

Where do we find the graphs from PT programmes showing measurement performance?

Paul De Bièvre¹

Published online: 22 January 2016
© Springer-Verlag Berlin Heidelberg 2016

Interlaboratory comparisons (ILCs), especially those aiming specifically at proficiency testing (PT), are intended to show measurement performance by participating laboratories. (For definitions of “ILC” and “PT”, see entry 7.3-1 in [1] and pp 255–274 in [2]). Comparative graphs of the results of such programmes are an excellent way to display such performance because any participant in the programme can easily and immediately recognize (and be recognized for) its performance [2]. This easy “readability” is also a particularly important feature for any user who needs to know—in a heavily regulated field—about the measurement performance of laboratories the services (s)he wants to use. These graphs have demonstrated in the past that the performance of “accredited” laboratories was not necessarily satisfying [3]. Already in 1998 that observation was reported by ILAC President Alan Bryden in his opening lecture for an ILAC seminar in Sydney [4]. He warned “that there exists documented evidence that accredited laboratories perform no better than non accredited laboratories” and Bernard King quoted in the same Seminar from “a major UK study” [5–7]: “The overall conclusion from this study is that laboratories with third party assessment performed no better than laboratories without such assessment. Although this is in line with some studies reported in the literature, it is at odds with the

high regard that the measurement community has for third party assessment schemes.”

Has the situation improved since?

To answer this question, it would be very useful to have such graphical displays regularly available for all who are in need of easy and unambiguous information in matters of measurement performance of laboratories making claims in measurement, with respect to the intended use of these results. As long as that is not the case, the answer to the above question is: we don't know.

One of the most useful characteristics comparative graphs enable us to see immediately is whether measurement uncertainties of participants' declared measurement results (entry 2.10 in [9]) do—or do not—overlap each other. In cases of good measurement performance, they should. When they do not, that leads to questions as formulated here, followed by our best answers:

1. do measurement results declared by the participants carry a measurement uncertainty? if they don't, they cannot be compared as measurement results because they are not measurement results as formulated by the highest authority in this domain: the Guide on the Expression of Uncertainty in Measurement (the GUM, published 1993/95), [8]:

“When reporting the result of a measurement of a physical quantity, it is obligatory that some quantitative indication of the reliability of the result be given” (Sect. 0.1 in the GUM) as well as in the International Vocabulary of Metrology (VIM), where measurement result is defined (entry 2.9 NOTE 2 in [9]); it follows that participants' *measurement results not carrying a measurement uncertainty statement, cannot be compared; they do not even come under the definition of*

Disclaimer The author is a member of the Joint Committee on Guides for Metrology (JCGM), Working Group 2 (VIM). The opinions expressed in this column do not necessarily represent the view of the Working Group or of ACQUAL.

✉ Paul De Bièvre
paul.de.bievre@skynet.be

¹ Kasterlee, Belgium

‘metrological comparability of measurement results’ (entry 2.46 in [9]) because they are not measurement results.

2. if they are stated, are the measurement uncertainties of the results as declared by the PT participants too small, i.e., too optimistic, thus suggesting “significant differences” between them which—in fact—are not real? closer analysis of the measurement uncertainties using GUM guidelines frequently shows that the answer is affirmative, e.g., when Type A uncertainty is equated to measurement uncertainty, whereas it is only part of that uncertainty; could a lack of knowledge by the participants about the GUM still be responsible for such underestimations 20 years after the release of the GUM? possibly yes
3. if the PT programme offers an “assigned” value [3], either an “average of participants’ values” or a “consensus” value, it must also be displayed with an uncertainty; that raises the question whether measurement uncertainties of the participants overlap the assigned value, or at least the uncertainty thereof? [again, note that measurement uncertainty is defined in entry 2.9 in [9] and the added NOTE 2: “A measurement result is generally expressed as a single measured quantity value and a measurement uncertainty”; this NOTE 2 makes a measurement uncertainty an inherent part of any measurement result]; the answer to the raised question seems to be: frequently there is no overlap; consequently, suspicion arises whether the stated measurement uncertainties were too optimistic
4. do participating laboratories in PT programmes still apply the definition of measurand given in the 1993 edition of the VIM as: “particular quantity subject to measurement” (entry 2.6 in [10] now superseded by the VIM 2008/2012 where it is redefined as “the quantity intended to be measured”) in which case just a measured value (i.e., without statement of measurement uncertainty) was considered sufficient? For chemists, the 1993 definition was an erroneous formulation: since most chemical measurements end up in the measurement of an electric current, only the uncertainty of that current measurement was to be taken into account as uncertainty for the end result of the measurement procedure; according to this—now antiquated—definition; there was no need to include the uncertainties of the unavoidable chemical operations necessary prior to the “electric measurement”; [in short, but of particular importance, is that uncertainties which of necessity are associated with chemical operations performed on the unknown sample before the measurement, did not have to become part of the measurement uncertainty of the final result: these operations were not included in the definition of measurand when its definition is limited to “quantities subject to measurement”].

PT programmes are almost universally claimed to serve the evaluation of the measurement performance of participating laboratories. One would think that such an evaluation needs criteria against which the evaluation is performed. But the “average” or “consensus” value of the participants is derived from the results of the participants, and hence is dependent from that “average” or “consensus.” Its use as criterion for evaluation of measurement performance is therefore dependent from the same measurement performance as the one which is being subjected to evaluation. This is a circular reasoning (or a “self-fulfilling prophecy”?): if every participant contributes to establishing the reference value, that essentially precludes the use of participants’ results (to contribute) to establishing a “reference value” serving the purpose of being a criterion for the evaluation.

Maybe there still are other “hidden structures behind the things we observe” which contribute in part to the inconsistency which we see in the pictured results of a number of PT programmes.

As usual, any comment, question, or amendment is welcome, preferably as a contribution to the Discussion Forum of this Journal.

References

1. De Bièvre P, Fajgelj A, Dybkaer R, Hibbert B (2011) Metrological traceability of measurement results in chemistry: concepts and implementation, IUPAC technical report. Pure Appl Chem 83:1873–1935. Section 9 at <http://iupac.org/publications/pac/83/10/1873>
2. De Bièvre P (2009) Comparing results of chemical measurements: some basic questions from practice, In: Pavese F, Forbes AB (eds.) Data modeling for metrology and testing in measurement science. doi:10.1007/978-0-8176-4804-6
3. ISO 17043 (2013) International Organization for Standardization, Geneva
4. Bryden A (1998) Opening lecture of ILAC seminar, NATA, Sydney, Australia
5. King B (1998) The ILAC 1998 seminar papers, 18–23 Oct 1998, NATA, Sydney, Australia
6. Lawn R (1994) An overview of the VAM 9 survey work and its principal findings, Laboratory of the Government Chemist, 1994
7. Lawn R, Thompson M, Walker RF (1997) Proficiency testing in analytical chemistry. The Royal Society of Chemistry, London
8. BIPM, IEC, IFCC, ILAC, IUPAC, IUPAP, ISO, OIML (1993/1995) Guide to the expression of uncertainty in measurement (GUM), JCGM 100:2008 at www.bipm.org/en/publications/guides/gum.html
9. BIPM, IEC, IFCC, ILAC, IUPAC, IUPAP, ISO, OIML (2008) The international vocabulary of metrology—basic and general concepts and associated terms (VIM), 3rd edn, Joint Committee on Guides for Metrology, JCGM 200:2008/2012 at <http://www.bipm.org/vim>
10. BIPM, IEC, IFCC, IUPAC, IUPAP, ISO, OIML (1993), International vocabulary of basic and general terms in metrology, 2nd edn, Joint Committee on Guides for Metrology (JCGM), ISO, Geneva