Calibration of a white light interferometer for the measurement of micro-scale dimensions

Carlo Ferri · **Julian Faraway** · **Emmanuel Brousseau**

the date of receipt and acceptance should be inserted later

Abstract Calibration is central to most measurement procedures. This is especially true in those cases where a large number of difficult-to-identify and difficult-to-control factors hinder the experimenters in their efforts to obtain reliable measurement results. Dimensional measurements of features on the micro- and nano-scale is one such case. A white light interferometer (WLI) microscope can perform measurements of a variety of measurands over a broad dimensional range: from surface texture characterisations on the nano-scale to measurements of step heights of several millimetres. Calibration methods based on the hypothesis of a linear calibration curve can be inadequate to express the relationship between measurement results and traceable reference materials (RM's). A calibration procedure built in a commercially available WLI microscope is critically compared with methods presented in an international standard. This comparison is enabled by a cost-effective procedure for establishing traceable RM's in the micro-range. Advantages of calibration procedures based on more than one RM are then demonstrated within the ranges from 180.5 to 219.5 μm and from 1.5 to 501.5 μm. Calibration methods involving regression modelling of transformed measurement results are considered for these two intervals to overcome the highlighted weaknesses of the calibration procedure built in the examined WLI.

Keywords Calibration · White light interferometry · WLI · regression analysis

Carlo Ferri

Department of Mechanical Engineering, University of Bath

Bath, BA2 7AY, UK

E-mail: c.ferri@bath.ac.uk

Julian Faraway

Department of Mathematical Sciences, University of Bath

Bath, BA2 7AY, UK

Emmanuel Brousseau

Manufacturing Engineering Centre, Cardiff University

Cardiff, CF24 3AA, UK

1 Introduction

The current trend towards product miniaturisation constitutes a driving force for an increasing research interest in micro- and nano-components manufacturing. Metrology has been acknowledged as central to this trend and research interest [18].

Calibration is ‘a set of operations which establish, under specified conditions, the relationship between values indicated by a measuring system and the corresponding accepted values of some “standards” ’ [3]. These ‘standards’ are referred to as reference materials (RM’s), which are substances or artefacts with one or more properties sufficiently well known to be associated with a numerical value called the ‘accepted value’ of the RM.

In standards from the International Organisation for Standardisation (ISO), it is recommended that the freedom from systematic error of measurements obtained using a measurement system should be methodically verified (cf. section 7.1 in BS ISO 11095:1996 [3] and, for a definition of systematic measurement error, section 2.17 in JCGM 200:2008 [13]). This especially holds in the verification of conformance to specifications of features on the micro- or nano-scale. In fact, the specification limits of functional characteristics of micro- and nano- structures are expected to be on the same micro- or nano-scale. Consequently, increasing efforts in the analysis of calibration methods become economically justified in order for a measuring system to discriminate reliably defective parts. They can in fact contribute to eschew erroneous rejection and/or acceptance of parts with their often severe economic implications (e.g. unnecessary reworking and delays).

White light interferometer microscopy is one of the measurement technologies having a prominent position in the miniaturisation trend mentioned above. In fact it provides measurements of surface texture and heights at both the micro and nano-metre scale. Currently, vertical scanning WLI techniques are used in a broad range of applications such as the determination of surface texture parameters [22], film thickness measurements [21], and form measurements [20].

Each of these areas of application presents problems that are specific to the area. For instance, measuring step heights less than the coherence length of the light source results in a problem known as batwings (cf. Harasaki and Wyant [12]). The top plane of the step in the proximity of its edge always appears higher than it actually is, whereas the bottom plane appears lower. The resulting shape of the profile caused by this false information justifies the name given to this problem. In the experiments performed during this investigation, the step heights have always been measured far from the curved edge of the steps beyond the area in which this effect appears. Harasaki and Wyant (cf. figure 1 in [12]) also observed that the batwings effect disappears when considering step heights larger than the coherence length of the source.

In the experiments carried out in this in investigation, the step heights are always larger than the coherence length of the light source, therefore this effect does not occur.

A number of influential factors must be considered in the design of a WLI, its controlling algorithms and its calibration procedure. Among these are the dispersive effects arising from small asymmetries in the interferometer components. For example, in a Linnik interferometer, deviations of a beam-splitter cube from the ideal cubic geometry is one such asymmetries and it may give rise to the so-called ghost steps. [17].

Harasaki *et al.* [11] and Park and Kim [16] identified the optical properties of the surface to be measured as a factor to be taken into account in the design of the calibration procedure. This especially holds when the object to be measured encompasses surfaces made of multiple materials.

The main thrust of this investigation is limited to the consideration of calibration procedures in already designed and commercially available instruments, when performing measurements of step heights on the scale of tens or hundreds micrometres.

According to Hansen *et al.* [10], 'surface interferometry can be considered as a kind of displacement interferometry, just in this case a whole array of photodetectors, i.e. a CCD camera, is used'. It is believed that this consideration helps to clarify the reason why RM's for measurement of lengths such as step heights rather than RM's of surface topography are quite common in calibration procedures suggested by WLI instrument manufacturers. Hansen *et al.* [10] have also reviewed and analised some of the difficulties of providing traceability via calibration in micro- and nanometrology. Particularly, they observed that on the micro- and nano-scale 'few national institutes offer calibrations artefacts or standards for the use in production environments'. General-purpose artefacts widely available in production environments are used in the calibration experiments performed in this investigation.

For the white light interferometer microscope examined in this study, a calibration procedure for performing dimensional measurements in the vertical direction is provided by the manufacturer. The procedure consists in measuring the height of a single RM. Then a value for a parameter called 'height correction' (HC) is calculated by the software of the instrument. This value is output automatically once the measured and accepted values of the RM are given to the software by the operator. The instrument manufacturer, however, does not provide any information about the physical meaning of HC. A need for interpretation is therefore apparent. In this study a comparison between this built-in procedure and good practice as recommended by standards such as the BS ISO 11095:1996 [3] is performed.

The objective of this paper is twofold: first, the limitations engendered by the current calibration procedure of the WLI under investigation are identified and demonstrated.

Second, alternative calibration techniques are considered and fully detailed in order to overcome such limitations.

2 Reference material and experimental set-up

A cost effective and versatile method for the realisation of traceable RM's for measurements of lengths in the micrometre scale along the vertical direction is presented in this study. Henceforth a RM is also referred to as step. The method hinges on the use of certified gauge blocks of grade 1 [4] traceable according to BS 4311-3:1993 [2] to the UK national realisation of the unit of length. According to this national standard, the compliance to pre-specified tolerances is verified for both the central length and the variation in length.

An auxiliary plate with a planar surface of the same texture as the measuring faces of the gauge blocks is necessary to meet the definition of length of a gauge block (cf. BS EN ISO 3650:1999 [4]). Like Decker and Pekelsky [6] and like Malinovsky et al. [15,14], quartz optical parallels were used as auxiliary plates. The parallelism and flatness tolerance specifications of these optical parallels (0.2 and 0.1 μm, respectively) are comparable with those imposed on the gauge blocks (cf. table 3 in BS EN ISO 3650:1999 [4]). The use of transparent optical parallels enabled the quality of the wringing procedure to be assessed by detecting the presence of interference colour fringes and bright spots on the two wrung faces by observing them through the quartz optical parallels.

Two blocks were used in preparing each step. They were wrung side by side onto a quartz optical parallel. A photograph of a generic artefact obtained with this method is displayed in Figure 1.

[Fig. 1 about here.]

For a RM a nominal value of length is defined by the following equation:

$$l_{n,s} = l_{n,1} - l_{n,2} \qquad (l_{n,1} > l_{n,2}) \tag{1}$$

In equation 1, $l_{n,1}$ and $l_{n,2}$ are the nominal lengths of the two gauge blocks as defined in BS EN ISO 3650:1999 [4].

If the actual length of the *i-th* block in a generic point on the unwrung measuring face is L_i ($i = 1,2$), then the actual step height in two generic points on the unwrung measuring faces of the two blocks, L_s, is given by the following equation:

$$L_s = L_1 - L_2 \tag{2}$$

In equation 2, the actual length of the step, L_s, is changing when calculated in relation to different points on the two blocks. In fact, the lengths L_1 and L_2 are expected to vary from point to point on the unwrung measuring face of the first and second gauge block, respectively. This circumstance is illustrated in Figure 2, where t_e is the semi-amplitude of the specification interval for the length of the

i-th gauge block, L_i, expressed as deviation from its nominal $l_{n,i}$ (cf. table 5 in BS EN ISO 3650:1999 [4]).

[Fig. 2 about here.]

The term 'step height' without further specification is hereafter used to identify a generic L_s of a given step. After a few passages, it can be shown that for every possible pair of points considered on the two measuring faces of the two blocks the following equation holds:

$$l_{n,s} - 2 \cdot t_e \leq L_s \leq l_{n,s} + 2 \cdot t_e \tag{3}$$

All the gauge blocks used in this investigation had nominal lengths in the range between 0.5 and 10 mm. For such a range and for blocks of grade 1, the international standard BS EN ISO 3650:1999 [4] in Table 5 prescribes $t_e = 0.2\,\mu$m. According to equation 3 therefore, for any calibrated blocks compliant with their specifications, the step height of the RM's built with the proposed method should always theoretically lie within $\pm 0.4\,\mu$m of the nominal step height $l_{n,s}$.

In practice, however, a number of events may occur during the process of building a RM that can cause the step height to deviate from its expected theoretical specification interval. Among these there are the handling of the blocks and the wringing process.

While building the RM, the blocks have never been touched or put in contact with bare hands. Prolonged contact of a block with a body at an average temperature of 34.5 °C, such as an human hand [5], can cause the block length to increase from its value at the reference temperature (usually 20 °C). An estimate of this change in length, for a generic gauge block, is given by

$$\Delta l = \alpha_{th} \cdot L \cdot \Delta T \tag{4}$$

In equation 4 L is the length of a block in a generic point on the unwrung measuring face and is measured in millimetres at 20 °C. The coefficient of linear thermal expansion α_{th} is given in the calibration certificate of the used set of blocks and is equal to $10.8 \cdot 10^{-6}\,\mathrm{K}^{-1}$. Therefore, for the worst case scenario mentioned above and expressing Δl in nanometres, it follows:

$$\Delta l = 10.8 \cdot 10^{-6} \cdot 10^{6} \cdot L \cdot (34.5 - 20) = 156.6 \cdot L \tag{5}$$

Decker and Peckelsky [6] suggested leaving the blocks untouched overnight to stabilise them thermally after handling. The study from Scarr [19], in which it is shown the cooling curve for a 25.4 mm slip gauge that was held in the hand for three minutes and then put in a stable ambient temperature, supports the idea that 15 minutes of thermal stabilisation would be sufficient. This cooling curve in fact showed that the deviation from the nominal length went from $2.032\,\mu$m to about $0.254\,\mu$m after 15 minutes and to only about $0.127\,\mu$m after 30 minutes. On the basis of these considerations, when building the RM's the blocks were always handled wearing thermally insulating gloves and waiting for at least 15 minutes before using them.

Wringing is the procedure by which a gauge block is made to adhere tightly to the surface of the auxiliary plate, so that the block is fixed firmly in position. The two units, the block and the plate, when wrung, can also be handled as a single part (cf. Decker and Pekelsky [6]). Although Decker and Pekelsky [6] suggested putting a very small amount of light textured oil on the surface of the plate and then cleaning it with a lint-free cloth, Malinovsky et al. [14] showed that interposition of oil between the plate and the block had the effect of increasing the deviation from the nominal length of the block. They also showed that the space between the plate and the block increases with time due to the presence of oil. In addition, neither BS 4311-3:1993 [2] in section 8.2 nor BS EN ISO 3650:1999 [4] in section A.4 mentioned the usage of oil. Thus, no oil or any other substance was used in the wringing of the gauge blocks on the optical parallel surface.

In this investigation the accepted value of a RM was set as the nominal value $l_{n,s}$.

Within the limits of the above observations, the expected actual values of step height is always inside the theoretical specification interval $l_{n,s} \pm 0.4\,\mu$m. This interval may be considered of negligible amplitude in a number of applications on the micrometre scale.

This method of building RM's on the micro-metre scale offers two main practical advantages over possible alternatives. On the one hand, it is cost-effective when compared with the purchase of expensive ad hoc calibrated artefacts. It is in fact based on gauge blocks, which are most common in traditional workshops. On the other hand, this method enables a large number of different steps on the micro-scale to be built.

The equipment used in this study is located in a thermostatically controlled laboratory with temperature set point 20 °C and with calm and dust free air conditions as typically expected in a generic industrial measurement laboratory. Moreover, the WLI has been placed on an ad hoc base that uses compressed air to reduce the transmission of vibrations from the surrounding environment to the instrument.

The WLI used in this study utilises a charge-coupled device (CCD) camera to record the intensity of bright and dark fringes for each pixel in the field of view while the object under observation is scanned perpendicular to its illuminated surface with a vertical movement of the interferometric objective (scanner element). The distance covered during this vertical movement is called scan length. In this study, unless otherwise specified, the scan length and the initial position of the scanner have been selected in a consistent manner in each of the measurement tests performed. In fact, Ferri and Brousseau [8] ascertained the presence of a significant effect of these two set-up parameters on the measurement results.

3 Current calibration procedure

The interpretation of the parameter HC provided in this section stems from the comparison between the calibration procedure built in the measurement system and the one-point calibration method described in the BS ISO 11095:1996 (cf. [3], section 8.2).

Repeated measurements of an RM are taken. In the one-point calibration method, a calibration curve is derived by fitting a first order linear regression model of the measured values on the accepted value of the RM and by assuming the intercept to be null. The only unknown regression coefficient β_1, i.e. the slope of a straight line, is estimated by using the ordinary least squares method (OLS), which results in the following estimate $\hat{\beta}_1$ of β_1:

$$\hat{\beta}_1 = \frac{\bar{h}_{\cdot}}{l_{n,s}} \tag{6}$$

In equation 6, the symbol $\bar{h}_{\cdot}$ is the average of all the measurement results performed on the RM with accepted value $l_{n,s}$. In this way, any subsequent measurement result, h, is obtained as follows:

$$h = \frac{h_{raw}}{\hat{\beta}_1} \tag{7}$$

where h_{raw} is the 'measurement result' before any 'correction' is applied (cf. sections 2.9 and 2.53 in JCGM 200:2008 [13] for a definition of 'measurement result' and 'correction', respectively).

To explain the meaning of HC, the built-in calibration procedure was assumed equivalent to the one-point calibration method described in the BS ISO 11095:1996 [3]. Therefore, HC is defined by the following equation:

$$\mathrm{HC} = \frac{1}{\hat{\beta}_1} \tag{8}$$

If equation 8 is true, then from the data of a calibration experiment and from equation 8, a value for HC, i.e. HC_0, is computed and a generic measurement result h is obtained from:

$$h = \mathrm{HC}_0 \cdot h_{raw} \tag{9}$$

where h is the value available to the operator and h_{raw} is internal to the system. Therefore, the uncorrected value h_{raw} is not directly accessible to the operator, unless $\mathrm{HC}_0 = 1$.

While measuring an RM of known length $l_{n,s,1}$, if the measurement

$$h_1 = \mathrm{HC}_0 \cdot h_{raw,1} \tag{10}$$

is significantly different from $l_{n,s,1}$, then the built-in calibration procedure is performed, i.e. a new value HC_1 for the HC parameter is obtained by imposing $h_1 = l_{n,s}$ in equation 10, namely:

$$l_{n,s,1} = \mathrm{HC}_1 \cdot h_{raw,1} \tag{11}$$

From equation 10 and 11 it follows

$$\mathrm{HC}_1 = \frac{l_{n,s,1}}{h_1} \cdot \mathrm{HC}_0 \tag{12}$$

The validity of equation 11 was verified by giving the same triplet $(l_{n,s,i}, h_i, \mathrm{HC}_0)$ as input to the software of the instrument and to an implementation of equation 12. The same

result for the HC parameter (HC_1) was obtained in both the procedures.

On the basis of the evidence presented in this section, the calibration procedure built in the WLI was identified with the one-point calibration described in the BS ISO 11095:1996 (cf. [3], section 8.2).

4 Limitations

An experiment was carried out in order to verify whether the one-point calibration method proposed by the manufacturer was suitable for measuring heights in the micrometer range.

The one-point calibration hinges on the assumption of linearity of the calibration function. Such a hypothesis was therefore experimentally tested. Four RM's with nominal dimensions 150, 183, 217, and 250 μm were built according to the procedure described in section 2 and were measured in these tests.

If the hypothesis of linearity is true, then the deviations of the measurement results from the nominal dimensions of the measured steps should exhibit a linear relationship.

Figure 3 shows the deviations of the measurement results, in excess of the semi-amplitude of the step specification interval (0.4 μm) and in absolute value, against the nominal dimensions of the step heights. Qualitatively, this figure shows experimental evidence for rejecting the hypothesis of linearity.

[Fig. 3 about here.]

To support this observation quantitatively, first and second order linear regression models were fitted to the data. The adequacy of the fitting was then assessed using the Akaike Information Criterion (AIC) [1]. The pair (AIC, R^2) was (195.45, 57.80 %) and (152.54, 79.08 %) for the first and the second order model, respectively, where R^2 is the coefficient of determination (cf. section 2.7 in Faraway [7]). The second order model has the lower AIC value. Therefore, it provides a better interpretation of the data.

This analysis provides experimental evidence to reject the linearity hypothesis.

To confirm the validity of the assumptions underlying the fitted models, their realised residuals were analysed graphically. No violations of the assumptions of constant variance and independence of the errors were observed.

In Figure 3 it is also apparent that a large contribution to the departure from linearity of the calibration function is due to the deviations associated with the RM of nominal length 183 μm.

The possibility that nuisance factors affected the process of building this step height cannot be excluded. This possibility is also supported by the circumstance that the second order model is not monotone on the interval of the reference materials considered. When measuring increasing nominal step heights, the occurrence of decreasing measurement results may in fact appear incongruous. A possible reason for

these unexpected measurement results may also be found with an investigation of the characteristics of motion of the WLI scanner element, rather than the build process of the RM's. More details regarding the centrality and criticality of the motion characteristics of the WLI scanner element can be found in Schmit and Olszak [20].

On the other hand, it appears unlikely that the vertical motion of the scanner element affected only the measurements of the RM 183 μm high, but not the other three. Also, as the RM's have been measured in a random sequence, it is unlikely that any lurking influential environmental condition such as the occurrence of vibrations would have affected solely the measurements taken on the RM 183 μm.

If the measurements of the step height with nominal length 183 μm are discarded, the first order model would have been preferable and the analysis presented in this section might have lead to confirming the hypothesis of linearity of the calibration function for the range of lengths and the WLI examined. However, even in such circumstances, the manufacturer's one-point calibration method would still be inadequate for measurement tasks on the micro-metre scale. In fact, the intercept term in the first order regression model of Figure 3 is at 6.37 μm, rather than being zero as the one-point calibration prescribes [3]. The presence of such a non zero intercept is therefore incompatible with the one point calibration.

In addition, a qualitative graphical observation, suggests that this intercept would increase in value if the measurement results of step height 183 μm were compliant with the linearity assumption.

On the basis of this experimental evidence, it is concluded that the built-in calibration procedure of the WLI studied is not satisfactory in the range of heights investigated. Alternative calibration procedures for a narrow range, i.e. the interval from 180.5 to 219.5 μm, and a wide range, i.e. the interval from 1.5 to 501.5 μm, respectively, are presented in the next two sections.

5 Calibration on a narrow range

Proper analysis of calibration data is essential to quantify bias and uncertainty performance of a measurement system (cf. section 2.18 and 2.26 in JCGM 200:2008 [13] for a definition of bias and uncertainty, respectively). Here an analysis is described in some detail as a demonstration of how the authors believe this should be done. Such an analysis does not contradict the recommendations of the BS ISO 11095:1996 [3], but is intended to complement it in providing examples of problems that may occur in practice.

As suggested in the BS ISO 11095:1996 [3], several RM's should be considered during a calibration experiment. The RM's should be selected so that they evenly span the region where the measuring system is to be deployed.

The number of step heights considered for this calibration is selected so that a $3\,\mu$m interval separates each of their accepted values. This separation corresponds to ± 3 times the estimate of repeatability standard deviation, i.e. $\hat{\sigma}_r$, provided in Ferri et al. [9]. A $6 \cdot \hat{\sigma}_r$ amplitude for an interval appears consistent with the rule of thumb, widespread in capability studies, of having 99.73 % of the expected WLI measurement results lying in this interval if they are normally distributed. At the same time, such an interval length, $3\,\mu$m, does not leave uncovered areas between adjacent RM's. The term 'uncovered' refers to sets of measurement results having infinitesimal probability of being originated by a particular RM.

A maximum of seven step heights could be built for an experiment. So the measurement test were carried out in two different experimental units (henceforth, also referred to as 'blocks of data' or 'blocks'). In the first, the RM's with nominal lengths 182, 185, 188, 191, 194, 197 and 200 μm were considered, whereas in the second the RM's with nominal lengths 203, 206, 209, 212, 215, 218 μm were used. The experiment was replicated three times. Given an experimental unit, the assignment of the RM's to the sequence of the tests was randomly chosen by assigning a label to each test in the sequence and randomly generating a permutation. All the tests belonging to one experimental unit were run first and then those of the other were also run. Overall 39 measurement tests were carried out. A plot of the data is shown in Figure 4

[Fig. 4 about here.]

If no calibration were necessary, the data would fall on the dashed line representing $y = x$. In this case, it can be seen that some calibration is clearly advisable. Also, from separate least squares fits to the two blocks of data it can be seen that there is an impression of discontinuity. Let y represent the actual measurement and x the nominal value. Let $b = 0, 1$ label the two blocks. Consider a model of the form:

$$y_i = \beta_0 + \beta_1 x_i + \gamma_0 b_{(i)} + \gamma_1 b_{(i)} x_i + \varepsilon_i \tag{13}$$

where $i = 1, \ldots, 39$ and the subscript (i) indicates a mapping from the i-th row of the data to corresponding block. The ε_i is the error term. γ_1 represents the difference in the slopes between the two blocks. This term was found to be not statistically significant and so it was removed from the model. Also $\hat{\gamma}_0 = -2.13\,\mu$m, which represents the estimated discontinuity between the two blocks. Since the estimated standard deviation (standard error) for this estimate is 0.79 μm, there is a statistically significant evidence of a real discontinuity between the two blocks.

From a practitioner's perspective, such a discontinuity is quite difficult to explain. In fact, it is not expected that measuring an RM just before or immediately after the point of discontinuity, e.g. 200 and 203 μm, would entail a system-

atic decrement in the measured values. A possible cause for such a discontinuity may be attributed to modality of running the experiment in two blocks. The occurrence of some lurking influential factor when running one of the two experimental units may have exerted an influence on the whole measuring result within the block. To limit this possibility, care has been taken to ensure that the environmental conditions were as close as possible between the two blocks of tests. The experimental effort requested to run a fully randomised experiment, which may have prevented the discontinuity, was not a viable possibility within the scope of this investigation.

The estimated intercept $\hat{\beta}_0 = -24.0\,\mu$m with a standard error of 6.7 μm indicates clearly that the linear fit does not pass through the origin. This implies that any one-point calibration procedure, which assumes that the calibration line passes through the origin, would be unsuitable with the amount of error depending on where that single point was chosen.

[Fig. 5 about here.]

Figure 5 shows a residual plot for the parallel lines model that omits the interaction term of (13), i.e. $\gamma_1 b_{(i)}$. For 9 of the 13 reference points, all three residuals are either positive or negative, which engenders the suspicion that some unexplained structure in the model may exist. The significance of this effect is tested by fitting the model:

$$y_i = r_{(i)} + \varepsilon_i \qquad (14)$$

where $i = 1,\ldots 39$ and the subscript (i) indicates the mapping from ith data point to one of the 13 nominal reference lengths represented by r. It should be noted that such a model has no value in calibration itself but does represent the best possible fit to the data. The lack of fit test was performed by using an F-test to compare this model to the proposed calibration model, obtaining a very small p-value. This indicates a lack of fit in the calibration model of equation 13.

There are two possible explanations for this lack of fit. Firstly, it is possible that the hypothesis of a linear calibration curve is incorrect. As it will be discussed later, it is the authors' opinion that the linear relationship does not hold globally, that is across a wide range of lengths. However, in this experiment, a relative narrow range of lengths have been examined where it may be thought that the local linear fit would be more than adequate. Nevertheless, roughness in the true calibration curve could explain the detected lack of fit. Roughness in the calibration function may be explained by irregularities of motion in scanner element, which may be subject to frequent accelerations and decelerations when covering the scan length instead of moving at constant speed, as it is usually assumed [20]. A second explanation for the lack of fit relates to possible problems in the reference materials, as already discussed in section 4, and in the strategy of the experiment. The run order of the experiment was completely randomised within each experi-

mental unit, so that each measurement result during the experiment constitutes a true replicate and not a repetition. So this aspect is not a cause of the lack of fit.

There is no way to distinguish between these competing explanations for the lack of fit, without resorting to a further, more refined, experiment.

Finally, other diagnostics such as quantile plots of the residuals to check for normality and a plot of the residuals against run order to check serial correlation and that the process was in control were performed. Nothing of concern was found in these plots and so these are not shown. Although there was no problem here, it is important that such diagnostics should be routinely checked.

The recommended calibration curve is:

$$\hat{y} = \hat{\beta}_0 + 0.5\hat{\gamma}_0 + \hat{\beta}_1 x \tag{15}$$

The common slope from the model of equation 13 was chosen and the line would pass half way through the discontinuity between the two blocks. This choice of curve does not depend on which explanation it is given for the lack of fit. Either way, this curve would be chosen unless we could obtain better experimental data.

The standard error for the curve itself is estimated to be around $0.3\,\mu$m in the middle of the range rising to around $0.4\,\mu$m at the edges. Extrapolation to measurements made outside the range would become less reliable and anything too far from this range would be subject to concerns about the global linearity of the calibration curve.

6 Calibration on a wide range

A calibration experiment has been considered on an interval about $500\,\mu$m wide with seven almost evenly distributed RM's with nominal lengths 3, 50, 100, 200, 300, 400 and 500 μm. The experiment was replicated 16 times. The RM's were randomly assigned to the sequence of the tests so that each measurement test was a true replicate rather than a repetition. Overall 112 tests were carried out.

In Figure 6, the measurement results from this calibration experiment are shown. The dashed line is $y = x$ while the solid line represents the least squares fit to the data. It is clear from this that at least some calibration is advisable.

[Fig. 6 about here.]

However, although the line appears to fit the data well, it is not possible to see the fine structure of the fit without the residual diagnostic plot show in Figure 7.

[Fig. 7 about here.]

It is observed that the variability in the residuals increases with the fitted values. Such behaviour is typical of most measurement systems considered over a wide range of measurands (cf. section 2.3 in JCGM 200:2008 [13] for a definition of measurand). This has consequences for the fitted model. Firstly, the fit will be weighted towards those measurements with the least variance and secondly, the accuracy of the fitted calibration curve will lessen in regions of higher variance.

There is also some indication of a repetition of the problems seen in the calibration on a narrow range where groups of residuals are either mostly positive or negative, again indicating an evidence of lack of fit or nonlinearity of the calibration curve. The error in this case is around 2-3 μm. This is not so much for the largest nominal length of 500 μm, but is crippling for the smallest nominal length of 3 μm. Some of the lack of fit may also be the result of the limits of the proposed method for building RM's. In fact this method enables the experimenter to build step heights with actual length theoretically lying always within $l_{n,s} \pm 0.4\,\mu$m. And this limit becomes more apparent for small nominal lengths. Yet it is also worthwhile noting that at the present time the cost of a single purpose built reference material with nominal length less than two micrometres is about four times the cost of a set of about 90 gauge blocks, which exceeds by far the needs of the whole proposed calibration procedure.

The poor performance of the model is not obvious from Figure 6, nor from considering the R^2 from the linear model which is hardly distinguishable from 1, which would represent a perfect fit.

The usual solution to the problem of increasing variability is to consider a transformed model using natural logarithms:

$$\log y_i = \beta_0 + \beta_1 \log x_i + \varepsilon_i \tag{16}$$

This has the advantage of modelling error in a relative (multiplicative) rather than absolute (additive) sense. In fact equation 16 is equivalent to

$$y_i = e^{\beta_0} \cdot x_i^{\beta_1} \cdot e^{\varepsilon_i} \tag{17}$$

Unfortunately, residuals diagnostics indicate a lack of fit in this model also, but the addition of a quadratic term offers considerable improvement:

$$\log y_i = \beta_0 + \beta_1 \log x_i + \beta_2 (\log x_i)^2 + \varepsilon_i \tag{18}$$

The residuals against fitted values for this model is shown in Figure 8. Since the model is fitted on a log scale, it is not possible to show sensibly units for the quantities in this figure. However, it is possible to interpret the residuals as relative errors, that, as shown in Figure 8, amount to no more than approximately 2%. It can be seen that the variability of these residuals for the smaller lengths is somewhat higher although not substantially so. This calibration model provides superior performance over the whole range of the data. Like for equation 16, the model of equation 18 is also provided in its equivalent multiplicative form

$$y_i = e^{\beta_0} \cdot x_i^{\beta_1 + \beta_2 \cdot \log x_i} \cdot e^{\varepsilon_i} \tag{19}$$

Some evidence of lack of fit due to the presence of some one-sided groups of residuals can still be seen. The same issues from the discussion of this problem in the narrow range case apply here.

[Fig. 8 about here.]

In contrast to the calibration on a narrow range, the data in this case are spread over a wide range. It is apparent that a linear model on the untransformed data might produce a misleading calibration curve with particularly poor performance for short lengths. The use of the log transformation and openness to considering a quadratic relationship produces a superior result.

7 Conclusions

Several issues are raised by the analysis of the two examples of calibration on a narrow and a wide range that have been presented:

- Further investigation into both the global and local shape of calibration curves would be helpful. It seems clear that globally linear calibration curves for lengths over a wide range are inappropriate, but more detailed data are needed to suggest a good alternative class of curves. There is also some uncertainty about the local behaviour of true calibration curves. Does roughness of the calibration curve truly exist or is it reasonable to assume that these should be smooth and monotone? If evidence of a corrugated calibration curve were confirmed, it would raise suspicions of the integrity and/or adequacy of the design of the hardware and software system producing and controlling the motion of the scanner element. However, finer scale data is necessary to resolve this.
- Proper use of diagnostic methods is essential for developing a calibration curve. Some human supervision is essential to interpret these diagnostics to avoid serious mistakes. Designing and implementing fully automated calibration procedures may be an unrealistic aim.
- Although it would be desirable to economise on the amount of data necessary to perform a calibration, this would be risky. This is what can be argued from the analysis of the built-in calibration procedure and from its comparison with the two cases presented. In both the narrow and large interval calibrations the usage of a wide range of reference materials contributed to identify otherwise undetectable characteristics of the calibration curve. A method to overcome the economic barrier that may limit the availability of reference materials on the micro-scale has also been presented.
- A calibration that uses only one or two reference materials could provide measurement performances utterly inadequate to the needs of the measurement and inspection tasks that a measuring system could economically and reliably carry out if carefully calibrated.

Acknowledgements The whole experimental activity and only part of the theoretical study connected with this investigation were carried out at the Manufacturing Engineering Centre of Cardiff University. The authors would also like to thank the Engineering and Physical Sciences Research Council (EPSRC) of the United Kingdom. This work was partially performed within the research initiative of the EPSRC

program Innovative Manufacturing Engineering Centres and especially 'The innovative Design and Manufacturing Research Centre at the University of Bath'.

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **AC-19**(6), 716–723 (1974)
2. BSI - British Standards Institution: BS 4311-3:1993. Specification for gauge blocks and accessories - Part 3: gauge blocks in use. (1993)
3. BSI - British Standards Institution: BS ISO 11095:1996. Implementation of ISO 11095:1996. Linear calibration using reference materials (1996)
4. BSI - British Standards Institution: BS EN ISO 3650:1999. Geometrical product specification (GPS) - length standards - gauge blocks. (1999)
5. CIRD-Centro interdipartimentale di Ricerca Didattica-Università di Udine: The temperature of the human body (2002). URL http://web.uniud.it/cird/secif/termo/stati/es6.htm. Last accessed on 13 June 2008. This document is in Italian
6. Decker, J., Pekelsky, J.R.: Gauge block calibration by optical interferometry at the national research council of canada. In: Measurement Science Conference. Pasadena, U.S.A. (1997)
7. Faraway, J.: Linear models with R. Chapman & Hall, Boca Raton (2005)
8. Ferri, C., Brousseau, E.: Variability in measurements of micro lengths with a white light interferometer. Quality and Reliability Engineering International **24**(8), 881–890 (2008)
9. Ferri, C., Brousseau, E., Dimov, S., Mattsson, L.: Repeatability analysis of two methods for height measurements in the micrometer range. In: 4M2006 - Conference on Multi-Material Micro Manufacture, pp. 165–168. Grenoble, France (2006)
10. Hansen, H., Carneiro, K., Haitjema, H., Chiffre, L.D.: Dimensional micro and nano metrology. CIRP Annals - Manufacturing Technology **55**(2), 721 – 743 (2006)
11. Harasaki, A., Schmit, J., Wyant, J.C.: Offset of coherent envelope position due to phase change on reflection. Applied Optics **40**(13), 2102–2106 (2001)
12. Harasaki, A., Wyant, J.C.: Fringe modulation skewing effect in white-light vertical scanning interferometry. Applied Optics **39**(13), 2101–2106 (2000)
13. JCGM - Joint Committee for Guides in Metrology: JCGM 200:2008. International vocabulary of metrology – Basic and general concepts and associated terms (VIM), third edn. (2008)
14. Malinosvsky, I., Titov, A., Dutra, J., Belaidi, H., Franca, R., Massone, C.: Towards subnanometer uncertainty in interferometric length measurements of short gauge blocks. Applied Optics **38**(1), 101–112 (1999)
15. Malinosvsky, I., Titov, A., Massone, C.: High-precision method for gauge block measurements and comparison of gauge block interferometers. In: 1998 Conference on Precision Electromagnetic Measurements, pp. 50–51. Washington D.C., U.S.A. (1998)
16. Park, M.C., Kim, S.W.: Compensation of phase change on reflection in white-light interferometry for step height measurement. Opt. Lett. **26**(7), 420–422 (2001)
17. Pförtner, A., Schwider, J.: Dispersion error in white-light linnik interferometers and its implications for evaluation procedures. Applied Optics **40**(34), 6223–6228 (2001)
18. Royal Society and Royal Academy of Engineering: Nanoscience and nanotechnologies: opportunities and uncertainties (2004). URL: http://www.nanotec.org.uk/finalReport.htm Last accessed on 14 June 2008
19. Scarr, A.J.T.: Metrology and precision engineering. McGraw-Hill, New York (1967)
20. Schmit, J., Olszak, A.: High-precision shape measurement by white-light interferometry with real-time scanner error correction.

Appl. Opt. **41**(28), 5943–5950 (2002)

21. Sun, C., Yu, L., Sun, Y., Yu, Q.: Scanning white-light interferometer for measurement of the thickness of a transparent oil film on water. Appl. Opt. **44**(25), 5202–5205 (2005)
22. Vallance, R.R., Morgan, C.J., Shreve, S.M., Marsh, E.R.: Microtool characterization using scanning white light interferometry. Journal of Micromechanics and Microengineering **14**(8) (2004)

List of Figures

Fig. 1 A reference material

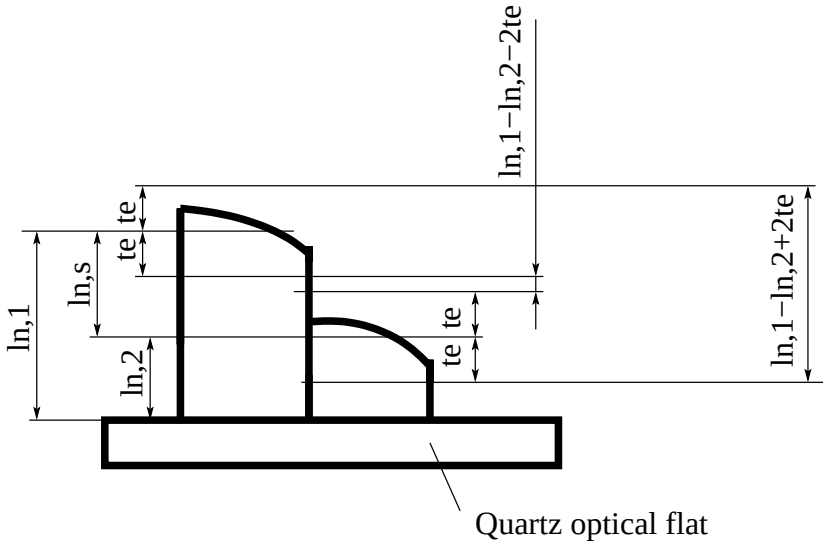

Fig. 2 Exhaustive range of values for the length L_s of a RM

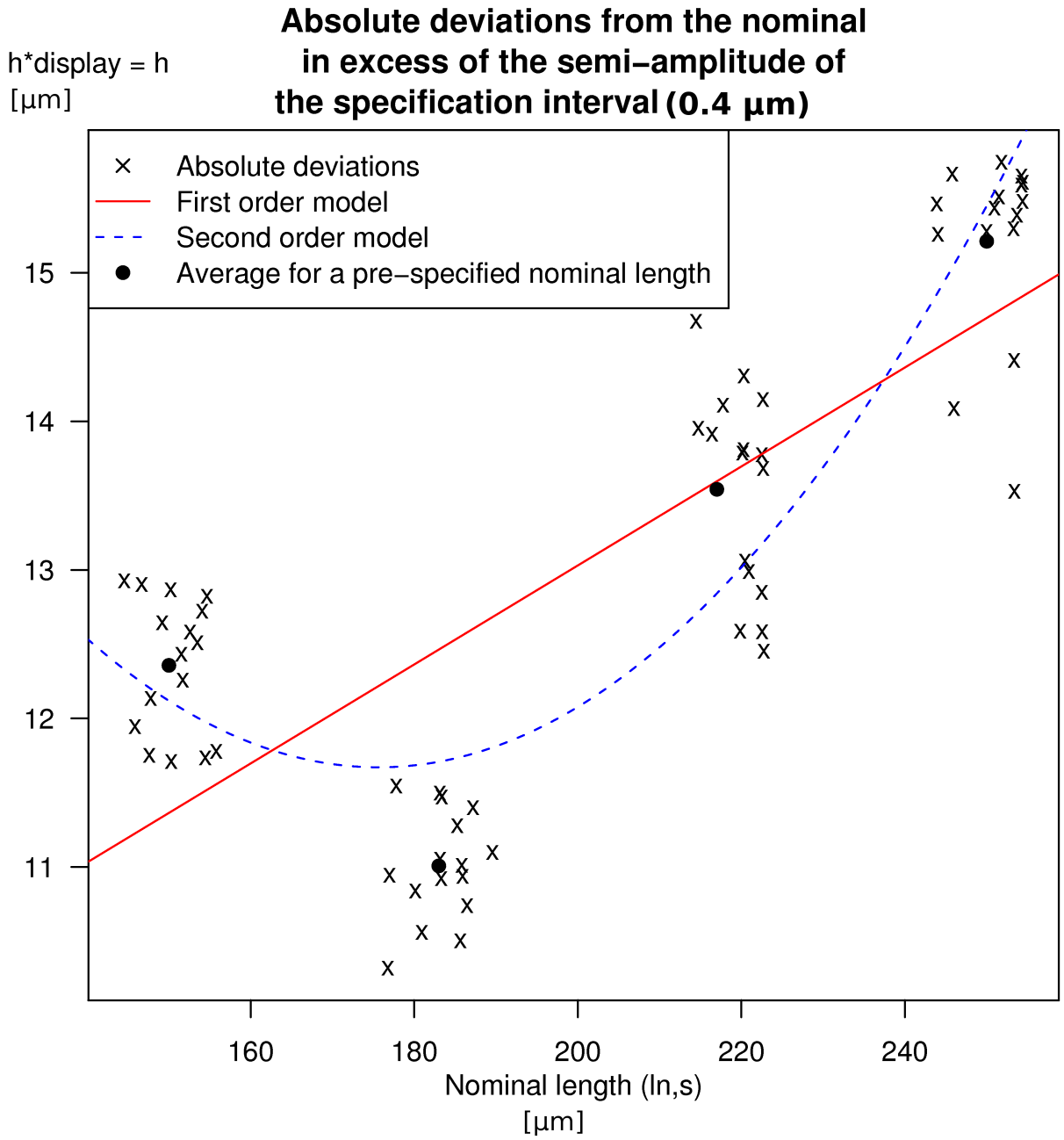

Fig. 3 Absolute deviations from the accepted values

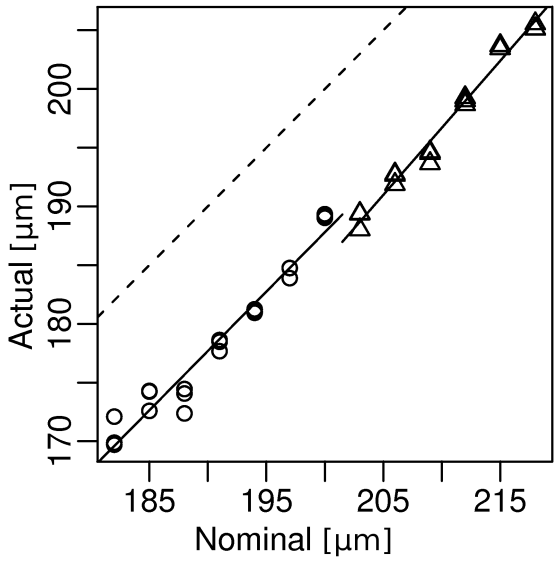

Fig. 4 Calibration data from the narrow range experiment. The dashed line is $y = x$ while the solid lines are least squares fits to the two blocks of data

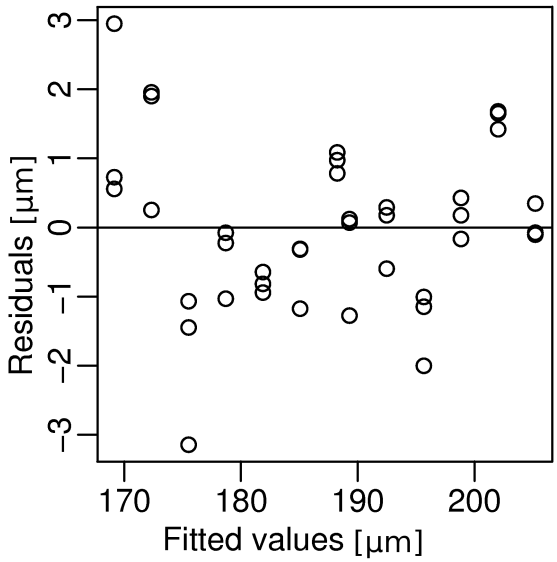

Fig. 5 Residual plot for the calibration model.

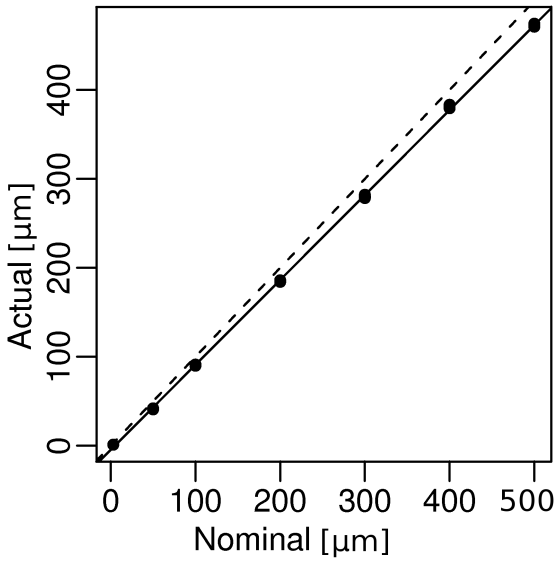

Fig. 6 Measurement results from the calibration experiment on a wide range

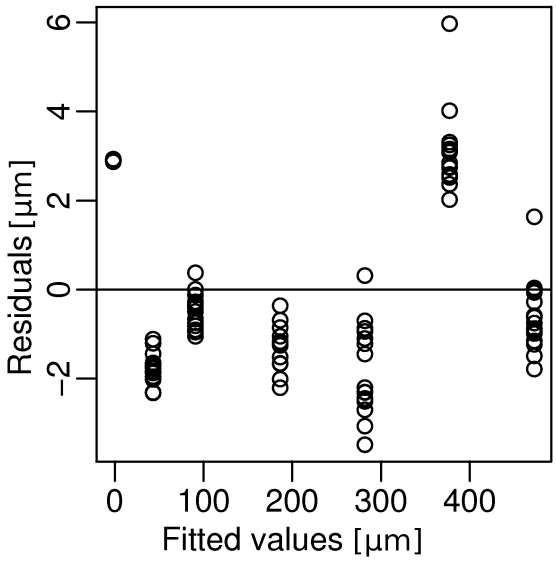

Fig. 7 Residuals versus fitted values for the linear calibration curve on a wide range

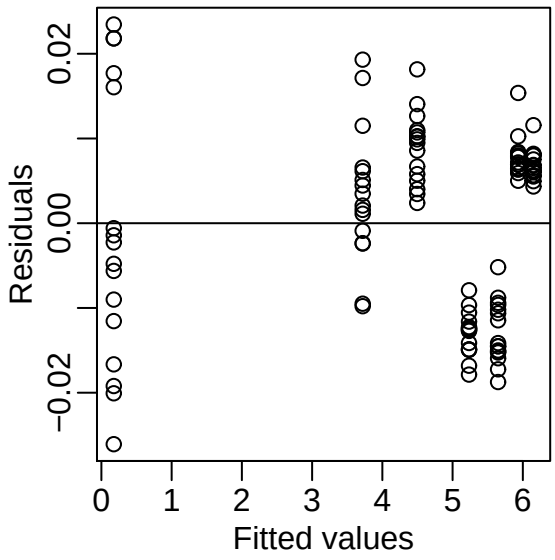

Fig. 8 Residuals vs. fitted values for the logged fit of the calibration experiment on the wide range.