern is random enough to have distinguishable patterns at different regions, ensuring that the speckles have an average size of 3-7 pixels and variability in speckle size. It is also recommended that the speckle pattern is high contrast, random and isotropic. The technique used to apply the speckle pattern, i.e. airbrush, spray can, can also be assessed within this attribute.

Software attributes can include the software used, the facet size and the step size [92]. Hardware attributes will focus mainly on the cameras – i.e. for stereoscopic DIC, the angle between the cameras, light intensity and associated camera settings [93]. These attributes can be combined with the National Physical Laboratory attributes used in this thesis, to produce a quality factor value which incorporates the quality of the acquisition process and the quality of the digital image correlation data. Incorporating DIC tailored attributes would allow for a better comparison between datasets obtained using DIC. However, when comparing datasets obtained using different techniques, the National Physical Laboratory attributes would be more appropriate as they are more general.

Socioeconomic consequence matrix

The quality factor-socioeconomic consequence matrix shown in Table 5 has been developed to allow the problem-owner to interpret the implications of the quality factor, relative to its problem of interest. This matrix provides guidance to the decision-maker regarding if their dataset is fit for purpose and it leads to four possible recommendations regarding the quality of the data: Sufficient quality, some quality issues, serious quality issues and insufficient quality, for its intended use. For data of sufficient quality, it can be employed for its intended use without concern. However, for data of insufficient quality, it is recommended that the data is re-acquired or used for a different application where it would be suitable.

The socioeconomic consequence has been divided into four rankings: High, medium, low and very low. For the purposes of this matrix, it is the problem-owner's responsibility to decide which of these rankings is applicable. One potential improvement would be to define and characterise the socioeconomic consequence rankings for the benefit of the problem-owner.

There are existing matrices used in risk management which assess consequences according to critical safety factors such as public safety, environment, lifestyle, economy and public administration [94]. For example, the most severe socioeconomic consequence would be associated with loss of life, catastrophic environmental damage, large repair costs and severe impacts on services. One of the issues that arises from using consequence matrices is the interpretation of terms that are found in the matrix to describe the amount of damage – i.e. 'numerous' and 'catastrophic' [94]. These terms are subjective and there will be variation in the interpretation of the terms. Therefore, it is recommended to use a consequence matrix which uses quantitative descriptors which can be interpreted objectively. Such quantitative descriptors could include numbers related to economic cost and the cost to human life. For the matrix described in this thesis, to determine the suitability of data, the socioeconomic consequences will be characterised within the industry by decision-makers.

9.4 Summary of discussion

Methodologies for performing validation can be found in existing guides such as the ASME and CEN guides. The quality of predictions from a model are assessed by comparing them to corresponding measurement data obtained from an experiment. The ADVISE project provided a calibration methodology for ensuring the acquisition of high quality full-field data for validation. However, a measure of the quality of the measurement data has not previously been incorporated into the validation process. This gap has been addressed in this thesis through the development of the data quality methodology in Chapter 4, and the incorporation of the resulting quality measure, QF, into an existing validation metric. The development of a new validation metric which incorporates a measure of data quality also represents a significant advance as such metrics have not been reported in the literature. This provides a validation outcome which moderates the quality of predictions with the quality of the measurement data used to validate the model.

The methodology has been successfully applied to sparse data and field data obtained using DIC. The aim of the research was to develop a quantitative technique for validation which could be used to perform validation with sparse data. The new methodology addresses this aim by relying on an expert panel of participants who use their knowledge, judgment and data to determine how well the data possesses a series of good measurement practice attributes.

The work described in this thesis also connects to recent work which explores establishing credibility in models when measurement data is sparse or absent; where credibility is the willingness of people to make decisions based on data from a model. The methodology provides a decision-maker with a quality factor and a portfolio of evidence which will allow the decision-maker to judge whether the data is suitable for an intended use. Further, the development of a matrix which assesses the quality factor against the socioeconomic consequence of misusing the data allows the decision-maker to ascertain if the data will be fit for its intended purpose.

The assessment of data quality also ties to the work described in the MOTIVATE project, described in Chapter 3. In this project, validation flowcharts which assess the suitability of historical data for validation were developed. The quality factor obtained from the new methodology can be incorporated into this flowchart to better support the assessment of historical data.

Although the aim of this thesis focused on validation using sparse data, it has been extended to include full-field data. Further work has been described in Chapter 11, but the aim and objectives of the thesis have been met successfully.

Chapter 10: Conclusions

The goal of the PhD was to develop a reliable and transferable quantitative validation technique, which could be employed in industry and academia, to validate models using sparse data. This goal has been achieved through the development of a new methodology which allows decision-makers to determine if measurement data, such as sparse and historical data, is suitable for validation purposes.

The quality of the data is assessed by an expert panel of participants who decide how well the dataset possesses good measurement practice attributes. The methodology presents two key advancements:

1. Quantitatively determining the suitability of a dataset for validation
2. Measuring the quality of the measurement data used to validate the model

The methodology has been successfully demonstrated using three case studies from the discipline of solid mechanics: a sparse set of strain gauge data from a loaded specimen, impact loading data from a bonnet liner, and modal excitation data from a plate subject to thermoacoustic loading. To guide decision-makers from industry and academia, a socioeconomic matrix was developed which can be used to assess the obtained quality factor against the model's socioeconomic consequences.

By determining if existing data is suitable for validation, it removes the requirement for performing additional validation experiments; thus, saving time and money. If the dataset is found to be unsuitable, it is also easy to identify and highlight which of the attributes require refinement. Although the attributes presented in this thesis have been taken from a National Physical Laboratory beginner's guide to measurement, the methodology can incorporate a different set of attributes which may be more relevant to the problem of interest.

The resulting measure of quality, the quality factor, has also been incorporated into an existing validation metric which has been applied to the latter two case studies. When the quality factor is incorporated into a validation metric, the validation outcome is weighted by the quality of the measurement data used to validate the model; thus, providing more information to decision-makers who may be making decisions of high socioeconomic consequences. If the quality of the data is low, this may decrease confidence in predictions, and if the quality of the data is found to be high, confidence in the predictions will be established.

The quality factors obtained from the three case studies are in line with expectations; the sparse strain data case study was found to have the lowest quality and the highest quality was observed for full-field data which was collected over the period of a PhD project. This indicates that employing a panel of participants with relevant expertise to assess the quality of the data has worked successfully.

The work described in this thesis has been presented at the British Society for Strain Measurement 2022 conference and a journal paper based on the work is pending submission.

Chapter 11: Future work

The novel data quality methodology presented in this thesis enables the quality of datasets to be assessed and the resulting quality measure, QF, can be incorporated into existing validation metrics. In Chapter 9, a number of improvements that could be made to the new methodology are proposed, some of which are discussed below.

It has been mentioned in this thesis that validation outcomes are hindered when sparse data is used as validation data. Although the new data quality methodology has been demonstrated to work using sparse data, it would still be useful for the decision-maker to have a measure of sparsity which can also be incorporated into validation metrics. This future work can be divided into three actions: 1. defining data sparsity, 2. developing a method for measuring sparsity and 3. incorporating the resulting sparsity measure into a validation metric. For the first action, the definition of sparsity will be highly dependent upon the industry. For example, what might be considered data rich in one industry, may be considered sparse in another. Additionally, a dataset can be considered to be sparse if it does not contain data from an important region of interest – i.e. this can arise in situations where data acquisition is not possible due to obstructions. In regards to developing a method for measuring sparsity, the method will be dependent upon the definition of sparsity. In computer science applications, the sparsity of a matrix is defined as the ratio of the number of zero-elements, to the number of total elements. This concept can be applied to engineering applications if the total number of elements, i.e. the maximum number of obtainable measurements, is known. If the resulting sparsity measure is expressed as a number between 0 and 1, this can be easily incorporated into the modified validation metric (equation 14).

In this thesis, the methodology's applicability in validation has been demonstrated by incorporating the quality factor into an existing probabilistic validation metric. The quality factor can also be incorporated directly into the validation framework shown in Figure 8. This would be particularly useful for the sub-flowchart which assesses the suitability of historical measurement data (Figure 9).

As discussed in Chapter 9, the quality factor is heavily influenced by the expertise of the panel of participants (the technical panel members and the subject matter experts) employed in the case studies. The usefulness of the quality factor would be enhanced if it was weighted by a quantity which measures the expertise of the participants. However, this is difficult to achieve in practice because as mentioned in Chapter 9, a methodology for measuring the expertise of an individual does not exist in literature.

In an industrial environment, the methodology can be used in conjunction with the socioeconomic-quality factor matrix to guide the acquisition of data and help identify and support the most appropriate use of the data.

References

[1] M. Calder and et al, “Computational modelling for decision-making: where, why, what, who and how,” *R. Soc. open sci,* vol. 5, no. 6, p. 172096, 2018.

[2] ASME V&V 10-2019, “Standard for verification and validation in computational solid mechanics,” American Soc. of Mech. Engineers, New York, 2020.

[3] P. Sanders, “DoD modelling and simulation (M&S) verification, validation and accreditation (VV&A),” Office of the under secretary of defense (acquisition and technology), Washington DC, 1996.

[4] W. L. Oberkampf, M. Sindir and A. T. Conlisk, “Guide for the verification and validation of computational fluid dynamics simulations,” American Institute of Aeronautics and Astronautics, no. AIAA G-077-1998, 1998.

[5] CEN, “Validation of computational solid mechanics models,” *Comite Europeen de Normalisation (CEN),* no. CWA16799(2014), 2014.

[6] SPOTS, “Standardisation Project for Optical Techniques of Strain Measurement,” EU contract no. G6RD-CT-2002-00856. See www.opticalstrain.org, 2003.

[7] ADVISE, “Advanced Dynamic Validations using Integrated Simulation and Experimentation,” Project No. SCP7-GA-2008-218595. See www.dynamicvalidation.org, 2005.

[8] E. A. Patterson, E. Hack, P. Brailly, R. L. Burguete, Q. Saleem, T. Siebert, R. A. Tomlinson and M. P. Whelan, “Calibration and evaluation of optical systems for full-field strain measurement,” *Optics and Lasers in Engineering,* vol. 45, no. 5, pp. 550-564, 2007.

[9] M. P. Whelan, D. Albrecht, E. Hack and E. A. Patterson, “Calibration of a speckle interferometry full-field strain measurement system,” *Strain,* vol. 44, no. 2, pp. 180-190, 2008.

[10] R. L. Burguete, G. Lampeas, J. E. Mottershead, E. A. Patterson, A. Pipino, T. Siebert and W. Wang, “Analysis of displacement fields from a high-speed impact using shape descriptors,” *The Journal of Strain Analysis for Engineering,* vol. 49, no. 4, pp. 212-223, 2014.

[11] A. S. Silva, C. M. Sebastian, J. Lambros and E. A. Patterson, “High temperature modal analysis of a non-uniformly heated rectangular plate: Experiments and simulations,” *Journal of Sound and Vibration,* vol. 443, pp. 397-410, 2019.

[12] R. Sargent and O. Balci, “History of Verification and Validation of Simulation Models,” in *Proceedings of the 2017 Winter Simulation Conference*, 2017.

[13] G. Fishman and P. Kiviat, “The statistics of discrete-event simulation,” *Simulation,* vol. 10, no. 4, p. 185, 1968.

[14] T. Naylor and J. Finger, “Verification of computer simulation models,” *Management Science,* vol. 14, no. 2, pp. 92-100, 1967.

[15] W. Schrank and C. Holt, “Critique of: "Verification of Computer Simulation Models",” *Management Science,* vol. 14, no. 2, pp. 104-106, 1967.

[16] K. Popper, The Logic of Scientific Discovery, Routledge, 2005.

[17] G. Kleindorfer, “The philosophy of science and validation in simulation,” in *Proceedings of the 1993 Winter Simulation conference*, 1993.

[18] G. Kleindorfer, L. O'Neill and R. Ganeshan, “Validation in simulation: various positions in the philosophy of science,” *Management Science,* vol. 44, no. 8, pp. 1087-1090, 1998.

[19] R. Van Horn, “Validation of simulation results,” *Management Science,* vol. 17, no. 5, pp. 247-257, 1971.

[20] R. Sargent, “Some subjective validation methods using graphical displays of data,” in *Proceedings of the 28th conference on Winter Simulation. IEEE Computer Society,* 1996.

[21] T.H. Naylor, J.L. Balintfy, D.S Burdick and K. Chu, “Computer Simulation Techniques,” in *John Wiley & Sons,* New York, 1966.

[22] ASME, “Guide for verification and validation in computational solid mechanics,” American Society of Mechanical Engineers, no. ASME V&V 10-2006, 2006.

[23] R. B. Norman and S. R. Blattnig, “A comprehensive validation methodology for sparse experimental data,” Technical Report TP-2010-216200, NASA, 2010.

[24] O. Balci, “How to assess the acceptability and credibility of simulation results,” in *Proceedings of the 21st conference on Winter Simulation* , 1989.

[25] A. Saygin, I. Cicekil and V. Akman, “Turing test: 50 years later,” *Minds and Machines,* vol. 10, no. 4, pp. 463-518, 2000.

[26] W. Oberkampf and M. Barone, “Measures of agreement between computation and experiment: validation metrics,” *Journal of Computational Physics,* vol. 217, no. 1, pp. 5-36, 2006.

[27] O. Balci, “Validation, verification, and testing techniques throughout the life cycle of a simulation study,” *Annals of Operations Research,* vol. 53, no. 1, pp. 121-173, 1994.

[28] O. Balci and R. Sargent, “Validation of multivariate response models using Hotelling's two-sample T2 test,” *Simulation,* vol. 39, no. 6, pp. 185-192, 1982.

[29] R. Sargent, “Verification and validation of simulation models,” in *Proceedings of the 2009 Winter Simulation Conference*, 2009.

[30] W. Oberkampf, M. Sindir and A. Conlisk, “Guide for the verification and validation of computational fluid dynamics simulations,” American Institute of Aeronautics and Astronautics, no. AIAA G-077-1998, 1998.

[31] C. Sebastian, E. Hack and E.A. Patterson, “An approach to the validation of computational solid mechanics models for strain analysis,” *The Journal of Strain Analysis for Engineering Design,* vol. 48, no. 1, pp. 36-47, 2013.

[32] G. Lampeas, V. Pasialis, X. Lin and E.A. Patterson, “On the validation of solid mechanics models using optical measurements and data decomposition,” *Simulation Modelling Practice and Theory,* vol. 52, pp. 92-107, 2015.

[33] A. Pakti, W. Wang, J. Mottershead and E.A. Patterson, “Image decomposition as a tool for validating stress analysis models,” *In EPJ Web of Conferences. EDP Sciences,* vol. 6, p. 46005, 2010.

[34] E. Hack and E.A. Patterson, “Experimental validation of simulations using full-field measurement techniques,” *AIP Conference Proceedings,* vol. 1253, no. 1, pp. 405-409, 2010.

[35] W. J.R. Christian, A. D. Dean, K. Dvurecenska, C. A. Middleton and E. A. Patterson, “Comparing full-field data from structural components with complicated geometries,” *R. Soc. Open Sci.,* vol. 8, no. 9, p. 210916, 2021.

[36] W.J.R Christian and A.D. Dean, "Theon Version 1.01". Available online: www.experimentalstress.com/software, 2021.

[37] E. Anderson, Z. Bai and J. Dongarra, “Generalized QR factorization and its applications,” *Linear Algebra and its Applications,* vol. 162, pp. 243-271, 1992.

[38] A. Alexiadis, S. Ferson and E. A. Patterson, “Transformation of measurement uncertainties into low-dimensional feature vector space,” *R. Soc. open sci,* vol. 8, no. 3, p. 201086, 2021.

[39] S. Ferson, W. Oberkampf and L. Ginzburg, “Model validation and predictive capability for the thermal challenge problem,” *Computer Methods in Applied Mechanics and Engineering,* vol. 197, no. 29-32, pp. 2408-2430, 2008.

[40] W. Oberkampf and C. Roy, Verification and Validation in Scientific Computing, Cambridge University Press, 2010.

[41] Y. Liu, W. Chen, P. Arendt and H. Z. Huang, “Toward a better understanding of model validation metrics,” *Journal of Mechanical Design,* vol. 133, no. 7, pp. 071 005-071 005-13, 2011.

[42] K. Dvurecenska, S. Graham, E. Patelli and E. A. Patterson, “A probabalistic metric for the validation of computational models,” *Royal Society Open Science,* vol. 5, no. 11, p. 180687, 2018.

[43] G. Gigerenzer, R. Hertwig, E. van den Broek, B. Fasolo and K. V. Katsikopoulos, “"A 30% chance of rain tomorrow": How does the public understanding probablistic weather forecasts,” *Risk Analysis,* vol. 25, no. 3, pp. 495-781, 2005.

[44] W. D. Rowe, “Understanding uncertainty,” *Risk Analysis,* vol. 14, no. 5, pp. 743-750, 1994.

[45] M. Plebani, L. Sciacovelli, D. Bernardi, A. Aita, G. Antonelli and A. Padoan, “What information on measurement uncertainty should be communicated to clinicians, and how?,” *Clinical biochemistry,* vol. 57, pp. 18-22, 2018.

[46] E.A. Patterson, “On the credibility of engineering models and meta-models,” *J Strain Analysis,* vol. 50, no. 4, pp. 218-220, 2015.

[47] L. W. Schruben, “Establishing the credibility of simulations,” *Simulation,* vol. 34, pp. 101-105, 1980.

[48] E.A. Patterson and M. P. Whelan, “On the validation of variable fidelity multi-physics simulations,” *Journal of Sound and Vibration,* vol. 448, pp. 247-258, 2019.

[49] E. A. Patterson, M. P. Whelan and A. P. Worth, “The role of validation in establishing the scientific credibility of predictive toxicology approaches intended for regulatory application,” *Computational Toxicology,* vol. 17, p. 100144, 2021.

[50] O. S. Vaidya and S. Kumar, “Analytic hierarchy process: An overview of applications,” *European Journal of Operational Research,* vol. 169, no. 1, pp. 1-29, 2006.

[51] L. G. Vargas, “An overview of the analytic hierarchy proess and its applications,” *European journal of operational research,* vol. 48, no. 1, pp. 2-8, 1990.

[52] E. H. Forman and S. I. Gass, “The analytic hierarchy process - an exposition,” *Operations research,* vol. 49, no. 4, pp. 469-486, 2001.

[53] R. W. Saaty, “The analytic hierarchy process - what it is and how it is used,” *Mathematical modelling,* vol. 9, no. 3-5, pp. 161-176, 1987.

[54] A. Ishizaka and A. Labib, “Review of the main developments in the analtic hierarchy process.,” *Expert systems with applications,* vol. 38, no. 11, pp. 14336-14345, 2011.

[55] F. Uzonwanne, “Rational model of decision making,” in *Global encyclopedia of public administration, public policy, and governance.,* Springer International. https:/doi. org/10.1007/978-3-319-31816-5_2474-1., 2016.

[56] C. Wild, “Ethics of resource allocation: instruments for rational decision making in support of a sustainable health care,” *Poiesis & Praxis* , vol. 3, pp. 296-309, 2005.

[57] R. Bhui, L. Lai and S. J. Gershman, “Resource-rational decision making.,” *Current Opinion in Behavioural Sciences,* vol. 41, pp. 15-21, 2021.

[58] E. J. Olden and E. A. Patterson, “A rational decision making model for experimental mechanics,” *Experimental Techniques,* vol. 24, pp. 26-32, 2000.

[59] L. Heracleous, “The rational decision making: myth or reality?,” *Management development review,* vol. 7, no. 4, pp. 16-23, 1994.

[60] B. Whitworth, B. Van de Walle and M. Turoff, “Beyong Rational Decision Making,” in *Group Decision and Negotiation Conference,* Scotland, 2000.

[61] L. Learning, “Rational Decision Making vs. Other Types of Decision Making,” Lumen, 2022. [Online].Available:https://courses.lumenlearning.com/wmopen-principlesofmanagement/chapter/rational-decision-making-vs-other-types-of-decision-making/. [Accessed Feb 2023].

[62] E. A. Patterson, I. Diamantakos, K. Dvurecenska, R. J. Greene, E. Hack, G. Lampeas, M. Lomnitz and T. Siebert, “Validation of a structural model of an aircraft cockpit panel: An industrial case study,” *J Strain Analysis,* vol. 57, no. 8, pp. 714-723, 2022.

[63] MOTIVATE, “Matrix Optimization for Testing by Interaction of Virtual and Test Environments, H2020 Clean Sky 2 Project (Grant Agreement No. 754660).,” [Online]. Available: www.engineeringvalidation.org. [Accessed January 2020].

[64] E. Hack, K. Dvurecenska, G. Lampeas, E. A. Patterson, T. Siebert and E. Szigeti, "Incorporating historical data in a validation process," in *14th Int. Conf. on Advances in Experimental Mechanics*, Belfast, 2019.

[65] P. Yan, L. Delannay, J. Payne and N. Tzelepi, "Micromechanistic modelling of the polycrystalline response of graphite under temperature changes and irradiation," *Carbon,* vol. 96, pp. 827-835, 2016.

[66] L. Delannay, P. Yan, J. Payne and N. Tzelepi, "Predictions of inter-granular cracking and dimensional changes of irradiated polycrystalline graphite under plane strain," *Computational Materials Science,* vol. 87, pp. 129-137, 2014.

[67] J. Brocklehurst and B. T. Kelly, "Analysis of the dimensional changes and structural changes in polycrystalline graphite under fast neutron irradiation," *Carbon,* vol. 31, no. 1, pp. 155-178, 1993.

[68] G. B. Neighbour, "Ageing studies and lifetime extension of materials," in *Modelling dimensional change with radiolytic oxidation in AGR moderator graphite,* Boston, Springer, 2001, pp. 419-427.

[69] G. Haag, Properties of ATR-2E graphite and property changes due to fast neutron irradiation, Forschungszentrum Jülich, 2005.

[70] J. Ahlf and J. Schinkel, "Upgrading and modernization of the high flux reactor Petten," *Nuclear engineering and design,* vol. 137, no. 1, pp. 49-56, 1992.

[71] M. Goldsmith, "A beginner's guide to measurement. Measurement Good Practice Guide No. 118," *National Physical Laboratory,* pp. 22-27, 2010.

[72] K. K. Lilja, K. Laakso and J. Palomaki, "Using the Delphi method.," *2011 Proceedings of PICMET'11: Technology Management in the Energy Smart World. IEEE,* pp. 1-10, 2011.

[73] P. Hopkins and M. Under, "What is a 'subject-matter expert?'," *Journal of Pipeline Engineering,* vol. 16, no. 4, 2017.

[74] G. Gopalakrishna, R. A. Mustafa, C. Davenport, R. J. Scholten, C. Hyde, J. Brozek, H. J. Schunemann, P. M. Bossuyt, M. M. Leeflang and M. W. Langendam, "Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but dooable," *Journal of clinical epidemiology,* vol. 67, no. 7, pp. 760-768, 2014.

[75] H. Balshem, M. Helfand, H. J. Schunemann, A. D. Oxman, R. Kunz, J. Brozek, G. E. Vist, Y. Falck-Ytter, J. Meerpohl, S. Norris and G. H. Guyatt, "GRADE guidlines: 3. Rating the quality of evidence," *Journal of clinical epidemiology,* vol. 64, no. 4, pp. 401-406, 2011.

[76] R. Siemieniuk and G. Guyatt, "What is GRADE?," BMJ Best Practice, [Online]. Available: https://bestpractice.bmj.com/info/toolkit/learn-ebm/what-is-grade/. [Accessed 2023].

[77] J. A. Kolb, S. Jin and J. Hoon Song, "A model of small group facilitator competencies," *Performance improvement quarterly* , vol. 21, no. 2, pp. 119-133, 2008.

[78] V. Pierce, D. Cheesebrow and L. M. Braun, "Facilitator competencies," *Group Facilitation* , vol. 2, p. 24, 2000.

[79] A. Van De and A. L. Delbecq, “Nominal versus interacting group processes for committee decision-making effectiveness,” *Academy of management journal,* vol. 14, no. 2, pp. 203-212, 1971.

[80] L. Bauto, “Factor of Safety: Ratio for Safety in Design and Use,” Safety Culture, Nov 23 2022. [Online]. Available: https://safetyculture.com/topics/factor-of-safety/. [Accessed Nov 2021].

[81] H. H. Friedman and T. Amoo, “Multiple biases in rating scale construction,” *Journal of international marketing and marketing research,* vol. 24, pp. 115-126, 1999.

[82] E. Limpert, W. A. Stahel and M. Abbt, “Log-normal distributions across the sciences: keys and clues: on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into v,” *BioScience,* vol. 51, no. 5, pp. 341-352, 2001.

[83] V. Solutions, “Risk Matrix Calculations - Severity, Probability, and Risk Assessment,” 23 Apr 2018. [Online]. Available: https://www.vectorsolutions.com/resources/blogs/risk-matrix-calculations-severity-probability-risk-assessment/. [Accessed Nov 2022].

[84] T. Nachazel, “What is a Strain Gauge and How Does it Work?,” Michigan Scientific Corporation, 12 Aug 2020. [Online]. Available: https://www.michsci.com/what-is-a-strain-gauge/. [Accessed Jan 2023].

[85] G. Origgi, “A Social Epistemology of Reputation,” *Social Epistemology,* vol. 26, no. 3-4, pp. 399-418, 2012.

[86] W.J.R Christian and E.A Patterson, "Euclid Version 1.01". Available: http://www.experimentalstress.com/software. [Accessed 2023]

[87] J. Shanteau, D. J. Weiss, R. P. Thomas and J. C. Pounds, “Performance-based assessment of expertise: How to decide if someone is an expert or not,” *European Journal of Operational Research,* vol. 136, no. 2, pp. 253-263, 2002.

[88] C. Martini, “The epistemology of expertise,” in *The Routledge handbook of social epistemology,* New York, Routledge, 2019, pp. 115-122.

[89] R. S. Grenier and M. L. Germain, “An Introduction to Expertise at Work: Current and Emerging Trends,” in *Expertise at Work*, Springer, 2021, pp. 1-13.

[90] D. Lecompte, A. S. Smits, S. Bossuyt, H. Sol, J. Vantomme, D. Hemelrijck and A. M. Hadbraken, “Quality assessment of speckle patterns for digital image correlation,” *Optics and lasers in Engineering,* vol. 44, no. 11, pp. 1132-1145, 2006.

[91] G. Quino, Y. Chen, K. R. Ramakrishnan, F. Martinez-Hergueta, G. Zumpano, A. Pellegrino and N. Petrinic, “Speckle patterns for DIC in challenging scenarios: rapid application and impact indurance,” *Measurement Science and Technology,* vol. 32, no. 1, p. 015203, 2020.

[92] K. Nwanoro, P. Harrison and F. Lennard, “Investigating the accuracy of digital image correlation in monitoring strain fields across historical tapestries,” *Strain,* vol. 58, no. 1, p. 12401, 2022.

[93] H. Zhu, X. Liu, L. Chen, Q. Ma and S. Ma, “Influence of imaging configurations on the accuracy of digital image correlation measurement,” *Measurement Science and Technology,* vol. 29, no. 3, p. 035205, 2018.

[94] R. Farrar, “Risk Tip #3 - Developing a Consequence Matrix,” Paladin risk management services, 16 Jan 2017. [Online]. Available: https://paladinrisk.com.au/risk-tip-3-developing-consequence-matrix/. [Accessed Nov 2022].

Appendix A: Analytical Hierarchy Process example

For the hierarchy presented in Figure 6 (taken from ref [53]), an example of how to determine the global priority vector is outlined below. This example has been adapted from ref [53].

The table below contains possible outcomes of pairwise comparisons of the criteria, using the 1-9 scale shown in Table 1, with respect to the overall objective:

Criteria	Location	Ambience	Reputation	Academics
Location	1	1/4	1/5	1/5
Ambience	4	1	3	3
Reputation	5	1/3	1	2
Academics	5	1/3	1/2	1

In this example, ambience is judged to be four times more important than location, and reputation and academics are each judged to be 5 times more important than location. The reciprocals of these values are then placed in the corresponding transpose positions. This would lead to the following positive reciprocal matrix, *A*:

$$A = [a_{ij}] \ = \begin{bmatrix} 1 & 4 & 1/5 & 1/5 \\ 1/4 & 1 & 3 & 3 \\ 5 & 1/3 & 1 & 2 \\ 5 & 1/3 & 1/2 & 1 \end{bmatrix}$$

The normalised matrix *B* is then given by:

$$B = [b_{ij}] \ = \begin{bmatrix} 0.067 & 0.130 & 0.043 & 0.032 \\ 0.267 & 0.522 & 0.638 & 0.484 \\ 0.333 & 0.174 & 0.213 & 0.323 \\ 0.333 & 0.174 & 0.106 & 0.161 \end{bmatrix}$$

The weights of the criteria, w_j are found by calculating the mean of each row:

w_j = (Location, Ambience, Reputation, Academics) = (0.068; 0.477; 0.261; 0.194)

From this vector, it can be seen that the criterion with the highest relative importance is ambience.

In the next step, pairwise comparisons are of each of the alternatives (i.e. college options) are made with respect to each of the criteria. For example, for the location criterion, the pairwise comparisons and resulting weight may look like:

Location	SWARTH	NORTHW	UMICH	VANDERB	CMU	Weight
SWARTH	1	1/4	1/3	1/3	7	0.115
NORTHW	4	1	2	3	7	0.402
U.MICH	3	1/2	1	3	6	0.284
VANDERB	3	1/3	1/3	1	4	0.163
CMU	1/7	1/7	1/6	1/4	1	0.037

This will lead to a local priority vector of:

Location: (SWARTH, NORTHW, U.MICH, VANDERB, CMU) = (0.115, 0.402, 0.283, 0.163, 0.037)

Following the same process for the remaining three criteria:

Ambience: (SWARTH, NORTHW, U.MICH, VANDERB, CMU) = (0.034, 0.539, 0.250, 0.121, 0.056)

Reputation: (SWARTH, NORTHW, U.MICH, VANDERB, CMU) = (0.521, 0.235, 0.147, 0.038, 0.059)

Academics: (SWARTH, NORTHW, U.MICH, VANDERB, CMU) = (0.564, 0.209, 0.132, 0.040, 0.055)

These vectors form the local priority *l_{ij}* and the global priority p_i of each alternative i, is then given by [54]:

$$p_i = \sum_j w_j . l_{ij}$$

Using the example above, pi is given by:

$$p_i = \begin{bmatrix} 0.115 & 0.034 & 0.521 & 0.564 \\ 0.402 & 0.539 & 0.235 & 0.209 \\ 0.284 & 0.250 & 0.147 & 0.132 \\ 0.163 & 0.121 & 0.038 & 0.040 \\ 0.037 & 0.056 & 0.059 & 0.055 \end{bmatrix} . \begin{bmatrix} 0.068 \\ 0.477 \\ 0.261 \\ 0.194 \end{bmatrix} = \begin{bmatrix} 0.269 \\ 0.386 \\ 0.203 \\ 0.086 \\ 0.056 \end{bmatrix}$$

Each element of p_i will represent the global priority of each of the alternatives. The alternative with the highest global priority will be considered as the most favourable option for the overall objective. In this example, it can be concluded that the second alternative is the most favourable option as it has a global priority of 0.386 – i.e. Northwestern would be the preferred school in the proposed example.

Appendix B: Case study #1 documentation

This appendix contains the form that was sent to the technical panel for data quality case study #1 – tensile plate with a hole. The form was sent to the technical panel members with the case study data and a document which explained the methodology.

Assigning scores to the attributes:

	Level 1	Level 2	Level 3	Level 4	Level 5
Fundamental attributes:	**Very dissatisfied**	**Dissatisfied**	**OK**	**Satisfied**	**Very satisfied**
Attribute 1	The measurement is not related to the problem				The measurement is fully appropriate for the problem being studied
Attribute 2	The instruments have not been calibrated				The instruments have been fully calibrated
Attribute 3	No training and instruction provided				The 'operator' has received the correct training and instructions
Attribute 4	No instrument check prior to use				Instruments are regularly checked and assessed
Attribute 5	Measurement is only valid in one place				Measurement has been replicated in different environments
Attribute 6	Measurement not carried out in accordance with written procedure				Measurement carried out in accordance with documents provided by manufacturer

Fundamental attributes:

Following the six guiding principles that should be followed in order to obtain a good result

1. The right measurement
2. Using the correct calibrated instruments
3. The right people
4. Regular review of measurement instruments
5. Demonstrating consistency – measurement cannot only be valid at the place it is made.
6. The right procedures

	Level 1	Level 2	Level 3	Level 4	Level 5
Desirable attributes:	**Very dissatisfied**	**Dissatisfied**	**OK**	**Satisfied**	**Very Satisfied**
Attribute 1	No measures have been in put in place to minimise electrical noise disruptions				Proper earthing of equipment ensured
Attribute 2	The object to be measured varies and no corrections have been applied				N/A or environmental conditions have been controlled, corrections and averages applied or dynamic measurements taken
Attribute 3	Measurement technique has not been well-designed and is not representative				Measurement technique is representative of important variations and there is knowledge of any expected changes
Attribute 4	The operator is not skilled				The operator has correctly set up the measuring equipment and prepared the thing to be measured
Attribute 5	Environmental factors not taken into account				Corrections applied to take account of any environmental factors

Desirable attributes:

Additional factors which can affect the measurement result:

7. Instruments – i.e. electrical measuring instruments can be affected by electrical noise.
8. The object to be measured – awareness of environmental conditions that may significantly change the measurement.
9. Sampling – Knowledge of expected changes is required.
10. Operator skill
11. Environmental factors

Case study sample data:

A Hounsfield tensometer was used to apply a uniaxial load to an aluminium specimen with a 20 mm diameter hole at its centre. A force of 8 kN was applied to the specimen and this was measured using a 1000 kgf load cell; where a 1kgf = 9.81 N.

The values of strain at the distances shown in Table B1 are measured by strain gauges and the readings are displayed on the DAQ in units of micro strains.

Distance/mm	Strain/µƐ
13.7	259.3617
21.8	180.9961
29.8	178.2972
37.4	165.4492
45.2	151.0977
200	141.0578

Table B1: *Sample data from case study #1 - Tensile plate with a hole.*

Appendix C: Case study #2 documentation

This appendix contains the documentation that was sent to the facilitator, technical panel members and subject matter experts for the second data quality case study #2 – impact on bonnet liner. There are three forms listed: one for the facilitator, one for the technical panel members and one for the subject matter experts.

Role: Facilitator

Name:

Profession:

Dates:

Technical panel meeting: Click or tap to enter a date.

Meeting with subject matter experts: Click or tap to enter a date.

Facilitator responsibilities:

1. Set-up a meeting with the technical panel
2. Takes notes during the meeting and record the final consensus technical panel scores
3. Send over consensus scores to the subject matter experts for their review
4. Set-up a meeting with subject matter experts to discuss their review and final comments

Form directions:

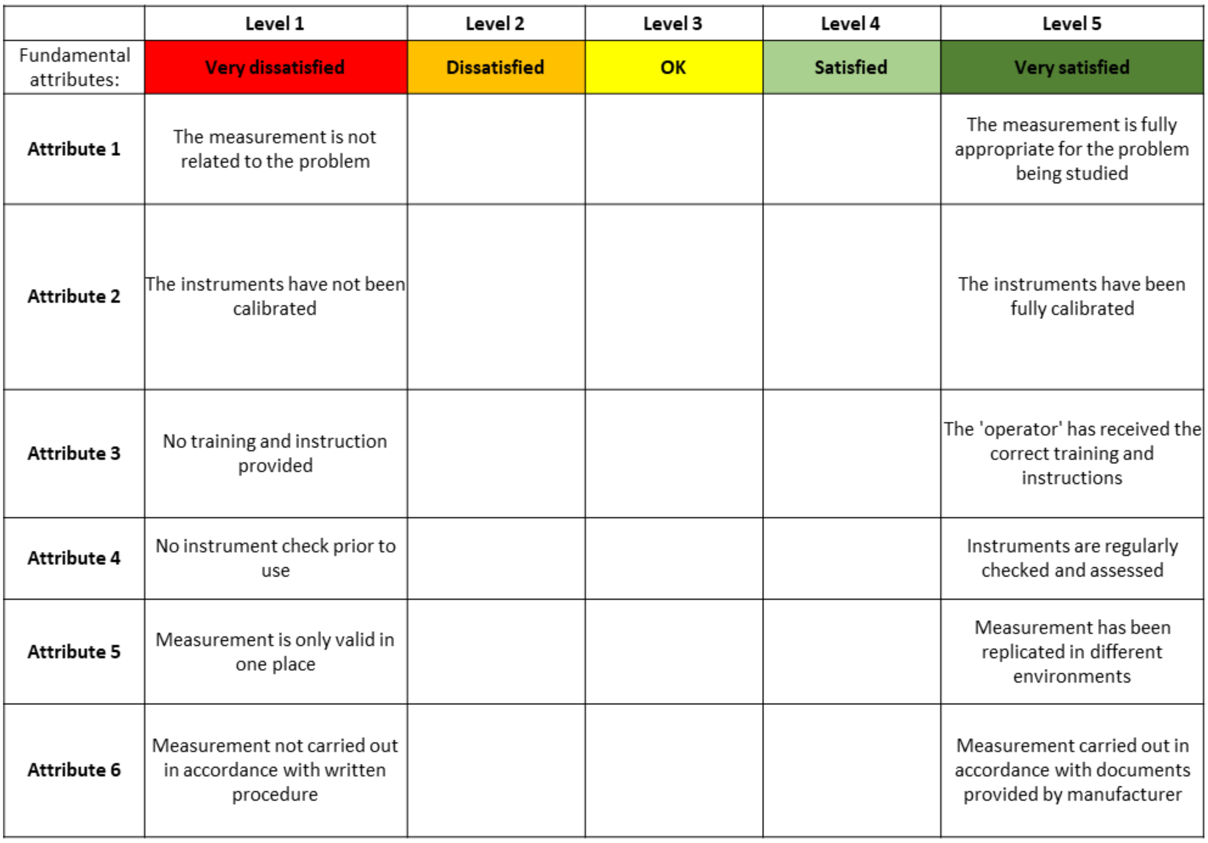
Please fill out the following –

1. Table 2 – Scores for the consensus fundamental attributes
2. Table 4 – Scores for the consensus desirable attributes
3. Please provide notes in the text boxes provided

Consensus Technical Panel Scores:

Please note down the consensus technical panel scores:

Fundamental:



	Level 1	Level 2	Level 3	Level 4	Level 5
Fundamental attributes:	Very dissatisfied	Dissatisfied	OK	Satisfied	Very satisfied
Attribute 1	The measurement is not related to the problem				The measurement is fully appropriate for the problem being studied
Attribute 2	The instruments have not been calibrated				The instruments have been fully calibrated
Attribute 3	No training and instruction provided				The 'operator' has received the correct training and instructions
Attribute 4	No instrument check prior to use				Instruments are regularly checked and assessed
Attribute 5	Measurement is only valid in one place				Measurement has been replicated in different environments
Attribute 6	Measurement not carried out in accordance with written procedure				Measurement carried out in accordance with documents provided by manufacturer

Table 1: *Guidance for assigning the fundamental attributes. The full list of attributes and their descriptions can be found in the manual. The scale is a measure of how well the attribute possesses the attribute. A score of 1 indicates 'very dissatisfied' and a score of 5 indicates 'very satisfied'.* ***You will not need to fill in this table.***

Attributes	**Score 1-5**	**Justification**
1		
2		
3		
4		
5		
6		

Table 2: *Score table for the fundamental attributes. Please note any justifications provided by the technical panel for the scores assigned, i.e. personal knowledge and expertise, references, etc.*

Desirable attributes:

	Level 1	Level 2	Level 3	Level 4	Level 5
Desirable attributes:	**Very dissatisfied**	**Dissatisfied**	**OK**	**Satisfied**	**Very Satisfied**
Attribute 1	No measures have been in put in place to minimise electrical noise disruptions				Proper earthing of equipment ensured
Attribute 2	The object to be measured varies and no corrections have been applied				N/A or environmental conditions have been controlled, corrections and averages applied or dynamic measurements taken
Attribute 3	Measurement technique has not been well-designed and is not representative				Measurement technique is representative of important variations and there is knowledge of any expected changes
Attribute 4	The operator is not skilled				The operator has correctly set up the measuring equipment and prepared the thing to be measured
Attribute 5	Environmental factors not taken into account				Corrections applied to take account of any environmental factors

Table 3: *Guidance for assigning the desirable attributes. The full list of attributes and their descriptions can be found in the manual. The scale is a measure of how well the attribute possesses the attribute. A score of 1 indicates 'very dissatisfied' and a score of 5 indicates 'very satisfied'.* ***You will not need to fill in this table.***

Attributes	Score 1-5	Justification
7		
8		
9		
10		
11		

Table 4: *Score table for the desirable attributes. Please note any justifications provided by the technical panel for the scores assigned, i.e. personal knowledge and expertise, references, etc.*

Please use this space to make note of any of the following things which may have occurred during the Technical Panel meeting:

- Difficulties
- Significant disagreements
- Associated issues
- Length of meeting

Subject matter expert meeting:

- The aim of this meeting is to discuss the subject matter experts review of the scores assigned to the attributes by the technical panel.

Notes from the subject matter expert meeting:

Final conclusion set by the subject matter experts:

Initial quality statement:

The decision-maker has set the following quality specification:

Attribute	Quality statement
Fundamental	
1. The right measurements	
2. The right tools	
3. The right people	
4. Regular review	
5. Demonstrable consistency	
6. The right procedures	
Desirable	
7. Instruments	
8. The object to be measured	
9. Sampling	
10. Operator skill	
11. Environmental factors	

This specification is not to be viewed by any member of the technical panel or the subject matter expert. This is to avoid the introduction of positive bias. You do not need to do anything with this quality statement.

(End of form)

Role: Technical Panel

Please review the manual document before filling out this form

Name:

Profession:

Expertise Rating:

Which of the following describes the level of expertise you have regarding this DIC case study?

Choose an item.

How confident are you in the expertise rating you have allocated?

☐ High

☐ Medium

☐ Low

Please provide some details of your relevant experience:

………

Technical panel member responsibilities:

1. Assign scores to attributes individually.
2. Meet with other technical panel members to discuss individual scores and reach a group consensus.

Form directions:

Please fill out the following –

1. Attribute Weightings; Table 2
2. Table 4 – Scores for the fundamental attributes
3. Table 6 – Scores for the desirable attributes

Attribute Weightings:

The table below contains the list of attributes which you will be scoring in this case study. The 1-5 score you assign will be a measure of how well the dataset possesses that particular attribute. Please review the case study information which is outlined in the manual document provided and use this information to rate the importance of the attributes:

Fundamental attributes – *Six guiding principles that should be followed in order to obtain a good result.*	**Desirable attributes –** *Additional factors which affect measurements*
1. The right measurements	7. Instruments
2. The right tools	8. The object to be measured
3. The right people	9. Sampling
4. Regular review	10. Operator skill
5. Demonstrable consistency	11. Environmental factors
6. The right procedures	

Table 1: *Table containing six fundamental attributes and six desirable attributes. Both sets of attributes have been taken from the National Physical Laboratory Good Measurement Practice Guide.*

Please review the list of attributes and rate how important you feel they are in the context of the case study:

1. Not important
2. Preferred
3. Important
4. Highly desirable
5. Essential

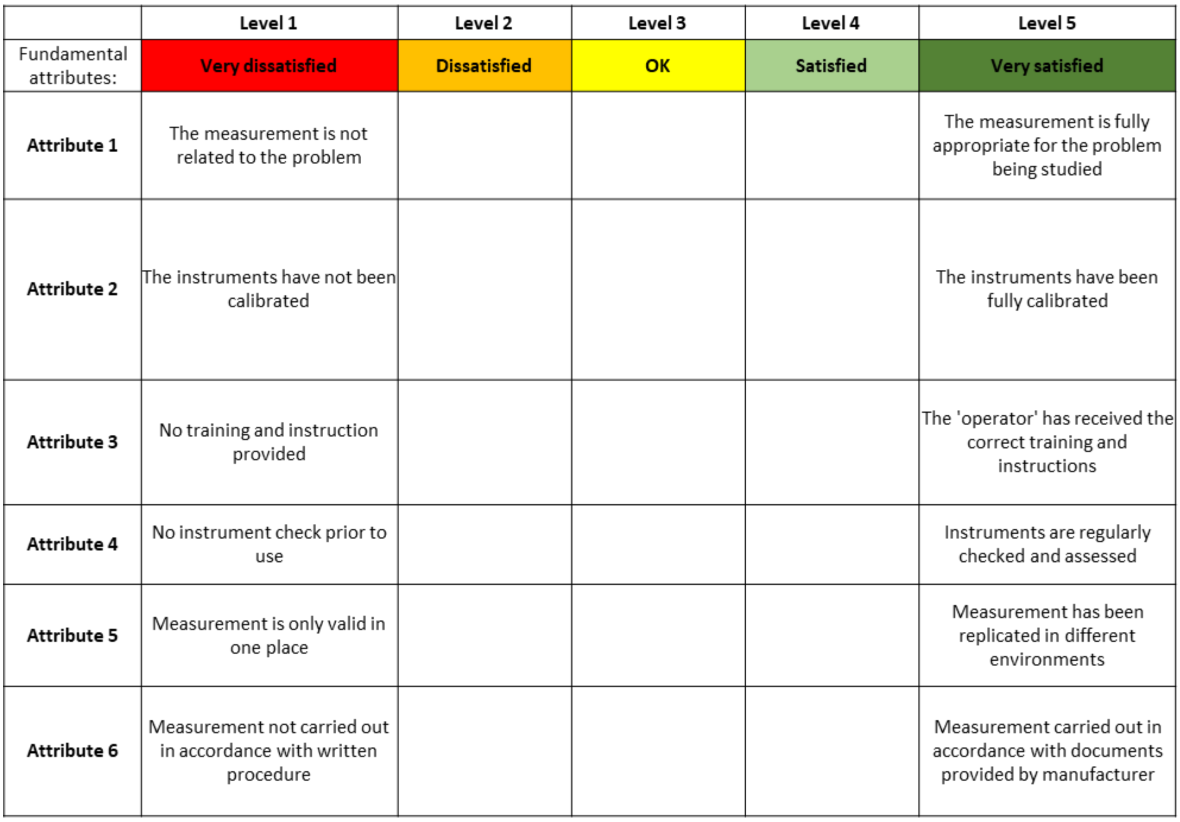
Attribute	**Importance**
Fundamental	
1. The right measurements	Choose an item.
2. The right tools	Choose an item.
3. The right people	Choose an item.
4. Regular review	Choose an item.
5. Demonstrable consistency	Choose an item.
6. The right procedures	Choose an item.
Desirable	
7. Instruments	Choose an item.
8. The object to be measured	Choose an item.
9. Sampling	Choose an item.
10. Operator skill	Choose an item.
11. Environmental factors	Choose an item.

Table 2: *Table for importance ratings for the fundamental and desirable attributes.*

Initial attribute scores:

These scores are to be assigned individually, prior to your meeting with the technical panel. Please use the DIC data PDF file for this exercise. This will contain all the information you need to assign the scores.

Fundamental attributes:



	Level 1	Level 2	Level 3	Level 4	Level 5
Fundamental attributes:	Very dissatisfied	Dissatisfied	OK	Satisfied	Very satisfied
Attribute 1	The measurement is not related to the problem				The measurement is fully appropriate for the problem being studied
Attribute 2	The instruments have not been calibrated				The instruments have been fully calibrated
Attribute 3	No training and instruction provided				The 'operator' has received the correct training and instructions
Attribute 4	No instrument check prior to use				Instruments are regularly checked and assessed
Attribute 5	Measurement is only valid in one place				Measurement has been replicated in different environments
Attribute 6	Measurement not carried out in accordance with written procedure				Measurement carried out in accordance with documents provided by manufacturer

Table 3: *Guidance for assigning the fundamental attributes. The full list of attributes and their descriptions can be found in the manual. The scale is a measure of how well the attribute possesses the attribute. A score of 1 indicates 'very dissatisfied' and a score of 5 indicates 'very satisfied'.* ***You will not need to fill in this table.***

Attributes	**Score 1-5**	**Justification**
1		
2		
3		
4		
5		
6		

Table 4: *Score table for the fundamental attributes. Please note any justifications for the scores assigned, i.e. personal knowledge and expertise, references, etc.*

Desirable attributes:

	Level 1	Level 2	Level 3	Level 4	Level 5
Desirable attributes:	**Very dissatisfied**	**Dissatisfied**	**OK**	**Satisfied**	**Very Satisfied**
Attribute 1	No measures have been in put in place to minimise electrical noise disruptions				Proper earthing of equipment ensured
Attribute 2	The object to be measured varies and no corrections have been applied				N/A or environmental conditions have been controlled, corrections and averages applied or dynamic measurements taken
Attribute 3	Measurement technique has not been well-designed and is not representative				Measurement technique is representative of important variations and there is knowledge of any expected changes
Attribute 4	The operator is not skilled				The operator has correctly set up the measuring equipment and prepared the thing to be measured
Attribute 5	Environmental factors not taken into account				Corrections applied to take account of any environmental factors

Table 5: *Guidance for assigning the desirable attributes. The full list of attributes and their descriptions can be found in the manual. The scale is a measure of how well the attribute possesses the attribute. A score of 1 indicates 'very dissatisfied' and a score of 5 indicates 'very satisfied'.* ***You will not need to fill in this table.***

Attributes	**Score 1-5**	**Justification**
7		
8		
9		
10		
11		

Table 6: *Score table for the desirable attributes. Please note any justifications for the scores assigned, i.e. personal knowledge and expertise, references, etc.*

Please take your individual scores and justifications with you to the group technical panel meeting.

(End of form)

Template:

Subject Matter Expert

Name:

Profession:

Expertise Rating:

Which of the following describes the level of expertise you have regarding this DIC case study?

Choose an item.

How confident are you in the expertise rating you have allocated?

☐ High

☐ Medium

☐ Low

Please provide some details of your relevant experience:

………

Subject matter expert member responsibilities:

1. Assign scores to attributes individually.
2. Review scores assigned to attributes by the technical panel and debate if they are acceptable.

Form directions:

Please fill out the following –

1. Attribute Weightings; Table 2
2. Table 4 – Scores for the fundamental attributes
3. Table 6 – Scores for the desirable attributes

Attribute Weightings:

The table below contains the list of attributes which you will be scoring in this case study. The 1-5 score you assign will be a measure of how well the dataset possesses that particular attribute. Please review the case study information which is outlined in the manual document provided and use this information to rate the importance of the attributes:

Fundamental attributes – *Six guiding principles that should be followed in order to obtain a good result.*	**Desirable attributes –** *Additional factors which affect measurements*
1. The right measurements	7. Instruments
2. The right tools	8. The object to be measured
3. The right people	9. Sampling
4. Regular review	10. Operator skill
5. Demonstrable consistency	11. Environmental factors
6. The right procedures	

Table 1: *Table containing six fundamental attributes and five desirable attributes. Both sets of attributes have been taken from the National Physical Laboratory Good Measurement Practice Guide.*

Please review the list of attributes and rate how important you feel they are in the context of the case study:

1. Not important
2. Preferred
3. Important
4. Highly desirable
5. Essential

Attribute	**Importance**
Fundamental	
1. The right measurements	Choose an item.
2. The right tools	Choose an item.
3. The right people	Choose an item.
4. Regular review	Choose an item.
5. Demonstrable consistency	Choose an item.
6. The right procedures	Choose an item.
Desirable	
7. Instruments	Choose an item.
8. The object to be measured	Choose an item.
9. Sampling	Choose an item.
10. Operator skill	Choose an item.
11. Environmental factors	Choose an item.

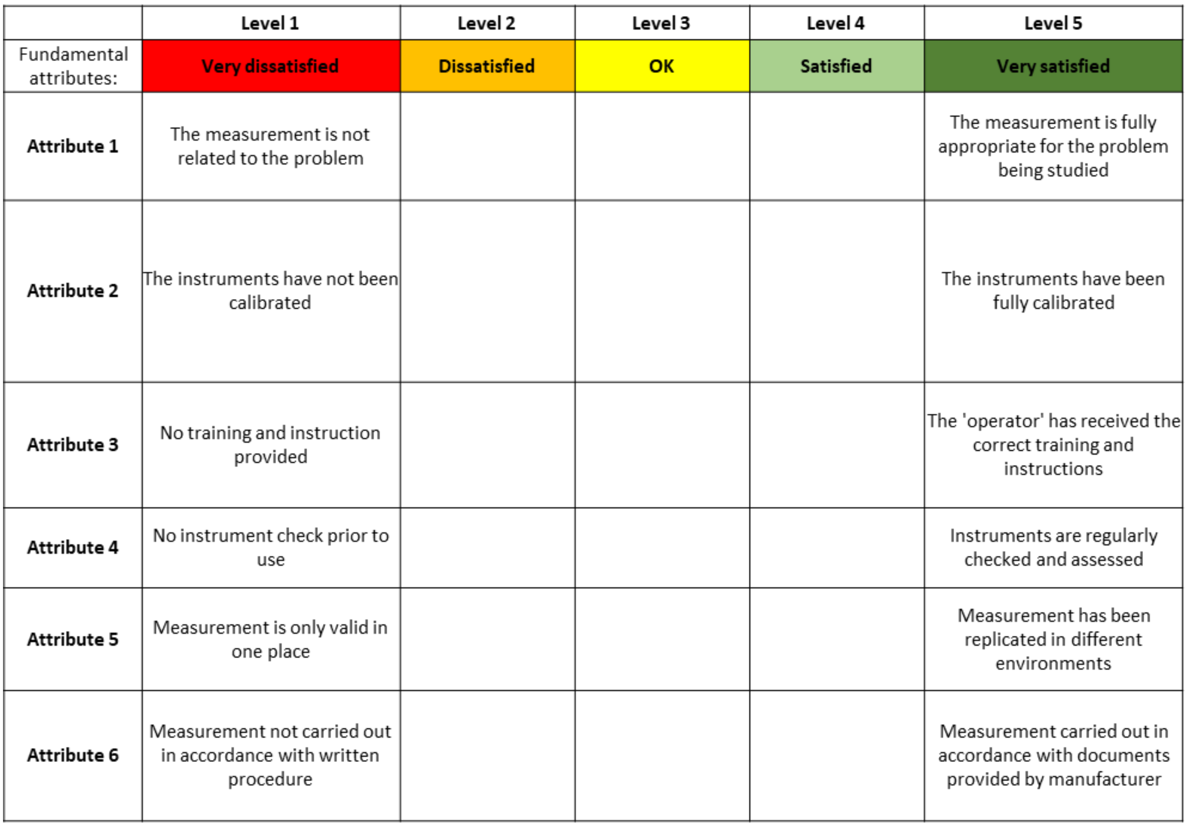
Table 2: *Table for importance ratings for the fundamental and desirable attributes.*

Attribute scores:

These scores are to be assigned individually. During your subject matter expert meeting, you can compare your scores with those obtained by the technical panel to determine if they are acceptable.

Please use the DIC data PDF file for this exercise. This will contain all the information you need to assign the scores.

Fundamental attributes:



	Level 1	Level 2	Level 3	Level 4	Level 5
Fundamental attributes:	Very dissatisfied	Dissatisfied	OK	Satisfied	Very satisfied
Attribute 1	The measurement is not related to the problem				The measurement is fully appropriate for the problem being studied
Attribute 2	The instruments have not been calibrated				The instruments have been fully calibrated
Attribute 3	No training and instruction provided				The 'operator' has received the correct training and instructions
Attribute 4	No instrument check prior to use				Instruments are regularly checked and assessed
Attribute 5	Measurement is only valid in one place				Measurement has been replicated in different environments
Attribute 6	Measurement not carried out in accordance with written procedure				Measurement carried out in accordance with documents provided by manufacturer

Table 3: *Guidance for assigning the fundamental attributes. The full list of attributes and their descriptions can be found in the manual. The scale is a measure of how well the attribute possesses the attribute. A score of 1 indicates 'very dissatisfied' and a score of 5 indicates 'very satisfied'.* ***You will not need to fill in this table.***

Attributes	Score 1-5	Justification
1		
2		
3		
4		
5		
6		

Table 4: *Score table for the fundamental attributes. Please note any justifications for the scores assigned, i.e. personal knowledge and expertise, references, etc.*

Desirable attributes:

	Level 1	Level 2	Level 3	Level 4	Level 5
Desirable attributes:	**Very dissatisfied**	**Dissatisfied**	**OK**	**Satisfied**	**Very Satisfied**
Attribute 1	No measures have been in put in place to minimise electrical noise disruptions				Proper earthing of equipment ensured
Attribute 2	The object to be measured varies and no corrections have been applied				N/A or environmental conditions have been controlled, corrections and averages applied or dynamic measurements taken
Attribute 3	Measurement technique has not been well-designed and is not representative				Measurement technique is representative of important variations and there is knowledge of any expected changes
Attribute 4	The operator is not skilled				The operator has correctly set up the measuring equipment and prepared the thing to be measured
Attribute 5	Environmental factors not taken into account				Corrections applied to take account of any environmental factors

Table 5: *Guidance for assigning the desirable attributes. The full list of attributes and their descriptions can be found in the manual. The scale is a measure of how well the attribute possesses the attribute. A score of 1 indicates 'very dissatisfied' and a score of 5 indicates 'very satisfied'.* ***You will not need to fill in this table.***

Attributes	**Score 1-5**	**Justification**
7		
8		
9		
10		
11		

Table 6: *Score table for the desirable attributes. Please note any justifications for the scores assigned, i.e. personal knowledge and expertise, references, etc.*

Please take your individual scores and justifications with you to the subject matter expert meeting.

(End of form)

Appendix D: Case study #3 documentation

This appendix contains the documentation that was sent to the facilitator, technical panel members and subject matter experts for the third data quality case study #3 – thermoacoustic plate. There are three forms listed: one for the facilitator, one for the technical panel members and one for the subject matter experts.

=

Role: Facilitator

Dates:

Meeting with the technical panel: Click or tap to enter a date.

Meeting with subject matter experts: Click or tap to enter a date.

Form directions:

Please fill out the following –

1. Table 1 – Consensus scores provided the technical panel for the National Physical Laboratory attributes
2. Table 2 – Final scores set by the subject matter experts

Please provide notes in the text boxes provided

Group meeting #1 – Meeting with the technical panel

Aim:

- To discuss individual scores assigned to the attributes by the technical panel
- Arrive at a consensus

Agenda:

- Group introductions
- Discussion of individual attribute scores and justifications
- Record consensus scores for attributes, with justifications

Consensus Technical Panel Scores [Action]:

Please note down the consensus technical panel scores:

Please provide any justifications provided by the technical panel for the scores assigned, i.e. personal knowledge, expertise, references etc.

Attributes	Score 1-5	Justification
1. The right measurements		
2. The right tools		
3. The right people		
4. Regular review		
5. Demonstrable consistency		
6. The right procedures		
7. Instruments		
8. The object to be measured		
9. Sampling		
10. Operator skill		
11. Environmental factors		

Table 1: *Consensus scores for the eleven National Physical Laboratory attributes. Please provide consensus scores with justifications. The scores are a measure of how well the data possesses the attribute. A score of 1 indicates 'very dissatisfied' and a score of 5 indicates 'very satisfied'.*

Please use this space to make note of any of the following things which may have occurred during the Technical Panel meeting:

- Difficulties
- Disagreements
- Length of meeting
- Any other comments

Group meeting #2 – Meeting with subject matter experts

Aim:

- Subject matter experts review consensus scores assigned by the technical panel
- They can choose to alter the consensus, with their own justifications

Agenda:

- Group introductions
- Discussion of individual attribute scores and justifications
- Present consensus scores provided by the technical panel, for the subject matter experts to discuss and review
- Record final SME consensus scores

Final Subject Matter Expert Scores [Action]:

The subject matter experts will review the technical panel consensus scores to determine if they are suitable.

Please note down the final consensus scores provided by the subject matter experts.

Please note down any justifications provided.

Attributes	Score 1-5	Justification
1. The right measurements		
2. The right tools		
3. The right people		
4. Regular review		
5. Demonstrable consistency		
6. The right procedures		
7. Instruments		
8. The object to be measured		
9. Sampling		
10. Operator skill		
11. Environmental factors		

Table 2: *Final scores set by the subject matter experts, following review of the technical panel consensus.*

Please use this space to make note of any of the following things which may have occurred during the subject matter expert meeting:

- Difficulties
- Disagreements
- Length of meeting
- Any other comments

When you have completed this form, please email it to cmcel@liverpool.ac.uk

(End of form)

Role: Technical Panel

Please review 'Doc 1 – TP background' before filling out this form

Name:

Profession:

Expertise Rating:

Which of the following describes the level of expertise you have regarding this DIC case study?

Choose an item.

How confident are you in the expertise rating you have allocated?

☐ High

☐ Medium

☐ Low

Please provide some details of your relevant experience:

…….

Technical panel member responsibilities:

1. Assign attribute scores and judged importance scores independently
2. Meet with other technical panel members to discuss individual scores and reach a group consensus.

Form directions:

Please fill out the following –

1. Table 1 – Judged importance scores
2. Table 2 – Attribute scores

[Independent review]

Judged importance scores:

Please review the list of attributes and rate how important you feel they are in the context of the case study:

1. Not important
2. Preferred
3. Important
4. Highly desirable
5. Essential

Attribute	Importance
1. The right measurements	Choose an item.
2. The right tools	Choose an item.
3. The right people	Choose an item.
4. Regular review	Choose an item.
5. Demonstrable consistency	Choose an item.
6. The right procedures	Choose an item.
7. Instruments	Choose an item.
8. The object to be measured	Choose an item.
9. Sampling	Choose an item.
10. Operator skill	Choose an item.
11. Environmental factors	Choose an item.

***Table 1:** Table for judged importance ratings*

Attribute scores:

- Please refer to **Table 2** and **Table 3** in 'Doc 1 – TP background' for guidance when assigning attribute scores.
- Please provide justifications for the scores you have provided.

Attribute	Attribute score (1-5)	Justification
1. The right measurements		
2. The right tools		
3. The right people		
4. Regular review		
5. Demonstrable consistency		
6. The right procedures		
7. Instruments		
8. The object to be measured		
9. Sampling		
10. Operator skill		
11. Environmental factors		

Table 2: *Table for attribute scores. A score of 1 indicates that we are very dissatisfied that the dataset possesses the attribute and a score of 5 indicates that we are very satisfied.*

- When you have completed this form, please send it to the facilitator, Name - Email

Please take your individual attribute scores and justifications with you to the group technical panel meeting.

(End of form)

Role: Subject Matter Expert

Please review 'Doc 1 – SME background' before filling out this form

Name:

Profession:

Expertise Rating:

Which of the following describes the level of expertise you have regarding this DIC case study?

Choose an item.

How confident are you in the expertise rating you have allocated?

☐ High

☐ Medium

☐ Low

Please provide some details of your relevant experience:

……….

Subject matter expert member responsibilities:

1. Assign attribute scores and judged importance scores independently
2. Meet with the other subject matter expert to review consensus scores provided by the technical panel

Form directions:

Please fill out the following –

1. Table 1 – Judged importance scores
2. Table 2 – Attribute scores

[Independent review]

Judged importance scores:

Please review the list of attributes and rate how important you feel they are in the context of the case study:

1. Not important
2. Preferred
3. Important
4. Highly desirable
5. Essential

Attribute	Importance
1. The right measurements	Choose an item.
2. The right tools	Choose an item.
3. The right people	Choose an item.
4. Regular review	Choose an item.
5. Demonstrable consistency	Choose an item.
6. The right procedures	Choose an item.
7. Instruments	Choose an item.
8. The object to be measured	Choose an item.
9. Sampling	Choose an item.
10. Operator skill	Choose an item.
11. Environmental factors	Choose an item.

Table 1: *Table for judged importance ratings*

Attribute scores:

- Please refer to **Table 2** and **Table 3** in 'Doc 1 – SME background' for guidance when assigning attribute scores.
- Please provide justifications for the scores you have provided.

Attribute	Attribute score (1-5)	Justification
1. The right measurements		
2. The right tools		
3. The right people		
4. Regular review		
5. Demonstrable consistency		
6. The right procedures		
7. Instruments		
8. The object to be measured		
9. Sampling		
10. Operator skill		
11. Environmental factors		

Table 2: *Table for attribute scores. A score of 1 indicates that we are very dissatisfied that the dataset possesses the attribute and a score of 5 indicates that we are very satisfied.*

- When you have completed this form, please send it to the facilitator, Name - Email

Please take your individual attribute scores and justifications with you to the group subject matter expert meeting.

(End of form)

bv

Table E1: Scores to describe possession of attributes assigned by the technical panel for **case study #1– tensile plate with a hole.**

Case study #1 - Tensile plate
Scoring to describe possession of the attributes:

Attribute:	TP #1	TP #2	TP #3	TP #4	TP #5	Average	Std	TP Consensus
The right measurements	5	3	4	3	3	3.60	0.89	4.00
The right tools	3	4	5	3	4	3.80	0.84	3.00
The right people	4	4	5	4	5	4.40	0.55	4.00
Regular review	4	3	4	4	4	3.80	0.45	4.00
Demonstrable consistency	3	3	3	3	3	3.00	0.00	3.00
The right procedures	4	4	4	3	4	3.80	0.45	4.00
Instruments	2	3	3	2	2	2.40	0.55	2.00
Object to be measured	3	4	5	2	3	3.40	1.14	4.00
Sampling	4	3	4	3	3	3.40	0.55	4.00
Operator skill	4	3	4	4	4	3.80	0.45	3.00
Environmental factors	3	3	3	3	2	2.80	0.45	3.00

Table E2: Scores to describe possession of attributes and judged importance scores assigned by panel of participants for **case study #2 – impact on bonnet liner.**

Case study #2 - Impact on bonnet liner

Scoring to describe possession of the attributes:

Attribute:	TP #1	TP #2	TP #3	TP #4	Average	Std	TP Consensus	SME #1	SME #2	Average	Std	SME consensus
The right measurements	5	5	4	4	4.50	0.58	4.00	5	4	4.50	0.71	5.00
The right tools	4	5	5	4	4.50	0.58	4.00	3	4	3.50	0.71	4.00
The right people	5	4	5	4	4.50	0.58	5.00	3	3	3.00	0.00	3.00
Regular review	3	3	5	5	4.00	1.15	4.00	3	3	3.00	0.00	4.00
Demonstrable consistency	4	3	3	4	3.50	0.58	3.00	5	2	3.50	2.12	3.00
The right procedures	5	5	5	4	4.75	0.50	4.00	3	3	3.00	0.00	3.00
Instruments	3	5	5	5	4.50	1.00	5.00	3	4	3.50	0.71	3.00
Object to be measured	5	5	4	5	4.75	0.50	5.00	5	4	4.50	0.71	5.00
Sampling	4	4	5	5	4.50	0.58	4.00	5	4	4.50	0.71	5.00
Operator skill	5	3	5	4	4.25	0.96	4.00	4	3	3.50	0.71	3.00
Environmental factors	3	3	3	4	3.25	0.50	3.00	5	2	3.50	2.12	3.00

Importance of attributes:

Attribute:	TP #1	TP #2	TP #3	Tp #4	Average	Std	SME #1	SME #2	Average	Std
The right measurements	4	4	5	5	4.50	0.58	5	5	5.00	0.00
The right tools	5	4	5	5	4.75	0.50	5	5	5.00	0.00
The right people	5	5	4	4	4.50	0.58	4	4	4.00	0.00
Regular review	3	4	3	2	3.00	0.82	4	4	4.00	0.00
Demonstrable consistency	5	1	4	4	3.50	1.73	3	5	4.00	1.41
The right procedures	4	5	5	4	4.50	0.58	3	3	3.00	0.00
Instruments	3	4	4	4	3.75	0.50	5	4	4.50	0.71
Object to be measured	4	5	5	3	4.25	0.96	2	3	2.50	0.71
Sampling	4	5	5	2	4.00	1.41	2	5	3.50	2.12
Operator skill	4	5	4	4	4.25	0.50	4	3	3.50	0.71
Environmental factors	3	3	4	3	3.25	0.50	1	2	1.50	0.71

Table E3: Scores to describe possession of attributes and judged importance scores assigned by panel of participants for **case study #3 – thermoacoustic plate.**

Case study #3 - Thermoacoustic plate:

Scoring to describe possession of the attributes:

Attribute:	TP #1	TP #2	TP #3	Average	Std	TP Consensus	SME #1	SME #2	Average	Std	SME consensus
The right measurements	5	3	5	4.33	1.15	4.00	5	5	5.00	0.00	5.00
The right tools	3	3	3	3.00	0.00	3.00	5	5	5.00	0.00	5.00
The right people	3	5	2	3.33	1.53	3.00	4	3	3.50	0.71	3.00
Regular review	3	1	2	2.00	1.00	2.00	3	1	2.00	1.41	3.00
Demonstrable consistency	5	1	2	2.67	2.08	2.00	5	1	3.00	2.83	5.00
The right procedures	3	5	3	3.67	1.15	3.00	3	5	4.00	1.41	3.00
Instruments	5	2	2	3.00	1.73	3.00	1	3	2.00	1.41	3.00
Object to be measured	5	5	4	4.67	0.58	5.00	5	2	3.50	2.12	5.00
Sampling	5	5	2	4.00	1.73	5.00	5	5	5.00	0.00	5.00
Operator skill	3	5	4	4.00	1.00	4.00	4	2	3.00	1.41	4.00
Environmental factors	5	4	4	4.33	0.58	4.00	5	3	4.00	1.41	4.00

Importance of attributes:

Attribute:	TP #1	TP #2	TP #3	Average	Std	SME #1	SME #2	Average	Std
The right measurements	5	5	3	4.33	1.15	5	5	5.00	0.00
The right tools	5	5	4	4.67	0.58	5	3	4.00	1.41
The right people	5	3	3	3.67	1.15	5	1	3.00	2.83
Regular review	3	2	2	2.33	0.58	3	1	2.00	1.41
Demonstrable consistency	5	3	3	3.67	1.15	1	2	1.50	0.71
The right procedures	4	5	4	4.33	0.58	4	5	4.50	0.71
Instruments	4	5	3	4.00	1.00	2	1	1.50	0.71
Object to be measured	5	3	2	3.33	1.53	2	1	1.50	0.71
Sampling	5	3	4	4.00	1.00	5	5	5.00	0.00
Operator skill	5	3	4	4.00	1.00	5	1	3.00	2.83
Environmental factors	3	5	5	4.33	1.15	2	3	2.50	0.71