Should the repeatability of the instrument under test be included in test uncertainty?

Speaker/Author: Craig M Shakarji
Physical Measurement Laboratory
National Institute of Standards and Technology
100 Bureau Dr., MS 8210
Gaithersburg, Maryland 20899, USA
Phone: (301) 975-3545, Email: craig.shakarji@nist.gov

Author: Steven D Phillips
Physical Measurement Laboratory
National Institute of Standards and Technology
100 Bureau Dr., MS 8210
Gaithersburg, Maryland 20899, USA
Phone: (301) 975-3565, Email: steven.phillips@nist.gov

Abstract

Should the repeatability of the instrument under test be included in test uncertainty? The answer—which surprises many—is no, at least for the range of cases documented in this paper. This has been an area of confusion even among experts around the world. Some ramifications of this confusion have been known cases where companies have had to greatly and artificially inflate their accuracy specifications for their measuring instruments in order to meet decision rule requirements for acceptance testing.

The purposes of this paper are to (1) Clearly identify a large set of common instrument testing/calibration situations where instruments are verified to their accuracy specifications where the repeatability of the instrument under test should not be included, (2) Take the reader through a detailed tutorial path to gain understanding as to why it must be that such a component should not be included in those cases, and (3) Identify four common mistakes that lead even experts to incorrectly include the repeatability component: These being (i) misunderstanding that there are two measurands involved in testing, (ii) misunderstanding the difference between the test value uncertainty and uncertainty in the overall test, (iii) misunderstanding the role of test value uncertainty by envisioning its use in subsequent measurements made by the tested instrument, and (iv) misapplying the inclusion of the repeatability of the calibrating system performing the test to include the instrument under test.

1. Introduction

The question of whether the repeatability of the instrument under test should be included in test uncertainty has caused confusion in the metrological world since the publication of the Guide to the Expression of Uncertainty in Measurement (GUM) [1] and likely earlier still. In “normal” measurements, i.e., on objects that are not indicating instruments, the repeatability of the measuring instrument must be included in the uncertainty evaluation somehow—even if it is incorporated as part of another component with another name. But when an

indicating instrument is being tested, and the measurand is the error in the instrument's measurement, how should the repeatability of the instrument under test be handled?

The answers to just what does and does not get included in test uncertainty is complicated by the fact that there are various types of testing with fundamentally different test measurands. Without going into further detail, we simply state that it is too large a subject for this paper to address all possible testing scenarios. However, there is a large swath of instrument testing per national and international standards that will be addressed here. For these, at least, the repeatability of the instrument should not be included in the test value uncertainty.

In 2016, Salsbury presented an NSCL International paper [2] highlighting some aspects of the recently published standard on test uncertainty related to dimensional metrology ISO 14253-5 [3], which included the fact that—for scope of testing relevant to that standard—the repeatability of the instrument under test is not included in the uncertainty of an individual test value. However, *why* the repeatability is not included involves subtle concepts that need explanation in light of confusion even among experts. The purpose of this paper is to take the reader through a course of thought to help explain why the repeatability should not be included in these cases.

For the purposes of this paper, we assume an indicating instrument has an accuracy specification defined by a documentary standard and stated by a maximum permissible error (MPE) that is a constant across its rated operating conditions. (An MPE specified by a non-constant function over its rated operating conditions is a straightforward extension of the concepts in this paper, but the constant MPE is used to keep examples and concepts simpler.) The rated operating conditions typically span several measuring conditions (e.g., range of ambient temperatures during measurement, range of allowable measurands, etc.).

Section 2 will further delineate the scope of testing we are considering in this paper. It describes a test protocol based on a system of a finite number of spot-checks. Section 3 then reasons from two example cases that the variation in time of the errors of the instrument under test have no bearing on the uncertainty of any individual test value for one of the spot-checks; thus, the repeatability of the instrument under test is not included as a component of the test value uncertainty. Our conclusions are given in Section 4.

2. A careful delineation of the scope of testing considered

Much of the confusion can be cleared up by a careful delineation of the scope of the testing being discussed. In our discussions in this paper we are assuming the following:

1) There is a written test protocol that has been agreed upon that includes sufficient definitions and instructions that—when executed—produces an unambiguous overall pass or fail. What action is taken as a result of an overall pass or fail result depends on the agreement. For example, in an acceptance test, a buyer and seller have agreed on the sale of a measuring instrument contingent upon the instrument's passing a particular documented test
2) The test protocol produces multiple (or possibly even just one) individual pass/fail test results, which then produce the overall test result of the test protocol. For example, a test protocol may require 35 measurements, each repeated 3 times for a total of 105 individual tests. If every one of them results in a pass, then the overall test result is a pass, and if any one (or more) individual test fails, then the overall test result is a fail.
3) Each individual test consists of obtaining a test value and comparing that test value against a threshold value. An uncertainty associated with the test value is taken into account by means of a decision rule to

produce an unambiguous pass or fail result for the individual test. For example, an instrument under test has a claimed maximum permissible error (MPE) and (following the test protocol) an individual test value is obtained by having the instrument perform a measurement of a calibrated object. The test value is (in this example) the estimated error (the measured value produced by the instrument under test minus the calibrated value). The test value is then compared against the MPE where—in the comparison—the uncertainty associated with the test value is taken into account by means of a predetermined decision rule.

The reader should take care to note the difference between the two concepts of the overall test and the individual tests that are contained within it. This difference may seem straightforward, but the careful distinction is important. This is because the test uncertainty discussed for this type of test is the uncertainty associated with the test value from an individual test. It is not the uncertainty associated with the overall test outcome. For this reason, the label "test value uncertainty" is more specifically descriptive than "test uncertainty" for the cases currently under consideration. Thus (as will be discussed later) the test value uncertainty is not a measure of how thoroughly the overall test interrogates an instrument but rather what uncertainty is associated with an individual test value. And various test values within an overall test could have different test value uncertainties.

Once the test outcomes are known for the individual tests, it is a matter of counting to determine if the overall test is a pass or fail, which is a step that incurs no additional uncertainty.

The type of testing contained in this scope has been employed in numerous national and international documentary standards. Some dimensional metrology examples include various ANSI/ASME B89 standards and the ISO 10360 series of standards on various coordinate measuring technologies. ISO 14253-5 [3] covers the topic of test uncertainty for the area of verification of dimensional metrology instrumentation, and the concepts covered in this paper are consistent with that standard.

3. Reasoning from two example cases

In this section, we use two example cases to guide our thinking through the concepts of the overall test, the test values, and the test value uncertainty, as well as to address the question of whether the repeatability of the instrument under test should be included in the test value uncertainty. The first example involves the testing of a laboratory that has been equipped with an environmental control system that regulates the temperature in the room and the measurand is the room's temperature. The second example involves a coordinate measuring machine (CMM) being tested against its accuracy specifications where the measurand is the CMM's measurement error. The second example is the most relevant to the topic of this paper (since it involves the testing of an indicating measuring instrument) but the first has been specifically chosen because we readily grasp the concept that seeking to measure the temperature at different times involves different measurands. We immediately know that the temperature outside at noontime is a different measurand than the temperature outside at midnight. We use this first example even though it does not involve the testing of a measuring instrument.

Example 1 involving temperature measurements

Consider an example where a company has been contracted to install an environmental control system in a laboratory such that the temperature within the room to always be kept between 19 °C and 21 °C (i.e., 20 °C ± 1 °C). The work has been completed, but before the payment is made the room must pass a temperature test—per the agreement—according to the specific test protocol identified.

Before going further, we apologize to the experts who specialize in environmental control; this is a completely invented example of a test protocol for the simple purpose of guiding our thinking on a general issue. Having said that, suppose this fictitious test protocol states that the temperature (at some well-defined point in the room) shall be measured 24 times in the span of about 24 hours. Each temperature reading shall be taken at some instance of time in the interval between 10 minutes before and 10 minutes after the hour (e.g., between 9:50 and 10:10, between 10:50 and 11:10, etc.).[1] While the test protocol allows the exact time of measurement (within the interval) to be up to the tester, the tester cannot use any test-specific information (e.g., watching the temperature) and can only arbitrarily pick some instant of time within the interval to take a reading. For each of these temperature readings, the measured temperature value is compared to see if it is indeed within the threshold 20 °C ± 1 °C. This comparison (as part of the written protocol) must be done, taking into account a decision rule of guarded acceptance [4] (also known as stringent acceptance [5]), such that the temperature and its entire expanded uncertainty interval (say at the $k = 2$ level) must be wholly contained within the specified threshold interval. Thus, the overall test consists of 24 individual tests, each of which must pass for the overall test to pass.[2]

Now suppose such a test were executed, and the thermometer chosen to perform the test is known to produce each measurement result with a $k = 2$ expanded uncertainty of 0.01 °C. The tester who (for some reason) manually records the temperature (and who stays awake for 24 hours!) makes use of an imprecise clock on the wall to determine when to take a temperature reading. Since (we will suppose) it is known the wall clock could be incorrect by a maximum of two minutes, the tester records the temperature at some instance between 8 minutes before and 8 minutes after each hour (per the time on the clock) recording the wall clock time at each measurement. See fig. 1 showing an example of the temperature graph over the 24 hours and the 24 measurement results. The 24 expanded uncertainty intervals (of 0.01 °C each) about the measured values are not shown, but are smaller than the dots representing the readings. The tester concludes that all 24 individual tests have passed (using an expanded test value uncertainty of 0.01 °C for every test value) and thus also concludes that the overall test result is a pass.

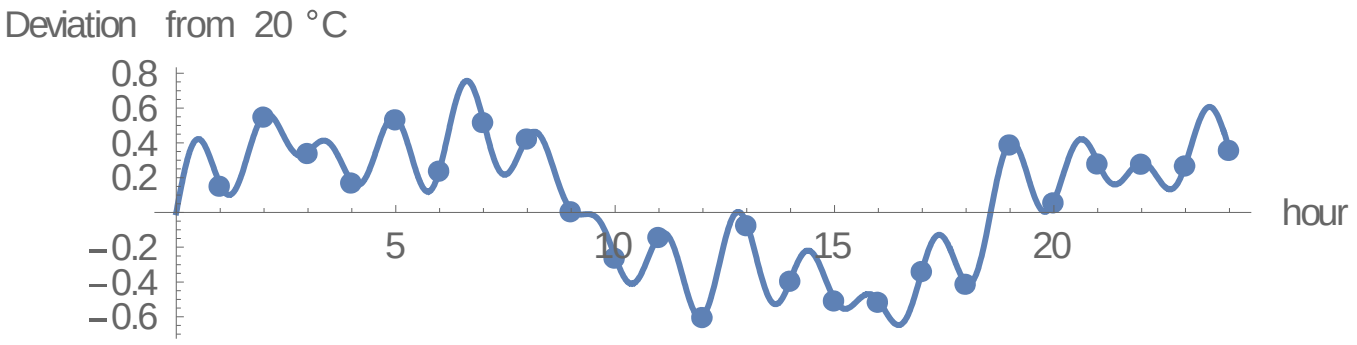

Fig. 1. An example graph showing the actual temperature variation and the points of temperature readings

We will now imagine the tester is interrogated on the matter and is called into question for not including the proper repeatability component in the test uncertainty. The first objection is that a component of the test uncertainty should be the standard deviation of the 24 measured temperatures and that such "repeatability" should be included in each uncertainty. In this example, this would increase the $(k = 2)$ expanded test value uncertainty from 0.01 °C to roughly 0.5 °C with the consequence that the overall test would not pass since the acceptance zone of the decision rule, which is maximum allowed temperature reduced by the coverage interval (VIM 2.36, [6]) also

[1] While fine for our fictitious example, a cyclical test like this would have a weakness, since the air system could also be similarly cyclical.
[2] In reality, temperature readings can be taken so easily that a test protocol would not be written with such sparse sampling.

known as the uncertainty interval [4], would not contain some of the observed temperature values. (We note that if the number of samples were increased, the standard deviation would not change much, meaning that the overall test would still not pass even if there were thousands of added readings, none of which exceeded 0.8 °C.)

The tester (correctly) rebuts that each test involves its own measurand. The temperature at hour 1 is a different measurand than the temperature at hour 2 just like we all know the temperature outside at 1:00 pm is a different quantity than the temperature at 2:00 pm. The fact that the quantity (temperature) varies with time has no bearing on the ability to measure the temperature at one specific time. If it did, then one could not measure the temperature on a summer day with much accuracy at all since the temperature would be so different next winter!

When we remember that each of the 24 tests is treated as an individual, isolated test with its own measurand this objection may seem like nonsense to the reader (and it should) but that is partly because this example has been carefully crafted to highlight the fallacy. In fact, this exact mistake is often made due to the fact that the concept of test uncertainty in other contexts is more subtle. The authors are aware of at least one case where this mistake had led to confusion in the marketplace for certain instrument specifications. And although the confusion is being cleared up for the case mentioned, we are emphasizing that this is no mere academic problem and that much confusion remains.

The objector concedes that point but now tries now another approach: “You indicate that at time 11:03 am you took a temperature reading. In fact, you were using an inaccurate clock. Due to the maximum two-minute clock error possible, your measurement could have actually taken place anytime between 11:01 and 11:05. Thus, you need to take into account the variation in temperature that can occur over a couple-of-minute time period in your test value uncertainty. This variation would be much greater than your claimed 0.01 °C test value uncertainty.” This objection demands that repeatability over a shorter time period be included in the test uncertainty.

This objection is answered again by the tester who indicates that the test protocol allows for any time within the stated interval to be an acceptable time to take a reading. Therefore, even though the tester does not know what that time exactly was, it was certainly a time within the acceptable interval and therefore an allowed point in time to take the reading. The tester insists that the temperature one minute or two minutes before or after the actual measurement time is a different quantity and has no bearing on the accuracy of the recorded temperature even if the tester does not know precisely when it was. So, while it is true that the tester does not know very well what the temperature was at the true time of 11:03, the tester certainly knows that at some instance in time between 11:01 and 11:05 the temperature was at its recorded value with a 0.01 °C expanded uncertainty.

The objector tries a third time and indicates that for some short period of time, the temperature is constant enough to be considered a single quantity having an essentially unique value. During that time, the variation of temperature readings—if they were taken in rapid succession—would be a repeatability that needs to be included. The tester replies in agreement, but indicates that that repeatability is the repeatability of the thermometer, which is already taken into account by the 0.01 °C expanded uncertainty attributed to the thermometer readings.

The objector then tries lastly to argue that 24 measurements (one per hour) is entirely too small a number to ensure that the temperature of the room never exceeded the required limits. It is argued that the number should be more like 2400 or 24 000. And since 24 measurements are too few, uncertainty needs to be taken into account to reflect the lack of knowledge of the variation of the temperature in the times between readings. To this the tester sympathizes that there is in fact a lack of knowledge of the temperature of the room between readings. The tester even goes as far as to agree that it would be a more robust test if the test protocol should have included more

individual tests. However, those issues are ones that have to be considered in the writing of the test protocol and in the agreement made. Once the agreement has been made that the overall pass or fail would be determined by the 24 temperature readings, the "rules of the game" have been set, and the overall test is simply comprised of however many individual tests are in the agreed-to protocol. The accuracy of any individual temperature reading cannot be dependent on how many other readings will be taken afterwards. A good test protocol should weigh the benefits of how many individual tests are performed against the cost and time burden incurred in performing the tests.

Example 2 involving CMM testing

The use of the contrived temperature example will help us in analyzing a more pertinent example of a CMM. Suppose a buyer and seller have agreed on the purchase of a CMM that has an accuracy specification (MPE) of 1 μm for any point-to-point length in any position within its measurement volume as defined in detail in ISO 10360-2 [7]. We also suppose that, for this example, an agreement was made that the purchase is contingent on the CMM—once installed—passing an acceptance test as document in the test protocol ISO 10360-2. While that standard includes various testing, for this example we will restrict ourselves to the fact that the overall "E_0" test requires that the CMM make 35 measurements, each repeated 3 times for a total of 105 measurements of calibrated test lengths (e.g., gage blocks) of various sizes and in various positions. Each measured value is compared with the calibrated length to estimate the error of the CMM measurement. These test values are compared with the MPE, taking into account the test uncertainty using the same decision rule described in the previous example.

For our specific example, we will assume that a CMM with claimed MPE of 1 μm has been installed and that the 105 measurements have been carried out using calibrated test lengths (e.g., gage blocks) each of which has an expanded ($k = 2$) uncertainty in their lengths of 0.1 μm. (This 0.1 μm is assumed to account for all things affecting the length as presented to the CMM, e.g., the calibration uncertainty, the nonrigid fixturing of the test length, etc.) The test was performed, and indeed, each test value was within the MPE, even taking into account the 0.1 μm uncertainty per the decision rule. The test protocol yields an overall test result of "pass" because each individual test result was a "pass."

The objections and their responses that arose during the temperature example apply here. We do not blindly inflate the test value uncertainty by taking the standard deviation of the 105 test values observed. Nor do we inflate the uncertainty due to imperfect knowledge of the exact operating conditions at the time of testing, provided they are within the rated operating conditions (remembering that the MPE is constant across the rated operating conditions in this case). Nor do we include an uncertainty component to take into account the lack of coverage, due to the agreement made.

However, this example of the CMM is different from the temperature example in this regard: for the case of the temperature measurements, repeated measurements taken over a short time period would yield little observed variation, but the 3 CMM measurements taken on the same gage block in the same position can have significant variation, even over the short term. This variation should still not be taken into account in the test value uncertainty, since the measurement error the CMM commits at time t1 is a different measurand than the error the CMM commits at time t2, even if the times t1 and t2 are close together. Just like in the temperature example, the fact that the temperature was significantly different at 1:00 than at 2:00 did not affect the test value uncertainty, so also the fact that the CMM error at a point in time (t1) is different than it is a few seconds later does not affect the test value uncertainty for the error at time t1.

The fact that the measurands are different at times t1 and t2 may need further explanation, since it is easy to misunderstand the measurand in this case. The test measurand—the quantity we intend to measure—in this case is the error the CMM makes when making a particular measurement. In order to ascertain this error, the CMM is asked to perform a measurement on a gage block, for example. We thus have two measurands involved. The key measurand we are interested in is the *error* the CMM makes. The other measurand (the sub-measurand, if you will) is the *length* of the gage block. When two successive measurements are made at close times t1 and t2, we agree that the "sub-measurand" has not changed, i.e., the length of the gage block hasn't measurably changed. However, the error the CMM makes from t1 to t2 can change dramatically. Indeed, they are two different measurands. When seeking to ascertain the error the CMM makes at t1, we are seeking a different quantity than when we are seeking the error the CMM makes at time t2.

The previous example involving temperatures can help us here. We readily saw that the temperature at one time was a different quantity than the temperature at a later time. If we think of the CMM as a black box that produces an error value at the press of a button, it becomes easier to see that the error at t1 is a different measurand than that at t2. If I press the button to view the CMM error at time t1 and then again at time t2, I would expect different numbers (since CMM errors normally have a seemingly random component to them). They are different measurands akin to the case where we readily realize that the temperature at 1:00 and at 2:00 are different measurands. In the CMM case, it does not matter that times t1 and t2 are close together, since the error the CMM makes can change significantly over a very short time.

If it appears that the physical phenomenon being measured at times t1 and t2 has not changed, it is probably because it is the "sub measurand" (the length of the gage block) that is incorrectly being considered. The measurand at hand is the error the CMM makes, and the inner workings of the CMM are in different states at times t1 and t2. Thus, one measurand was the error the instrument makes when its inner workings were in the state that they were in at time t1 and the other measurand is the error the instrument makes when its inner workings are in the state they were in at time t2. The fact that we cannot describe and do not even know the states of the inner workings does not matter. This is akin to the case of the temperature reading earlier: even though a temperature reading was taken at approximately 11:03, the tester could not say exactly when it was except that it was an allowable time per the test protocol. If the tester had to put the measurand into words, it would be "the temperature at the defined location in the room at whatever the actual time was when the reading was taken." The measurand sounds oddly defined, but it is perfectly sound and analogous to the inability to know or express the state of the instrument in the CMM case.

As in the case of the temperature, the repeatability of the thermometer was included in the test uncertainty (it was already incorporated into the 0.01 °C expanded uncertainty). Likewise, we would include the repeatability of the gage block length presented to the CMM (which is very small and already included in the 0.1 µm expanded uncertainty).

We further note that even though the test value uncertainty was 0.1 µm in every case, this value is not intended to and in fact does not represent the accuracy of the CMM when it is used in subsequent measurements. That accuracy is more correctly conveyed by the MPE of 1 µm.

The writers of the ISO 10360-2 standard purposefully included certain testing configurations that are known by experts to be likely to reveal the largest errors of the CMM. They also included some user-selectable configurations, which means a CMM manufacturer does not know a priori all the locations and orientations that

will be tested. ISO 23165 [8] covers the subject of test uncertainty in the specific context of CMM testing per ISO 10360-2 and is consistent with the concepts discussed here.

4. Conclusions

The subject of test uncertainty is complex and efforts have been made in this paper to reduce issues to simplest cases. Even with such efforts, the concepts remain subtle but we have demonstrated that—at least for the scope of testing considered here—the repeatability of the instrument under test should not be included in test uncertainty. We have reduced example cases to guide through the reasoning for this, and sought to highlight areas where confusion can likely occur. In the end, the thinking has been consistent with published international standards that have been written for the kind of testing considered in the scope of this paper.

References

1. JCGM 100:2008, Evaluation of measurement data — Guide to the expression of uncertainty in measurement (GUM)

2. Salsbury, James, G., Understanding the Test Measurand and the Profound Impact on Calibration, Verification, and Uncertainty, NCSL International Workshop and Symposium, 2016.

3. ISO 14253-5:2015, Geometrical product specifications (GPS) -- Inspection by measurement of workpieces and measuring equipment -- Part 5: Uncertainty in verification testing of indicating measuring instruments

4. JCGM 106:2012, Evaluation of measurement data – The role of measurement uncertainty in conformity assessment

5. ISO 14253-1, Geometrical product specifications (GPS) -- Inspection by measurement of workpieces and measuring equipment -- Part 1: Decision rules for proving conformity or nonconformity with specifications.

6. JCGM 200:2012, International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM 3rd edition)

7. ISO 10360-2:2009 Geometrical product specifications (GPS) -- Acceptance and reverification tests for coordinate measuring machines (CMM) -- Part 2: CMMs used for measuring linear dimensions

8. ISO/TS 23165:2006, Geometrical product specifications (GPS) -- Guidelines for the evaluation of coordinate measuring machine (CMM) test uncertainty