Validating Automatic Consistence Measurements

Rev	Date	Author	Reviewer	Comments
01	03/03/2026	Dr Mark Steadman	Dr Ed Waugh	
02				

Contents

Executive summary

This report describes a method of analysis for determining if an automatic consistence measurement system is a suitable replacement for the manual test. It uses the best-practice technique for testing equivalence between different measurement methods where no ground truth value is available (Bland-Altman Limits of Agreement [1]). Drawbacks of the existing methods in the Good Concrete Guide and the performance of the new method are described.

The method has been designed to ensure that if it were applied to a manual test, it would be determined valid 90% of the time and achieving this requires 20 measurements, each consisting of two tests from three testers.

This document provides a detailed rationale for the design of this method and examples of reporting. For a practical guide to using the technique in practice please see the Cloud Cycle white paper: Procedure for Validating Automatic Consistence Measuring Systems [2].

1. Introduction

With the introduction of automatic consistence measurement systems, it has become necessary to define a validation test to ensure that these systems are suitable replacements for the existing manual tests. Although consistence tests provide a valuable indication of concrete workability, individual test results are inherently variable. This variability is well characterised and documented in the relevant standards.

Consequently, it is not appropriate to treat the result of a single manual test as a "ground truth" against which an automatic system can be directly compared. Determining whether the two measurement methods can be considered equivalent is therefore not as straightforward as it might initially appear.

Fortunately, the problem of assessing a novel measurement method against an established reference method in the absence of a true ground-truth is well known and has been addressed in many other fields, see [3] for a recent discussion.

One common source of confusion, specific to concrete consistence, is that the measured quantity is self-defined (that is, "the slump is the slump"). For assessing agreement between two measurement methods, this distinction is not relevant. It is sufficient to recognise that each batch of concrete has an underlying (but unobservable) "true" consistence, which can only be estimated through one or more measurement methods.

This report outlines the principles of a validation test procedure for automatic consistence measurement systems based on this established methodology. It presents the development of appropriate validation criteria to demonstrate that an automatic system is at least as accurate as the manual test. Automatic systems that are shown to be valid may then be used as direct replacements for the manual test.

1.1. Definition of Terms

Calibration
A process by which a measurement system is adjusted to remove systematic errors. This normally involves making a measurement with the system alongside a reference measurement and calculating an error correcting function. That function is then embedded in the system and used in future calculations.

Validation
A process that ensures that the measurement system meets the needs of the user. This shall include comparing the system output to a reference value and assessing it against a set of criteria. This process may be repeated periodically.

Verification
A process that ensures that the measurement system meets a set of design requirements. This is normally an internal process where non-customer facing parameters may be tested.

Consistence
Property of freshly mixed concrete that determines the ease at which it can be mixed, placed, consolidated, and finished to a homogeneous condition.

Consistence Test
An assessment of the consistence of fresh concrete. A single test performed by one tester.

Consistence Measurement
Multiple consistence tests performed simultaneously by multiple testers on concrete from the same sample.

Consistence Class
A range of consistence measurements grouped to provide a meaningful short-hand description.

Technology Provider
Any provider of an automatic system that can be used to monitor or control the consistence of fresh concrete.

Accuracy
Both Trueness and Precision

Precision
The spread of the results of repeated measurements determined by random errors.

Trueness
The closeness of the result to the true value, determined by systematic errors.

Standard Error
A measure of how much a sample statistic, such as the mean, is likely to differ from the true population value.

Variance
A measure of how widely spread a set of values are from their mean. Defined as the average squared difference from the mean.

Standard Deviation
A measure of how widely spread a set of values are from their mean. Defined as the square root of the variance, it is in the same units as the underlying value.

Normal Distribution
Also known as a Gaussian distribution or bell curve, a continuous probability distribution defined by a mean and variance.

Tester
A person trained in performing the manual test method.

True Consistence
A notional value that is estimated by different measurement methods.

Reproducibility
Reproducibility is the spread of the differences between two Testers, with different sets of test equipment, performing the same test.

Repeatability
Repeatability is the spread of the differences between successive measurements performed by the same Tester with the same equipment.

Bias
A constant difference between the reference value and the measured value caused by systematic errors.

Concrete Mix
A combination of a set of constituents designed to achieve concrete for use in a specific application

Concrete Family
A grouping of similar concretes that have the same behaviour when measured automatically. These are technology dependent and determined by the technology provider.

True Positive
In this context, a true positive is the case where the measurement system is genuinely at least as precise as the manual test and passes validation criteria.

False Positive
In this context, as false positive is the case where a measurement system is less precise than the manual test but is erroneously determined to be valid.

2. Accuracy of manual consistence tests

Consistence measurements (slump, flow table or slump-flow) provide an estimate of the "true" consistence of a batch of concrete. The results of a single consistence test can differ from the true value for a number of reasons. These include tester bias (caused by systematic differences in technique such as a tendency to raise a slump cone faster or slower on average compared to other testers), and random variation (from multiple other sources such as the shape and distribution of coarse aggregates in the sample). In addition, rounding error is introduced by the practice of reporting results to the nearest 10 mm.

Sample bias, whereby the test sample is not representative of the average composition of the batch, can contribute a further source of error. In this report, the batch of concrete is assumed to be well mixed, such that any sample bias is negligible. This assumption is a prerequisite for carrying out a valid consistence test. Under this assumption, rounding error may be characterised by assuming that the measured value lies within ±5 mm of the reported result.

The expected tester bias and random variation can be measured though trials where samples are taken from a batch of concrete and tested simultaneously by multiple testers, multiple times. These tests have been carried out and are the basis of the precision tables reported in the standards defining the procedure for each manual test. These tables therefore provide a reference for the accuracy of a manual consistence test.

2.1. Understanding precision tables

The standards defining the procedure for carrying out manual consistence tests (slump, flow and slump-flow) contain precision tables, often with separate entries for different consistence ranges. They include entries for repeatability (single operator) and reproducibility (multi-laboratory) conditions.

2.1.1. Reproducibility

Reproducibility (R in the British standards or *d2s* in ASTM) is twice the standard deviation of the differences between many pairs of measurements reported by different testers where samples were taken from the same batch of concrete and tested at the same time:

$$R = 2\sqrt{\frac{\sum_{i=1}^{N}\left(d_i - \bar{d}\right)^2}{n-1}} \qquad 4.1$$

Where d_i is the difference between the measurements for pair i, $\bar{d}$ is the average difference between all pairs of tests and n is the number of batches tested.

BS EN 12350-2 [4] indicates that it comprised test results from 16 testers, which were a subset of BS 1881. The data for ASTM C-143 [5] are described in [6] and comprises test results from 15 testers, each carrying out 18 tests on a single load to which water was added to target different slump ranges.

We can interpret this value as the range within which we can expect pairs of measurements reported by different testers to agree 95% of the time (for samples taken from a batch of concrete at the same time). This means that 95% of all data points will lie within ±2 standard deviations of the mean for normally distributed data. This is put in simpler terms in [4, 7, 8]:

> *"Test results on the same sample obtained within the shortest feasible time interval by two operators each using their own apparatus will differ by the reproducibility value R on average not more than once in 20 cases in the normal and correct operation of the method."*

2.1.2. Single-test standard deviation

The precision tables also report a "reproducibility standard deviation" (S_R in the British standards and $1s$ in the ASTM). This is the standard deviation of a single test result that would, when two such results are compared, produce the observed spread of differences characterised by R.

We model each consistence measurement as a normally distributed random variable, with the reported result as the mean and S_R as the standard deviation. This allows us to make a useful practical statement: there is approximately a 95% chance that the true consistence lies within $\pm 2 S_R$ of any single test result.

Figure 1 illustrates this concept for three reported slump measurements of 30 mm, 85 mm and 160 mm, using the single-test standard deviations given in ASTM C-143 [5]. Each curve represents uncertainty about the true consistence given a single test result. The shaded region contains 95% of the probability, spanning approximately $\pm 2 S_R$ either side of the reported value. The distributions become narrower at lower slump values, reflecting the improved reproducibility of the test at lower consistence as per [5].

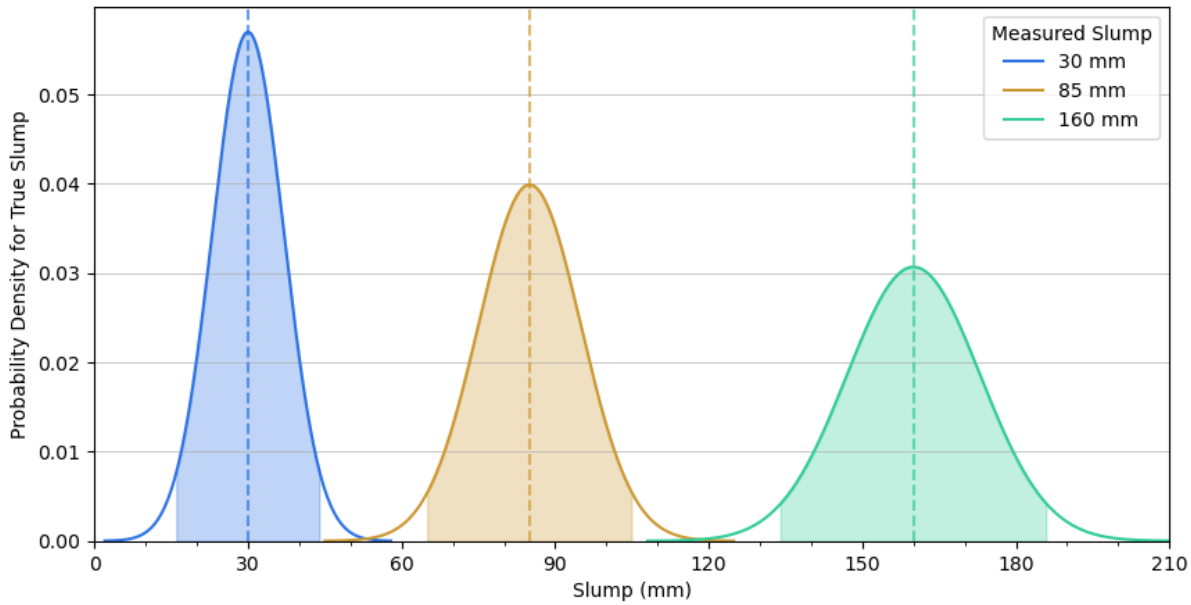

Figure 1 – Probability density curves showing the probability that the true slump takes a given value given each of three reported measurements; 30 mm, 85 mm and 160 mm. The standard deviations are taken from [5]. Shaded regions contain 95% of the probability for each curve (i.e. given the reported measurement, the true slump is likely to be within the shaded range in 95% of cases).

S_R cannot be measured directly from a single test. Instead, we observe reproducibility (the spread of differences between pairs of tests) and work backwards. Because reproducibility reflects the combined variability of two independent measurements, we can separate the contribution of each individual test to arrive at S_R.

The normal distribution model is a simplification and breaks down when measurements are close to the upper or lower limits of the test (since the true slump can never be lower than 0 mm or higher than 300 mm, for example). However, it is a useful and accurate approximation across the range within which each test is considered sensitive, as specified in Annex L in [9].

Mathematically, the reproducibility R can be written as:

$$R = 2\sqrt{S_{R1}^2 + S_{R2}^2} \qquad 4.2$$

Where S_{R1} is one standard deviation of the uncertainty associated with the first consistence test and $S_{R,2}$ is that of the second. If we assume these are the same, we can re-write this as:

$$R = 2\sqrt{2S_R^2} \quad 4.3$$

And rearrange to get S_R as a function of R:

$$R = 2\sqrt{2}S_R \quad 4.4$$

$$S_R = \frac{R}{2\sqrt{2}} \quad 4.5$$

Test type	Test range	R (mm)	S_R
Slump	S1	20 [5]	7
	S2	25 [4]	9
	S3 - S4	37 [5]	13
Flow	F2 – F5	91 [7]	32
Slump-flow	SF1 – SF2	43 [8]	15
	SF3	28 [8]	10

Table 1 – Summary of reproducibility and single test standard deviation from the standards

2.1.3. Tester bias

The expected range of tester bias (a tendency to report slightly higher or lower than the average tester) is not given explicitly in the standards. However, it can be inferred from the precision tables if we consider repeatability and reproducibility together. Repeatability conditions consider differences between measurements reported by a single tester, which are determined by random variation alone. Reproducibility conditions consider differences between measurements reported by different testers, which are determined by the combination of random variation *and* tester bias. The difference between repeatability and reproducibility is therefore due to tester bias.

When combining multiple independent sources of variability, it is convenient to work in terms of variance rather than standard deviation, since variance is additive. The single-test variance associated with tester bias can then be expressed as:

$$\sigma_{bias}^2 = \frac{\sigma_R^2 - \sigma_r^2}{2} \quad 4.6$$

Where σ_R^2 is the variance of the difference distribution under reproducibility conditions and σ_r^2 is the variance under repeatability conditions. Since both are components of the variance of the difference distribution, these are the result of the combination of variances from both tests, so the division by 2 makes σ_{bias}^2 in terms of a single test rather than the contribution of tester bias to the spread of difference.

The quantities σ_R^2 and σ_r^2 are obtained from the reproducibility measure R and repeatability measure r, respectively. Both R and r are equal to twice the standard deviation of the corresponding difference distribution, giving:

$$\sigma_R^2 = \left(\frac{R}{2}\right)^2, \sigma_r^2 = \left(\frac{r}{2}\right)^2 \quad 4.7$$

As an example, ASTM C143 [5] reports a repeatability measure of 28 mm and a reproducibility measure of 37 mm for slump at 160 mm. Substituting this into equations 4.6 and 4.7 gives a standard deviation of 8.5 mm. This means that tester bias relative to the true value is expected to be less than ±17 mm, 95% of the time.

In practical terms, this means a bias of 10 mm or more relative to the true consistence is common, even for experienced testers. Unless the results reported by a given tester are compared against those of a group of

other testers, such bias cannot be identified. This underlines the importance of collecting measurements from multiple testers for each sample when assessing measurement performance.

2.2. Rounding error

The precision metrics reported in the standards were determined empirically using consistence measurements reported to the nearest millimetre. In practice, consistence measurements are rounded to the nearest 10 millimetres. This adds an additional source of error, since the true consistence could have been anywhere within ±5 mm of the recorded value.

The rounding error increases the variance associated with a single test result. Since the error is equally likely to be any value in the range ±5 mm, it is characterised by a uniform distribution over that range. The variance of a uniformly distributed random variable is given by:

$$\sigma_{rounding}^2 = \frac{w^2}{12} \tag{4.8}$$

Where w is the width of the rounding interval (10 mm) (see [10] for a derivation). This can be added together with the variances due to the tester bias and random variation and converted to modified single test standard deviation $S_R^{rounded}$. The variance associated with tester bias is given in equation 4.6. The single-test variance due to random variation can be calculated from the repeatability r as follows:

$$\sigma_{random}^2 = \frac{\sigma_r^2}{2} \tag{4.9}$$

Where σ_r^2 is given in equation 4.7, being the variance of the distribution of differences between pairs of tests under repeatability conditions. Combining the effects of tester bias, random variation and rounding, a modified single-test standard deviation can be calculated as follows:

$$S_R^{rounded} = \sqrt{\sigma_{bias}^2 + \sigma_{random}^2 + \sigma_{rounding}^2} \tag{4.10}$$

For slump at 160 mm where r is reported as 28 mm and R is reported as 37 mm [5], using equations 4.6 – 4.10 gives a single-test standard deviation of 13.4 mm. Excluding the effect of rounding, S_R is 13.1 mm.

The rounding error only marginally increases the single-test standard deviation but given its relationship with reproducibility R (equation 4.4), the effect of rounding is non-negligible. This is relevant when assessing the range within which pairs of tests agree.

2.3. Improving accuracy through averaging

Individual measurements are subject to tester bias and random error, but both can be mitigated through averaging. Averaging tests carried out by one tester can only reduce random error. Averaging tests carried out by different testers reduces both random error and tester bias. To get the best estimate of the true consistence for a given number of tests, it is necessary to use multiple testers.

The standard deviation (or uncertainty) of an averaged measurement σ_{avg} comprising N tests, each carried out by a different tester can be expressed in terms of the single-test standard deviation S_R as follows:

$$\sigma_{avg} = \frac{S_R}{\sqrt{N}} \tag{4.11}$$

It is also possible to improve the accuracy of a consistence measurement further by having each tester also carry out multiple tests. Random and rounding errors decrease as a function of the total number of tests (the number of testers multiplied by the number of repeats), while tester bias decreases as a function of the number of testers only. If N_T is the number of testers and N_R is the number of repeated tests carried out by each tester, the standard deviation of the averaged result is:

$$\sigma_{avg} = \sqrt{\frac{\sigma_{bias}^2}{N_T} + \frac{\sigma_{random}^2 + \sigma_{rounding}^2}{N_T N_R}} \quad 4.12$$

Comparing to equation 4.10, we can see that for a single test σ_{avg} is equal to S_R (for a rounded measurement). Using the example of slump at 160 mm, where samples are tested by three testers each carrying out two tests, σ_{avg} reduces from the single test value of 13.4 mm to 6.5 mm, more than doubling the accuracy.

Figure 2 shows the effects of averaging on the measurement standard deviation using equation 4.12. All data points assume the true slump is around 160 mm, such that values of 28 mm and 37 mm can be taken from the standards for repeatability and reproducibility respectively [5].

Adding more testers has diminishing returns, with the largest benefit seen when using two or three testers compared to one. In addition, averaging tests across multiple testers is much more beneficial than having a single tester carry out repeated tests. This is because the tester bias will never decrease no matter how many times an individual tester repeats a test.

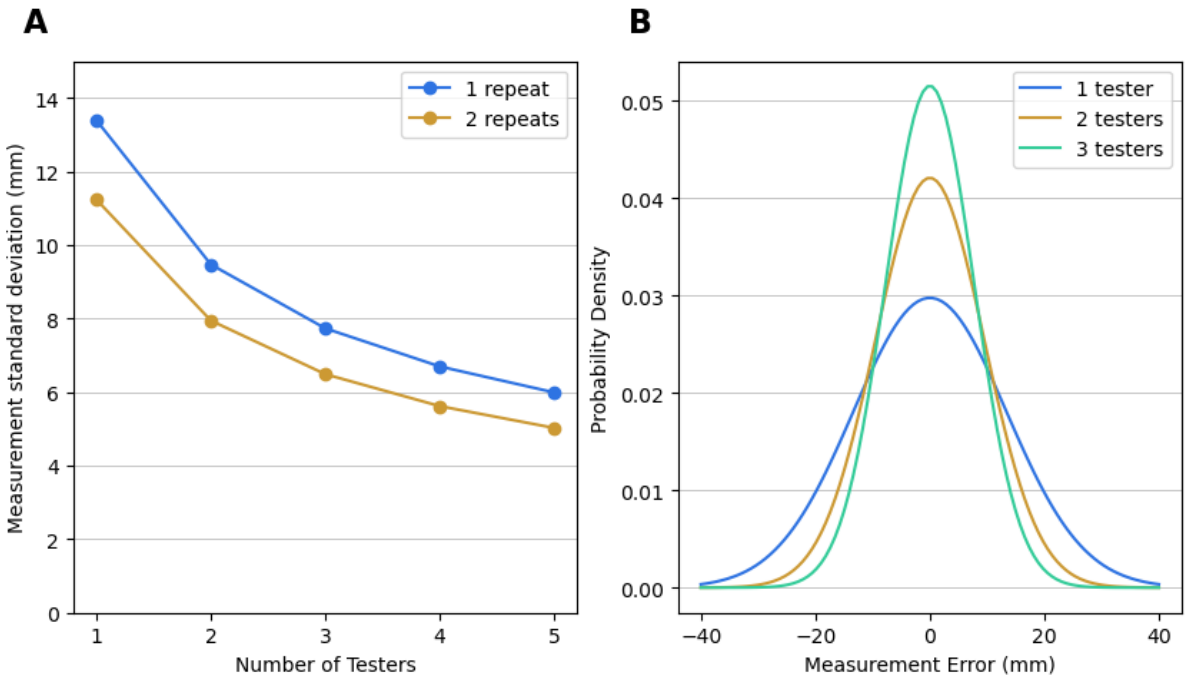

Figure 2 – Effect of averaging across multiple slump tests for a true slump of 160 mm. A) measurement standard deviation as a function of the number of tests carried out by different testers. The blue line assumes a single test per tester and the orange assumes each tester carries out two tests in quick succession. B) Probability density of measurement error for averaged measurements for 1 – 3 testers each carrying out a single test.

2.4. Summary

The standards describe the expected range of differences between manual tests, and the reproducibility values can be used to calculate the single test standard deviation. Rounding to 10 mm is a non-negligible part of the variation and tester bias can be calculated which is a large component of the variation. The manual measurement can be significantly improved by using multiple testers and repeats.

3. Design of the Validation Procedure

Validating automatic consistence measurement is done by gathering a set of manual and automatic measurement pairs. The range of differences between the automatic and manual measurements are then assessed against the range of differences expected if the automatic system were equivalent to a manual test. This document describes the method to determine how many pairs need to be collected, how to quantify the range of differences and how to determine equivalence to a manual test.

3.1. Validation procedure requirements

A validation procedure should have the following properties:

1. The outcome should be a binary decision (i.e. valid vs not valid) supported by quantitative data
2. The procedure should be sufficiently constrained such that it is not possible to manipulate the outcome (for example through choice of sampling method, consistence range or number of data points).
3. We propose that if the manual consistence test itself was assessed using the validation procedure, it should be determined valid at least 90% of the time. This ensures that the likelihood of a false negative outcome (i.e. a valid system is incorrectly assessed as not valid) is low.
4. We propose that the decision will be based on 20 data points, which are feasible to collect over the course of one or two days of testing. This sets the false positive rate for a system that is just non-compliant to 41% and can only be improved by increasing the amount of testing (section 4.3).

For a detailed description of the procedure, see the Cloud Cycle white paper: Procedure for Validating Automatic Consistence Measuring Systems [2]

3.2. Comparing different validation procedures

Any validation procedure will be based on a dataset comprising pairs of measurements; a manual measurement (which may be an averaged measurement based on more than one manual test) and an automatic measurement. The efficacy of a validation procedure may be tested by simulating such datasets under various assumptions. The steps for generating a simulated dataset are outlined below.

For each data point $i = 1, \dots, N$, a value for true consistence C_i^{true} is sampled from a uniform distribution over the target range of the slump class:

$$C_i^{true} \sim \mathcal{U}(\min, \max) \tag{5.1}$$

For example, for slump measurements in the S3/S4 range, this corresponds to a uniform distribution between 100 mm and 210 mm [9].

Manual measurements can be simulated for multiple testers and repeats. Each tester must be assigned a bias b_j, randomly drawn from a normal distribution with the expected scale (see 2.1.3). Each individual measurement is then generated based on the sum of the true consistence, the tester bias, and a random error term $e_{i,j,k}$:

$$C_{i,j,k}^{phys} = C_i^{true} + b_j + e_{i,j,k} \tag{5.2}$$

Where $C_{i,j,k}^{phys}$ represents the manual measurement for data point i for the k-th test carried out by tester j. The random error is also drawn from a normal distribution with the expected scale.

For each data point i, an automatic measurement must be simulated. Like the manual tests, this is the true consistence plus a bias and random error term.

$$C_i^{auto} = C_i^{true} + b + e_i^{auto} \tag{5.3}$$

The magnitude of e_i^{auto} can be chosen to represent scenarios where the automatic system performs better than, equivalent to, or worse than the manual test.

In this report, validation procedures are tested against five simulated automatic consistence measurement systems with different degrees of accuracy. All simulated systems have zero bias. All examples assume that the validation test is for slump in the S3/S4 range, but the same process can be applied for any consistence test type and range by setting appropriate values for the expected tester bias and random error (see Chapter 2 for details).

The five systems are characterised as follows:

System name	Description	Valid
System 1	True Consistence measurement	Yes
System 2	Better than the manual test (S_R = 8 mm)	Yes
System 3	Equivalent to the manual test (S_R = 13 mm)	Yes
System 4	Worse than the manual test (S_R = 18 mm)	No
System 5	Reports randomly within the target range	No

Table 2 - Description of the 5 automatic measurement systems that were simulated and used to assess the efficacy of validation test procedures

An example of a simulated validation test dataset for System 3, which has an accuracy equivalent to the manual test for S3/S4 slump, comprising 100 data points is shown below:

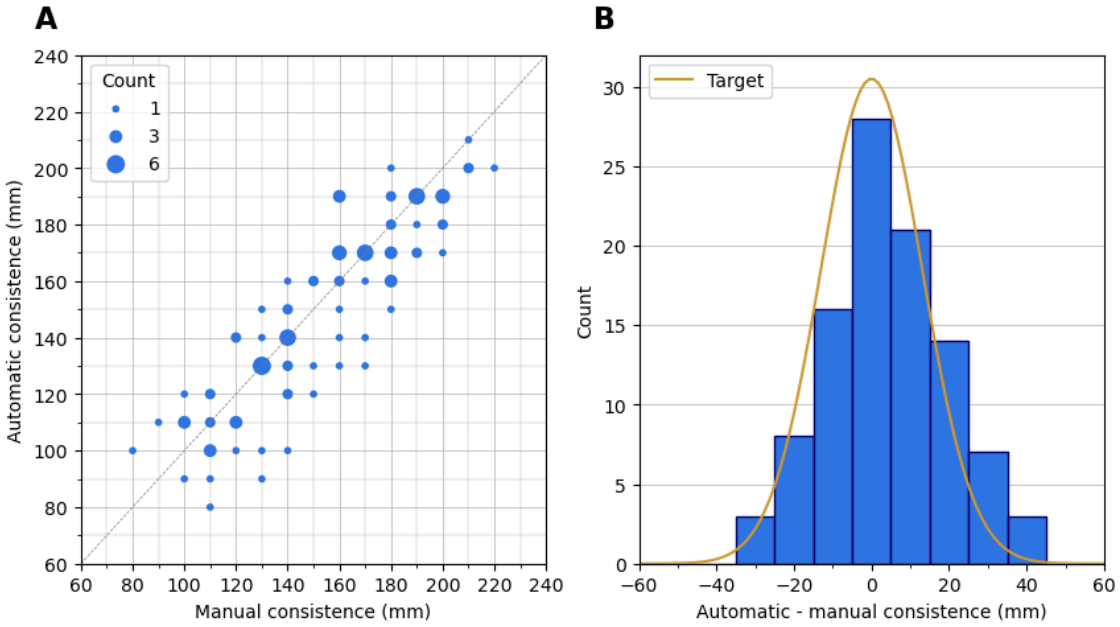

Figure 3 – A) scatter plot of simulated manual (horizontal) and automatic (vertical) consistence measurements for a validation test carried out on System 3 (equivalent to a manual test). All measurements are rounded to the nearest 10 mm for plotting, and the diameter of each point represents the count. B) histogram of the differences between automatic and manual measurements. Orange line indicates the expected asymptotic distribution (i.e. when $n \to \infty$) when the automatic system is equivalent to the manual test.

Many datasets can be simulated to assess the performance of the validation procedure statistically. In this report, 10,000 datasets comprising 100 rows of true consistence, six manual measurements (for three testers each carrying out two tests) and measurements for each of the five automatic systems were simulated.

Proposed validation procedures can be applied to each of the simulated datasets to assess the performance of the procedure in terms of whether they validate each of the proposed systems. An ideal procedure would

validate systems 1 to 3, which are better than, or equivalent to the manual test, but not systems 4 and 5, which are less accurate than the manual test.

4. Validation procedure

Bland-Altman analysis [11, 1] is an established method to assess the agreement between two measurement techniques of the same underlying quantity. The process defines estimates of the range within which the two measures are likely to agree based on a dataset of pairs of test results. This range is defined by lower and upper Limits of Agreement (LoA). These can be compared against the range that would be expected if the automatic system were equivalent to the corresponding manual test.

4.1. Data collection and analysis

A recommended procedure for data collection and analysis is described in [2], and that is the document to use if you are planning to perform your own testing. The important details are summarised here for clarity and are used in the subsequent analysis.

The validation test procedure involves collecting paired manual and automatic consistence measurements. Samples of concrete are taken from well-mixed batches, and consistence tests are carried out simultaneously by three testers. Using three testers is the minimum required to identify extreme biases, while providing a practical balance between improved measurement accuracy, logistical complexity, and cost.

Each tester carries out two tests in quick succession. This further increases expected measurement accuracy for little cost, and the consistence can be considered constant.

The measurement reported by the automatic system is also recorded. If the system provides continuous monitoring, the measurement corresponding to the average time of the manual tests should be used. Otherwise, the automatic measurement should be carried out immediately after all testers have completed the first of their two tests.

The number of paired measurements determines the sensitivity of the validation test (i.e. its ability to correctly identify whether an automatic measurement system may be considered valid). To meet the requirement that the outcome must be determined based on a limited number of samples, this has been fixed at 20.

The Limits of Agreement (LoA) are calculated from the dataset of paired measurements (an averaged manual measurement and an automatic measurement) as described in [11]:

$$LoA = \bar{d} \pm 2s \tag{7.1}$$

Where $\bar{d}$ is the mean difference between the manual and automatic measurements and s is the (unbiased) standard deviation of the differences. That is:

$$s = \sqrt{\frac{\sum_{i=1}^{N}\left(d_i - \bar{d}\right)^2}{n-1}} \tag{7.2}$$

The test for validity is if the upper and lower LoA are within the expected LoA for the manual test. Limits are specific to the consistence test type (i.e. slump, flow or slump-flow) and range.

4.2. Criteria for validation

If the reference measurement was a single manual test, $2s$ from equation 7.1 would be equal to the reproducibility (with adjustment for rounding error). Since the reference measurement is averaged across three testers, each reporting two tests, the limit must be adjusted as the variation is reduced. If the maximum absolute value for either the lower or upper LoA is Δ, this is:

$$\Delta = 2\sqrt{\sigma_{avg}^2 + \sigma^2} \tag{7.3}$$

Where σ_{avg} is the standard deviation of the manual (averaged) measurement calculated using equation 4.12, and σ is the single-test standard deviation, adjusted for rounding error. Using the example of slump concrete at 160 mm, σ_{avg} for three testers each repeating a test is 6.5 mm and σ is 13.4 mm (see 2.2). This gives a limit value of 30 mm, rounding to the nearest millimetre.

To demonstrate the performance of the validation test procedure using this limit, the simulated datasets were assessed against this criterion. For each dataset the LoA is calculated and compared to the limit value. To be valid both the upper and lower LoA must be within the limit.

To demonstrate the benefit of averaging across multiple tests, results are presented for three scenarios; using reference measurements from a single tester only, averaging across three testers, and averaging across three testers each reporting two measurements.

The proportions of simulated validation tests where each of the five systems described in 3.2 were determined to be valid are shown in Figure 4. The benefit of averaging multiple manual tests is clear. For example, System 2 (more accurate than the manual test) passes validation in 75% of simulated validation tests where a single manual test is used. This increases to 93% if each of the three testers carries out two tests.

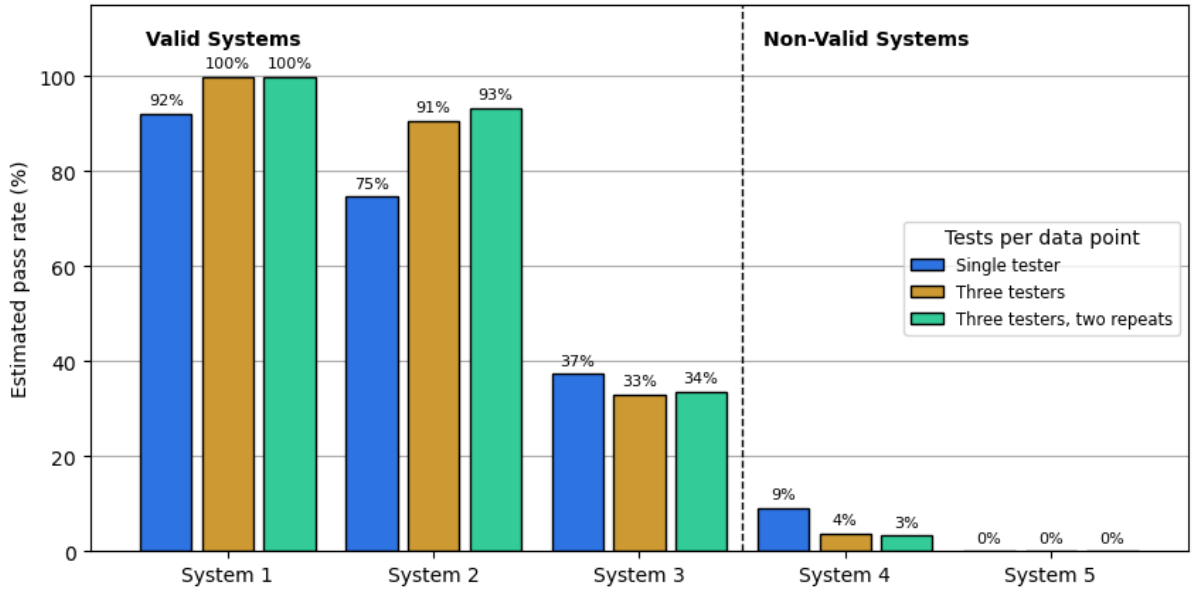

Figure 4 - Result of validation testing assessment using limits-of-agreement criteria based on accuracy exactly equivalent to the manual test for each of five simulated automatic consistence measurement systems (Table 2). Systems 1-3 are valid systems with accuracy at least equivalent to the corresponding manual test. Systems 4 and 5 report inferior measurements that should fail a validation test.

One issue with this method is that it does not meet the requirement that a system with accuracy equivalent to that of the manual test (System 3) should pass validation at least 90% of the time (section 3.1). Whilst the number of data points, as well as the number of manual tests that are averaged per data point, limit the sensitivity of the validation (i.e. its ability to reliably discriminate valid vs non-valid systems), the outcome can be biased such that this requirement is met. This is at the cost of increased false positives (i.e. cases where a system that is less accurate than the manual test nonetheless passes the validation criteria due to chance).

The bias is implemented by introducing a tolerance to the limit value Δ. This tolerance is based on the theoretical 90% Confidence Interval (CI) of the LoA, if the dataset comprised averaged manual measurements as the reference and a single manual test as the automatic system.

The first step is to calculate the standard error of the LoA, assuming equivalence (i.e. the standard deviation of the distribution of sample limits of agreement if the validation test were run multiple times). Bland & Altman used the following approximation [11, 1].

$$SE_{LOA} = \sqrt{\frac{3s^2}{n}} \quad 7.4$$

Where s is the expected standard deviation of the distribution of differences, given by the following (refer to equation 7.3 for symbol definitions):

$$s = \sqrt{\sigma_{avg}^2 + \sigma^2} = \frac{\Delta}{2} \quad 7.5$$

This then needs to be multiplied by the value corresponding to a 90% confidence interval. This is the t-score for the 95th percentile of the Student's t-distribution for $n - 1$ degrees of freedom [1]. An adjusted maximum for the limits of agreement is then:

$$\Delta_{adj} = \Delta + t_{n-1,0.95}\sqrt{\frac{3s^2}{n}} \quad 7.6$$

Note that in the literature $t_{n-1,0.975}$ (giving the value for the 97.5th percentile) is often used. This is appropriate for the upper limit of the 95% confidence interval. In this case, the requirement is that automatic system must pass at least 90% of the time where it is equivalent to the manual test, so we use $t_{n-1,0.95}$ to consider the 90% confidence interval.

Test type	Test range	r (mm)	R (mm)	Δ (mm)	Δ_{adj} (mm)
Slump	S1	17 [5]	20	16	21
	S2	16 [4]	25	20	27
	S3 - S4	28 [5]	37	29	39
Flow	F2 – F5	69 [7]	91	72	96
Slump-flow	SF1 – SF2	42 [8]	43	33	44
	SF3	22 [8]	28	22	29

Table 3 - Repeatability, reproducibility and the calculated expected absolute value of the limits of agreement based on the assumption that the reference test is an average of measurements reported by three testers, each carrying out two tests and the automatic system is equivalent to a single manual test, rounded to the nearest 10 mm. Δ_{adj} is the threshold criterion for calculated limits of agreement including the tolerance of 90% from (3.1).

4.3. Performance against the simulated data set

Figure 5 shows the proportion of simulated validation tests where each of the five systems are assessed for validity using the adjusted thresholds for the limits of agreement. The adjustment has worked as expected with System 3 (equivalent to the manual test) passing validation in 91% of trials.

System 4 (less accurate than a manual test) also passes validation in 41% of trials, representing an increase in false positive validation decisions compared to the non-adjusted thresholds. The only way to reduce this rate is through additional data points to increase the sensitivity of the test.

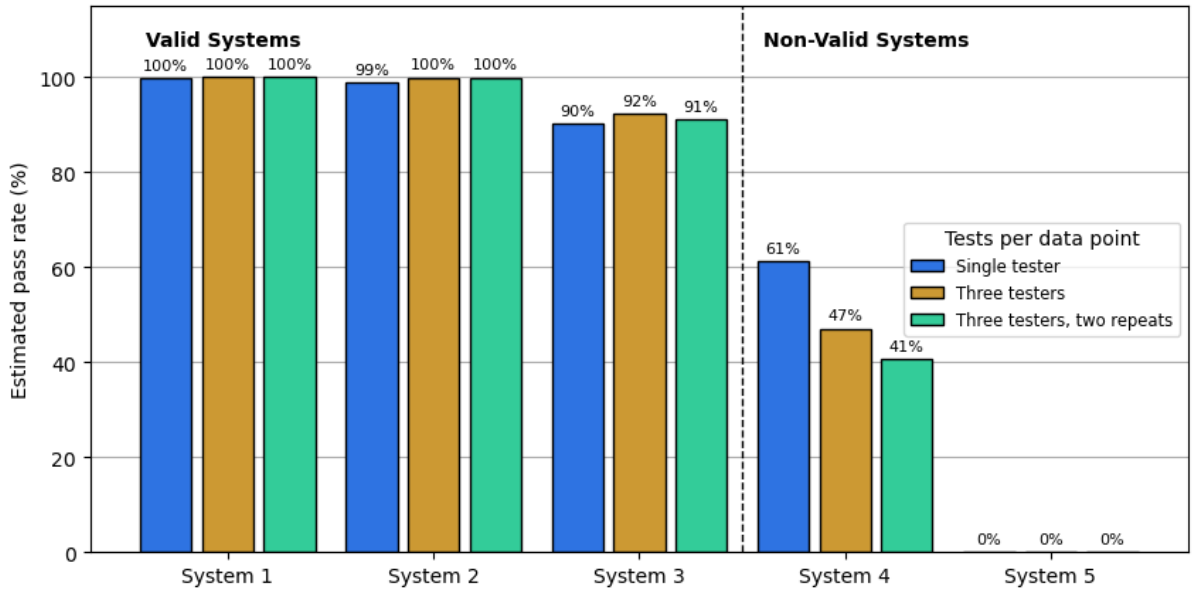

Figure 5 - Result of validation testing assessment using limits-of-agreement criteria including tolerance for each of the five simulated automatic consistence measurement systems. Systems 1-3 are valid systems with accuracy at least equivalent to the corresponding manual test. Systems 4 and 5 report inferior measurements that should fail a validation test.

5. Reporting validation outcomes

Validation outcomes shall be reported in a standardised way for each campaign.

The required components of a validation report are:

1. A scatter plot of the manual vs automatic measurements and a line of perfect agreement
2. A histogram of the differences showing the upper and lower limits of agreement and the target (adjusted) limits of agreement.
3. A statement of validation including the average manual consistence and its standard deviation, the average difference (bias), and the upper and lower Limits of Agreement.

Below is an example of a report based on validation testing of the Cloud Cycle flow measurement system.

"For the validation test carried out on 19th and 20th March, 2026 the Cloud Cycle system was found to be a valid replacement for the manual flow table test across the F2-F5 range. The mean manual flow across all tests was 586 mm with a standard deviation of 67 mm. The mean difference (bias) between automatic and manual measurements was 5 mm. The limits of agreement were -46 mm to 56 mm against a target of ±96 mm."

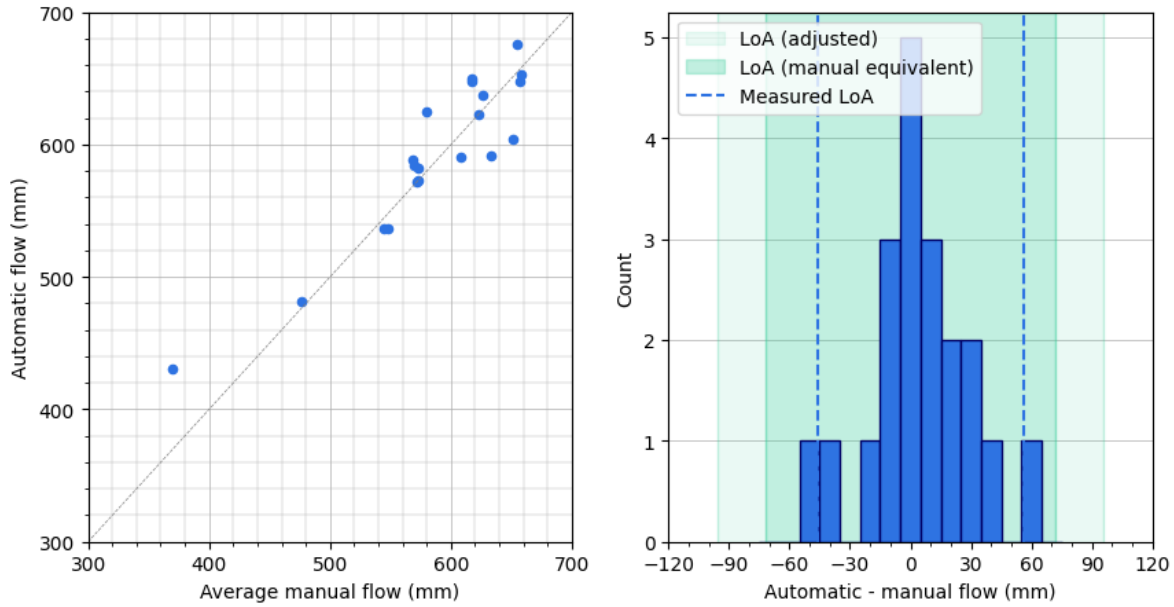

Figure 6 - Simulated validation of System 3, (A) example scatter plot (B) example histogram

6. Good Concrete Guide 11 validation guidance

The Concrete Society has published guidance on the use of digital monitoring and measurement of fresh concretes, which includes criteria with which to assess agreement between manual and automatic consistence measurements (referred to as digital measurements in the document) [12]. The proposed criteria have some undesirable properties that we believe make them unsuitable for assessing the validity of these systems.

In this section, the proposed criteria are assessed using the same analytical and simulation-based approach applied in the previous sections, considering automatic measurements with better, worse or equivalent performance than the corresponding manual test.

6.1. Criteria for validation

The guidance document discusses “correlation” and “evidence of consistence correlation”. It should be emphasized that correlation is not an appropriate quantity to use to assess agreement, since an automatic measurement method could exhibit a strong bias or be on an entirely different scale while still showing a strong correlation with the manual measurement. However, the validation criteria subsequently presented in the GCG are based on agreement rather than correlation.

This criteria for validation are presented in terms of a tolerance (an allowable maximum difference between a manual and digital test), and a compliance level, expressed as a percentage. There are different values for both tolerance and compliance depending on the sampling method used for the manual test. The table is reproduced below:

Test method	Sample method	Tolerance	Compliance level
Slump	Composite	±30 mm	95%
	Spot sample	±40 mm	90%
Flow table	Composite	±50 mm	95%
	Spot sample	±60 mm	90%
Slump-flow	Composite	To be determined from trials	
	Spot sample		

Table 4 - Validation criteria for digital consistence measurement systems published in Good Concrete Guide 11 [12]

Note that while the guidance does reference the importance of considering reproducibility and repeatability to give context to validation results, the criteria given in Table 4 are not derived from these values.

As shown in Table 4, different tolerance and compliance criteria are applied depending on whether spot or composite samples are taken, with significantly higher allowances for spot samples. While composite sampling can mitigate sample bias for a batch that is not well mixed, there is no reason to expect that the sampling method should affect the precision of the test itself. It will therefore be significantly easier for systems to meet the validation requirements using the spot sample criteria.

Compliance levels are reported as a percentage, but since the guidance allows for validation to be based on 15 samples, these can be interpreted as a simple rule that no pair of measurements may differ by more than the given tolerance if composite sampling is used, but one difference is permitted if spot sampling is used.

6.2. Expected validation outcomes

6.2.1. Analytical method

The probability of an automatic measurement system meeting validation requirements using the criteria given in the Good Concrete Guide 11 can be determined analytically. The slump test will be used as an example. For concrete with slump of 160 mm, the single-test standard deviation accounting for rounding error is

13.4 mm (see 2.2). For an automatic measurement system with accuracy equivalent to the manual test (System 3, Table 2), the standard deviation of the distribution of differences can be calculated by combining the variability of the two tests for each measurement pair:

$$\sigma_d = \sqrt{2\sigma^2} \tag{6.1}$$

For the example, this is 19.0 mm. This analysis will assume the decision will be based on the minimum number of data points, 15, and that the composite test criteria will be used (i.e. 95% within ±30 mm). The probability that the system passes validation can be stated as the probability that of 15 draws from a normal distribution with a mean of 0 and a standard deviation of 19.0 mm, none of the values have an absolute value greater than 30.

For a single draw, the probability that the result is within range is given by:

$$P(|d| \leq 30) = P(-30 \leq d \leq 30) = \Phi\left(\frac{30}{\sigma_d}\right) - \Phi\left(\frac{-30}{\sigma_d}\right) \tag{6.2}$$

Where Φ is the standard normal cumulative density function. This evaluates to 88.7%. For 15 samples, assuming they are independent, this must be raised to the power of 15.

$$P(|d| \leq 30 \mid n = 15) = \left[\Phi\left(\frac{30}{\sigma_d}\right) - \Phi\left(\frac{-30}{\sigma_d}\right)\right]^{15} \tag{6.3}$$

Which evaluates to 16.4%. The probability of correctly determining that an automatic slump measurement system with accuracy equivalent to the manual test in the 160 mm range is valid is less than 20% using the composite sample criteria.

For spot sample criteria, the tolerance is significantly larger (40 mm rather than 30 mm in this example). In addition, the compliance level is 90%, which allows for a single difference to be outside the tolerance assuming the minimum number of data points are collected. We can formulate the probability of an automatic system passing validation as the addition of the probability that all differences are within the tolerance and the probability that only one difference is larger than it. If the tolerance is T, the probability p of a single result being with the tolerance is:

$$p = \Phi\left(\frac{T}{\sigma_d}\right) - \Phi\left(\frac{-T}{\sigma_d}\right) \tag{6.4}$$

And the probability that no more than one of 15 differences is larger than T is:

$$P = p^{15} + 15(1-p)p^{14} \tag{6.5}$$

Since there are 15 different ways to observe exactly one difference outside the tolerance range. In this case there is a 90% chance of the automatic system being determined valid. A very different result than if the composite limits are used.

6.2.2. Simulation method

To evaluate how effectively these proposed validation criteria distinguish between automatic consistence measurement systems that are equivalent or better than the manual test and those that are less accurate, a series of validation tests were simulated using realistic parameters. A detailed description is given in 3.2.

The results for each of the systems described in Table 2 are shown in Figure 7. Systems 1 – 3 are valid, in that they measure consistence with an accuracy better than (Systems 1 and 2) or equivalent to (System 3) that of the manual test.

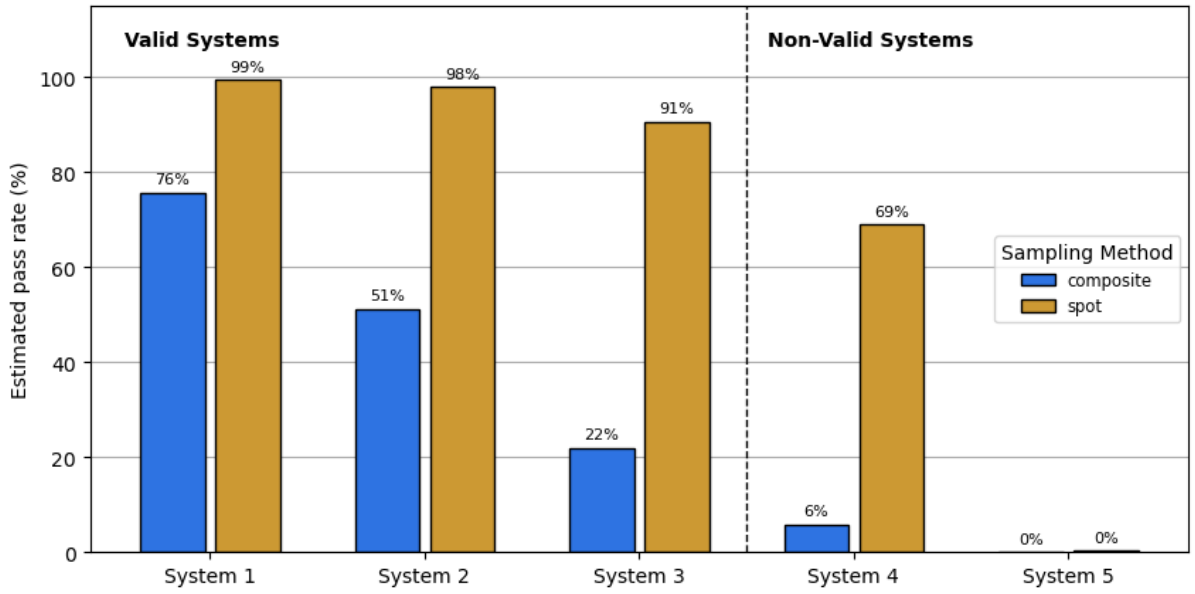

Figure 7 - Result of validation testing assessment using Good Concrete Guide 11 criteria for each of five simulated automatic consistence measurement systems. Systems 1-3 are valid systems with accuracy at least equivalent to the corresponding manual test. Systems 4 and 5 report inferior measurements that should fail a validation test.

System 1 is an idealised system that reports the true consistence exactly every time. Whilst this system passes validation 99% of the time when applying spot sample criteria, even this system has a 76% pass rate using the composite sample criteria, due to the limited precision of the manual test.

System 4 has significantly worse accuracy than the manual test but still has a 69% pass rate when assessed against the spot test criteria. In addition, it is significantly easier for the non-valid system (System 4) to pass validation using spot sample criteria than it is for valid systems 2 and 3 to pass when using the composite sample criteria.

This creates a high incentive to use the spot sample criteria, which allows systems that are significantly less accurate than the manual test to pass validation relatively easily. Conversely, the composite test criteria are very difficult to meet even in an ideal case where the automatic measurement system can report true consistence with zero error (System 1). This is due to the random variation and expected tester bias inherent in the manual test.

The guidance defines 15 data points as a minimum but allows for more. Since the compliance level is based on percentages of numbers smaller than 100, there are boundary effects that make the outcome highly sensitive to the specific number of data points collected.

For example, when validating an automatic slump measurement system using composite samples, the criterion is that 95% of the differences between automatic and manual tests are within ±30 mm. For 19 or fewer data points, this is only true if none of the differences are outside the limit. However, by adding one additional data point, any one of the differences is allowed to be outside the limit.

The effect of the number of samples on the validation pass rate for each of the simulated systems is shown in Figure 8. This demonstrates the boundary effect, whereby the proportion of simulated validation tests in which systems 2 and 3 are assessed as valid increases significantly for 20 data points compared with 19 or fewer. The same effect is seen going from 39 to 40 data points.

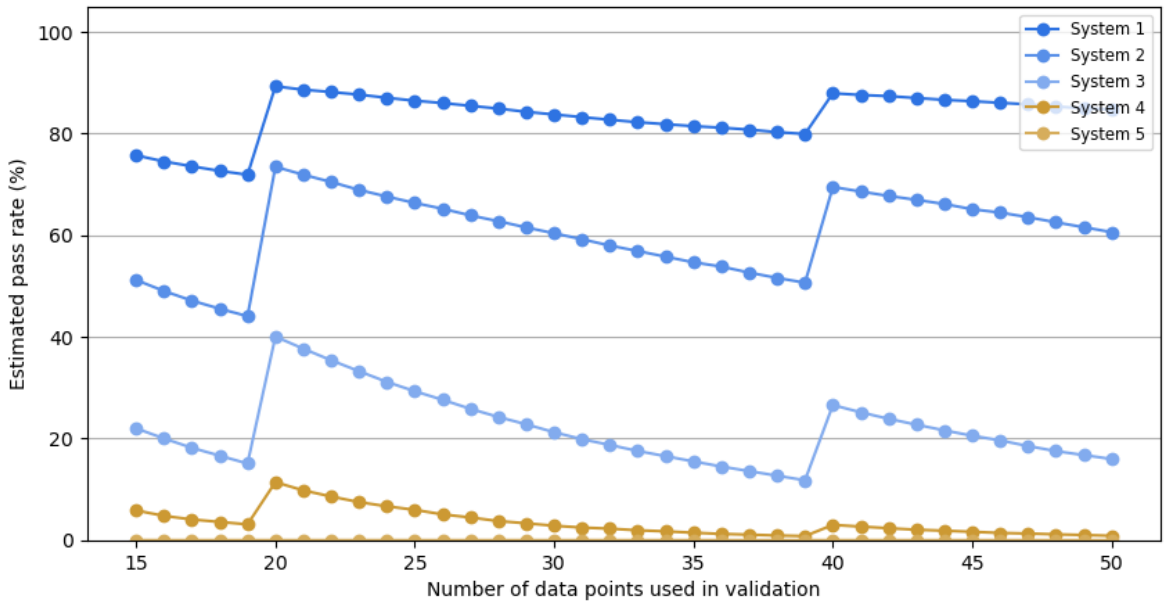

Figure 8 - The proportion of simulated validation tests where automatic slump measurement systems are assessed as valid using the Good Concrete Guide 11 [12] criteria for composite samples. Blue lines denote the three systems with accuracy better than or equal to the manual test. Orange lines denote the system with accuracy less than the manual test.

In summary, this approach has significant drawbacks and permits validation tests to be structured in ways that make the criteria either too lenient or unduly restrictive. Consequently, it does not provide a reliable basis for determining whether an automatic consistence measurement system is a valid replacement for the manual test.

7. FAQ

Can an automatic consistence measurement system be more accurate than the manual test?

Yes. There is no reason why an automatic system cannot be more accurate than a manual test. An automatic system with a low-bias and a low standard deviation can outperform a single tester and reach the performance of multiple averaged tests.

Would more data points improve the accuracy of the validation test outcome?

Yes, they would. It is possible to design the validation test procedure based on different requirements such as an acceptable limit of false positives and false negatives under certain conditions. The validation test procedure has been designed pragmatically such that it can be carried out within a reasonable time at a reasonable cost (section 3.1).

Can I carry out more than the specified number of tests, if this gives a more accurate outcome?

No. The tolerance applied to the limits of agreement is based on a defined number of data points. Adding more without adjusting the tolerance would increase the risk of a false positive.

Would it be better for the outcome to be the result of a statistical test?

Yes, this would provide information about how confident we are in the outcome based on the data. A common approach is to estimate confidence intervals for the limits of agreement and carry out a "two one-sided tests" analysis to determine whether both limits are within the target range with sufficient degree of confidence. Bayesian approaches could enable statements like "based on the data, there is a 70% chance that the system is valid". These are not described here as they require expert interpretation and therefore a higher barrier for adoption.

The automatic system reports consistence to the nearest 10 mm. Do I need to adjust the analysis?

No. It is not necessary to adjust the analysis. If the system only reports results to the nearest 10 mm this is a design choice that introduces rounding error to that system's measurements. This will reduce the accuracy of the system, but the benchmark against which to validate it is based only on the properties of the manual test and remains the same.

For compliance, the standards allow greater tolerance for spot tests vs. composite tests. Why are they treated the same here?

There is no reason the method of sampling should affect the precision of the manual test. Correspondingly, the precision tables in the standards do not distinguish between the two sampling methods. However, if the load is not well mixed then it is reasonable to expect that the composite test could be more representative of the average for a batch of concrete, which is likely the justification for allowing different tolerances for consistence tests carried out on spot vs composite samples. When validating automatic consistence measurement, it is important that the samples used in both the automatic and manual tests are well matched, and the best way to do this is to ensure that the load is well mixed prior to carrying out any testing.

Why are some consistence classes missing from the table of limit values?

Annex L of [9] states:

Due to a lack of sensitivity of the test methods beyond certain values of consistence it is recommended to use the indicated tests for:

Slump ≥ 10 mm and ≤ 210 mm

Flow diameter > 340 mm and ≤ 620 mm

Slump flow diameter > 550 mm and ≤ 850 mm

Measurement comparison should only be made where the manual test results are within the sensitive range. This means that tests within the F1, F6 and S5 ranges are not included. These consistence classes are defined

by having a consistence lower or higher than the sensitive ranges, this does not matter when using the tests to determine compliance. However, the precision metrics are not defined so no comparison can be made.

Can I use the limits of agreement to compare performance of multiple automatic measurement systems?

No. In a validation study comprising 20 tests, the calculated limits of agreement could have a high level of uncertainty associated with them. It would be misleading to conclude that one system is more accurate than another based on the limits of agreement determined from this procedure.

8. References

[1] M. J. Bland and D. G. Altman, “Measuring agreement in method comparison studies,” *Statistical methods in medical research,* vol. 8, no. 2, pp. 135-160, 1999.

[2] Cloud Cycle, “Procedure for validation automatic consistence measurement systems,” 2026.

[3] M. A. Mansournia, R. Waters, M. Nazemipour, M. Bland and D. G. Altman, “Bland-Altman methods for comparing methods of measurement and response to criticisms,” *Global Epidemiology,* vol. 3, 2021.

[4] British Standards Institution, “BS EN 12350-2 Testing fresh concrete - Slump test,” 2019.

[5] ASTM International, “C143/C143M-12 Standard Test Method for Slump of Hydraulic-Cement Concrete,” 2015.

[6] ASTM International, “Interlaboratory Study to Establish Precision Statements for ASTM C143/C143M, Standard Test Method for Slump of Hydraulic-Cement Concrete,” 1999.

[7] British Standards Institution, “BS EN 12350-5 Testing fresh concrete - Flow table test,” 2019.

[8] British Standards Institution, “BS EN 12350-8 Testing fresh concrete - Self-compacting concrete. Slump-flow test,” 2019.

[9] British Standards Institution, “BS EN 206 Concrete. Specification, performance, production and conformity,” 2013.

[10] S. M. Ross, “Introduction to probability and statistics for engineers and scientists,” Academic Press, 2020.

[11] M. J. Bland and D. G. Altman, “Statistical methods for assessing agreement between two methods of clinical measurement,” *The lancet,* vol. 327, no. 8476, pp. 307-310, 1986.

[12] The Concrete Society, “Good Concrete Guide 11: Digtial Monitoring and Measurement of Fresh Concrete,” 2025.

[13] ISO, “5725-1 Accuracy (trueness and precision) of measurement methods and results,” 2023.

[14] A. Carkeet, “Exact parametric confidence intervals for Bland-Altman limits of agreement,” *Optometry and Vision Science,* vol. 92, no. 3, pp. 71-80, 2015.