Report on Global Consistency Testing for CCPR-K6.2010

Firstly, regarding outliers – CCPR-G2 identifies an outlier when the "deviation from the KCRV is larger than 3 times of its associated expanded uncertainty with k=2".

Fortunately there are no results in this category and therefore there will be no exclusion of data on this basis.

Secondly, we must consider whether the data are consistent with the model we have used to determine the parameters of interest.

Determining consistency requires first an agreement on the model we are using for the data. There are two distinct models that have been used in CCPR comparisons over the years.

The first model, which I will call Model A, is that each measurement can be written as (assuming a single artefact for simplicity):

$$y_{ir} = T + \varepsilon_{ir} \tag{1}$$

where y_{ir} is the rth measurement result by participant i, T is the true value of the artefact and ε_{ir} is the measurement error. The model is solved for the value of T and degrees of equivalence are the residuals of the fit for each participant. If the data are consistent with the model, then the errors should be distributed normally – this can be tested with the χ^2 test at an agreed upon p-value (usually 0.05) for the known degrees of freedom. For model A,

$$\chi^2 = \sum_{ir} \frac{\varepsilon_{ir}^2}{\sigma_{ir}{}^2} = \sum_{ir} \frac{(y_{ir}-T)^2}{u_{ir}{}^2} \tag{2}$$

where σ_{ir} is the standard deviation of the error distribution and u_{ir} is the standard uncertainty of the difference $y_{ir} - T$ and is an estimate of the unknown σ_{ir}.

The second model, Model B, is that each measurement can be written as

$$y_{ir} = T + \Delta_i + \varepsilon_{ir} \tag{3}$$

where Δ_i is the unknown bias of each participant, the degree of equivalence. The model is solved for both T and Δ_i. Then the consistency test is

$$\chi^2 = \sum_{ir} \frac{\varepsilon_{ir}{}^2}{\sigma_{ir}{}^2} = \sum_{ir} \frac{(y_{ir}-\Delta_i-T)^2}{u_{ir}{}^2} \tag{4}$$

Whether we use Model A or Model B depends on the purpose of the comparison.

- Model A

The purpose of the comparison is to estimate the KCRV, all submitted measurements are estimates of that value. The consistency check determines whether all participants measurements are consistent with the KCRV – if not, uncertainty is added to all participants until consistency is achieved (using Mandel-Paule).

Advantage: The meaningfulness of the KCRV is retained and validated.

Disadvantages: We are masking bias in participants' measurements which is arguably one purpose of the comparison. The meaning of the uncertainty component associated with the Mandel-Paule is unclear – it is 'making up for' differences between scales and could therefore be associated with 'international scale uncertainty' or 'uncertainty in the definition of the KCRV'. Internal inconsistency of participant data is not checked.

- Model B

The purpose of the comparison is to estimate degrees of equivalence as participants' unknown biases using the definition of the KCRV as a tool only. The consistency check identifies whether participant measurements are 'well distributed' around their own degree of equivalence – the χ^2 test identifies where there is a severe internal inconsistency in the results of participants, or where many participants are close to the limits of internal consistency.

Advantages: We retain the meaningfulness of the degrees of equivalence. The Mandel-Paule added uncertainty increases internal consistency of participants' results and can be associated with artefact and/or system stability.

Disadvantage: The definition of the KCRV is taken as defined and not validated by this test.

Attached are the results of the χ^2 tests carried out for all filters at all wavelengths using both models for the data. Note that for model A, the data has been reduced to a single measurement per participant as recommended in CCPR G2 document.

Some points to note from the data:

- As expected, there is a much higher rate of global inconsistency using Model A as we are attempting to make all participants' measurements consistent with the KCRV.
- The amount of uncertainty required to be added by Mandel-Paule to achieve global consistency is much larger for Model A than for Model B
- The number of participants for whom the degree of equivalence is greater than the expanded uncertainty (k=2) is often reduced after global consistency is achieved according to model A, but it doesn't change using Model B (it could change, but not in any of these instances).
- At some comparison points, only Model B shows inconsistency – this occurs at points where one or more labs have submitted internally inconsistent results.

Summary and decision to make:

The choice of model is not inconsequential – it will affect the CMC claims made on the basis of this comparison.

The CCPR G2 document does not explicitly state the model being used for the comparison problem, but the calculation of χ^2 in equation (19) implies the use of Model A. This step was included in the guidelines during the reporting of CCPR K2c in order to provide some guidance during a particularly problematic analysis. However I think that Model B more correctly serves the purpose of the comparison and that the large uncertainties added during Mandel-Paule using Model A hide cases where CMC claims should be adjusted.

Perhaps we have two options:

1. To comply with the guidelines, the results of this comparison will be reported using Model A. The results using Model B can be included in an appendix.
2. Wait until a discussion has been held with the CCPR as to the purpose of comparisons and therefore the method of solution before completion and publication of the comparison results.

Although it does mean some delay, I prefer option 2. This is because K6 is the first comparison in the 'second round' and it is preferable for all comparisons in this round to be carried out the same way. Also, for comparisons linked to this one, it will be preferable that a single set of unambiguous results is presented in the final report.