Int. J. Metrol. Qual. Eng. **1**, 7–9 (2010)
© EDP Sciences 2010
DOI: 10.1051/ijmqe/2010004

# On hierarchical vs. non-hierarchical comparisons in metrology and testing

F. Pavese*

Istituto Nazionale di Ricerca Metrologica (INRIM), Strada delle Cacce 73-91, 10139 Torino, Italy

Received: 15 November 2009 / Accepted: 5 February 2010

**Abstract.** The type of data treatment is different depending on whether the comparison, in particular a key comparison of the MRA (mutual recognition agreement), is of the hierarchical or non-hierarchical type. This term does not mean a possible hierarchy among the participant laboratories; nor, in the opposite sense, a non-hierarchy among them like in the MRA key comparisons, but an intrinsic characteristic of the comparison measurand or design. It is a typical *hierarchical comparison* when the comparison involves artefact standards. In this case, the summary parameters of the comparison are hierarchically higher than the input dataset. In case of *non-hierarchical comparisons*, the summary parameters are generally not of a hierarchically higher level than the input dataset, because the comparison dataset can be considered drawn from a single super-population. This happens, when a single standard is circulated for measurement; when the measured samples are all drawn from a single batch of a reference material; when the standards are all realisations of a single condition – namely a physical or chemical state. This paper will discuss in detail these two categories.

**Keywords:** Comparison; hierarchy; metrology; testing; CAFMET2010; papers

## 1 Introduction

The type of comparison data treatment is different depending on whether a comparison is of the hierarchical or non-hierarchical type.

This term does not mean a possible hierarchy among the participant laboratories, typically the National primary laboratory and accredited laboratories in metrology comparisons, or in most similar exercises in the field of testing (often called proficiency tests or round-robin exercises); nor, in the opposite sense, a lack of hierarchy among them, like in the MRA key comparisons or some top-level studies for a new method or a new reference material in testing. It is discussed here, instead, as an *intrinsic characteristic* of the comparison measurand or design [1].

The typical and generic observation model of the comparison data, for the $n$th laboratory providing data to the comparison as *its comparison input*, is the following [1]:

$$x_n = a + b_n + e_n \, n = 1, ..., N, \tag{1}$$

where $x_n$, an estimate of the measurand value, is drawn from a random variable $X_n = f(Q_1, ..., Q_M)$, $Q_m$ being the 'measurable quantities', and $e_n$ is the zero-mean random error associated with the observation under repeatability conditions. Most obviously, a *laboratory* bias term $b_n$ is included in the model because its statistical estimate is the main purpose of the comparison. The exact meaning of $a$ and $b_n$ depends on the type of comparison.

In general, $x_n$ is a single value supplied to the comparison by each $n$th laboratory, assumed to be an estimate of the *summary value* $E(X_n)$ of the typical population of laboratory $n$. For this reason, an index $i$ for the observations of the $n$th laboratory is not necessary[1]. The associated estimate of the random error $e_n$ is usually in the form of $\mathrm{Var}(X_n)$. The set of equation (1) for the $N$ participant laboratories forms the comparison input set that must be analysed and summarised.

## 2 Hierarchical comparisons

A typical *hierarchical comparison* is when it involves artefact standards [2].

We define a Class 1 [2] (type 2 in [3]) comprising standards that are commonly classified as 'artefacts', meaning that their value is specific of each standard (e.g., for mass standards, the mass value of the specific piece of metal used for a laboratory standard; a specific mixture prepared in a laboratory; a biological sample collected in a laboratory; etc.). Sometimes they can be *constructed* so that their values are made close to each other as much as

* Correspondence: f.pavese@inrim.it

[1] Generally, each laboratory performs a single observation when testing. However, in this case, the laboratory is assumed to work under repeatability conditions and use a standard method, this implying that the single observation pertains to the typical population of observations of the laboratory. In other cases, this value can be the mean of several repeated observations.

Article published by EDP Sciences

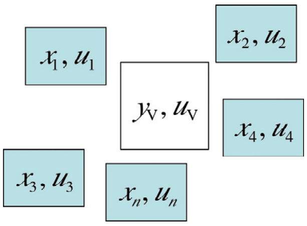

**Fig. 1.** Summarising comparison results for artefact standards: a ‘Virtual standard’ $Y_V$ represents a group of $X_n$ standards based on artefacts: $x$ or $y$ are samples drawn from $X$ or $Y$, respectively, and $u$ are the associated uncertainties.

is technologically or practically possible, or to fit the intended use. However, there are never intrinsic constraints (physical, chemical, biological, ...) for any pair of them to carry identically the same value. When a comparison of these standards is carried out, the measurement on each of them should be considered to concern an independent and *distinct* variable $X_n$.

This case is depicted in Figure 1.

Any summary of the group of standards should be considered as an attribute of the ‘virtual standard’, by consensus representative of the group and of a higher hierarchical status. It is characterised by a value $y_V$ carrying a variability characterised by $u_V$, arising from the variability of the measured values of each and all the standards probability distribution of the chosen summary statistics, plus the additional uncertainty introduced by the comparison. Value $y_V$ is usually not constrained to the value of any of the participating (real) standards $x_n$, though there are some cases where constraints apply, namely when one standard has a hierarchically higher status. The variance $s_V$ of the probability density distributions (pdf) of the virtual standard is computed by combination of the variances $s_n$ of the probability density distributions of each standard, performed using a *convolution* in the linear case, or the *product* of the pdfs in the general case, plus the additional uncertainty introduced by the comparison.

In this case, the summary parameters of the comparison are *hierarchically higher* than the input dataset of equation (1), whence the attribute of ‘hierarchical comparison’. Here, $a$ becomes $a_n$ since the values of the standards for the quantity intended to be measured are intrinsically different for each standard; $a$ is not representing the ‘true value’ of any of the standards – but by chance – but a *hierarchically higher summary* or a conventional value $a_V$.

As to bias, it should be indicated here as ${}^{(a)}b_n$ – where ${}^{(a)}$ means that it refers to artefacts – being ${}^{(a)}b_n = \sum^{(a)} b_{nh}$, with $h = 1, ..., H$, i.e. models the effects of $H$ influence quantities, and is a sample from a random variable $B_n$ modelling the *laboratory component* of bias, typical of each laboratory[2], with, in general, $E(B_n) \neq 0$. Each ${}^{(a)}b_n$ is the sum of two effects:

$$
{}^{(a)}b_n = \Delta a_n + b_n \, n = 1, ..., N, \tag{2}
$$

where $\Delta a_n \approx a_n - a_V(a_V$ being the general mean or a reference value) is different from zero since there is no intrinsic reasons for all artefacts (e.g., pieces of mass) to exactly carry the same value[3]; $b_n$ is modelling the usual occurrence of the fact that the laboratory may not exactly be measuring $\Delta a_n$. The relation between $\Delta a_n$ and $a_V$ is only approximate, since the definition of $a_V$ is not necessarily univocal – e.g., choosing the summary of $a_n$ is not univocal.

It should be noted that it is generally possible to get separate evidence of the values of the two terms on the right in equation (2), i.e. is to have a measure of $b_n$ or of the differences between pairs of laboratories $(b_K - b_H)$, *if and only if* at least one of the standards is hierarchically higher in level (i.e., having an *additional different* meaning of ‘hierarchical’, the more common one) and assigned a stipulated value.

## 3 Non-hierarchical comparisons

In case of *non-hierarchical comparisons*, the summary parameters are generally *not* of a hierarchically higher level than the dataset in equation (1) because the comparison dataset can be considered drawn from a single *super*-population.

We define a Class 2 [2] (type 1 in [3]) comprising standards of different categories or different modalities for performing a comparison:

1. Standards defined in terms of a univocal ‘condition’ (physical, chemical, biological, ...). Quoting from [4], “standards of this type are, for example, the fixed points defined in terms of a thermodynamic state (phase transition) of an ideal substance in thermometry or high-pressure manometry, or of a stipulated physical law or of a physical constant, or a batch of a substance (including reference materials). The value of the state or condition is one and the same for each and all standards established (‘natural value’) and its numerical value is attributed by the written standard according to the state-of-the-art knowledge: the experimental knowledge of this numerical value is unavoidably affected by an uncertainty but, usually, no uncertainty is associated with the numerical value in the written standard (defined or stipulated value). Each standard of this type aims at providing an accurate replication of the experimental conditions that are necessary to attribute the expected value of the performance, the stipulated value of the physical or chemical condition, the same for each and all laboratories”. In a comparison of such standards, there is a *single* random variable, $X$, involved.
2. Comparisons where, (i) the same artefact (or set of artefacts) is circulated and measured by all participant laboratories, or (ii) samples from the same homogeneous artefact (e.g., samples from a reference material)

[2] It would reduce to a fixed value in the case the laboratory is assumed to work under repeatability conditions, as it commonly happens in the field of testing : a fixed value, not a random variable.

[3] Incidentally, the first term of equation (1) can be $a_n(t)$ for unstable standards.

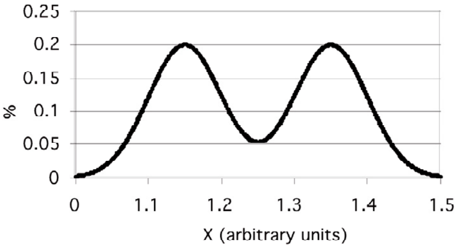

**Fig. 2.** Summarising comparison results for standards from two different laboratories achieving the same 'condition' $X$ (physical, chemical, biological, ...): pooled pdf (in arbitrary units).

are distributed and measured by all participant laboratories. Also in this case, there is a *single* random variable, $X$.

For this Class, Figure 2 shows the case for a pair of laboratories, e.g., each using a different sample from a single batch of a reference material. The standard of each laboratory is intended to achieve the *same 'condition'* according to the state-of-the-art level and the *within*-laboratory knowledge of each laboratory.

The summary of the comparison results is a *single pooled pdf* (i.e., a *sum* of pdfs), since a single variable is involved: in the figure an evident occurrence of non-zero-mean systematic effect is depicted, ruling out the possibility to be in a condition of 'random effects'.

In this case, in the dataset of equation (1) one must drop the index $n$ from $x$, since one deals with a single variable $X$. As to $a$, it now represents the single value of the 'condition' (true value in metrology, conventional value in testing), while $b_n$ is here the *laboratory* bias and footnote 2 can also apply in the indicated cases.

$E(X)$ and $\mathrm{Var}(X)$ have the same meaning as in the hierarchical case, but refer here to the distribution of the *super*-population, called *mixture distribution* [5]. In fact, the pooled distribution of the $N$ populations distributions should be considered and the pdfs *summed* up, being the *super*-population the only random variable under analysis[4].

## 4 Conclusion

The basic differences for the summary of a non-hierarchical comparison, with respect to the hierarchical one, are:

1. The summary parameters are not of a higher hierarchical level than the input data.
2. A *super*-population mixture pdf should be used, which can be multimodal if some of the $B_n$ have $E(B_n)$ sufficiently far apart with each other.
3. A '*random effect*' for the laboratory bias can occur for the chosen set of laboratories, *only* if their random choice is effective, in which case $E({}^{(comp)}B) \equiv 0$, where ${}^{(comp)}B = \{E(B_n)\}$, the distribution of the $E(B_n)$.
4. Should the mixture pdf still be closed to a Normal, the effect of the distinct contributions to $\mathrm{Var}(X)$ of the single laboratories could be analysed for occurrence of '*fixed effects*' by using, e.g., ANOVA.
5. Should instead the mixture pdf not be such or even be multimodal, the value of the parameter characterising the dispersion of the results can be considerably larger than, say, the variance of the mean, because one does not resort to any hierarchical summary; however, it accurately describes the *to-date* available knowledge[5].
6. The location parameter of the mixture pdf represents the best guess one can make of the value of the 'condition' with the available dataset supplied by the participant laboratories. Its choice generally is, as in the case of hierarchical comparisons, subjective to some extent (e.g., see [6,7]).

Incidentally, a different case is the non-hierarchical use of the comparison dataset occurring when only the differences between laboratories are used for the summary (e.g., see [8,9]).

## References

1. F. Pavese, An introduction to data modeling principles in metrology and testing, in *Data Modeling for Metrology and Testing in Measurement Science*. Series: *Modeling and Simulation in Science, Engineering and Technology*, edited by F. Pavese, A.B. Forbes (Birkhäuser, Boston, 2009), Chap. 1, pp. 1–30
2. F. Pavese, A metrologist viewpoint on some statistical issues concerning the comparison of non-repeated measurement data, namely MRA key comparisons, Measurement **39**, 821 (2006)
3. R.N. Kacker, R.U. Datla, A.C. Parr, Statistical analysis of CIPM key comparisons based on the ISO, Metrologia **41**, 340 (2004)
4. F. Pavese, The Definition of the Measurand in Key Comparisons: lessons learnt with thermal standards, Metrologia **44**, 327 (2007)
5. F. Pavese, Metrologia **42**, L10 (2005)
6. P. Ciarlini, M.G. Cox, F. Pavese, G. Regoliosi, The use of a mixture of probability distributions in temperature interlaboratory comparisons, Metrologia **41**, 116 (2004)
7. D.L. Duewer, A comparison of location estimators for interlaboratory data contaminated with value and uncertainty outliers, Accred. Qual. Assur. **13**, 193 (2008)
8. R.J. Douglas, A.G. Steele, Pair-differences chi-squared statistics for Key Comparisons, Metrologia **43**, 89 (2006)
9. A.G. Steele, R.J. Douglas, Establishing confidence from measurement comparisons, Measur. Sci. Technol. **19**, 064003 (2008), doi:10.1088/0957-0233/19/6/064003

[4] The single populations should not be considered distinct random variables, similarly to the familiar case of application of ANOVA.

[5] The fact that this will in most cases imply further work for the laboratories to remove inconsistency, should not be confused with the fact that the *to-date* knowledge is accurately affected by a large variance.