Uncertainty: Part 3 The Role and Estimation of Bias

By John Carson

Q What is bias and can it be addressed?

A Bias is defined by the *International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM)* as the "estimate of a systematic measurement error."[1] Systematic measurement error is further defined as a "component of measurement error that in replicate measurements remains constant or varies in a predictable manner."

Form and Style for ASTM Standards (A22.1) states:

"Measurement uncertainty is an estimate of the magnitude of systematic and random measurement errors that may be reported along with the measurement result. An uncertainty statement relates to a particular result obtained in a laboratory carrying out the test method, as opposed to precision and bias statements, which are mandatory parts of the method itself and normally derived from an interlaboratory study conducted during development of the test method."

Both the *VIM* and *Form and Style* make it clear that bias is a component of measurement error. Bias should be tested for and monitored. If present, bias should be estimated and reduced as much as practicable. Unmanaged bias represents a risk to impartiality of measurement results. ISO/IEC 17025:2017 requires evaluation and monitoring of risks to impartiality (International Organization for Standardization/International Electrotechnical Commission 17025, general requirements for the competence of testing and calibration laboratories).

Bias "remains constant or varies in a predictable manner." The simplest way to view bias is in connection with additive and multiplicative error. This does not cover all possibilities but does address the majority of practical cases. Additive error is the simplest type of error. It is a random error term added to the "true" value of the measurand (property being measured) when a measurement is made. If the average of this additive error term is fixed and non-zero, it is called additive or constant bias. The vast majority of simple statistical methods, like t-tests and ordinary least squares regression, assume additive error with zero bias.

Additive bias may be detected by using reference materials (RMs) with known accepted reference values (ARVs) and uncertainties and with a quality control program that requires regular measurements of the RM. The practice for conducting equivalence testing in laboratory applications (E2935) ignores bias below a threshold representing a practically significant difference or equivalence limit, defined by the context of the application.[2] E2935 may be used to test bias equivalence, whether the mean of the measurements of a particular RM is equivalent (within the limit) to its ARV.[2] If the test rejects the null hypothesis of non-equivalence, then the test method is regarded unbiased for practical purposes at the level of the ARV and for test items that are comparable to the RM. If equivalence is rejected, then the test method is assumed to have a bias equal to the difference between the mean of measurements and the ARV.

Another method that is very powerful and also useful for monitoring for bias or change in bias is the tabular cumulative sum (CUSUM) chart (practice for use of control charts in statistical process control, E2587) for the RM using the ARV or the ARV + known bias as the process target value. The tabular CUSUM also has a threshold or "slack band" that allows it to be used for monitoring equivalence of the process mean to a target.

Having detected a bias for measurement of a given RM, the question must be asked, "Is it constant across varying levels of the measurand and various types of items tested?" The first step in answering this question is repeated routine testing of other RMs with different ARVs and possibly representing different types of test items (like different types of materials or sample matrices). These sequences of QC data are also tested and monitored for bias versus the ARVs of their respective RMs as described above. If only one type of material is being tested, and the RMs vary only by the level of the measurand, then linear regression (practice for regression analysis, E3080) of the measurements onto the ARVs may be used to test for constant bias. If the p value for the regression is small (for instance, <0.01), that is a good indication that the bias may be constant or nearly constant.

In this case, the RM data may be combined by subtracting the corresponding ARVs from the measurements. An equivalence test for mean 0 in the combined data may be used to test for practically zero bias, and a CUSUM chart with process target 0 may be used to monitor for bias. Combining the QC data in this way greatly increases the power to detect existing and practically significant bias using statistical tests and reduces the time to detect new bias if the testing process has an upset.

Multiplicative error is more complex than additive error. It is the type of error induced when a factor or divisor in a measurement formula has error in it; it also occurs naturally in many measurement processes. Multiplicative or proportional bias is a component of multiplicative error. If RMs with ARVs at multiple levels are used, linear regression, regressing the measurements onto the ARVs, simultaneously detects and estimates both additive and multiplicative bias. The constant or intercept in the regression model is the estimator of additive bias. A slope of 1 indicates no multiplicative bias. Therefore, the multiplicative bias is estimated by the regression slope minus 1.

Through the use of regression in linear calibration, both additive and multiplicative bias is reduced and managed.[3] In this case, the ARVs are regressed onto the raw indications from the measurement instrument (voltage, peak areas, current, deflection, counts, etc.). The regression equation then directly provides a calibration function, with the raw indication from an unknown test item as input and an estimate of the measurand as output. There may also be adjustments for quantities like the mass or volume of the test increment or dilutions.

The conditions of the problem and characteristics of the data dictate the most appropriate form of the regression. However, calibration using appropriate regression methods minimizes additive and multiplicative bias, allows estimation of their likely magnitude, and at the same time maximizes the precision of estimate of the measurand. This is also true for nonlinear regression applied to nonlinear calibration.

The reader may wonder why this calibration using appropriate regression methods only minimizes bias and does not completely eliminate it. The answer is that the regression coefficients are always estimated with error and that this error is known probabilistically but not exactly. Therefore, it is impossible to further correct for. In the long run, over many calibrations, the parameter estimates and predictions are unbiased and have minimum error variance. However, over the period in which a particular calibration is in use, the errors in its parameter estimates become biases. Nevertheless, the bias is as small as possible as long as the calibration remains valid. Through the use of calibration intervals and control charting the response to a mid-level intermediate check standard or calibration standard, the validity of calibration can be maintained. This responds to the need to minimize and manage bias, as mentioned in the previous paragraph.

Use of inappropriate regression methods in calibration or the failure to monitor and to ensure the continuing validity of calibration increases bias and hinders attempts to manage it. For instance, for many test methods, the standard deviation of repeated measurements is an increasing function of the measurand level. In this setting, the use of ordinary least squares regression in calibration increases the error in estimating the intercept.[3] This inflates the additive bias during the period in which these inappropriate calibrations are in use. It is similar to using 0-intercept (slope or span only) calibrations when additive bias is present and would be detected using regression with an intercept. A related issue is that different types of test items or sample matrices may have very different measurement biases, both additive and multiplicative. This may require the use of RMs that vary in type as well as in measurand level.

In summary, to ensure and monitor the validity of test results and impartiality of the testing process, bias must be assessed, minimized, monitored, and managed. This work requires that several elements be effectively implemented in the laboratory's quality control program: RMs with known ARVs; routine testing of these RMs; statistical tests for bias; monitoring RM test results using control charts; appropriate use of regression in calibration; appropriate calibration intervals; and monitoring the continuing validity of calibrations.

REFERENCES

1. *VIM3: International Vocabulary of Metrology – Basic and General Concepts and Associated Terms,* 3rd Ed., Joint Committee for Guides in Metrology, 2008.
2. Bzik, Tom, "Data Significance," Data Points, ASTM *Standardization News,* Vol. 44, No. 4, July/August 2016, pp. 40-41.
3. Al-Ghamdi, K.S., "Calibration and Regression," Data Points, ASTM *Standardization News,* Vol. 46, No. 3, May/June 2018, pp. 40-41.

John Carson, Ph.D., a senior statistician for Neptune and Co., is the Data Points column coordinator. He is chairman of the subcommittee on statistical quality control (E11.30), part of the committee on quality and statistics (E11), and a member of the committee on petroleum products, liquid fuels, and lubricants (D02), air quality (D22), cannabis (D37), and environmental assessment, risk management, and corrective action (E50).