Int. J. Metrol. Qual. Eng. **11**, 16 (2020)

https://doi.org/10.1051/ijmqe/2020010

International Journal of Metrology and Quality Engineering

Available online at:
www.metrology-journal.org

RESEARCH ARTICLE OPEN ACCESS

Variable data measurement systems analysis: advances in gage bias and linearity referencing and acceptability

Mahjoub Abdelgadir*, Chris Gerling[1], and Joel Dobson

Texas Instruments Inc. Quality Standards and Statistical Services, Dallas, TX 75243, USA

Received: 12 September 2020 / Accepted: 23 October 2020

Abstract. Measurement systems analysis (MSA) is a set of requirements and procedures adopted by the automotive industry and other disciplines to evaluate the accuracy and precision of measurement systems through assessing and quantifying the random and systematic errors and assigning appropriate dispositions for tolerance and performance acceptance. The methodology of variable data MSA comprises studies of a system's stability, bias, linearity and gage repeatability and reproducibility (GR&R). This paper describes advances in referencing and criteria for estimation of uncertainty errors, dispositions, and acceptability of MSA bias and linearity, proposing an extension to the basic statistical zero null-hypothesis to include overlap between confidence intervals and uncertainty associated with the reference standards used in bias and linearity studies.

Keywords: Measurement system / bias / linearity / traceable standard / consensus standard

1 Introduction

A measurement system may be defined collectively as the gage instrument hardware, software and tooling; the standards or reference parts; the procedures, personnel and measurement environment; and the statistical assumptions, hypotheses and data analysis. Measurement systems analysis (MSA) aims to estimate the accuracy and precision of measured, tested, and inspected characteristics of manufactured products; ensuring the inherent variabilities from all elements of a measurement system are understood and controlled, side by side with the product manufacturing process variability which is controlled within set limits. Variable data MSA study for a given characteristic comprises collecting data on stability, bias, linearity, and gage repeatability and reproducibility (GR&R); then – based on statistical hypothesis and disposition criteria – deciding acceptability of the measurement system. Bias and linearity studies expose any systematic errors and validate the accuracy of the measurement system over the operating range. GR&R studies, on the other hand, expose random errors and validate precision of the gage. Stability charts track normal random variation of measurements over usage time, flagging any drift or other special cause effects in the system.

The approach in this paper aligns with the guidance provided in the automotive Measurement Systems Analysis (MSA) reference manual 4th Edition [1], with acceptance set at 95% confidence ($\pm 2\sigma$ statistics). All relevant requirements and procedures are captured in Texas Instruments Inc. internal specifications, including formulated Excel worksheets for calculations and dispositions. Additionally, the paper proposes an extension to acceptance of bias and linearity by the statistical zero null-hypothesis to include quantified overlap between the bias confidence intervals and the uncertainty associated with the reference standards used in bias and linearity studies.

Section 1 of the paper introduces the types of reference standards used in MSA studies, which include traceable, consensus and check standards. We derive the formulae estimating expanded uncertainty for calculated values of consensus and check standards, using a nested ANOVA method. Section 2 outlines the method for evaluating the amount of bias in a measurement system using repeatability trials, and the acceptance condition by null-hypothesis statistical zero bias condition (statzero). We then propose extending acceptance by a new criterion which we call statzero proxy, based on the degree of overlap between the confidence interval for the bias data fit at 95% confidence, and the uncertainty associated with the reference standard used in the bias evaluation experiment. We also include the Student's t-test for small repeatability sample.

Section 3 deals with evaluation of the measurement system's bias linearity over the gage operating range. First, we derive the simple linear regression formulae that are

* Corresponding author: m.a.gadir53@gmail.com
[1] Retired in 2017. Consultant

needed for computing the best fit line, its slope and intercept, and the confidence interval hyperbolae of the regression analysis. Then we set up the statzero conditions needed for acceptance of linearity, applicable to the regression best fit line as well as to the slope and intercept. The Student's t-test is also deployed to justify acceptance for a small sample. Furthermore, we extend the acceptance of linearity to the statzero proxy criteria based on the degree of overlap between the confidence interval hyperbolae curves and the uncertainty bars associated with the reference standards used in the linearity evaluation experiment.

Section 4 introduces examples to demonstrate calculation of a check standard and a consensus standard. It also contains examples demonstrating evaluation and acceptance of bias and linearity by the basic statzero conditions and the extended statzero proxy criteria.

Figure 1 shows a typical flow for the reference standard (s), the setup of the measurement trials for single-point bias and multi-point bias linearity studies, and the decision tree for acceptance.

2 Materials and methods

2.1 Standards

2.1.1 Traceable standard

MSA studies ab initio require reference standards with known values and uncertainties that are traceable to National Measurement Institute (NMI)-accepted values, such as NIST or equivalent. This prerequisite is essential for assessment of accuracy and precision of the measurement system by repeatability trials of a known standard value. Nonetheless, NMI-traceable standards may not be available for all measurement situations, e.g. could be non-existent for a unique measurement characteristic and/or a unique metrology system; or maybe too expensive to purchase, e.g. for destructive test systems. In such cases, MSA may be performed using in-house reference specimens or master parts, referred to as *check standards* in the MSA reference manual [1]. However, whereas check standards are certainly suitable for stability and GR&R studies which do not require accuracy, we believe they should not be the go-to for usage in bias and linearity studies due to self-traceability limitation and lack of independently verified accuracy; unless no other option (see example under Results & Discussion). A better alternative to NMI-traceable standards are consensus-generated standards, referred to as *consensus standards* in [1]. We will present these types in § 2.1.2 and § 2.1.3, and propose methods for estimating their values and uncertainties.

2.1.2 Check standard (in-house reference part)

The MSA reference manual [1] defines a check standard as "a measurement artifact that closely resembles what the process is designed to measure, but is inherently more stable than the measurement process being evaluated." Accordingly, we define it as an in-house reference specimen or master part created and verified at a production site or laboratory under controlled conditions at least similar to or better than normal processing conditions. We offer an evaluation method to estimate the check standard value, $\mathbf{R}chk$, and uncertainty, $Uchk$, that includes correlation to NMI-traceable standard generic with the check standard but with different value, if available.

The evaluation starts by running repeatability measurement trials Ri on the check standard using a calibrated gage having as much precision as possible, preferrably 10× the resolution of the systems under MSA study (rule of thumb). $\mathbf{R}chk$ is taken as the mean value:

$$\mathrm{R}chk = \frac{1}{m}\sum_{i=1}^{m} Ri \tag{1}$$

To estimate uncertainty, given that repeatability sample size is typically small ($10 \le m \le 20$), we start by using t-distribution statistics whereby T_{stat} is expressed as:

$$T_{\mathrm{stat}} = (\overline{x} - \mu)/(s/\sqrt{m}) \tag{2}$$

$\overline{x}$ and s are the sample mean and standard deviation, normally distributed about the true mean μ, and $s/\sqrt{m}$ is the familiar standard deviation of the mean (also called standard error of the mean). Tstat characterizes a wider spread and shift of mean for the t-distribution of random small samples relative to normal distribution of population at large ($N \gg 30$, std. dev. $= \sigma$) (see e.g. [5], § 2.7.3.). For t-curve with $(m-1)$ degrees of freedom (df $= m-1$ since $\overline{x}$ is already decided), equation (2) is expressed at $(1-\alpha)\%$ confidence by the critical value $t(\alpha/2, m-1)$, which we call Tcrit and rearrange:

$$\overline{x} - \mu = T_{\mathrm{crit}}/(s/\sqrt{m}) \tag{3}$$

The left hand side of (3) represents uncertainty $U(\overline{x})$ as a delta between $\overline{x}$ and the true mean; thus estimated by calculating the sample standard deviation and using Tcrit from standard T-statistic tables or by the Excel function = TINV($\alpha/2$, $m-1$). In this paper we use 95% confidence, $\alpha = 0.05$.

The standard deviation is found from the variance of the m repeatability measurements for $\mathbf{R}chk$:

$$V_{chk} = \frac{1}{m-1}\sum_{i=1}^{m}(R_i - \mathrm{R}_{chk})^2 \tag{4}$$

Hence,

$$U(\overline{x}) = T_{\mathrm{crit}}\sqrt{(V_{chk}/m)} \tag{5}$$

Next, we estimate the 'combined uncertainty'. This item is discussed in many literature references, but we limit our referencing to the MSA manual [1], NIST [2], and for more details the JCGM Guide to Uncertainty [3]. Here we include the gage calibration uncertainty tolerance Ug as specified by the equipment manufacturer or supplied by a calibration house, and the limit of its resolution ρ as a capability error component. We combine these in quadrature with $U(\overline{x})$ to obtain the combined uncertainty uc, and multiply by 95%-confidence 2-tail coverage factor $k = 2$ to

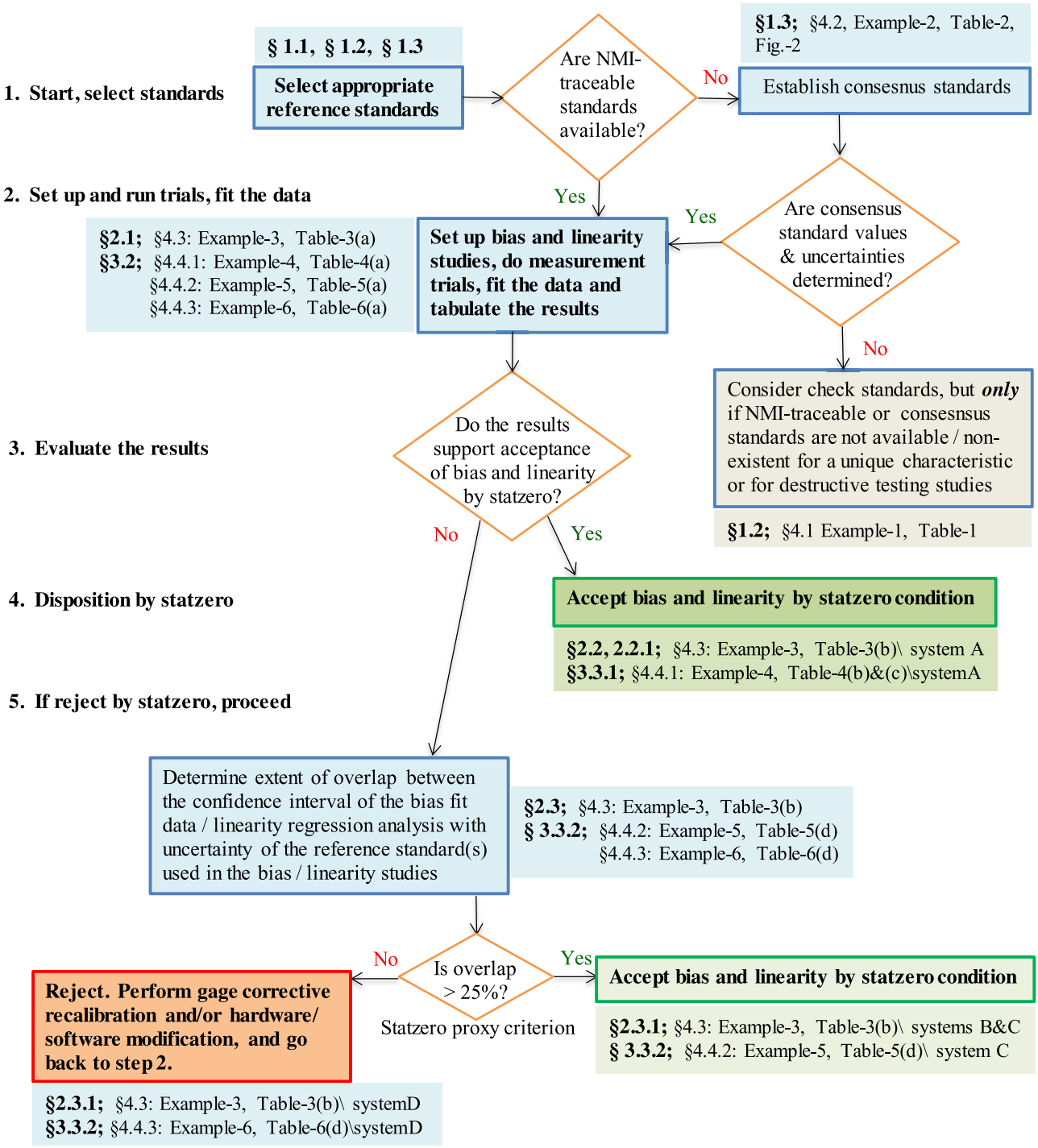


Fig. 1. Gage bias and linearity flow chart/decision tree (paper layout).

obtain the measurement 'expanded uncertainty' $U = 2uc$ (using the terminology and symbols in [2]):

$$U = 2u_c = 2[[U(\overline{x})]^2 + (U_g)^2 + \rho^2]^{1/2}$$
$$= 2[(T_{\text{crit}})^2 V_{chk}/m + (U_g)^2 + \rho^2]^{1/2} \quad (6)$$

Even though (6) ensures a reasonable estimate of measurement uncertainty by combining the standard error of the mean with fixed errors due to the quoted gage calibration uncertainty and the resolution limit of the instrument, this only goes to validate precision of the gage with a high degree of confidence. Not the same degree of confidence can be inferred regarding accuracy of the gage, i.e. how close is $\mathbf{R}chk$ really to true value within the calculated measurement uncertainty. The biggest concern is whether the gage has a hidden 'offset' that the repeatability measurement method would not be able to

uncover. To help address this, we propose adding variance statistics for a generic NMI-traceable standard, if available at the site owning the gage, (generic as being similar type to the check standard, e.g. a thin film wafer standard, having thickness different from the check standard.). Let the generic standard be characterized by $\mathbf{R}t \pm U_t$, where $\mathbf{R}t$ is the traceable value (closer to true value, whatever its value is), and U_t is the quoted uncertainty. Running m repeatability trials $R'i$ on the traceable standard using the same gage, the mean value is:

$$\mathrm{R}_m = \frac{1}{m}\sum_{i=1}^{m} R'i \tag{7}$$

The delta between $\mathbf{R}t$, the quoted value of the traceable standard, and $\mathbf{R}m$, its mean value as determined by repeatability using the in-house gage, can be considered a systematic offset error:

$$\Delta\mathrm{R} = |R_m - R_t| \tag{8}$$

The uncertainty associated with $\mathbf{\Delta R}$ may be composed additively from the expanded uncertainty of the repeatability trials on the generic standard, and U_t the uncertainty quoted for it. The variance of the repeatability trials $R'i$ is:

$$V_t = \frac{1}{m-1}\sum_{i=1}^{m}(R'i - \mathrm{R}_t)^2 \tag{9}$$

The measurement expanded uncertainty for the generic standard would be (similar to (6)):

$$U' = 2[(T_{crit})^2(V_t/m) + (U_g)^2 + \rho^2]^{1/2} \tag{10}$$

Hence, the uncertainty in ΔR (in quadrature)

$$U(\Delta\mathrm{R}) = \sqrt{\{(U')^2 + (U_t)^2\}} \tag{11}$$

Finally, applying quadrature combination of the uncertainty components (6) and (11) and the offset error $\mathbf{\Delta R}$ of (8), we obtain the total estimated uncertainty for the check standard:

$$U_{chk} = [(U)^2 + (U')^2 + (U_t)^2 + (\Delta\mathrm{R})^2]^{1/2} \tag{12}$$

The $\pm Uchk$ uncertainty represents self-traceable estimated accuracy error bar around the value estimated for the check standard.

Even though the formula (12) represents a reasonably good estimate of a 'simulated' accuracy of the gage by including an offset factor relative to a generic traceable standard, there is no guarantee that the offset is a constant, i.e. can be applied as is across the measurements range that the gage is used for. Because of this, and the evident self-traceability handicap, we do not recommend a check standard as alternate to NMI-traceable standard for use in bias and linearity studies unless no other option. See Example 1 in the Results & Discussion section for a quantitative illustration supporting the counter-recommendation. On the other hand, check standards are quite useful and handy for GR&R studies and ongoing stability tracking via SPC control/monitor charts.

For bias and linearity assessment, a more acceptable alternate to NMI-traceable standard, if unavailable or cost-prohibitive, is the consensus standard. This features better traceability than just self-traceability, as discussed next.

2.1.3 Consensus standard

The MSA reference manual [1] describes consensus value as "based on collaborative experimental work under the auspices of a scientific or engineering group, defined by a consensus of users such as professional and trade organizations." Accordingly, a consensus standard may start as a check standard belonging to one site (factory, laboratory), then gets evaluated by consensus measurement trials across three or more independent sites that have measurement systems compatible with the system in the site which generated the check standard. Additionally: (i) the participant sites' gages used in generating the consensus information should be calibrated and have at least equivalent or greater resolution (preferably 10×, rule of thumb) than the gages for which the consensus standard is to be used in MSA studies; and (ii) the gages' calibration uncertainty tolerances Ug, as quoted by equipment manufacturers or by calibration vendors, should be available to be included in assessing the combined uncertainty. Based on these criteria, successful generation of a consensus standard would assure reasonable confidence in the accuracy of reference value within uncertainty limits established by independent subgroup data sets and augmented by available gage calibration and resolution errors.

A consensus standard is characterized by consensus value and combined uncertainty. Each site participating in consensus standard evaluation would run m repeatability measurement trials Ri on the characteristic feature(s) of the check standard/reference part at the same reference point(s), and calculate their subgroup sample average $Rp(s)$ similar to equation (1). With carefully executed trials and assuming samples with normal distribution, the estimated consensus value $\mathbf{R}con$ is the mean of the subgroup samples' averages:

$$\mathrm{R}_{con} = \frac{1}{k}\left[\sum_{s=1}^{k} R_p(s)\right] \tag{13}$$

where k is the number of participating sites (subgroups), and $R_p(s) = \frac{1}{m}\sum_{i=1}^{m} R_i$.

Estimation of the combined uncertainty needs more work by assembling independent errors from the significant components of variation: *viz.* random standard deviation errors associated with analysis of variance (ANOVA) of independent sample means, and – as in § 2.1.2 – systematic errors due to equipment calibration uncertainty and instrument resolution limit.

Each site calculates the variance $Vp(s)$ in their subgroup repeatability sample using equation (4):

$$V_p(s) = \frac{1}{m-1}\sum_{i=1}^{m}(R_i - R_p(s))^2 \tag{14}$$

And calculates subgroup measurement expanded uncertainty U(s) according to equation (6):

$$U(s) = 2[(T_{crit})^2 V_p(s)/m + (U_g)^2 + \rho^2]^{1/2} \tag{15}$$

where Ug is the gage calibration uncertainty tolerance and ρ is the gage discriminating resolution.

Next, the participating sites combine the measurement variance over all subgroups. This will have two components: {mean within-subgroup} sample variance Vms, and {subgroup ↔ subgroup} variance Vss:

$$V_c = V_{ms} + V_{ss} \tag{16}$$

Vms is estimated by averaging the repeatability sample variances $Vp(s)$ over all subgroups:

$$V_{ms} = \frac{1}{k}\left[\sum_{s=1}^{k} V_p(s)\right] = \frac{1}{k(m-1)}\cdot\sum_{s=1}^{k}\sum_{i=1}^{m}(R_i - R_p(s))^2 \tag{17}$$

To estimate Vss, we use nested random-effects ANOVA model treating subgroup average $Rp(s)$ as a sample-dependent statistic around the group mean $\mathbf{R}con$, with repeated measurement trials mathematically nested within the subgroups. Based on this, the expected value of the mean sum of squares from subgroup to subgroup is expressed by:

$$\varepsilon(\mathrm{Mss}) = \frac{1}{k-1}\sum\nolimits_{s=1}^{k}(Rp(s) - \mathbf{R}con)^2 = \mathrm{V}ss + \mathrm{V}ms/m \tag{18}$$

where Vms/m is the standard variance of the samples mean relative to the population mean; in this case it is a correction factor accounting for overestimation of the expected value of Vss due to the nested subgroups ANOVA structure (see e.g. [4] Ch.10 on theory of ANOVA). Hence, Vss is obtained from equation (18) by subtracting the correction factor from $\varepsilon(Mss)$, then substituting in (16) to get the combined variance:

$$V_c = \varepsilon(M_{ss}) + \frac{(m-1)}{m}V_{ms} \tag{19}$$

Additionally, we consider the systematic error due to the participant sites' gage calibration uncertainty tolerances, Ug. Treating this like a variance, we estimate an average of the calibration uncertainty tolerance over the group of gages using quadrature summation:

$$\overline{U}g = \left\{\sum\nolimits_{g=1}^{k}(Ug)^2/k\right\}^{1/2} \tag{20}$$

We also add the gage resolution ρ as a systematic capability error applicable to all measurements (assuming the participant gages have the same resolution.)

Hence, the combined uncertainty u_c over the group of all measurement trials is expressed by:

$$u_c = \left[Vc + \left(\overline{U}g\right)^2 + \rho^2\right]^{1/2} \tag{21}$$

Finally, using (19) in (21) gives the expanded uncertainty $U = 2u_c$ for the consensus standard:

$$Ucon = 2\left[\varepsilon(\mathrm{M}ss) + \frac{(m-1)}{m}\mathrm{V}ms + \left(\overline{U}g\right)^2 + \rho^2\right]^{1/2} \tag{22}$$

Vms and $\varepsilon(Mss)$ are calculated by equations (17) and (18), respectively. The $\pm Ucon$ expanded uncertainty represents consensus-traceable accuracy error bar around $\mathbf{R}con$, the estimated value of the consensus standard established by (13). See Example 2 for a quantitative illustration.

2.2 Gage bias

2.2.1 Bias measurement

Bias study requires using a NMI-traceable reference standard $\mathbf{R}t \pm Ut$. However if this is justifiably not available, then a consensus standard $\mathbf{R}con \pm Ucon$ may be used. For conciseness we will use $\mathbf{R}r$ (*reference*) to mean either $\mathbf{R}_t$ or $\mathbf{R}con$, and Ur for either Ut or $Ucon$.

The procedure starts by checking that the measurement system's gage is properly calibrated, then proceeding to repeatability measurement trials Ri of the reference standard by a qualified person or by automation as the case may require. The sample size should be $m \geq 10$ trials. The bias average Bav is then obtained by averaging the deltas between the trial values Ri and the reference value $\mathbf{R}r$ over the sample size:

$$B_{av} = \frac{1}{m}\sum_{i=1}^{m}(R_i - \mathrm{R}_r) \tag{23}$$

B_{av} may be expressed as a percentage of the reference value: $(B_{\mathrm{av}})\% = (B_{\mathrm{av}}/\mathbf{R}r) \times 100$.

Ideally Bav should be zero. However, this is not typically the case due to inherent variation in the measurement system and random normal variation in the repeatability trial runs. Most, if not all, measurement systems tend to show a small non-zero positive or negative bias. Acceptability is subject to non-rejection of the null hypothesis, as will be discussed below.

2.2.2 Statistical zero bias hypothesis (statzero)

Acceptance of bias is subject to testing the null hypothesis: $\{\mathbf{H_0}: \mathbf{B} = \mathbf{0}\}$, such that the bias error of a measurement system is acceptable if not statistically significantly different from zero [1], a condition referred to as 'statistical zero bias'. We will call this '**statzero**' for short. For validation, we take into account the standard deviation of

the trials' sample and the interval for normal 2-tail distributed bias at 95% confidence. We also validate the Student's *t*-test: *T*stat < *T*crit in accordance with small sample size in bias studies (typically $10 \leq m \leq 20$.)

The standard deviation of the bias repeatability trials, s_r, is given by:

$$s_r = \left[\frac{1}{m-1}\sum_{i=1}^{m}(R_i - \mathrm{R}_r)^2\right]^{1/2} \tag{24}$$

Unlike statistical systems in general where the population mean is unknown, the bias study case has a precise population mean, its target zero value. Hence, substituting $\overline{x} = B_{\mathrm{av}}$ and $\mu = 0$ in equation (2) gives:

$$T_{stat} = \frac{B_{av}}{s_r/\sqrt{m}} \tag{25}$$

Next, we determine the upper and lower limits of the confidence interval [UCL; LCL], for the small single sample bias study using the general formula for boundaries of a presumed normal *t*-distribution at $(1 - \alpha)\%$ confidence: $\overline{x} \pm (t\alpha/2,n-1)\ s/\sqrt{m}$ (see e.g. [5], § 7.3).

$$[\mathrm{UCL};\mathrm{LCL}] = B_{av} \pm T_{crit}(s_r/\sqrt{m}) = B_{av}[1 \pm (T_{crit}/T_{stat})] \tag{26}$$

The second equivalence in (26), obtained by substituting for $s_r/\sqrt{m}$ from (25), indicates the wider interval and shift in mean for the small sample subject to the Student *t*-test: $T_{\mathrm{stat}} < T_{\mathrm{crit}}$.

2.2.3 Bias acceptance by statzero condition

This is fulfilled by not rejecting the null hypothesis $\{H_0\colon B = 0\}$, subject to zero confined within the confidence interval about the bias average [1]:

$$B_{av} - \left[T_{crit}(s_r/\sqrt{m})\right] \leq \text{zero} \leq B_{av} + \left[T_{crit}(s_r/\sqrt{m})\right]$$
$$\text{and}: \quad T_{\mathrm{stat}} < T_{\mathrm{crit}}\ (\text{smallsample}) \tag{27}$$

2.2.4 Statistical zero bias proxy (statzero proxy)

Acceptance by statzero condition (27) does not take into account the factor of uncertainty spread around true value of the reference used in bias study; namely extent of overlap bewteen the 95% confidence interval of repeatability trials and the reference uncertainty bar. The MSA manual [1] did not include specific guidance or procedure to account for this overlap when making bias acceptance decisions. Henceforth, we propose an additional test of significance for non-zero bias, to include acceptance based on extent of the overlap. Disposition with the proposed criterion will be established by calculating ΔU_{ovrlp} as a ratio of magnitude of overlap between the width of the confidence interval, UCL − LCL, and the reference uncertainty bar, $\pm Ur$:

$$\Delta U_{ovrlp} = \{\mathrm{Min}(\mathrm{UCL}, +U_r) - \mathrm{Max}(\mathrm{LCL}, -U_r)\}/(\mathrm{UCL} - \mathrm{LCL}) \tag{28}$$

ΔU_{ovrlp} is a positive number between 0 and 1: $0 \leq \Delta U_{ovrlp} \leq 1$. A value <0 means no overlap.

Expression (28) represents how much of the repeatability 95% confidence interval lies within the reference uncertainty interval. See Figure 2 for cartoon diagrams depicting various overlap cases.

2.2.5 Bias acceptance by statzero proxy criterion

We propose extending acceptance of non-zero bias as still insignificant if the 95% confidence interval of the repeatability sample is overlapping the reference uncertainty bar $\pm Ur$ by more than 25%, i.e.:

$$\Delta U_{ovrlp} \times 100 > 25\% \tag{29}$$

The validity of >25% overlap as a general acceptability rule of thumb had been established in Statistics literature [4]. We adopt (29) as a criterion for incrementally extending bias acceptability beyond the statzero condition, and call this extended acceptance 'statistical zero bias proxy', or, for short, '**statzero proxy**'. It draws credence from appreciable probability that the *estimated uncertainty* for the reference value, as determined by repeatability and represented by the confidence interval, is sufficiently overlapping with the *traceable uncertainty* of the standard used; thus facilitating extension of not rejecting the null hypothesis. This makes sense in light of the basic definition of uncertainty in the MSA reference manual as the "estimated range of values about the measured value in which the true value is believed to be contained". The criterion (29) therefore safeguards that the bias can still be considered statistically zero by proximity of the estimated reference value to the true value within acceptable overlap of uncertainty values.

See Example 3 demonstrating statzero and statzero proxy dispositions.

2.3 Gage linearity

2.3.1 Regression analysis

The purpose of linearity study is to verify that the bias of a measurement system satisfies the primary null hypothesis statzero condition (27) over the system's applicable operating range. Based on the statzero proxy criterion (29) advanced in § 2.2.5 for single bias sample, we propose extending the acceptance by statzero proxy to the linearity case. Mathematical validation of linearity requires bivariate linear regression analysis in place of single bias univariate analysis. Acceptance requires applying the statzero, or its proxy, not just to the bias average but also to the slope and intercept of the regression best fit line.

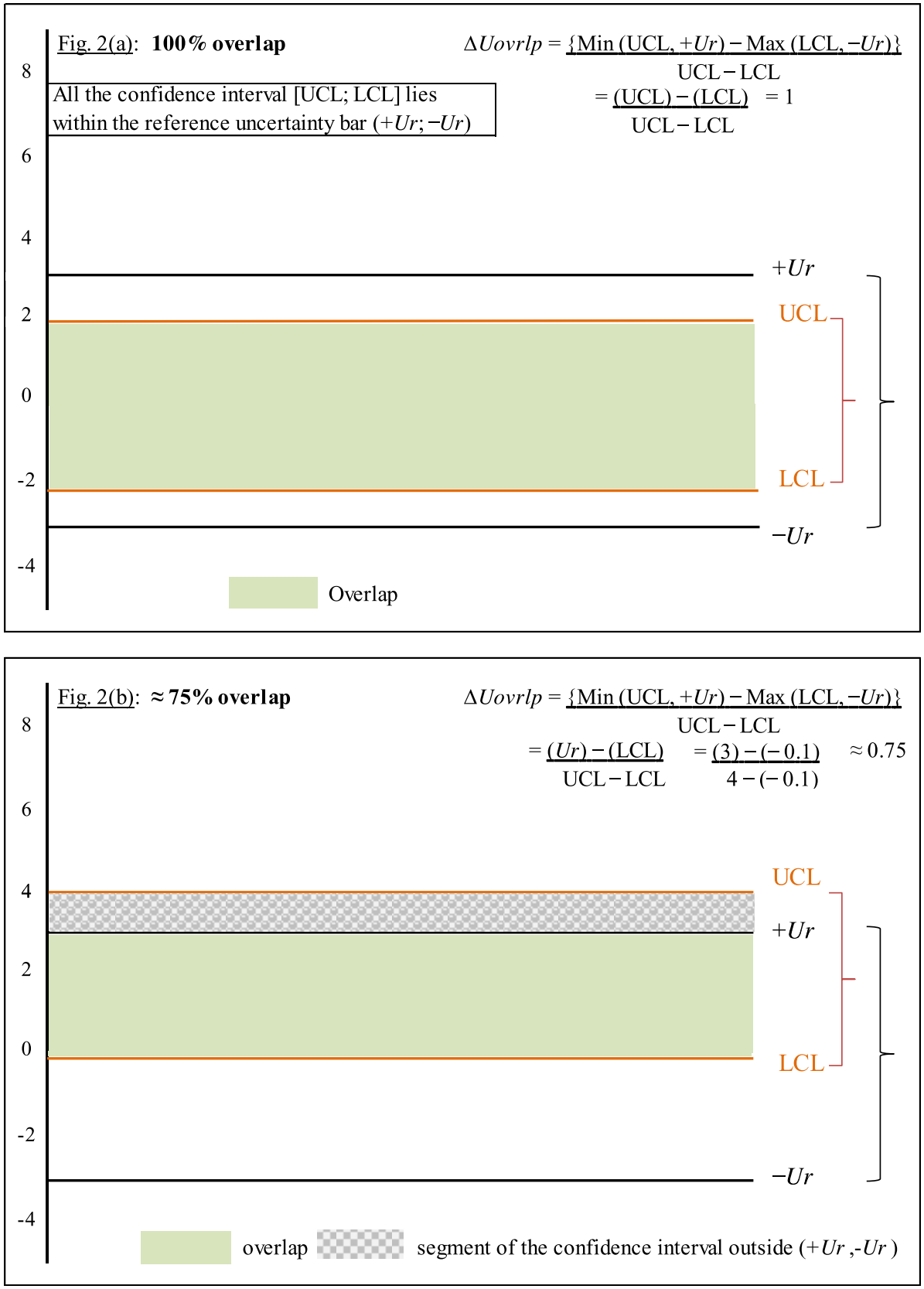
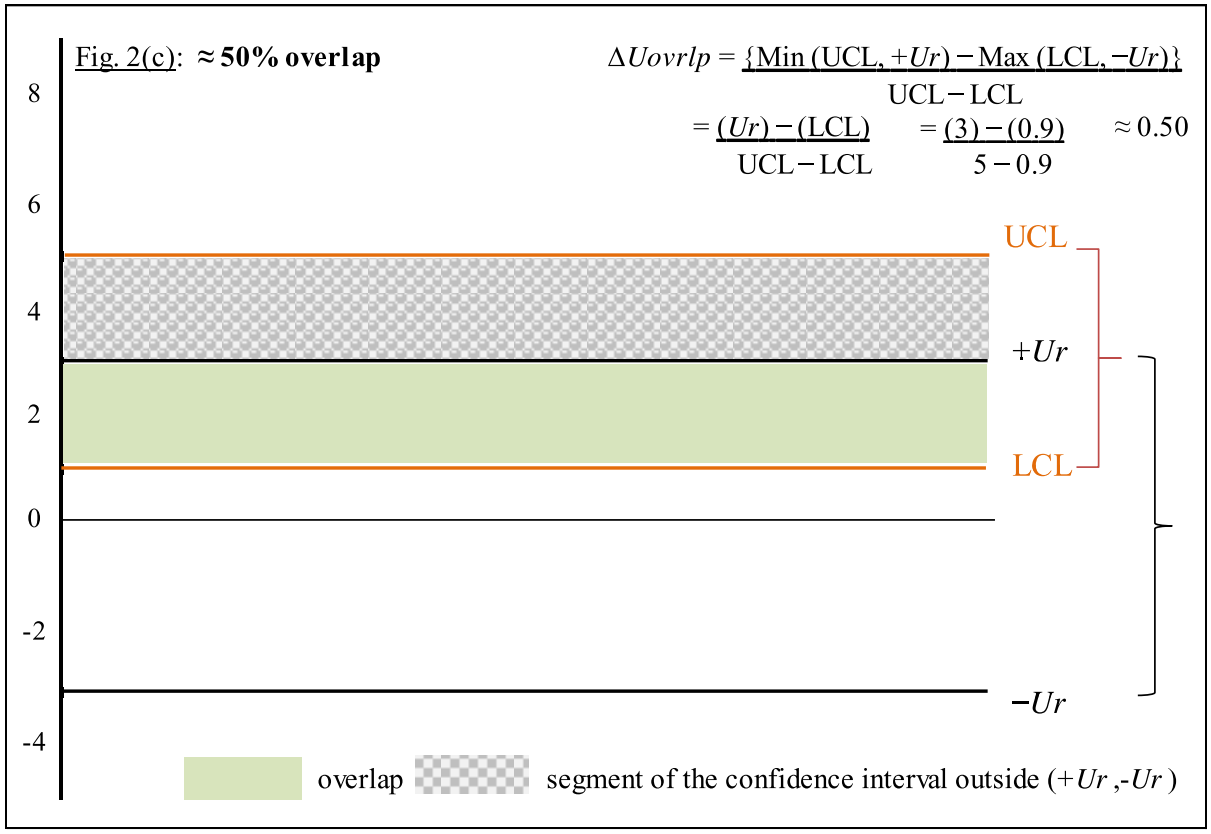


Fig. 2. Cartoon diagrams depicting 100%, 75%, 50%, 25% ΔU_{ovrlp} per equation (28) (Numbers on the y-axis are arbitrary for illustration purpose).



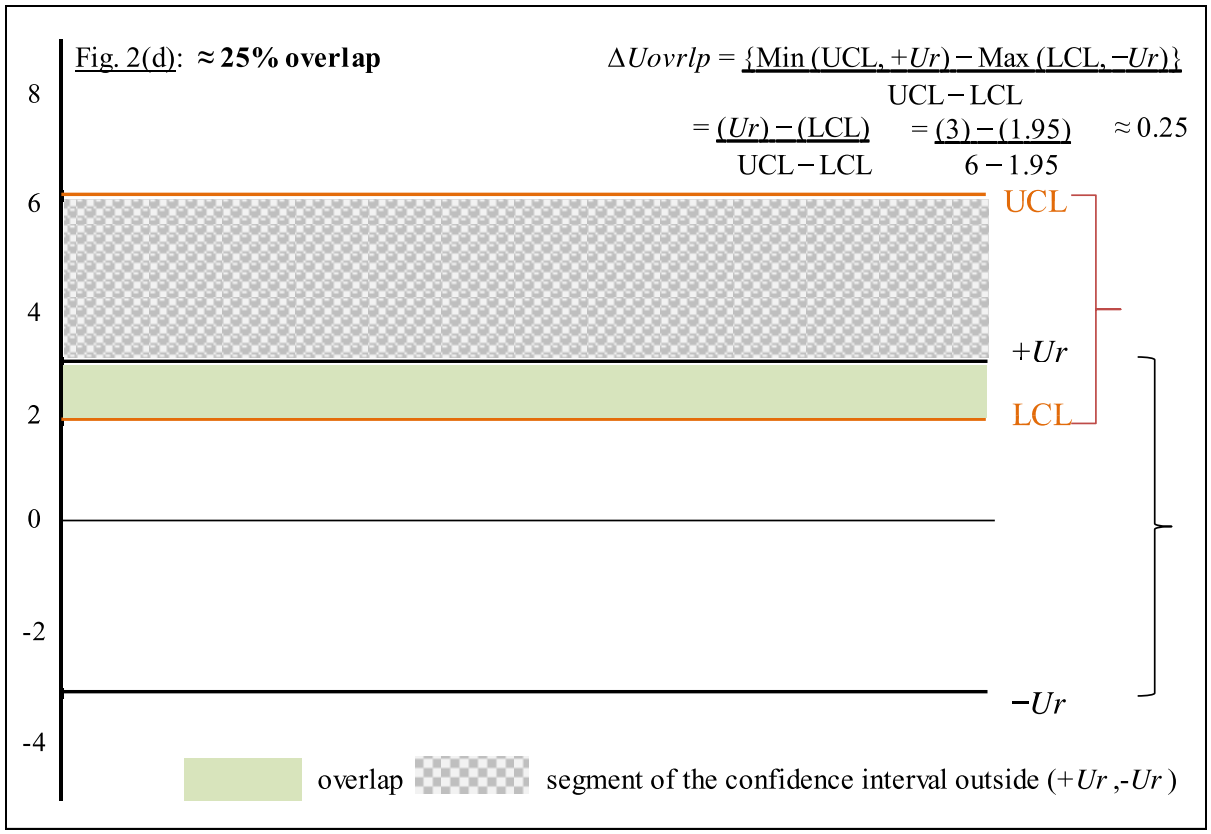


Fig. 2. Continued.

It is therefore needed to determine the confidence intervals for the regression slope and intercept, in addition to the confidence interval about the bias measurements scatter points. The basic formula (26) for confidence interval limits still applies; however the repeatability standard deviation by least squares regression is algebraically more complex due to the error sum of squares analysis and the slope and intercept statistics.

To proceed, we first present the general formulae of simple linear regression model, which are solutions to a linear equation in the parameters a and b: (for ref. we use [5–8]).

$$y = a + bx + \varepsilon \tag{30}$$

Given n scatter data points (y_i, x_i), the least squares estimators for the regression best fit line slope, $\boldsymbol{\beta}$, and intercept, $\boldsymbol{\alpha}$, are obtained by minimizing the sum of the squared deviations ε_i^2:

$$\boldsymbol{\beta} = S_{xy}/S_{xx} = \sum (x_i - \overline{x})(y_i - \overline{y}) / \sum (x_i - \overline{x})^2 \tag{31a}$$

$$\boldsymbol{\alpha} = \frac{1}{n}\left(\sum y_i - \boldsymbol{\beta} \sum x_i \right) \tag{31b}$$

where $\overline{x} = \frac{1}{n}\sum x_i$ and $\overline{y} = \frac{1}{n}\sum y_i$ are the samples' means of the x and y variables, respectively.

Working out the algebraic expressions yields the following decoupled formulae for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$:

$$\boldsymbol{\beta} = \left[\sum x_i y_i - \sum x_i \sum y_i / n \right] \Big/ \left[\sum {x_i}^2 - \left(\sum x_i \right)^2 / n \right] \tag{32a}$$

$$\boldsymbol{\alpha} = \left[\sum {x_i}^2 \sum y_i - \sum x_i \sum x_i y_i \right] \Big/ n \left[\sum {x_i}^2 - \left(\sum x_i \right)^2 / n \right] \tag{32b}$$

(Note: These formulae appear in the MSA manual with a and b interchanged ([1], p. 97). Here we use a and $\boldsymbol{\alpha}$ for intercept and b and β for slope, in alignment with [5–8]).

The regression best fit line points, $\hat{\boldsymbol{y}}$, would be expressed by the equation:

$$\hat{\boldsymbol{y}} = \boldsymbol{\alpha} + \boldsymbol{\beta} x \tag{33}$$

For repeated trial runs such as in bias linearity studies, the standard deviation for least squares repeatability residuals, s_{rr}, is estimated from variance of the y_i scatter points about the regressed best fit line $\hat{\boldsymbol{y}}$. In the so-called 'reduced major axis regression method' [6], this is done by summing the rectangles of deltas between the y_i data points and the expected values $\hat{\boldsymbol{y}}_i$ on the best fit line:

$$(s_{rr})^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2$$

$$= \frac{1}{n-2} \left[\sum {y_i}^2 - \boldsymbol{\alpha} \sum y_i - \boldsymbol{\beta} \sum x_i y_i \right] \tag{34}$$

where $(n - 2)$ are the degrees of freedom (df) associated with the bivariate analysis (since the estimator $\hat{\boldsymbol{y}}$ is dependent on *two* estimators: α and β). The right hand form of equation (34) is derived by expanding $\sum (y_i - \hat{\boldsymbol{y}}_i)^2$ and using $\hat{\boldsymbol{y}}$ from equation (33), then using $\boldsymbol{\beta} \sum x_i = (\sum y_i - n\boldsymbol{\alpha})$ and $\boldsymbol{\beta} \sum {x_i}^2 = (\sum x_i y_i - \boldsymbol{\alpha} \sum x_i)$ obtainable from the formulae (31b) and (32b), respectively.

Since $\hat{\boldsymbol{y}}$, α, β are not known, one needs to transform the formula (34) by substituting the formulae (32a) & (32b) into (34) followed by algebraic manipulation to obtain:

$$(s_{rr})^2 = \frac{1}{n-2} \left\{ \sum y^2 - n\overline{y}^2 - \left(\sum xy - n\overline{x}\,\overline{y} \right)^2 \Big/ S_{xx} \right\} \tag{35}$$

where for brevity we drop the subscript 'i', and use $S_{xx} = \sum (x_i - \overline{x})^2 = \sum x^2 - n\overline{x}^2$.

The estimated values of the covariate slope and intercept of the regressed best fit line are influenced by the scatter-dependence of the data points, on the premise that a set of simulated regression lines around the population's true best fit line represents a sampling-dependent statistic with slopes and intercepts distributed relative to the samples' means $\overline{x}$ and $\overline{y}$ [5]. Hence, variance components due to slope and intercept need to be considered in estimating the combined variance for the best fit line. Formulation can be simplified by realizing that: (a) all regression lines anchor at the point ($\overline{x}$,$\overline{y}$) such that $\overline{y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\overline{x}$ is valid; and (b) the intercept variance can be handled through a transformation at a specified x_0 value of the independent variable, such that $\hat{y} = \overline{y} + \boldsymbol{\beta}(x_0 - \overline{x})$. Assuming uncertainty is the same for all y_i measurements, the variance/standard error for the best fit line $\hat{y}$ is therefore a combination of the standard error of the repeatability mean $\overline{y}$, given by $V\overline{y} = (srr)^2/n$, and the variance of the estimated mean slope V_β multiplied by a factor. This is expressed by the following formula (for more details see [6] or [7]):

$$V\hat{y} = V\overline{y} + (x_0 - \overline{x})^2 V_{\boldsymbol{\beta}} = (s_{rr})^2 \left[\frac{1}{n} + (x_0 - \overline{x})^2 / S_{xx} \right] \tag{36}$$

where

$$V_{\boldsymbol{\beta}} = (s_{rr})^2 / S_{xx} \tag{37}$$

is obtainable from expression (31a) under the assumption of negligible uncertainty of the independent variable x; hence $V_{\boldsymbol{\beta}} = V_y / \sum (x_i - \overline{x})^2 = (s_{rr})^2 / \sum (x_i - \overline{x})^2$.

(See [8] p. 425 for proof of $V(cx) = c^2 V(x)$, where x is variable and c is a multiplication factor.)

The formula for the variance associated with the intercept of the best fit line is obtained additively from the equation $\overline{y} = \alpha + \beta\overline{x}$, and substitution of V_β from expression (37):

$$V_{\boldsymbol{\alpha}} = V_{\overline{y}} + \overline{x}^2 V_{\boldsymbol{\beta}} = (s_{rr})^2 \left[\frac{1}{n} + \overline{x}^2 / S_{xx} \right]$$

$$= (s_{rr})^2 \sum x^2 / n S_{xx} \tag{38}$$

Switching to standard deviation expressions since these will be used in the formulae of confidence intervals, we insert the formula (35) into (36) followed by algebraic manipulations and taking square root to obtain $S\hat{y}$, the calculable standard deviation of the best fit line $\hat{y}$:

$$S\hat{y} = \frac{1}{\sqrt{(n-2)}}\left\{\left[\sum y^2 - n\overline{y}^2 - \left(\sum xy - n\overline{x}\,\overline{y}\right)^2 \Big/ S_{xx}\right] \times \left[\frac{1}{n} + (x_0 - \overline{x})^2 / S_{xx}\right]\right\}^{1/2} \quad (39)$$

The calculable standard deviation associated with the slope of the best fit line is obtained by substitution of (35) in (37) and taking square root:

$$S_{\boldsymbol{\beta}} = \frac{1}{\sqrt{(n-2)}} \times \left\{\sum y^2 - n\overline{y}^2 - \left(\sum xy - n\overline{x}\,\overline{y}\right)^2 \Big/ \mathrm{S}xx\right\}^{1/2} \Big/ \sqrt{\mathrm{S}xx} \quad (40)$$

The calculable standard deviation associated with the intercept of the best fit line is obtained by substitution of (35) in (38) and taking square root:

$$S_{\boldsymbol{\alpha}} = \frac{1}{\sqrt{n(n-2)}} \times \left\{\sum x^2 \left[\sum y^2 - n\overline{y}^2 - \left(\sum xy - n\overline{x}\,\overline{y}\right)^2 \Big/ \mathrm{S}xx\right]\right\}^{1/2} \Big/ \sqrt{\mathrm{S}xx} \quad (41)$$

We are now in position to formulate the confidence intervals for $\hat{\boldsymbol{y}}$, β and α:

$$[\mathrm{UCL};\mathrm{LCL}]\hat{y} = \hat{\boldsymbol{y}} \pm (T_{crit})(S\hat{y}) = \boldsymbol{\alpha} + \boldsymbol{\beta}x_0 \pm (T_{crit})(S\hat{y}) \quad (42)$$

$$[\mathrm{UCL};\mathrm{LCL}]_{\boldsymbol{\beta}} = \boldsymbol{\beta} \pm (T_{crit})(S_{\boldsymbol{\beta}}) \quad (43)$$

$$[\mathrm{UCL};\mathrm{LCL}]_{\boldsymbol{\alpha}} = \boldsymbol{\alpha} \pm (T_{crit})(S_{\boldsymbol{\alpha}}) \quad (44)$$

Due to quadratic nonlinear components in the formulae above, the confidence interval points will trace hyperbolae curves at the lower and upper boundaries (see Figs. 4–6).

2.3.2 Linearity bias measurements and regression analysis

Gage linearity study requires a number of traceable reference standards or, in lieu consensus standards as appropriate, having accurate scalar values $\mathbf{R}r(1)$, $\mathbf{R}r(2)$, ..., $\mathbf{R}r(g)$, $(g \geq 5)$; such that the values cover the applicable operating range of the measurement system [1]. Information about the traceable or consensus-assessed uncertainty values $Ur(1)$, $Ur(2)$, ..., $Ur(\mathrm{g})$ must also be available.

Using a typical gage representing the measurement system, the reference standards are to be measured by a single qualified appraiser – or by automation, as applicable – using repeatability trials' sample size $m \geq 10$ for each reference subgroup $\mathbf{R}r(\mathrm{j})$. In what follows, we index the subgroup references by j and the m trials by i. To minimize appraiser memory recall, it is recommended to randomize the standards and trials [1], if practically feasible. Random number generator Excel sheet, for example, may be used to set up random sequences. (Note that random sequencing may not be practical for fully automated systems.)

After collecting the group $\{R\mathrm{ji}\}$ of reference measurement data for the g sets of repeatability trials, the bias value Bji for each individual trial is calculated, and all arranged in a matrix:

$$\{B_{ji}\}_{gm} = \{R_{ji} - \mathrm{R}r(j)\}_{gm}; \quad j = 1\,\mathrm{to\,g},\ \mathrm{i} = 1\ \mathrm{to\ m},\ (\mathrm{n} = \mathrm{gm}) \quad (45)$$

Using equation (23), the bias average is calculated for each subgroup j:

$$\mathrm{Bav(j)} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Bji} \quad (46)$$

The bias repeatability scatter data Bji (dependent variable y) and the bias averages per (46) are plotted against values of the reference standards (independent subgroup x). Simple least squares linear regression is applied using the formulae in § 2.3.1 to calculate the regression parameters and obtain and plot the best fit line. Calculations and plots may be performed with any desired package, e.g. Minitab, JMP, or recently the increasingly popular R [7,8]. However, we chose to set up the formulae and execute using Excel since it is widely used and gives users the opportunity to readily verify the formulae. Our linearity Excel worksheet calculates the bias scatter values Bji and the average Bav(j) for each subgroup; then computes, $\overline{x}$ $\overline{y}$, $\sum$x, $\sum$y, $\sum$xy, $\sum\mathrm{x}^2$, $\sum\mathrm{y}^2$ for the whole group (n = gm) and uses the formulae (32), (33), (39), (42) to determine the slope $\boldsymbol{\beta}$ and intercept $\boldsymbol{\alpha}$ of the best fit line, the regression's best fit points $\hat{\mathrm{y}}_i$, the standard deviation $\mathrm{s}_{\hat{\mathrm{y}}}$, and the 95%-confidence $[\mathrm{UCL};\mathrm{LCL}]_{\hat{\mathrm{y}}}$ points; plotting the best fit line and confidence hyperbolae curves. Moreover, it uses (40) and (41) to calculate the standard deviations $\mathrm{s}_{\boldsymbol{\beta}}$ and $\mathrm{s}_{\boldsymbol{\alpha}}$ and the t-stat values $\mathrm{Tstat}(\boldsymbol{\beta})$ and $\mathrm{Tstat}(\boldsymbol{\alpha})$; then uses (43) and (44) to calculate 95%–confidence $[\mathrm{UCL};\ \mathrm{LCL}]_{\boldsymbol{\beta}}$ and $[\mathrm{UCL};\ \mathrm{LCL}]_{\boldsymbol{\alpha}}$ limits. The value of $\mathrm{T}_{\mathrm{crit}}$ is obtained by the Excel function $= TINV$ (0.05, gm−2).

2.3.3 Linearity acceptance

The acceptance of gage linearity requires disposition by the null hypothesis statzero condition or, by extension as we propose to the statzero proxy criterion, at every reference point on the linearity range. We will use the disposition in § 2.2.2 for acceptance by statzero and the disposition in

§ 2.2.4 for acceptance by statzero proxy to establish the dispositions appropriate for linearity validation, and provide illustrative examples.

2.3.3.1. Statzero condition applied to linearity

This requires the null hypothesis $\{H_0: B = 0\}$ not to be rejected at *each* bias checkpoint corresponding to a reference standard in the linearity study, i.e. subject to validity of the statzero condition (27) over the operating range of the measurement system. Furthermore, the acceptance test includes the slope and the intercept also meeting statzero condition. This imposes the following requisites:

i) Zero is contained within the confidence interval around the regression's best fit points throughout the linearity range at every reference point j, whereby (42):

$$\boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{R}r(\mathrm{j}) - (\mathrm{T_{crit}})(\mathrm{S}\hat{y}) \leq \mathrm{zero} \leq \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{R}r(\mathrm{j}) + (\mathrm{T_{crit}})(\mathrm{S}\hat{y}) \tag{47}$$

$\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are calculated by (32a) & (32b) and $\mathrm{s}_{\hat{y}}$ is calculated by (39), using the substitutions:

$$\sum x = \sum \mathbf{R}r(\mathrm{ji}); \sum y = \sum \mathrm{Bji};$$
$$\sum xy = \sum \mathbf{R}r(\mathrm{ji})\mathrm{Bji}; \sum x^2 = \sum [\mathbf{R}r(\mathrm{ji})]^2;$$
$$\sum y^2 = \sum (\mathrm{Bji})^2; \overline{x} = \sum \mathbf{R}r(\mathrm{ji})/\mathrm{gm};$$
$$\overline{y} = \sum \mathrm{Bji/gm}; \; x_0 = \mathbf{R}r(\mathrm{j}); \text{ and } \mathrm{n} = \mathrm{gm}$$

ii) The null hypothesis is also applicable to the slope and intercept statistics, such that by (43) and (44):

$$\boldsymbol{\beta} - (\mathrm{T_{crit}})(\mathrm{S}_{\boldsymbol{\beta}}) \leq \mathrm{zero} \leq \boldsymbol{\beta} + (\mathrm{T_{crit}})(\mathrm{S}_{\boldsymbol{\beta}}) \tag{48}$$

$$\boldsymbol{\alpha} - (\mathrm{T_{crit}})(\mathrm{S}_{\boldsymbol{\alpha}}) \leq \mathrm{zero} \leq \boldsymbol{\alpha} + (\mathrm{T_{crit}})(\mathrm{S}_{\boldsymbol{\alpha}}) \tag{49}$$

iii) The Student t-test is valid for the slope and intercept statistics, such that:

$$\mathrm{T_{stat(\boldsymbol{\beta})}} < \mathrm{T_{crit}}; \quad \mathrm{T_{stat(\boldsymbol{\alpha})}} < \mathrm{T_{crit}} \tag{50}$$

Where

$$\mathrm{T_{stat(\boldsymbol{\beta})}} = \boldsymbol{\beta}/\mathrm{S}_{\boldsymbol{\beta}}; \quad \mathrm{T_{stat(\boldsymbol{\alpha})}} = \boldsymbol{\alpha}/\mathrm{S}_{\boldsymbol{\alpha}} \tag{51}$$

[Eqs. (51) are derived from the formula (2) by replacing $\overline{x}$ by the mean slope $\boldsymbol{\beta}$ or mean intercept $\boldsymbol{\alpha}$, applying $\mu = 0$ for the population of slopes and intercepts, and using the standard deviations of the mean slope and intercept, $\mathrm{s}_{\boldsymbol{\beta}}$ and $\mathrm{s}_{\boldsymbol{\alpha}}$, respectively.]

The validation of small sample linearity study is by default subject to fulfilling the null hypothesis statzero conditions (47), (48), (49), and the t-test (50). See the illustrative Example 4. On the other hand, if the result of a linearity study fails *any* of the conditions above, then the next step is to evaluate acceptance by the statzero proxy criterion which we have proposed in § 2.2.4 for single sample bias case; here to be tested for linearity validation at every reference point j of the linearity subgroup samples, as will be explained below.

2.3.3.2 Statzero proxy criterion applied to linearity

Based on the criterion developed in § 2.2.4 for single sample bias, the acceptability of linearity by statzero proxy is subject to assessing the amount of overlap, $\boldsymbol{\Delta}_{Uovrlp}$ as determined by expression (28), between the hyperbolae-bounded 95% confidence interval about the regression best fit line and the reference value uncertainty, at *each* of the linearity study reference values spaced across the gage applicable operating range. We consider linearity to be acceptable if $\boldsymbol{\Delta}\,Uovrlp$ is greater than 25% at every reference point, in alignment with the criterion (29). See the illustrative Examples 5 and 6.

3 Results & discussion

We will present generic examples and discuss them to illustrate the methods we proposed in § 2.1.2 check standard; § 2.1.3 consensus standard; § 2.2.2 single sample bias acceptance by statzero condition; § 2.2.4 single sample bias acceptance by statzero proxy criterion; § 2.3.3.1 linearity acceptance by statzero condition; and § 2.3.3.2 linearity acceptance by statzero proxy criterion.

3.1 Check standard evaluation

Example 1: A production site keeps a NMI-traceable thin film oxide wafer standard with quoted thickness and expanded uncertainty $\mathbf{R}\mathrm{t} \pm U_\mathrm{t} = (3000 \pm 5)$ nm. The site starts a new process that requires a film thickness of ≈1000 nm; however there is no available standard for this at the site so they decide to use in-house reference parts for MSA stability and GR&R. The thin film gage used by the site has resolution $\rho = 2$ nm and calibration uncertainty tolerance $Ug = \pm 2$ nm. The site metrology engineer proceeds to establish a check standard by best estimate of a 1000 nm target thermal oxide film on prime wafer using the procedure described in § 2.1.2, running repeatability measurement trials on the check wafer and on the available traceable standard wafer, obtaining the data sets in Table 1 resulting in $R_\mathrm{chk} \cong 1005$ nm and $Rm \cong 3010$ nm. Using the repeatability variance results from Table 1 and the values of ρ and Ug above with $T_\mathrm{crit} = 2.262$ ($m = 10$, $\alpha = 0.05$), equations (6) and (10) yield the measurement expanded uncertainty $U \cong 6.0$ nm for the check standard and $U' \cong 5.9$ nm for the traceable standard. By equation (8), the gage offset error $\boldsymbol{\Delta}\mathbf{R} = 3010–3000 = 10$ nm. Using this and the values of U and U', and half the value of the traceable standard expanded uncertainty (half $Ut = 2.5$ nm) into equation (12) gives the total estimated uncertainty for the check standard: $Uchk \cong 13.5$ nm. Hence the value of the in-house check standard is estimated to be $\mathbf{R}chk \cong (1005 \pm 14)$ nm. This is quite good for stability and GR&R studies. However, the gage offset of 10 nm will present an issue for bias and linearity studies since it represents a 'hidden' bias increment by ≈0.33% at 3000 nm which will not be accounted for if one uses the in-house check standard

Table 1. Example 1. Check standard trials (measurement unit = nm.).

Thin film gage	Trial index (m)	Check standard	Traceable standard
T	1	1004	3009
R	2	1007	3010
I	3	1005	3011
A	4	1003	3010
L	5	1005	3012
S	6	1004	3009
	7	1005	3012
	8	1006	3010
	9	1007	3009
	10	1005	3011
Trials mean (nm) ⊨		$\mathbf{R}chk \cong 1005$ {Eq. (1)}	$\mathbf{R}\mathrm{m} \cong 3010$ {Eq. (7)}
Repeatability ⊨ variance (nm^2)		$\mathrm{V}chk \cong 1.66$ {Eq. (4)}	$\mathrm{Vt} \cong 1.34$ {Eq. (9)}

whose assessed value is traceable only to the in-house gage. This demonstrates why using check standards for bias and linearity is not recommended unless there is no other option for a unique measurement characteristic and/or a unique gage system or for destructive testing, as already alluded to in § 2.1.1 and 2.1.2. In such cases, one may adjust the bias readings to account for the offset. For process control monitoring, applying the offset to collected process data in SPC charts − if known at the process target value − is reasonable provided the specified process tolerance is sufficiently accommodating to absorb any negative impact on process Cpk entitlement; otherwise one may consider adjusting the tolerance limits in correlation with the offset, if allowed. The MSA manual [1] advises that if a system has non-zero bias, the first thing to do is attempt to recalibrate or remodify it to remove the offset, i.e. reset the gage to zero bias. If this is not successful, the manual posits that the gage may still be used by correcting for the offset at every measurement reading.

3.2 Consensus standard evaluation

Example 2: Four factory sites of a company, FAC-1–FAC-4, need a dimensional measurements standard for a characteristic feature on new product with target pitch of ≈500 nm ± 1.0% tolerance, to be verified by contactless profilometry. Traceable standards of titanium alloy with micro-etched features are available commercially but too expensive to purchase. The sites decide to adopt a self-made 3D-printed reference block which includes a ≈500 nm trench as a consensus standard for their profilometry systems. The gages calibration uncertainty values are $Ug = 1.0$ nm, 1.0 nm, 1.5 nm, and 1.5 nm, respectively for FAC-1–FAC-4; and the gage resolution $\rho = 0.5$ nm as quoted by OEM manual. The sites then run repeatability measurement trials on the feature using the procedure described in § 2.1.3, obtaining 4 independent data sets shown in Table 2. This table also shows results per site of the trials means $Rp(s)$, the repeatability variance $Vp(s)$ by (14), and the measurement expanded uncertainty $U(s)$ by (15). Using equation (13) with the values of $Rp(s)$ in Table 2 yields the estimated consensus value $\mathbf{R}con = 501.9$ nm. Using equation (17) with the values of $Vp(s)$ in Table 2 and $k = 4$ yields $Vms = 2.1$ nm. Using equation (18) with the values of $\mathrm{R}p(\mathrm{s})$ in Table 2 yields $\varepsilon(\mathrm{M}ss) = 0.4$ nm. Using equation (20) with the values of Ug yields $\overline{U}_g = 1.27$ nm. And finally, using equation (22) with the numerical results above yields the expanded uncertainty for the consensus standard: $Ucon = 4.1$ nm. Hence the consensus standard value is best estimated to be $\mathbf{R}con \pm Ucon \cong (502 \pm 4)$ nm.

Graphically, Figure 3 shows the readings for each gage, the mean value of the measurements, and the error bars as calculated by equation (6) for expanded uncertainty of individual subgroup. It also shows the consensus value $\mathbf{R}con$ of 502 nm and its error bar of ±4 nm. It is a validation of our method that the ANOVA-estimated consensus expanded uncertainty error bar of ±4 nm encompasses the individual gage readings and error bars, within the target tolerance of ±5 nm.

The consensus standard round-robin method whereby samples of measurement trials for the same reference part are performed on independent measurement systems, coupled with ANOVA modeling, enhances the confidence in traceability and provides assurance that the estimated group mean $\mathbf{R}con$ represents a reasonably accurate value in the vicinity of the true population mean within the expanded uncertainty bar of ± $Ucon$.

3.3 Single sample bias disposition

Example 3: To illustrate the statzero and statzero proxy dispositions for single sample bias, suppose the factory site FAC-1 of Example 2 uses the established consensus standard reference of (502 ± 4) nm to run bias trials on four similar systems A, B, C, D in different processing areas of their factory, collecting the data in Table 3 and obtaining the results in Table 4. (Note that system A and system B are matched in precision by having similar expanded uncertainty of ±2.5 nm, while C and D are also matched at ±2.8 nm.) The results in Table 4 show that all four systems have lower means relative to the consensus reference value, with progressively negative bias offset

Table 2. Example 2. Consensus trials; $k = 4$ factories (FAC-1–FAC-4).

Factory site ⇛		FAC-1	FAC-2	FAC-3	FAC-4
Measurement gage (profilometer) ⇛		PFL-1	PFL-2	PFL-3	PFL-4
Gage calibration uncertainty Ug (nm) ⇛		1.0	1.0	1.5	1.5
	1	501	504	504	505
	2	500	503	500	504
	3	502	501	502	503
	4	503	503	503	502
T	5	502	500	502	503
	6	501	504	499	501
R	7	500	504	500	502
	8	499	502	504	503
I	9	502	504	500	500
	10	501	503	501	502
A	11	503	501	503	504
	12	501	502	500	503
L	13	503	504	503	500
	14	500	503	502	503
S	15	501	500	501	501
	16	500	501	502	504
	17	503	503	501	500
	18	502	504	503	501
Measurement unit = nm	19	498	502	500	503
	20	500	501	503	503
Trials mean per factory site (nm) ⇛ {Rp(s), Eq. (13)}		501.1	502.5	501.7	502.4
Repeatability variance per site (nm^2) ⇛ {Vp(s), Eq. (14)}		1.99	1.94	2.24	2.13
Expanded uncertainty per site (nm) ⇛ {U(s), Eq. (15)} ⁋⁑		2.60	2.59	3.46	3.45
Consensus reference value ⇛ {Eqs. (13) & (22)}		**R**$con \pm Ucon \cong (502 \pm 4)$ nm			

⁋ Gage resolution $\rho = 0.5$ nm.
⁑ Tcrit($\alpha/2$, m $-$ 1) = 2.09, (for $\alpha = 0.05$, $m = 20$).

and confidence intervals shifting to negative numbers. System A's mean of (501.5 ± 2.5) nm is the closest to the reference value and shows the smallest negative bias (0.09%). This is acceptable by statzero condition (27) since zero is contained within the confidence interval and Tstat is less than Tcrit, as seen in Table 4. System B shows 0.15% negative bias, slightly more than system A; however because the confidence interval slips below zero in negative territory and Tstat goes above Tcrit, system B is not accepted by statzero, even though it is matched in precision to system A, (note the sensitivity of the statzero hypothesis, there is only ≈0.25 nm difference between the trials means of systems A and B.) Applying equation (28) to system B's data gives 100% Δ_{Uovrlp}, [Table 4]; hence system B is acceptable by statzero proxy criterion (29). On the other hand, systems C and D exhibit bias an order of magnitude larger than system A, clearly away from the statzero zone. However, testing by the statzero proxy criterion shows that system C has 31% Δ_{Uovrlp}, so its bias is acceptable by proxy and can be tolerated. System D, which is matched in precision to system C but exhibits slightly more negative bias than system C, just fails the statzero proxy criterion (29) by having 23% overlap, and thus its bias error cannot be tolerated. Action must be undertaken to investigate the source of the intolerable negatively-offset bias problem of system D, and adjustments should be made to bring it back to statzero or at least statzero proxy status.

In general, if the size of bias offset is within the maximum permissible calibration error (uncertainty tolerance) set by the gage manufacturer, then one may, if possible, tune the gage by counter-offset to correct the bias problem. However, if the size of offset exceeds the maximum permissible calibration error and, we propose, fails the statzero condition *and* statzero proxy criterion, then the gage is not acceptable and should be subjected to corrective recalibration or hardware/software modification. In this illustrative example, the gage calibration

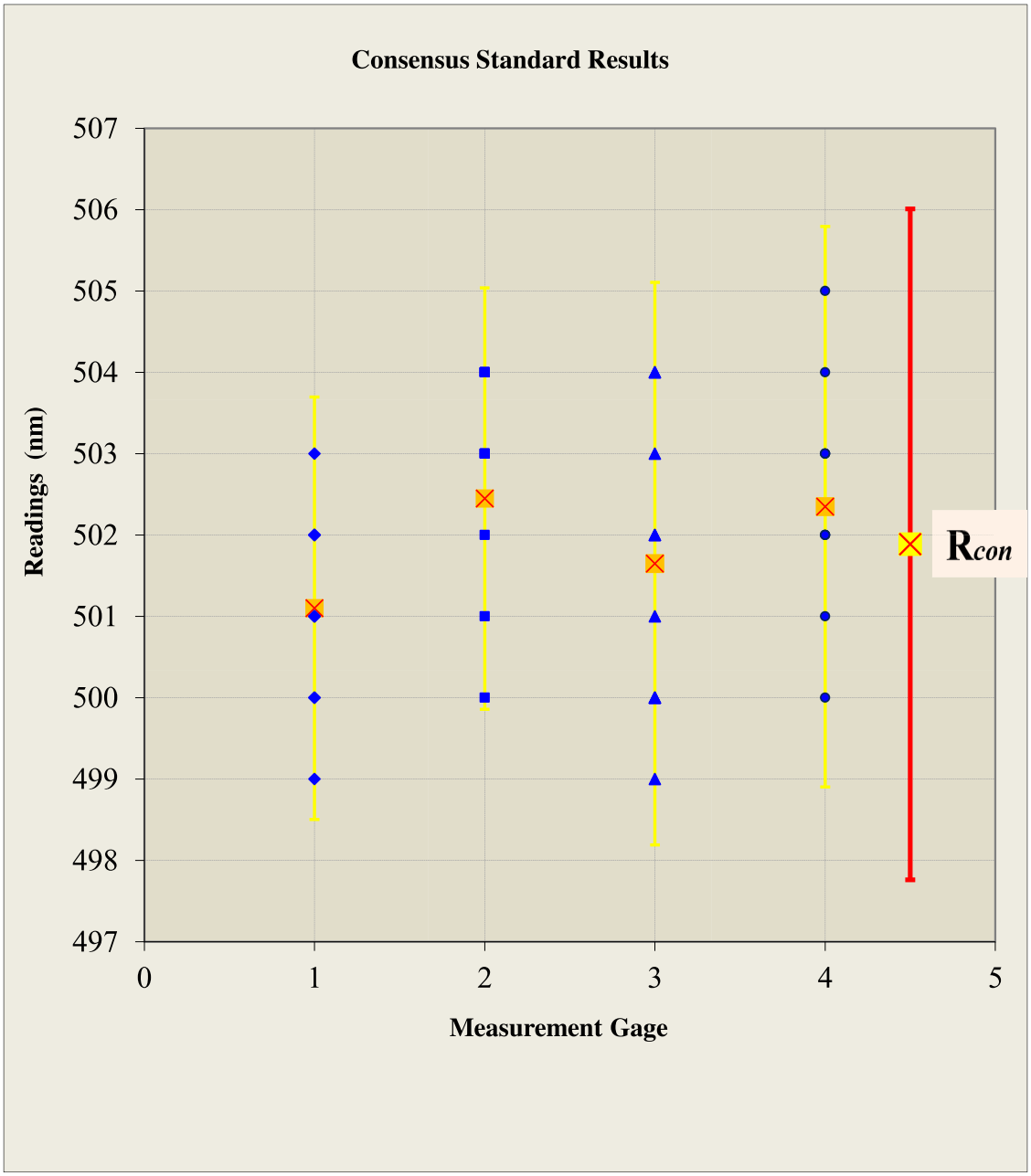


Fig. 3. Example 2. Results of consensus standard work: $R_{con} \cong (502 \pm 4)$ nm (red bar.) Yellow bars represent the ± expanded measurement uncertainty per Eq. (22). [See Example 2].

uncertainty $Ug = 1.0$ nm translates to a maximum permissible error of $\approx \pm 0.2\%$ (relative to the reference value 502 nm). Both systems C and D exceed this error; however system C passes by the statzero proxy criterion and so is considered still in the accuracy zone, i.e. acceptable for use in process/product measurements with attempt to counter the offset bias if possible. On the other hand, by failing the statzero proxy criterion, system D has drifted outside the accuracy zone, so attempting to tune the nonconforming offset bias back is not the best course of action since the system may have significant hardware/software issues that need to be investigated and addressed.

3.4 Gage linearity

To illustrate the statzero and statzero proxy dispositions for linearity acceptance, we present and discuss the following generic examples:

3.4.1 Linearity acceptable by statzero

Example 4: Suppose that in addition to the 500 nm feature in Example 2, other 3D-printed micro-etched blocks are patterned with features at target pitches ≈1000, 1500, 2250, and 3000 nm, and maximum tolerances

Table 3. Example 3. FAC-1; 4 measurement systems; Bias trials, Consensus standard = (502 ± 4) nm, (measurement unit = nm.).

Trial index (m)	System A	System B	System C	System D
1	503	501	498	499
2	500	502	497	496
3	502	501	500	499
4	501	503	497	496
5	503	500	500	499
6	503	501	496	498
7	501	500	497	497
8	502	502	498	499
9	502	503	495	495
10	501	502	498	498
11	502	501	499	499
12	500	500	498	495
13	501	500	495	499
14	502	501	498	496
15	500	502	499	498
Trials mean & expanded uncertainty ⇛ by system (nm) {Eqs. (13) & (15)}⁋§⁂	501.5 ± 2.5	501.3 ± 2.5	497.7 ± 2.8	497.5 ± 2.8

Table 4. Example 3 results (FAC-1, 4 systems).

system ▼	Bias average (nm) (Bav %)*	Confidence** interval (nm) [UCL; LCL]	Tstat⁂ {Eq. (25)}	Bav accepted by statzero? {condn. (25)}	(Δ_{Uovrlp}) {Eq. (28)}	Bav accepted by statzero proxy? {criterion (29)}
A	–0.47 (0.09%)	0.12; −1.05	1.70	Yes		
B	–0.73 (0.15%)	–0.16; −1.31	2.75	No	100%	Yes
C	–4.33 (0.86%)	–3.48; −5.19	10.9	No	31%	Yes
D	–4.47 (0.89%)	–3.61; −5.33	11.1	No	23%	No

*Bav % as percentage of the consensus reference value 502 nm.
**Calculated by equation (26).
⁋Gage resolution $\rho = 0.5$ nm.
§Gage calibration uncertainty tolerance $Ug = 1.0$ nm.
⁂Tcrit($\alpha/2$, $m - 1$) = 2.14 ($\alpha = 0.05$, $m = 15$).

of ±0.8%, ±0.6%, ±0.5%, and ±0.4% respectively. The four sites which participated in generating the consensus standard $\mathbf{R}con(1) \cong (502 \pm 4)$ nm now run trial measurements on the other four features and generate consensus reference parts with the following values and expanded uncertainty: $\mathbf{R}con(2) \cong (1012 \pm 5)$ nm, $\mathbf{R}con(3) \cong (1509 \pm 5)$ nm, $\mathbf{R}con(4) \cong (2262 \pm 6)$ nm, and $\mathbf{R}con(5) \cong (3015 \pm 6)$ nm. FAC-1 site then uses the five consensus standards for a linearity study on their measurement system A. The trials data shown in Table 5 are analyzed by simple linear regression analysis using Excel worksheet to obtain:

– the best estimated values of the regression slope and intercept, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, by equations (32a & 32b);
– the best fit line points by equation (33): $\hat{\mathbf{y}} = \boldsymbol{\alpha} + \boldsymbol{\beta}$xo, and using xo = $\mathbf{R}con$(j), j = 1 to 5;
– the points tracing the upper and lower confidence hyperbolae about the best fit line, calculated for xo = $\mathbf{R}con$(j) using equations (42) with (39);
– the upper and lower confidence limits for the slope and intercept, by equations (43) with (40) and equations (44) with (41), respectively; and
– the values of Tstat($\boldsymbol{\beta}$) and Tstat($\boldsymbol{\alpha}$) for the slope and intercept t-statistics, by equation (51).

{Tcrit is obtainable from standard statistics tables or by the Excel function *TINV*(0.05, gm − 2) at 95% confidence level.}

Table 6 shows the regression analysis results for the best-estimated slope and intercept, and Table 7 shows the results for the best fit line. Both tables validate the statzero conditions: (48) for slope, (49) for intercept, and (47) for best fit line, as well as the Student t-test (50), are

Table 5. Example 4. FAC-1; measurement system A; Linearity study.

Consensus Standard Ref. value (nm) ⇛		Rcon(1) 502	Rcon(2) 1012	Rcon(3) 1509	Rcon(4) 2262	Rcon(5) 3015
Uncertainty (nm) ⇛		4	5	5	6	6
	1	500	1011	1508	2261	3016
	2	502	1014	1510	2259	3013
T	3	501	1012	1509	2262	3018
R	4	503	1010	1512	2264	3016
I	5	498	1008	1507	2263	3012
A	6	501	1009	1510	2265	3014
L	7	502	1010	1506	2261	3016
S	8	504	1014	1509	2259	3014
	9	501	1011	1507	2263	3011
	10	502	1014	1510	2262	3012

Table 6. Example 4. FAC-1; system A; linear regression analysis results; slope & intercept.

Best estimated slope value (**β**) ▼	Upper confidence limit for **β**	Lower confidence limit for **β**	Tstat $_{(\boldsymbol{\beta})}$ ▼
{Eq. (32a)}	{Eq. (43) with (40)}⁂		(51)
– 8.6 E – 06	6.3 E – 04	– 6.1 E – 04	0.03
Best estimated intercept (**α**) ▼	Upper confidence limit for **α**	Lower confidence limit for **α**	Tstat $_{(\boldsymbol{\alpha})}$ ▼
{Eq. (32b)}	{Eq. (44) with (41)}⁂		(51)
– 0.49	0.67	– 1.66	0.85

⁂ Tcrit($\alpha/2$, gm − 2) = 2.01, (α =0.05; gm = 50, df = 48).

Table 7. Example 4. FAC-1; system A; linear regression analysis results; best fit line.

Reference value **R***con*(j) (nm) ⇛	502	1012	1509	2262	3015
Bias average, Bav (nm) ⇛	–0.60	–0.70	–0.20	–0.10	–0.80
(Bav %)* ⇛	(0.12%)	(0.07%)	(0.01%)	(0.004%)	(0.03%)
Best fit line $\hat{\mathbf{y}}$ values (nm) ⇛	–0.49	–0.49	–0.48	–0.47	–0.47
(Eq. (33) $\hat{\mathbf{y}} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{R}con(\mathrm{j})$★					
Best fit line hyperbolic upper confidence limit (nm) ⇛ {Eq. (42) with (39)}★	0.41	0.20	0.08	0.19	0.53
Best fit line hyperbolic lower confidence limit (nm) ⇛ {Eq. (42) with (39)}★	–1.39	–1.17	–1.04	–1.14	–1.47

* As a percentage of consensus reference value.
★xo = Reference value **R***con*(j).

all met at 95% confidence, with zero contained within the respective confidence intervals and both Tstat($\boldsymbol{\beta}$) and Tstat$_{(\boldsymbol{\alpha})}$ less than Tcrit. Accordingly, the linearity of measurement system A is acceptable by the statzero condition. Note that this is true even as the best fit line shows a slight negative bias intercept of ≈ -0.5 nm through the range studied, as seen in Table 7 and the plot in Figure 4.

3.4.2. Linearity acceptable by statzero proxy

Example 5: Suppose the FAC-1 site of Example 3 next uses the five consensus standards of Example 4 for a linearity study on their measurement system C. The measurement trials are shown in Table 8, and the linear regression analysis results are in Tables 9 and 10. These show that statzero condition is satisfied for the slope, with

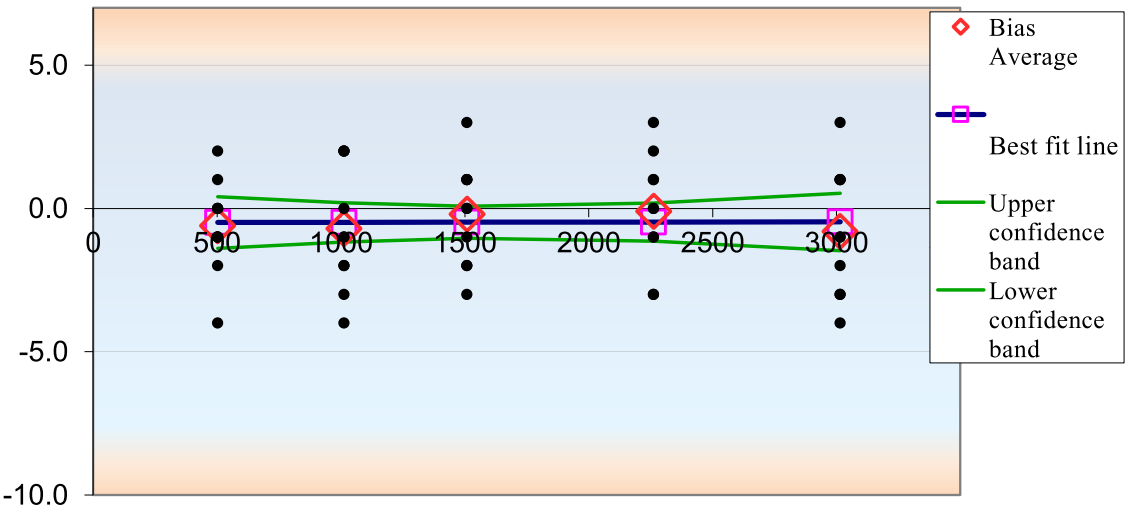


Fig. 4. Example 4, FAC-1, measurement system A: Linearity accepted by statzero conditions. Reference values are on the x-axis in nm units {xo = **R**con(j)}. Bias data are on the y-axis in nm units (solid circles).

Table 8. Example 5. FAC-1; measurement system C; linearity study.

Consensus Standard Ref. value (nm) ⇛		Rcon(1) 502	Rcon(2) 1012	Rcon(3) 1509	Rcon(4) 2262	Rcon(5) 3015
Uncertainty (nm) ⇛		4	5	5	6	6
	1	500	10080	15050	22570	30110
	2	497	10110	15020	22550	30080
T	3	500	10090	15060	22580	30130
R	4	496	10100	15050	22600	30110
I	5	497	10060	15040	22590	30120
A	6	495	10060	15080	22560	30150
L	7	496	10100	15030	22560	30140
S	8	500	10070	15060	22570	30090
	9	499	10080	15070	22610	30120
	10	498	10050	15040	22580	30100

Table 9. Example 5. FAC-1; system C; linear regression analysis results; slope & intercept.

Best estimated slope value (**β**) ▼	Upper confidence limit ▼	Lower confidence limit ▼	Tstat(**β**) ▼
{Eq. (32a)}	{Eq. (43) with (40)}⁂		(51)
1.8 E − 04	7.9 E − 04	–4.2 E − 04	0.61
Best estimated intercept (α) ▼	**Upper confidence limit ▼**	**Lower confidence limit ▼**	**Tstat(α) ▼**
{Eq. (32b)}	{Eq. (44) with (41)}⁂		(51)
–4.3	–3.2	–5.4	7.55

⁂ Tcrit(α/2, gm − 2) = 2.01, (α =0.05; gm = 50, df = 48).

zero contained within the slope's confidence interval and Tstat (**β**) < Tcrit, but is not satisfied for the best fit line nor for the intercept; hence system C linearity is not accepted by statzero hypothesis. On the other hand, applying the statzero proxy criterion (28) and (29) gives the results in Table 11, which validate that all overlaps are >25%. Hence, linearity of measurement system C is acceptable by statzero proxy. Note that the bias average over the linearity range is in the negative zone as evidenced by the results in Table 10 and the graph of Figure 5, showing

Table 10. Example 5. FAC-1; system C; linear regression analysis results; best fit line.

Reference value $\mathbf{R}con(j)$ (nm) ⇛	502	1012	1509	2262	3015
Bias average, Bav (nm) ⇛ (Bav %)* ⇛	–4.2 (0.84%)	–4.0 (0.40%)	–4.0 (0.27%)	–4.3 (0.19%)	–3.5 (0.12%)
Best fit line $\hat{\mathbf{y}}$ values (nm) ⇛ (Eq. 33) $\hat{\mathbf{y}} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{R}con(j)$★	–4.21	–4.12	–4.03	–3.89	–3.75
Best fit line hyperbolic upper confidence limit (nm) ⇛ {Eq. (42) with (39)}★	–3.32	–3.45	–3.48	–3.24	–2.77
Best fit line hyperbolic lower confidence limit (nm) ⇛ {Eq. (42) with (39)}★	–5.10	–4.79	–4.58	–4.54	–4.74

* As a percentage of the consensus reference value.
★xo = Reference value $\mathbf{R}con(j)$.

Table 11. Example 5 results (FAC-1, system C). Overlap of the confidence interval★ with uncertainty of the reference values (ΔU_{ovrlp}, {Eq. (28)}).

Reference value uncertainty (nm) ⇛	4	5	5	6	6
$\Delta Uovrlp$ {Eq. (28)} ⇛	38%	100%	100%	100%	100%
Bav* accepted by statzero proxy criterion (29)? ⇛	Yes	Yes	Yes	Yes	Yes

★ Upper and lower confidence limits in Table 10.
* Bias average values in Table 10.

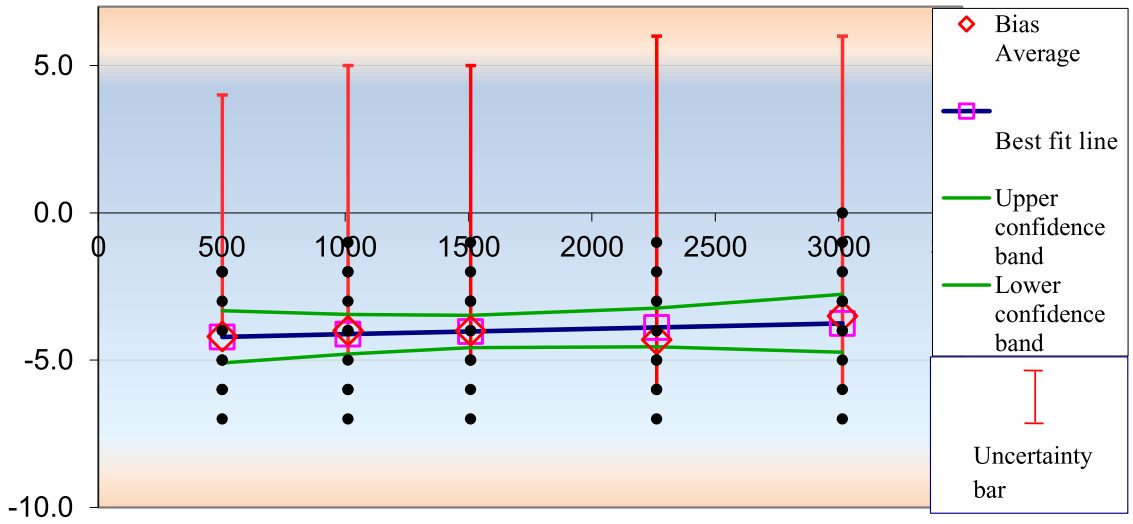


Fig. 5. Example 5, FAC-1, measurement system C: Linearity accepted by statzero proxy. Reference values are on the x-axis in nm units [xo = $\mathbf{R}con(j)$]. Bias data are on the y-axis in nm units (solid circles)

a small linear gradient from − 4.2 nm for $\mathbf{R}con(1)$ to − 3.5 nm for $\mathbf{R}con(5)$ at a small slope of 1.8 E − 4. Nonetheless, acceptance is justified by the amount of overlap between the confidence interval about the regression best fit line and the reference uncertainty bar being more than 25% for each of the five reference points [Tab. 11]), ensuring the gage is in the accuracy zone with acceptable linearity by regression analysis over the operating range. This facilitates tuning the gage back to statzero, if possible, by an amount equivalent to the linear regression's best line intercept, in this example approximately 4 nm. Alternatively, if practical, the offset may be applied to individual measurement points as the process/product data are being collected.

3.4.3 Linearity unacceptable

Example 6: Suppose the FAC-1 site of Example 3 next uses the five consensus standards of Example 4 for a linearity study on their measurement system D. The measurement

Table 12. Example 6. FAC-1; measurement system D; linearity study.

Consensus Standard Ref. value (nm) ⇛		Rcon(1) 502	Rcon(2) 1012	Rcon(3) 1509	Rcon(4) 2262	Rcon(5) 3015
Uncertainty (nm) ⇛		4	5	5	6	6
	1	500	10080	15050	22570	30110
	2	497	10040	15020	22550	30080
T	3	494	10090	15060	22610	30130
R	4	496	10100	15050	22530	30110
I	5	495	10060	15040	22590	30120
A	6	499	10060	15020	22560	30090
L	7	497	10040	15030	22560	30140
S	8	494	10070	15000	22540	30090
	9	498	10080	15010	22590	30120
	10	496	10050	15040	22580	30100

Table 13. Example 6. FAC-1; system D; linear regression analysis results; slope & intercept.

Best estimated slope value (**β**) ▼	Upper confidence limit ▼	Lower confidence limit ▼	Tstat(**β**) ▼
{Eq. (32a)}	{Eq. (43) with (40)}**		(51)
4.7 E − 04	1.1 E − 03	1.9 E − 04	1.44
Best estimated intercept (**α**) ▼	Upper confidence limit ▼	Lower confidence limit ▼	Tstat(**α**) ▼
{Eq. (32b)}	{Eq. (44) with (41)}**		(51)
–5.9	–4.7	–7.2	9.65

** Tcrit(α/2, gm − 2) = 2.01, (α =0.05; gm = 50, df = 48).

Table 14. Example 6. FAC-1; system D; linear regression analysis results; best fit line.

Reference value **R***con*(j) (nm) ⇛	502	1012	1509	2262	3015
Bias average, Bav (nm) ⇛ (Bav %)* ⇛	–5.4 (1.08%)	–5.3 (0.52%)	–5.8 (0.38%)	–5.2 (0.23%)	–4.1 (0.14%)
Best fit line $\hat{\mathbf{y}}$ values (nm) ⇛ (Eq. (33)) $\hat{\mathbf{y}} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{R}con(j)$★	–5.71	–5.47	–5.23	–4.88	–4.52
Best fit line hyperbolic upper confidence limit (nm) ⇛ {Eq. (42) with (39)}★	–4.75	–4.74	– 4.64	–4.17	–3.46
Best fit line hyperbolic lower confidence limit (nm) ⇛ {Eq. (42) with (39)}★	–6.67	–6.19	–5.83	–5.58	–5.59

* As a percentage of the consensus reference value.
★xo = Reference value **R***con*(j).

trials are shown in Table 12, and the linear regression analysis results are in Tables 13 and 14. These show that statzero condition is satisfied for the slope, with zero contained within the slope's confidence interval and $\text{Tstat}_{(\boldsymbol{\beta})} < \text{Tcrit}$, but is not satisfied for the best fit line nor for the intercept; hence system D linearity is not accepted by statzero hypothesis. Applying the statzero proxy criterion (28) and (29) gives the results in Table 15, which shows the >25% overlap criterion is valid for **R***con*(3) − **R***con*(5), but not valid for **R***con*(1) and **R***con*(2). Hence linearity of measurement system D is not acceptable by statzero proxy. The results in Table 14 show the bias average over the linearity range in the negative zone but, unlike in Example 5, it is nonlinear as it exhibits an inflexion point at **R***con*(3) = 1509 nm, graphically depicted in Figure 6 at the intersection of the two dashed lines. The linear regression results show a slope of 4.7 E − 4, which is 2.6 times the slope in example 5 (1.8 E − 4), and intercept of − 5.9 nm. These results indicate that system D is non-linear

Table 15. Example 6 results (FAC-1, system D). Overlap of the confidence interval* with uncertainty of the reference values (Δ_{Uovrlp}, {Eq. (28)}).

Reference value uncertainty ⊨ (nm)	4	5	5	6	6
Δ_{Uovrlp} {Eq. (28)} ⊨	0%	18%	31%	100%	100%
Bav* accepted by statzero proxy ⊨ criterion (29)?	No	No	Yes	Yes	Yes

* Upper and lower confidence limits in Table 14.
* Bias average values in Table 14.

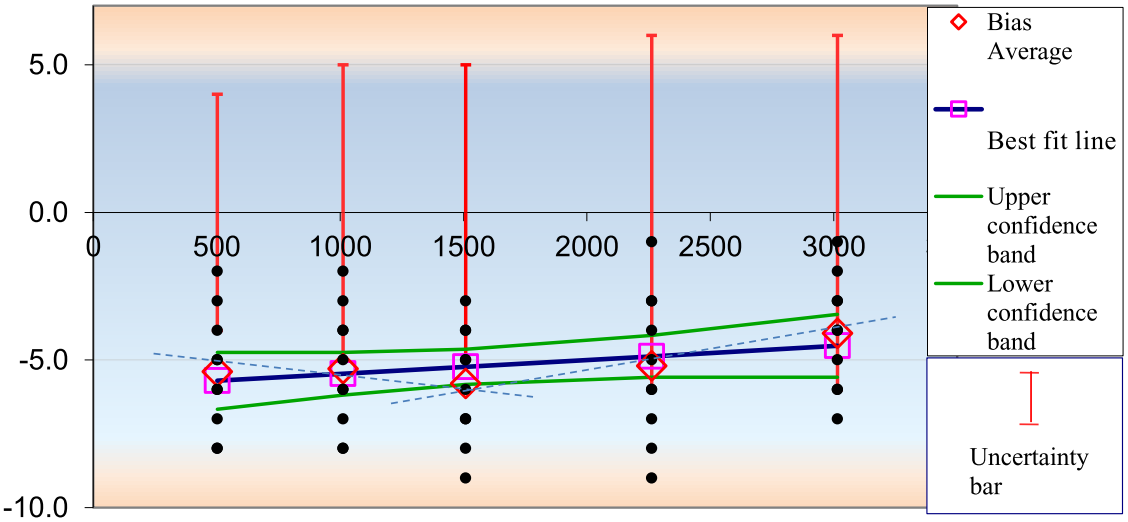


Fig. 6. Example 6, FAC-1, measurement system D: Linearity not acceptable. Reference values are on the x-axis in nm units [xo = **R***con*(j)]. Bias data are on the y-axis in nm units (solid circles).

and hence does not lend itself to simple tuning back to statzero, or applying a uniform offset to the data points. This system's gage has to be subjected to corrective recalibration and/or hardware/software modification to fix the bias non-linearity problem.

3.4.4. Range consideration

When the product manufacturing or test/inspection process spans a wide range of characteristic measurements, it is recommended to validate MSA linearity using three studies with three sets of reference parts, each set having at least five distinctly independent references representing the low end, mid-range, and high end of the production measurements. A similar approach may be adopted if the measured characteristic has ranges that differ widely by technology type.

4 Conclusions

This paper starts by introducing methods for establishing reference for MSA bias and linearity studies when there are no available traceable standards; in particular a method for establishing consensus and check standards values and expanded uncertainty using a nested ANOVA approach. The paper argues for unsuitability of check standards, however, for evaluating bias and linearity of measurement systems due to limitation of self-traceability (even though check standards are appropriate for stability and GR&R studies of gage systems). We then proceed to present the mathematical t-statistic based background for studies of gage bias and linearity, providing the appropriate formulae for the single reference bias case as well as deriving the formulae for simple linear regression analysis needed for multi-reference bias linearity validation. For acceptance, we primarily use the null-hypothesis statistical zero bias (statzero) condition, combined with the Student's t-test to justify acceptance of bias and linearity given the small samples normally used in such studies (typically $10 \leq m \leq 20$.). Moreover, we propose a novel idea of taking in consideration the degree of overlap between the confidence interval of bias fit data or the confidence hyperbole in case of linearity regression analysis, to extend acceptance of gage bias and linearity according to the criterion of >25% overlap between confidence intervals and the uncertainty bars of the reference standards used in bias and linearity studies. We call this extended test for significant overlap the statzero proxy criterion. We provide illustrative examples at the end to demonstrate the concepts and formulae used in this work, using calculated consensus standards.

References

1. Measurement Systems Analysis (MSA) reference manual – Chrysler, Ford, GM (under the auspices of the Automotive Industry Action Group, AIAG), 4th Edition (June 2010, ISBN#: 978-1-60-534211-5)
2. NIST Reference on Constants, Units, and Uncertainty, online access through: https://physics.nist.gov/cuu/Uncertainty/index.html
3. Evaluation of Measurement Data – An Introduction to the Guide to the Expression of Uncertainty in Measurement and Related Documents, JCGM 104, 1st Edition (July 2009)
4. G. van Belle, Statistical Rules of Thumb, 2nd Edition, 39–40 (Wiley Series in Probability and Statistics, 2008, ISBN#: 978-0-470-14448-0)
5. Jay L. Devore, Probability and Statistics for Engineering and the Sciences, 8th Edition (Cengage Learning Publisher, 2012, ISBN#: 978-81-315-1839-7)
6. Shalabh, Dept. of Mathematics & Statistics/Indian Institute of Technology, online Simple Linear Regression Analysis lecture notes. http://home.iitk.ac.in/~shalab/econometrics/Chapter2-Econometrics-SimpleLinearRegressionAnalysis.pdf
7. G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning with Applications in R, 7th Edition (Springer, 2017, ISBN#: 978-1-4614-7138-7 (eBook))
8. C. Heumann, M. Schomaker, Shalabh, Introduction to Statistics and Data Analysis (Springer International Publishing Switzerland, 2016, ISBN#: 978-3-319- 46162–5 (eBook))

Cite this article as: Mahjoub Abdelgadir, Chris Gerling, Joel Dobson, Variable data measurement systems analysis: advances in gage bias and linearity referencing and acceptability, Int. J. Metrol. Qual. Eng. **11**, 16 (2020)