Investigation of the Effects of Harmonising Diffusion Tube Methodology

Report for Defra and the Devolved Administrations

AEAT/ENV/R/3122
ED48673045
Issue Number 2
Date 05/07/2011

Customer:

Defra and the Devolved Administrations

Customer reference:

RMP2877

Confidentiality, copyright & reproduction:

AEA reference:

ID: AEAT/ENV/R/3122

Ref: ED48673045- Issue Number 2

Contact:

Alison Loader
AEA Technology plc
Gemini Building, Harwell, Didcot, OX11 0QR
t: 0870 190 6518
f: 0870 190 6377
e: alison.loader@aeat.co.uk
AEA is a business name of AEA Technology plc
AEA is certificated to ISO9001 and ISO14001

Author:

Alison Loader, Paul Willis, Jaume Targa

Approved By:

Paul Willis

Date:

5th July 2011

Signed:

Executive summary

This report summarises a study carried out by AEA on behalf of Defra, under the previous Defra contract RMP2877 for Support to Local Authorities, which ran from 1st Dec 2005 to 28th Feb 2011.

In early 2008 a working group operating on behalf of Defra and the Devolved Administrations produced a practical guidance document covering the preparation, use and analysis of Palmes-type diffusion tubes for indicative monitoring of nitrogen dioxide. The intention was that, in the absence (at that time) of a British or European standard method for the technique, the methodology set out in this guidance document would serve as a harmonised method for NO_2 diffusion tubes used in Local Authority Air Quality Management. It was intended that participating laboratories (of which there were around 20 at the time) would implement the harmonised method by 1st Jan 2009, and most did so over the period Feb 2008 – Jan 2009: it was not practical for all of them to make the changes at the same time.

This report describes an investigation of the effects of harmonisation, based upon the results of two ongoing, long-term UK quality assurance/quality control (QA/QC) activities –

- The Workplace Analysis Scheme for Proficiency (WASP), which tests analytical proficiency based on analysis of artificially prepared samples (diffusion tubes spiked with nitrite). This formed the bulk of the investigation: because the WASP scheme is based on analysis of artificial samples, it tests analytical performance only, and is unaffected by factors relating to the preparation or exposure of the diffusion tubes.
- The Field Intercomparison, in which diffusion tubes from each participating laboratory are exposed monthly in triplicate alongside an automatic chemiluminescent NO_X analyser, which provides a reference measurement. As the Field Intercomparison involves outdoor exposure and subsequent analysis of diffusion tubes supplied by the participants, it can provide information on all aspects of diffusion tube processes, including preparation, exposure and analysis.

By looking at the results from both these QA/QC activities, it may be possible to identify whether any changes in laboratory performance relate to analysis, preparation or exposure-related factors.

The main aim of the study was to establish whether harmonisation had improved the agreement of the participating laboratories' results in these activities, i.e. whether inter-laboratory variation has been reduced – both collectively and on an individual basis.

Results from the WASP scheme indicate that inter-laboratory agreement has improved *slightly* over the period investigated (Apr 2007 – Jul 2010), but the improvement appears to be gradual, rather than specifically linked with the timing of the harmonisation.

Performance in each quarterly round of WASP is quantified here in terms of a "Performance Index" (PI) score assigned to each participant. The average quarterly Performance Index (PI) score, for all participants in the WASP scheme who were involved in the harmonisation exercise, did not appear to show any improvement which could be directly attributed to the harmonisation.

Longer-term performance in WASP is quantified here in terms of a "Rolling Performance Index" (RPI), which is calculated as the mean of the best four of the last five rounds, and thus allows the worst result to be excluded. The average Rolling Performance Index has improved over the period investigated, indicating that typical performance is improving. However, there is no specific change associated with the timing of the harmonisation.

It should be noted that the Performance Index based scoring system used in this report is the system that HSL (the operators of the WASP scheme) were planning to implement, at the time this report was written. Since then (and for reasons not related to the work reported here) HSL decided not to adopt this system, instead keeping their existing system, based on the Z-score statistic. Therefore, the scoring system used in this report is not the one now used in the WASP scheme. As this change happened after the present report had been finalised, the report has not been re-written, but re-issued with notes clarifying the change, and explaining how the scores presented here relate to those now used.

Apparently random outliers (usually low rather than high) are still occurring in the WASP data, and the reasons for these should be investigated.

The effect of harmonisation was investigated for each individual participant in WASP, taking into account the actual time at which they completed the changes. Only those which carried out diffusion tube analysis for the purpose of UK Local Air Quality Management were included, as only these were obliged to adopt the harmonised method. There were twenty such laboratories. Harmonisation appeared to ***improve*** the performance of the worst performers in WASP, particularly four laboratories that had previously been performing relatively poorly compared to most others. It is not assumed that the improvement in performance is necessarily due to harmonisation, as other changes such as staff training, or upgrading laboratory equipment, may also be involved. However, an improvement was observed for these participants.

Five laboratories have shown ***worse*** performance in WASP since harmonisation. All were generally performing well before harmonisation, and it is important to note that subsequently their scores remained mostly in the “good” and “acceptable” bands (**according to the Performance Index scoring system**). However, they were typically *slightly* worse. In four of these five cases, it appeared that negative bias, i.e. under-estimation of the spiking level on the WASP sample, had been introduced. Two reported concerns regarding extraction: this should be investigated.

The remainder - four laboratories that were performing relatively well prior to harmonisation, and seven whose performance had been variable - showed no clear change. The performance of the seven “variable” performers remained variable after harmonisation.

Both “good” and “variable” laboratories continued to have occasional “poor” rounds, i.e. PI scores in the “unacceptable” range. These were usually due to substantial under-read of one or more of the samples, and the reasons should be investigated.

Analysis of data from the Field Intercomparison, 2006 to Jun 2010, showed that the average bias of the tubes exposed in this activity has been increasing year-on-year since 2006. The tubes in this study now typically show positive bias (i.e. over-read). This is consistent with the known sources of interference affecting diffusion tubes exposed in ambient air. This increase in bias is thought to be related to factors other than analysis, since no corresponding increase has been observed in the WASP results (which are based on artificially spiked samples so would not be expected to show exposure-related positive bias).

The precision of the triplicate tubes in the Field Intercomparison has improved by around 1% (from 6% to 5%), between 2006 and 2009. As this improvement is not seen in the WASP data, it is not thought to be analysis-related, but linked to some other factor - possibly improved consistency in the preparation of diffusion tubes.

As of July 2010, the Field Intercomparison had not shown any improvement in inter-laboratory agreement.

Data from the Field Intercomparison appear to show an association between large negative bias, and poor precision. This may be due to the presence of low outliers in some datasets, and the reasons for these should be investigated.

Data from Air Quality Consultants' database of combined bias adjustment factors (based on collation of UK Local Authority co-location studies) was also examined (although it should be noted that these co-location studies are not necessarily carried out under consistent conditions). This did not show any reduction in the spread of combined bias adjustment factors between 2007 and 2009. Data from this database show a similar association between large negative bias, and poor precision, as also seen in the Field Intercomparison data.

As of July 2010, the harmonisation of diffusion tubes methods within the group of laboratories supplying and analysing NO_2 diffusion tubes for LAQM had not delivered any clear improvement in inter-laboratory agreement. However, it will hopefully make the adoption of the forthcoming CEN standard method (expected later in 2011 or 2012) more straightforward.

The recommendations are as follows:

- Further investigation of the reasons why four laboratories' WASP results began to show negative bias after harmonisation. One aspect of diffusion tube analysis which may not have been satisfactorily optimised in all cases is extraction, and this should be investigated further.
- The reasons for the observed occasional (and apparently randomly occurring) low outliers in WASP and the Field Intercomparison should be investigated further.
- QA/QC of diffusion tubes used in Local Air Quality Management should continue.
- Laboratory performance should continue to be monitored, and advice provided to laboratories where appropriate.

Table of contents

1 Introduction

1.1 Background

This report summarises a study carried out by AEA on behalf of Defra, under the previous Defra contract RMP2877 for Support to Local Authorities, which ran from 1st Dec 2005 to 28th Feb 2011.

Palmes-type diffusion tubes are widely used by UK Local Authorities for indicative monitoring of nitrogen dioxide (NO_2), and around 20 UK laboratories currently supply and analyse NO_2 diffusion tubes for this purpose. However, historically there was no British or European standard method for diffusion tube preparation and analysis, and considerable variation existed in the methods used by the various laboratories (particularly in analysis). Diffusion tube performance, as measured in various QA/QC exercises, had also been observed to vary between laboratories, and it was suspected that the differences in laboratory procedures were a contributing factor.

Therefore, Defra and the Devolved Administrations commissioned AEA and Air Quality Consultants to set up and manage a Working Group on harmonisation of NO_2 diffusion tube preparation and analysis methods. This work was undertaken during 2006 and 2007 as part of the former Defra contract RMP 2877.

In February 2008, the Defra Working Group produced a Practical Guidance document[1]. aimed at both laboratories supplying and analysing diffusion tubes, and end users - in particular, Local Authorities using diffusion tubes for Local Air Quality Management purposes. This Guidance set out a harmonised method for preparation and analysis of diffusion tubes. It was Defra's intention that all suppliers and analysts of diffusion tubes used for Local Air Quality Management (LAQM) in the UK should implement the harmonised method by 1st January 2009.

However, it was not practicable for all laboratories to change their procedures at exactly the same time: many are UKAS-accredited and therefore needed to time the changes to be completed before their scheduled annual assessment visit. (They would otherwise have required an extra assessment visit, which for many would have prohibitively expensive). Therefore, the laboratories were given a "window" of 11 months – between Feb 2008 when the guidance document was released, and the deadline of 1st Jan 2009 - during which to implement the changes. Most completed their harmonisation by this deadline, although for a small number the changes were not completed until early 2009. It should therefore be noted that the laboratories did not harmonise their procedures simultaneously, but over a period of time.

Shortly before the publication of the Defra harmonised method, it was announced that a CEN (European) Standard Method would be developed for NO_2 diffusive samplers. However, this would take some time and was unlikely to be published before 2011. Several members of the Defra Working Group also participated in the CEN Working Group for this standard, and were able to bring the UK's input and experience to the process. It is expected that the new CEN standard will be published in late 2011 or early 2012. The 2008 Defra Practical Guidance document may need updating to take account of the CEN standard: while this may mean some changes to the harmonised UK method, its implementation is expected to be easier because the UK's diffusion tube laboratories have already adopted consistent approaches to key aspects.

1.2 Objectives

This report investigates whether the harmonisation of preparation and analysis methods has reduced inter-laboratory variation in performance of NO_2 diffusion tubes used in the UK.

If the harmonisation has been effective in reducing inter-laboratory variation in diffusion tube performance, it might be possible to observe this effect in the results of two ongoing quality assurance and quality control (QA/QC)schemes, in which the majority of UK laboratories that supply and analyse diffusion tubes participate.

These are as follows:

(i) the Workplace Analysis Scheme for Proficiency (WASP), an independent analytical performance testing scheme, operated by the Health and Safety Laboratory (HSL), which uses artificially spiked diffusion tubes to test each participating laboratory's analytical performance on a quarterly basis.
(ii) The NO_2 Field Intercomparison, a monthly field intercomparison exercise operated on behalf of Defra and the Devolved Administrations, in which diffusion tubes from each laboratory are exposed in triplicate, co-located with an automatic chemiluminescent NOx analyser (which is defined within Europe as the reference method for NO_2).

These two QA/QC activities complement each other: the WASP scheme tests analytical performance only, while the Field Intercomparison also investigates the performance of the tubes themselves under actual exposure conditions. These two QA/QC schemes together provide a good opportunity to investigate changes in performance of the participating laboratories, over time.

It might also be possible to see changes in the national database of combined bias adjustment factors (currently maintained by the National Physical Laboratory, developed and formerly maintained by Air Quality Consultants Ltd. This is available online at http://laqm.defra.gov.uk/bias-adjustment-factors/national-bias.html). If inter-laboratory variation has indeed decreased, the range of the bias adjustment factors reported for the various laboratories might also have decreased. However, as these studies are carried out at different Local Authority monitoring sites throughout the UK, and not under controlled conditions, any effects of harmonisation are likely to be more difficult to identify.

The objectives of this study were therefore:

(i) To investigate whether the results of the WASP scheme or the Field Intercomparison showed improved agreement between laboratories (i.e. a reduction in inter-laboratory variation) since harmonisation. (This comprised the largest part of the investigation.)
(ii) To investigate whether the results of the WASP scheme or the Field Intercomparison showed any evidence of changes in diffusion tube performance since harmonisation.
(iii) To briefly investigate whether the range of combined bias adjustment factors reported in Air Quality Consultants' on-line database has decreased since harmonisation.

The main focus of this investigation was on analytical performance, as prior to harmonisation, there had been a great deal of variation in analytical procedures. Therefore, the main part of the investigation concentrated on the WASP data. Being based on analysis of identical artificially spiked samples, the WASP scheme offered the best opportunity to observe any changes in analytical performance resulting from harmonisation.

2 Methodology

This investigation has used the results of the independent Workplace Analysis Scheme for Proficiency (WASP), and the Defra NO_2 Diffusion Tube Field Intercomparison, to assess the effectiveness of the introduction of harmonised procedures in improving inter-laboratory variation in diffusion tube analysis. Investigation of the WASP dataset comprised the largest part of the investigation. This section first describes the WASP scheme, and how the data were used. It then gives a brief overview of the Field Intercomparison, and summarises how the data from this activity were also used. Finally, it summarises the data available in the national database of bias adjustment factors and how these have been used here.

2.1 About WASP

The Workplace Analysis Scheme for Proficiency (WASP) is an independent analytical performance testing scheme, operated by the Health and Safety Laboratory (HSL). This performance testing scheme uses artificially spiked diffusion tubes to test each participating laboratory's analytical performance on a quarterly basis.

Every quarter, (in January, April, July and October each year) each laboratory (of which there are currently approximately 20) receives four diffusion tubes "spiked" with an amount of nitrite known to HSL but not the participants. The tubes are all prepared by one single independent supplier. At least two of the tubes are usually duplicates, which enables precision, as well as accuracy, to be assessed. The mass of nitrite on the spiked tubes is different each quarter, and reflects the range encountered in actual ambient monitoring. The participants analyse the tubes, and report the results to HSL. HSL assign a performance score to each laboratory's result, based on their deviation from the known mass of nitrite in the analyte (the "spike value"). The "spike value" used as a reference value is calculated based upon gravimetric and volumetric considerations.

The spiking repeatability (expressed as a coefficient of variation) is included in the results summaries provided to all participants by HSL: over the rounds covered by this report (rounds 96-110) the spiking precision ranged from 0.13% to 1.5% with a mean of 0.28%.

The above requirements have been in place since April 2007 (WASP round 97). Prior to this, a different system of monthly rounds involving single samples was used. Because of this major change in the WASP regime, only WASP rounds from April 2007 onwards are considered here.

Results are communicated to each participating laboratory by HSL, and during the period covered by the former Defra contract RMP2877 (Support to Local Authorities for Air Quality Monitoring) they were also communicated to AEA for purposes of work carried out under this contract.

2.1.1 WASP Performance Index

Note: ***At the time Issue 1 of this report was prepared, HSL were planning to move away from a scoring system based upon the Z-score, to one based on the Performance Index Statistic.***

Details of this had been communicated to participants, and HSL intended to move to the new scoring system as of Round 113, in April 2011. The new and old scoring systems had been used alongside each other in WASP reports for several rounds. Accordingly, the present report was based upon the proposed new scoring system.

However, HSL subsequently (and for reasons not connected with the work reported here) decided not to adopt the Performance Index scoring system, instead staying with their existing system, based on the Z-score statistic. Therefore, the Performance Index scoring system used here is <u>not</u> the one now used in the WASP scheme. This change happened after the present report had been finalised, so, rather than re-write it, the report has been re-issued with notes clarifying this change, and explaining how the scores presented here relate to those now used.

The Performance Index statistic is calculated from the four sample results in each round, in Equation 1:

$$Performance\ Index = \frac{\sum_{s=1}^{4}\left(\frac{x_s}{\bar{x}} - 1\right)^2 \times 10{,}000}{4} \qquad \text{Equation 1}$$

- where x_s is the result obtained by the laboratory for sample number s (of four), and $\bar{x}$ is the spike value for sample 's'. (The ratio $x_s / \bar{x}$ is the "standardised result", i.e. the result obtained by the participant, divided by the "spike value".) (The multiplication factor of 10,000 is arbitrary, to avoid having to deal with very small numbers).

The spike value is cross-checked by analysis (using ion chromatography) of 12 samples.

2.1.2 Rolling Performance Index

The Rolling Performance Index (RPI) allows long-term trends in performance to be monitored. It is calculated as the arithmetic mean of the best four Performance Index values from the most recent five rounds. (If a participant has participated in less than four of the last five rounds, it is not possible to calculate an RPI).

2.1.3 Performance Criteria

At the time Issue 1 of this report was written, HSL were planning to adopt the following performance criteria, based on the RPI:

- GOOD: Results obtained by the participating laboratory are on average within 7.5% of the spike value. This equates to an RPI of 56.25 or less.
- ACCEPTABLE: Results obtained by the participating laboratory are on average within 15% of the spike value. This equates to an RPI of 225 or less.
- UNACCEPTABLE: Results obtained by the participating laboratory differ by more than 15% of the spike value. This equates to an RPI of greater than 225.

These are the performance criteria used in the present report. However, as explained above, the scoring system that HSL are now using is based on the Z-score, calculated as in Equation 2:

$$Z = \frac{\left(x_s - \bar{x}\right)^2}{\sigma_{ref}} \qquad \text{Equation 2}$$

- where x_s is the result obtained by the laboratory for the sample, $\bar{x}$ is the theoretical spike value calculated from sample preparation parameters, and σ_{ref} is the reference standard deviation, fixed at 7.5% of the spike value. On this basis, the category "Good" above equates to a Z-score within ±1, "Acceptable" above equates to a Z-score within ±2, and "Unacceptable" equates to a Z-score outside ±2.

However, the performance categories used in the new WASP scoring system are as follows:

- Satisfactory: Z-score within ±2.
- Unsatisfactory: Z-score > 2 or < -2.

Prior to round 105, (and therefore during the harmonisation period), HSL used a system of performance scores based on Z-scores, but the criteria were more lenient in that the reference standard deviation was set at 13% of the spike value, rather than 7.5% as is now used. It should therefore be noted that some results in rounds up to 105, which were classified at the time as "good" or "acceptable" would now be given lower classifications under the current criteria.

Details of current performance criteria are given in the May 2011 edition of the WASP participants' handbook, available at http://www.hsl.gov.uk/media/111783/12th%20wasp%20participant%20handbook%20may%202011.pdf .

2.1.4 Laboratory Identification Code Numbers

The WASP scheme is covered by certain confidentiality provisions, and in the results summaries provided to the participants by HSL, the laboratories' results are identified only by identification numbers, not actual names. This allows the participant to see their own results, and to compare their performance with the other participants, without knowing the other participants' identities. Therefore, the same identification numbers are used here, rather than laboratory names. It is intended that the identities of the participants remain anonymous.

2.1.5 How the WASP Results Have Been Used

Harmonisation has required all the laboratories to adopt essentially the same procedures for diffusion tube analysis. The WASP scheme specifically tests laboratories' *analytical* performance. As WASP uses artificially spiked tubes, prepared by a single supplier, the results are not influenced by exposure-related factors, or the tube preparation process. Examination of the participants' WASP results may therefore reveal whether harmonisation has improved agreement in analysis of identical diffusion tubes.

This study investigated:

(i) Whether overall performance in WASP (based on the Performance Index and Rolling Performance Index) has changed over time, since the current WASP methodology was adopted in April 2007, and whether there are any observable differences particularly over the years 2008 -2009 when the harmonisation was implemented.
(ii) Whether the precision of WASP results (based on the duplicate samples) has changed – although it should be noted that harmonisation was not necessarily expected to improve precision.
(iii) Whether the range (or spread) or the WASP results has been reduced by the harmonisation, as might be expected if inter-laboratory agreement had been improved.

The study also investigated how individual laboratories' performance in WASP has changed, from the time at which they completed harmonisation. This is important, as it was not

practicable for all laboratories to make the changes required for harmonisation at the same time.

This investigation is based only on the results of those participating laboratories that analyse diffusion tubes for the purposes of Local Air Quality Management (LAQM) in the UK. There are a small number of other WASP participants (for example some overseas businesses) who are not involved with LAQM. Their WASP results are not included in this investigation as they have not been obliged to harmonise their methodology in any way.

2.2 About the Field Intercomparison

The NO_2 Field Intercomparison is operated on behalf of Defra and the Devolved Administrations. It consists of a monthly field intercomparison exercise, in which three diffusion tubes per month, from each laboratory, are exposed simultaneously, co-located with an automatic chemiluminescent NOx analyser (which is defined within Europe as the reference method for NO_2).

The objective is to estimate diffusion tube bias (accuracy relative to the reference chemiluminescent method) and precision, under normal field operating conditions, for diffusion tubes from each participating laboratory.

The Intercomparison is intended to supplement the WASP scheme by providing information on how the diffusion tubes themselves perform under actual exposure conditions.

It takes place at the London Marylebone Road automatic air quality monitoring site. This site is used because it fulfils Defra's specification to use a roadside or kerbside site with public exposure, and is one of the few such sites at which large numbers of diffusion tubes can be exposed securely, out of reach of the public. This kerbside site, on a busy city centre road, is part of Defra's Automatic Urban and Rural Network (AURN). This site is shown in Figure 2-1. shows an example of some tubes exposed in this intercomparison, and how the tubes are fixed in place. In line with normal practice in the UK, the tubes are not exposed in any kind of wind shelter.

Figure 2-1 London Marylebone Road Air Quality Monitoring Site

Each laboratory sends three tubes per month, plus a travel blank, for exposure at the automatic monitoring site. The tubes are exposed for a set period approximating to a calendar month (usually 4 or 5 weeks), before being returned (with the travel blank) to the supplying laboratories for analysis. The travel blanks are isolated in sealed sample bags, and stored in a cool place throughout the exposure period. The participating laboratories then analyse the exposed tubes and return their results for collation, and for comparison with the reference NO_2 concentration from the automatic chemiluminescent analyser.

Each participating laboratory is sent an annual summary of their performance in the last calendar year, including precision and accuracy. The spreadsheet is sent to the laboratory concerned and, with their permission, included in Air Quality Consultants' database of co-location studies.

Figure 2-2 Tube Exposure in Field Intercomparison

2.2.1 How the Field Intercomparison Results Have Been Used

Prior to harmonisation, results from the Field Intercomparison showed considerable inter-laboratory variation. This study has investigated:

(i) Whether there has been any change in the average accuracy (or bias) of diffusion tubes in the ongoing Field Intercomparison, relative to the automatic analyser (the reference method).
(ii) Whether there has been any change in the average precision of diffusion tubes in the ongoing Field Intercomparison (although there was no expectation of any such change, as a result of harmonisation).
(iii) Whether there has been any reduction in inter-laboratory variation (i.e. whether the agreement between the various participants' results has improved), during the period before, during and after harmonisation.

2.3 Database of Local Authority Bias Adjustment Factors

Many Local Authorities carry out co-location studies in order to evaluate the precision and accuracy of the diffusion tubes they use, and to calculate a *bias adjustment factor* which can be used to adjust the annual mean results from other diffusion tube sites. The bias adjustment factor is calculated by dividing the diffusion tube result by the automatic analyser result, on an annual mean basis: it is the reciprocal of the standardised result.

A national database of the results of Local Authority co-location studies is compiled and updated by Air Quality Consultants, and is available via the internet, and can be downloaded from http://laqm.defra.gov.uk/bias-adjustment-factors/national-bias.html .

The database shows the bias adjustment factors from each individual co-location study reported to Air Quality Consultants for the purpose, with details of the analytical laboratory, tube preparation technique, and site where the co-location study was carried out.

The database also includes combined bias adjustment factors for each laboratory and preparation technique combination. These are calculated from the results of the individual co-location studies, by an orthogonal regression technique explained at http://laqm.defra.gov.uk/documents/NO2-Diffusion-Tube-Collocation-Methodology.pdf . These combined bias adjustment factors can then be used by Local Authorities who have not carried out their own co-location study, to adjust measured annual mean NO_2 concentrations for diffusion tube "bias".

The data held in this database were used to investigate whether the spread of combined bias adjustment factors had changed, between 2006 (two years before the Practical Guidance was released) and 2009 (the most recent year for which data were available).

3 Results and Discussion

3.1 Changes in WASP Performance Index – All Participants

If the implementation of the harmonised method resulted in general improvement in analytical performance, the average Performance Index (PI) score of all participants in WASP might be expected to decrease – i.e. improve. As the laboratories did not complete their harmonisation simultaneously, a clear "step change" in performance at a particular time was not anticipated: but it may be possible to see a change over the relevant period.

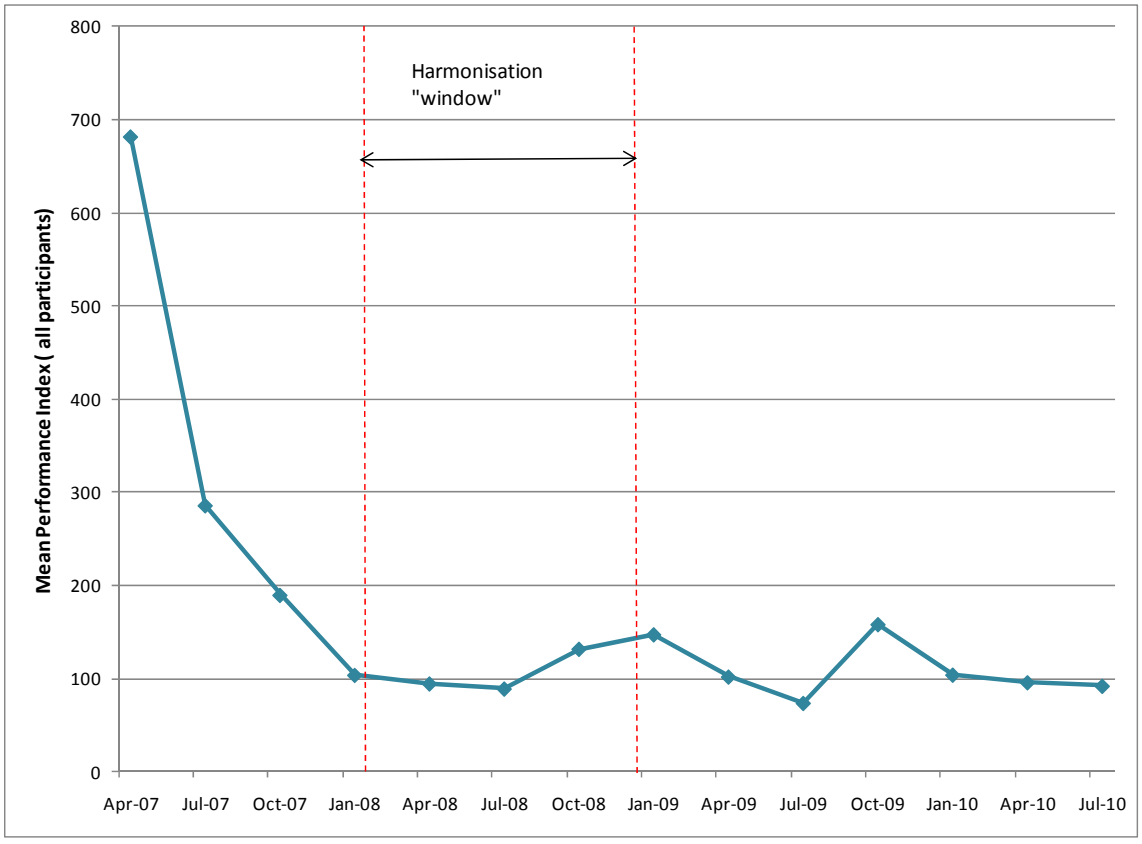
Figure 3.1 shows how the arithmetic mean of the PI score has varied over time, since the move to the current WASP procedures in April 2007. The most notable feature was the sharp decrease in the early rounds (Apr 2007 – Jan 2008). This possibly reflects the participants getting accustomed to the new WASP procedures.

Following this sharp decrease, there was no clear trend in the mean PI. The mean PI therefore did not appear to show any changes clearly attributable to harmonisation.

There was some fluctuation: the mean PI was higher than usual around Jan 2009 (the deadline for completion of the harmonisation). However, there was a similar peak in the Oct 2009 round, which was not a time at which any changes occurred.

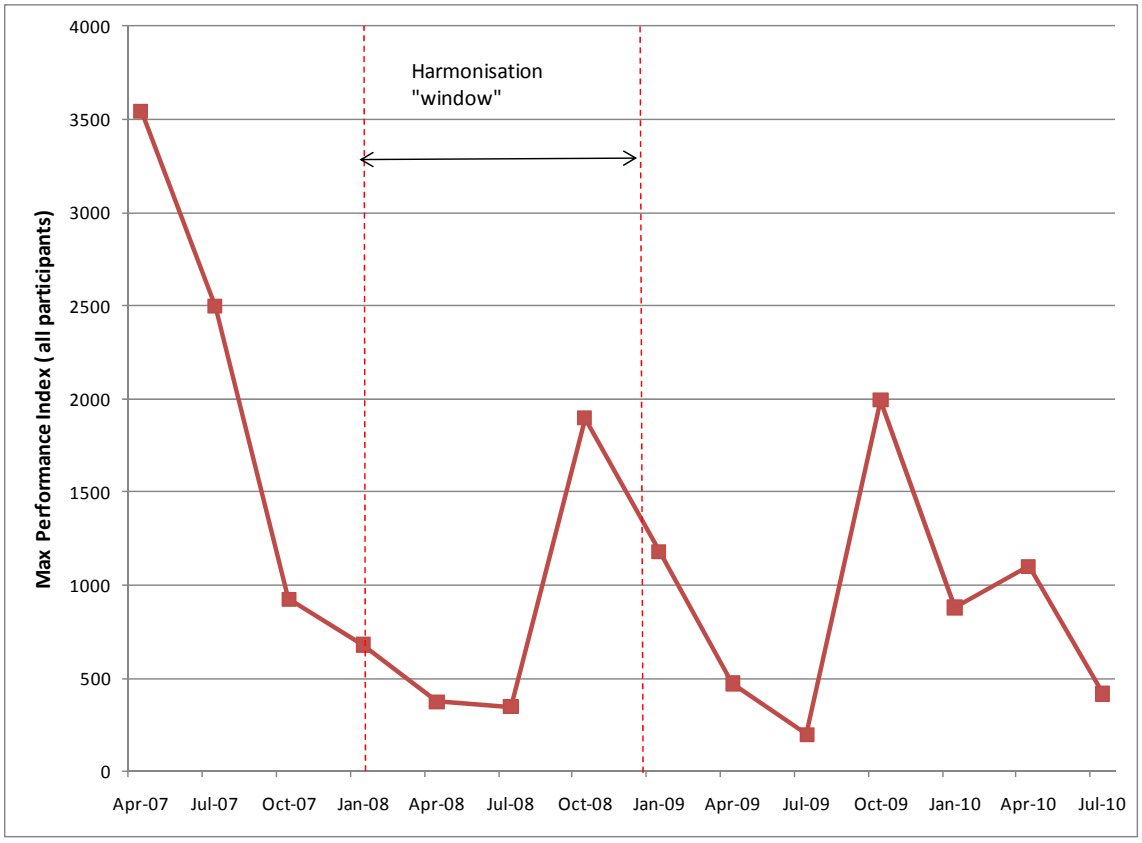
In fact, the arithmetic mean PI for any given round can be increased substantially by one or two extremely high (poor) scores, and the small peaks in Jan 09 and Oct 09 were caused by this.

Figure 3-1 Mean Performance Index, All LAQM Participants



The maximum PI in each round is shown in Figure 3-2. In both Jan 2009 and Oct 2009, one very high (poor) PI score in the region of 2000 occurred (these were obtained by a different participant in each case), and it is these that caused the fluctuations in the mean. Occasional poor results have continued to occur post-harmonisation and in recent rounds. These poor results appeared to be randomly distributed among the participants – it was not always the same laboratories that obtained poor results.

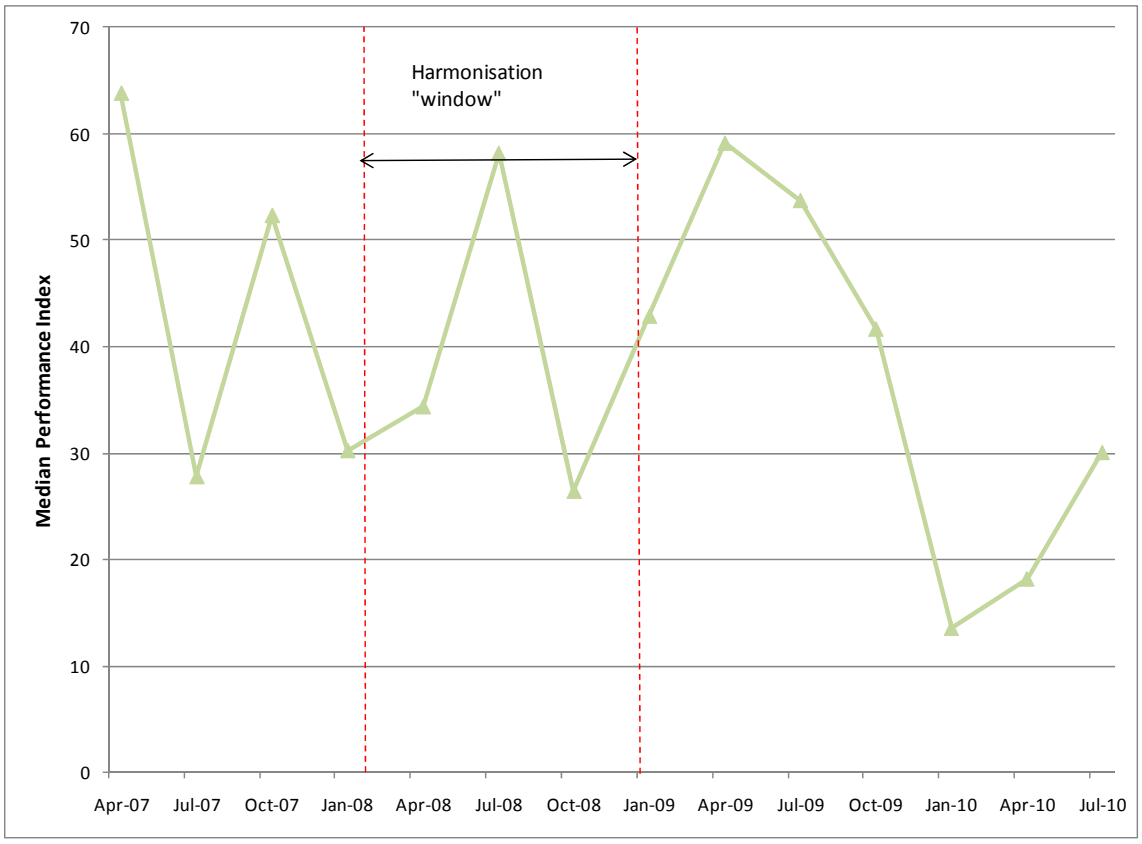
Figure 3-2 Maximum PI in Each Round



Because the arithmetic mean PI can be distorted by one or two very poor (high) scores, the median (50th percentile) PI may be a better indicator of any trends or changes, and this statistic is shown in

Figure ***3-3***. The median did not appear to show any consistent pattern or improvement associated with harmonisation. However, the median PI for the most recent three rounds (in 2010) has been relatively low (i.e. good): it remains to be seen whether this will continue.

Figure 3-3 Median Performance Index (All LAQM Participants)



As explained in section 2.1.2 above, in the present report, long-term performance in WASP is assessed on the basis of the Rolling Performance Index (RPI) – that is, the arithmetic mean of the best four of the last five rounds.

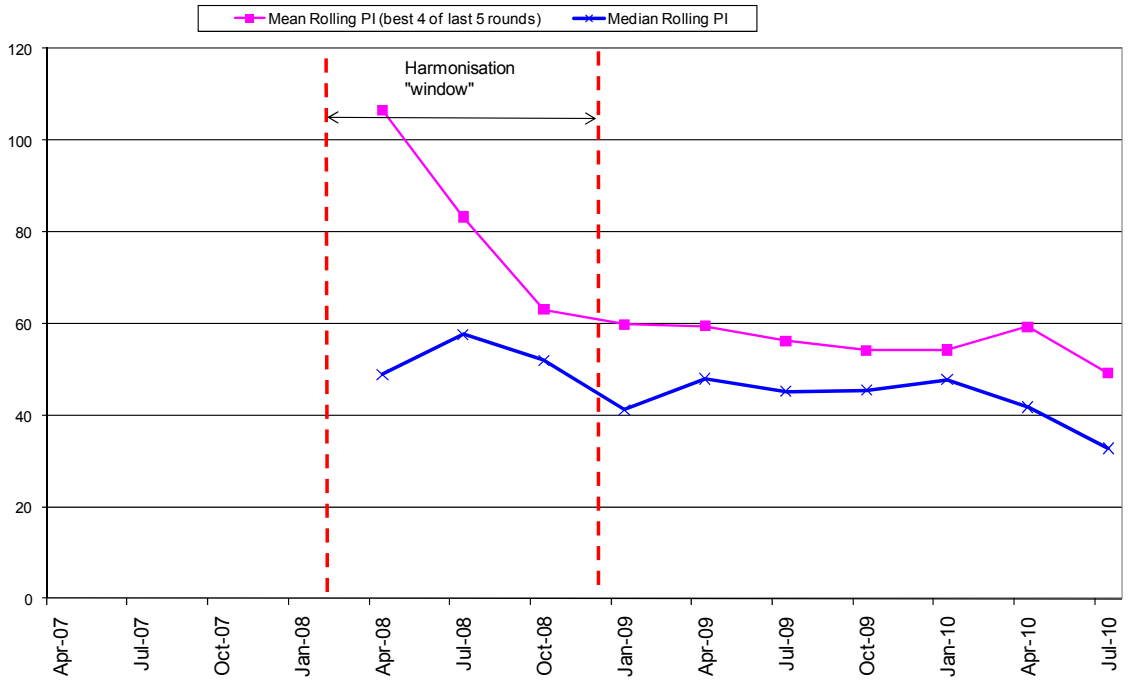
Figure 3-4 shows the mean Rolling Performance Index (RPI). This parameter shows a substantial improvement through the harmonisation “window” period. However, because the RPI is based on the *preceding* five rounds, the improvement during 2008 cannot be attributed to the harmonisation, but simply reflects the improvement that occurred in the first few rounds of the new WASP regime during 2007, as highlighted above.

Following the harmonisation deadline in January 2009, the mean RPI continued to decrease slightly in most subsequent rounds (the exception being Apr 2010), indicating gradual steady improvement.

Figure 3-4 also shows the median RPI: this also showed a general improvement since early 2008, although again there is no clear link with harmonisation.

It therefore appears that performance in WASP (as shown by the Rolling Performance Indicator) has generally showed a small but steady improvement over time, despite apparently random poor results still occurring on a regular basis.

Figure 3-4 Mean and Median Rolling Performance Index, All Participants



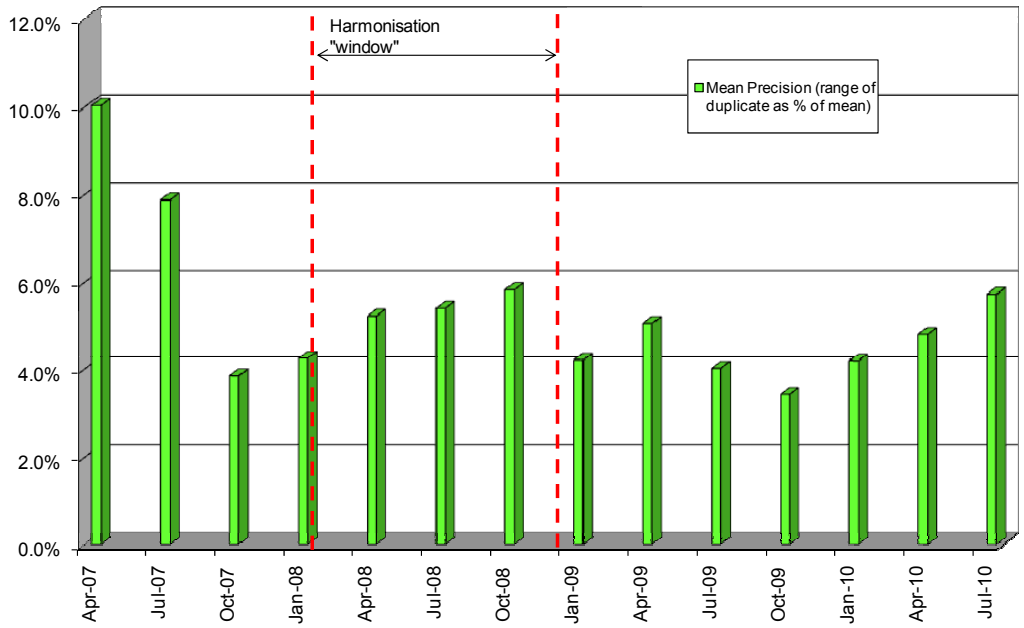
3.2 Changes in Precision of WASP Results

Whilst there was no expectation that harmonisation would necessarily improve diffusion tube precision, this has been investigated.

The four WASP samples analysed by each participant in each round comprise two pairs of duplicates. This allows an estimate of precision to be made.

Where the number of replicate measurements is three or more, diffusion tube precision is usually expressed in terms of the relative standard deviation (i.e. the standard deviation of the replicate measurements, expressed as a percentage of the mean) However, because the number of replicates in each case is only two, the standard deviation (and therefore the relative standard deviation) is of little use. Instead, the *difference* between the pair of replicate samples has been expressed as a percentage of the mean, and this value used as an *estimate* of precision. This value was calculated for each pair of duplicate tubes. The arithmetic mean of the two estimated precisions for the two pairs of duplicates has then been calculated, and this mean then used as an indicator of that participant's precision in that round. A time series of mean precision, averaged over all participants, is shown in Figure 3-5: there appeared to be no clear trends, so no evidence of any improvement in precision as a result of harmonisation.

Figure 3-5 Mean Precision of Duplicate Samples (All Participants)



3.3 Changes in Spread of WASP Results

If harmonisation has been effective in reducing inter-laboratory variation, it might be possible to identify a decrease in the scatter or spread of the results reported in WASP, after harmonisation. This might be identifiable in, for example –

- the interval between the maximum and minimum results
- the standard deviation
- the interval between a high and low percentile, for example the 90th and 10th.

Because different rounds use different spike levels, it is necessary to base this investigation on the mean *standardised* result. The standardised result for each sample is the ratio of the reported result (in microgrammes of nitrite) to the known spike value for that sample. The standardised results for the four samples in each round have been averaged, giving a mean standardised result for each participant in each round.

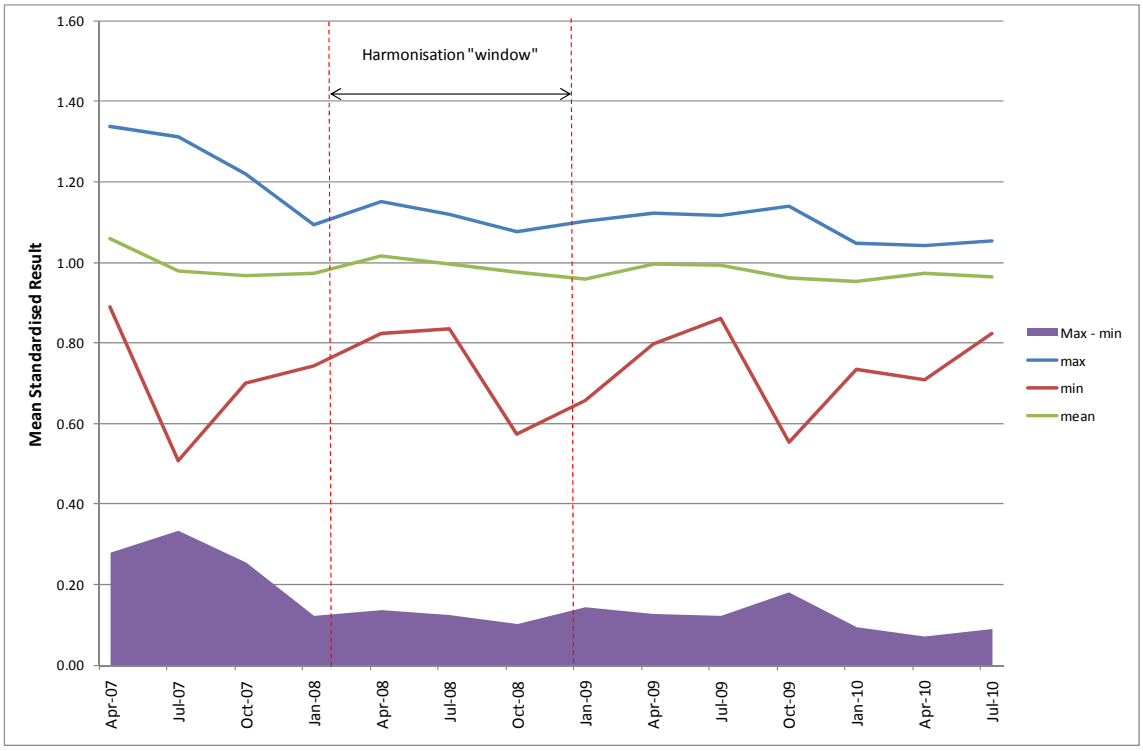
Figure 3-6 shows the maximum, minimum and arithmetic mean values of all participants' mean standardised results. Also shown is the difference between the maximum and minimum (as the shaded area).

The difference between the maximum and minimum standardised result does appear to show a slight downward trend since early 2008, indicating some decrease in inter-laboratory variation has decreased. However, the timing of this does not appear to be linked to the harmonisation.

The minimum standardised result varied much more than the maximum, indicating that occasional low outliers were more prevalent than high outliers.

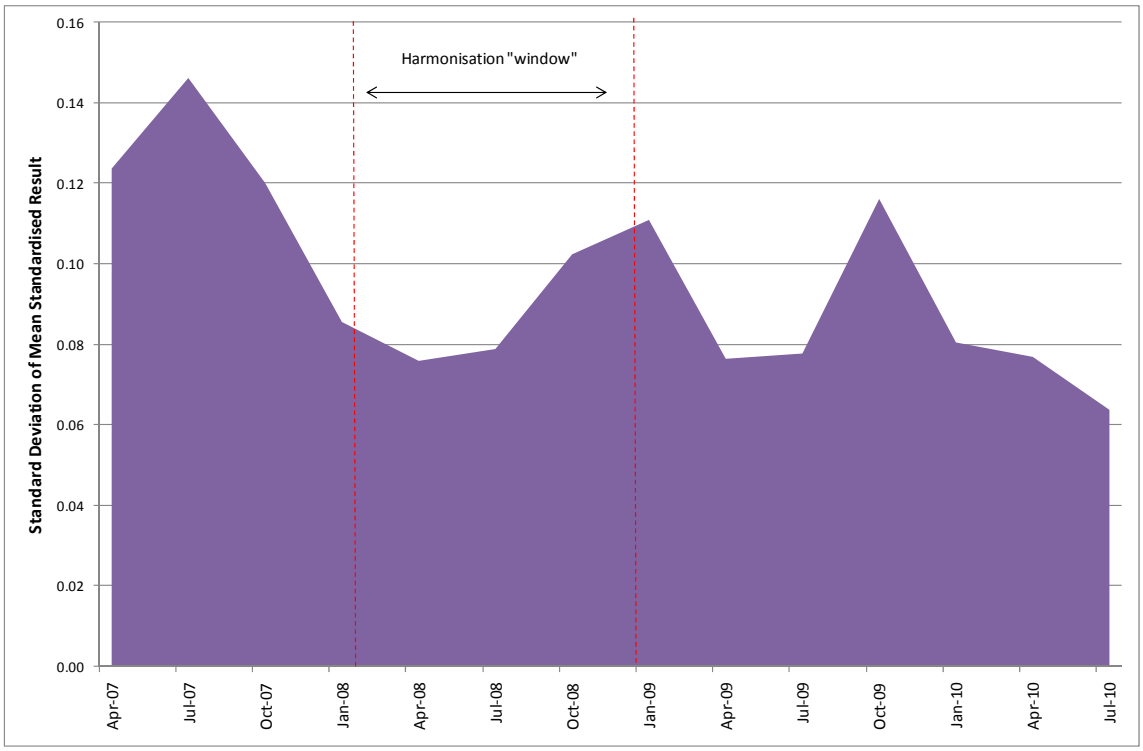
It also appears that the mean standardised result has decreased slightly since the early rounds, and is now typically slightly under 1.00 – but again there appears to be no clear link with the timing of the harmonisation.

Figure 3-6 Maximum, minimum and arithmetic mean values of all participants' mean standardised results, and Max-Min Difference



Another useful parameter in this context is the standard deviation. Figure 3-7 shows how the standard deviation of the mean standardised result has changed over time.

Figure 3-7 Standard Deviation of Mean Standardised Result



Although the standard deviation has reduced over time, again there is no clear link with the timing of the harmonisation. The standard deviation has continued to vary during and after the harmonisation period. In particular, the SD was unusually high in Oct 2008, Jan 2009 and Oct 2009: in each of these months, just one laboratory obtained outlying results – in this case, much lower than the other participants.

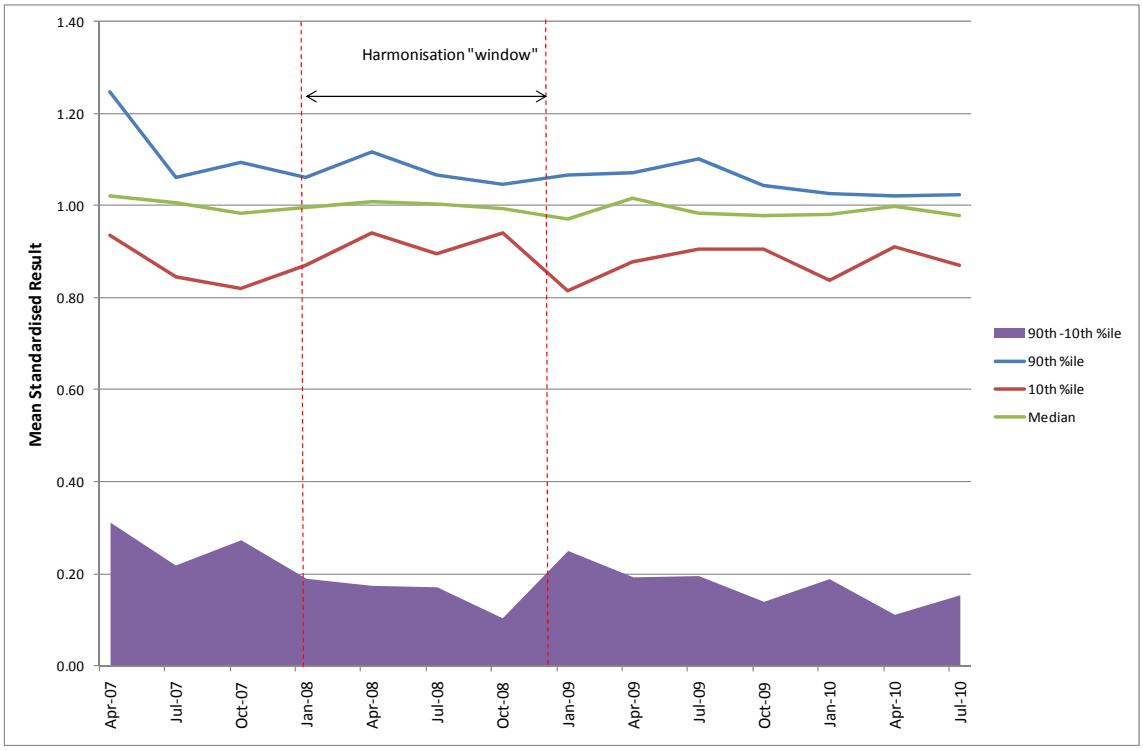
The 90th percentile and 10th percentile give an indication of the spread of the results, without the distorting effect of extreme outliers. Figure 3-8 shows the 90th percentile, 10th percentile, and 50th percentile (median) of the participants' mean standardised results in each round. Also shown is the difference between the 90th and 10th percentiles (the shaded area).

This graph also appears to show that some aspects of WASP performance have converged: the difference between the 90th percentile and the median (shown by the shaded area) has clearly decreased, indicating a reduction in high outliers.

The difference between the 90th and 10th percentile (which indicates the spread of the majority of the results, excluding outliers) decreased from Apr 2007 to Oct 2008, but this was followed by an increase around the harmonisation deadline of Jan 2009. It is possible that this indicates some initial disruption due to the changes. However, it then continued to decrease.

The difference between the median and 90th percentile has clearly decreased since harmonisation, indicating that high outliers are now less common. However, again there is no clear link with the timing of the harmonisation.

Figure 3-8 90th, 50th and 10th percentile of mean standardised result. Also difference between 90th and 10th percentile (shaded area) .



3.4 Summary of Findings from WASP

(i) On average, performance in WASP (as shown by the mean and median Rolling Performance Indicator) has improved over the past three years, but it is not possible to link this with the timing of the harmonisation.

(ii) Apparently random poor results – usually low outliers - still occur on a regular basis.

(iii) The WASP results do not show any evidence of improved precision (although this was not a specific aim of harmonisation).

(iv) Inter-laboratory variation of WASP results has reduced slightly, on the basis of the following statistics –
- Difference between maximum and minimum standardised result
- Standard deviation of all participants' standardised results
- Difference between 10th and 90th percentile

but the improvement appears to have been gradual, rather than a step-change that could be clearly linked with the timing of the harmonisation.

(v) The difference between the median and 90th percentile has clearly decreased, indicating that high outliers are now less common.

3.5 Changes in Performance of Individual Participants

In 2006, before the beginning of the harmonisation study, laboratories were asked to complete a questionnaire on their methods for tube preparation and analysis at the time. The information was treated as confidential, but confirmed that the laboratories used a range of different procedures for extraction and analysis of their diffusion tubes. Some of the variations included the following:

- reagents were sometimes added separately and sometimes pre-mixed,
- different volumes of reagent were used,
- the order in which the reagents were added varied,
- different methods of agitation were used - manual shaking, vortex mixers and vibrating trays – and the amount of agitation varied,
- some laboratories extracted using water only, before adding reagents,
- tubes were left to stand for different periods and at different stages between and after addition of reagents.

Procedures used to calibrate analytical instrumentation (colorimeters) also varied.

Therefore, the harmonisation process required each laboratory to make a different set of changes. This section investigates how WASP results have changed for each individual participant. In some cases, the information available regarding procedures before and after harmonisation has been used to assess the value of the specific changes made.

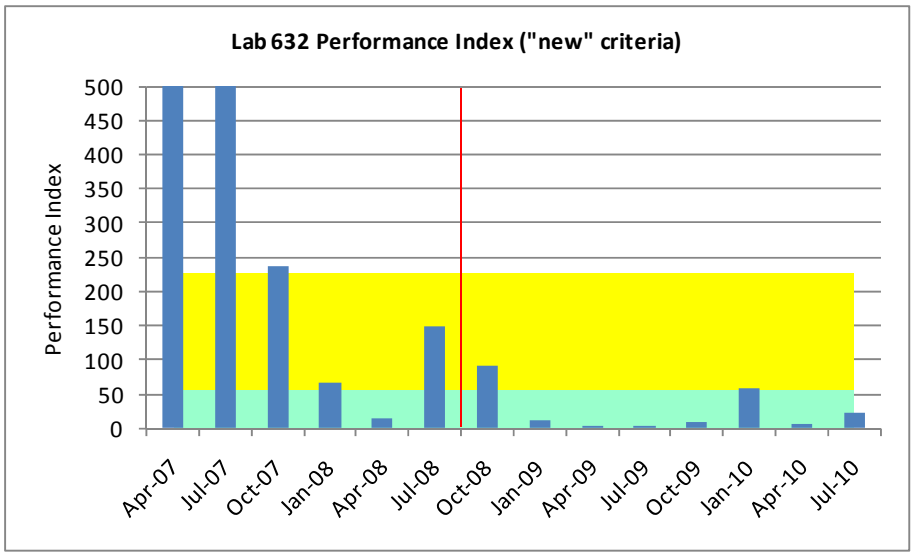
3.5.1 Laboratories Whose Performance Has Improved

Four laboratories appeared to show an improvement in their WASP performance after harmonisation. These are discussed below. **Please note that the "Good" and "Acceptable" ranges shown in the charts relate to the proposed WASP scoring system based on the Performance Index, which has not been adopted. Under the present WASP scoring system, the "Good" range (which equates to a Z-score within ± 1) and "Acceptable" range (which equates to a Z-score within ± 2) are together classed as "Satisfactory".**

3.5.1.1 Lab 632

Lab 632 was not performing particularly well in WASP prior to harmonisation (which they undertook in Oct 2008). Figure 3-9 shows their PI scores (very poor scores, PI in excess of 2000, were obtained in Apr 2007 and Jul 2007: these are above the maximum of 500 shown on the y-axis of this chart). Harmonisation appears to have **improved** performance, bringing their PI scores down predominantly into the "Good" range, as shown in Figure 3-9.

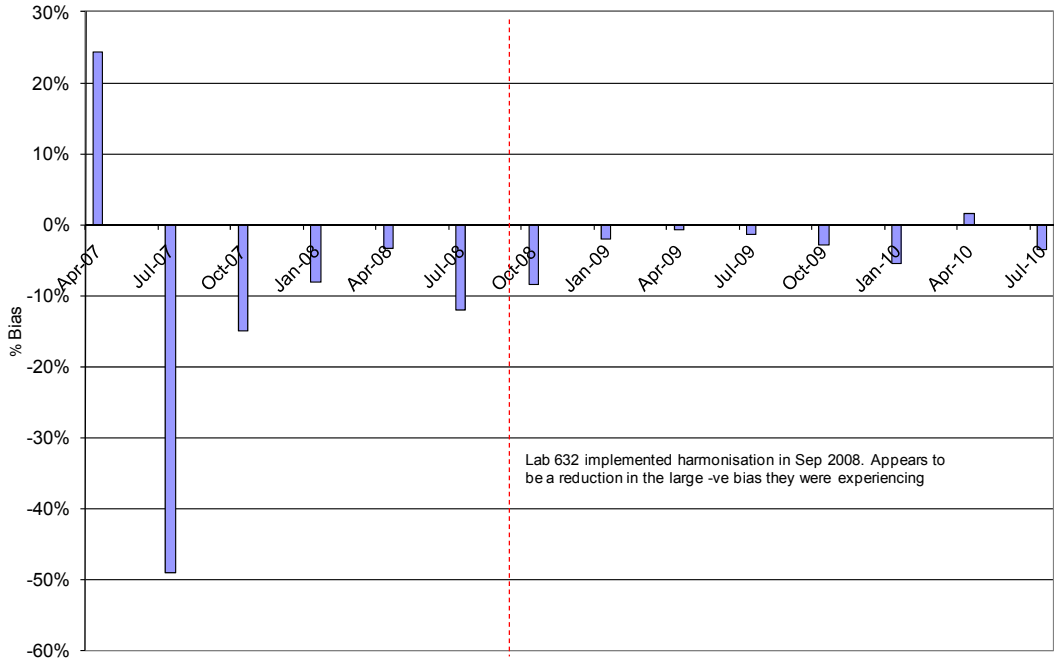
Figure 3-9 Lab 632 Performance Index (green band = "good", yellow band = "acceptable")



Examination of the accuracy of their results shows that before harmonisation they were typically underestimating the spiking levels (i.e. their results exhibited negative bias, Figure 3-10).

There appears to be a reduction in the negative bias this laboratory's WASP results were exhibiting before harmonisation. The changes made appear to have been beneficial. The laboratory says that the main change was that the colour reagent is now added as a mixed solution. Previously the two reagents were added separately. Staff changes had also occurred – which may have introduced changes that were nothing to do with methodology. Extraction is now carried out using a vortex mixer rather than manual shaking.

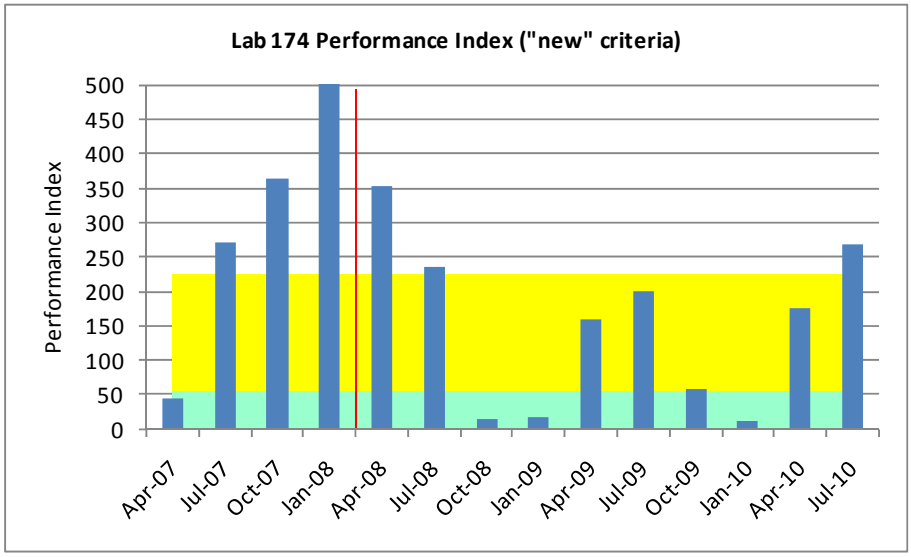
Figure 3-10 Lab 632 - Accuracy (% Bias) of WASP results relative to Spike Value



3.5.1.2 Lab 174

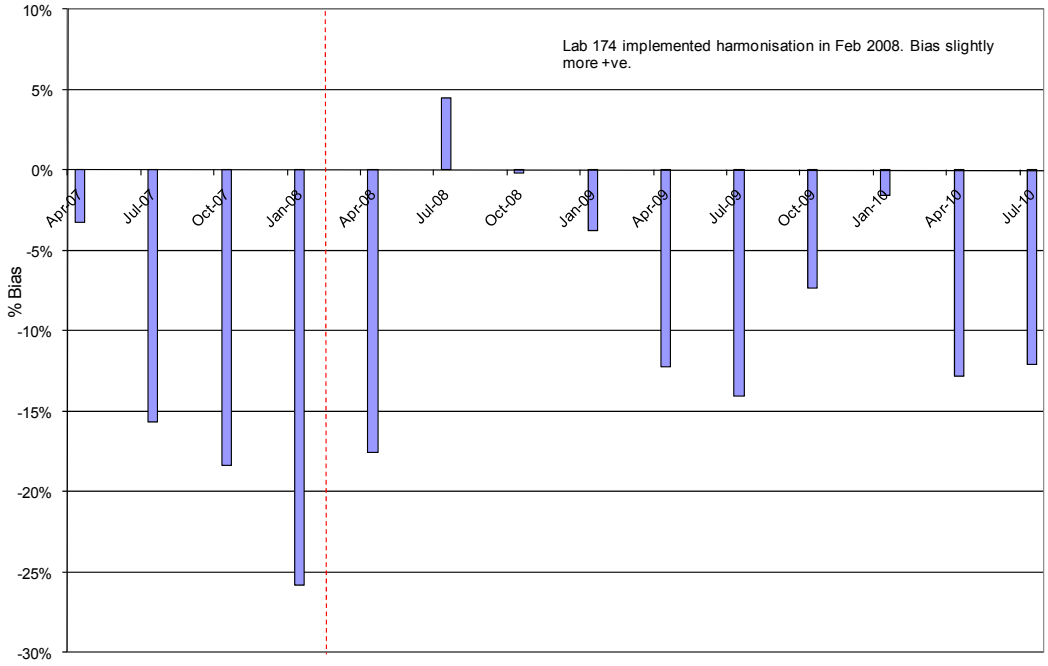
Lab 174 was previously a relatively poor performer in WASP. They implemented harmonisation in Feb 2008. Performance appears to have **improved** since harmonisation and (with the exception of a few poor rounds) is usually now in the current "acceptable" band (Figure 3-11).

Figure 3-11 Lab 174 Performance Index (green band = "good", yellow band = "acceptable")



Lab 174 formerly had a problem with large negative bias. This appears to have improved to some extent but is still a problem – see Figure 3-12. Harmonisation has been helpful to Lab 174, but negative bias has not been completely eliminated.

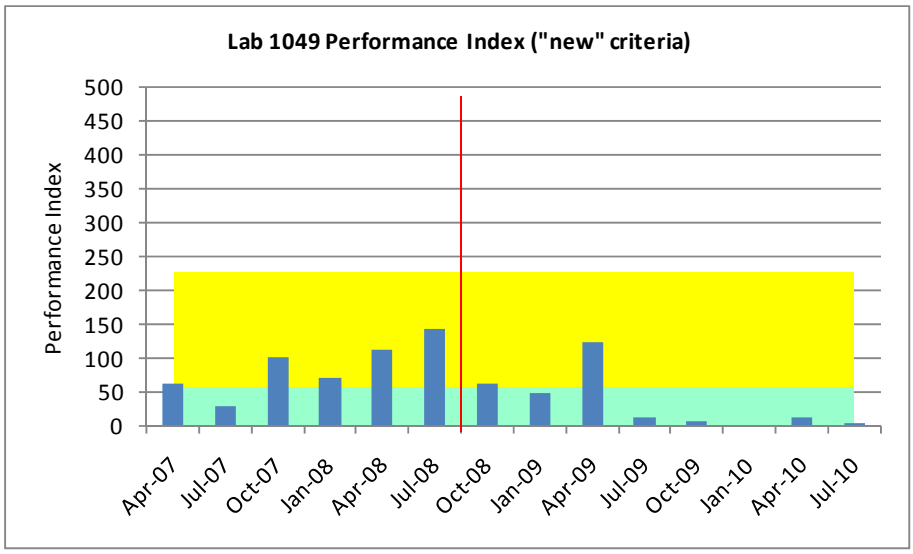
Figure 3-12 Lab 174 Accuracy (% Bias) of WASP results relative to Spike value



3.5.1.3 Lab 1049

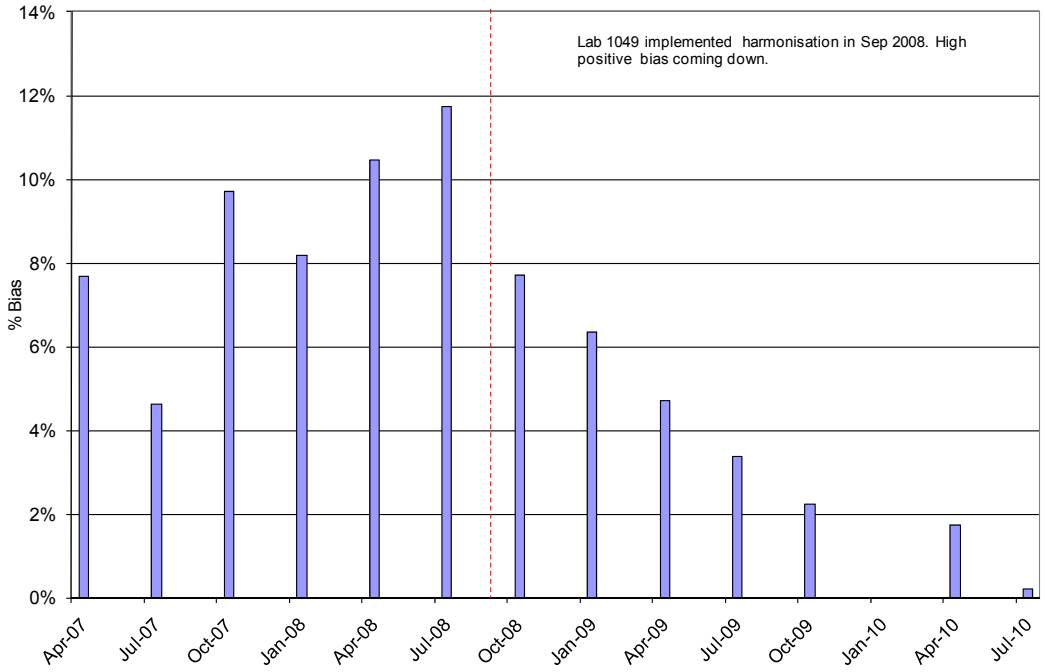
Lab 1049 was performing acceptably in WASP, prior to implementing the harmonisation measures in Sep 2008. However, since then (particularly since the July 2009 round) their performance has improved and has mostly been within the “Good” band (Figure 3-13).

Figure 3-13 Lab 1049 Performance Index (green band = "good", yellow band = "acceptable")



Prior to harmonisation, Lab 1049’s WASP results showed consistent high positive bias (overestimation of the spiking level). This clearly indicates something was wrong with the calibration or analysis – as it is not possible to recover more nitrite from the artificially spiked sample than was originally added. Since harmonisation in Sep 2008, the high bias has been steadily reducing - Figure 3-14.

Figure 3-14 Lab 1049 - Accuracy (% Bias) of WASP results relative to Spike value



Lab 1049's main changes in the process of harmonisation were:

1. from preparing the calibration standards by serial dilutions of the standard nitrite stock solution, to doing just one dilution from the 1000mg/L stock for each of the five standards.

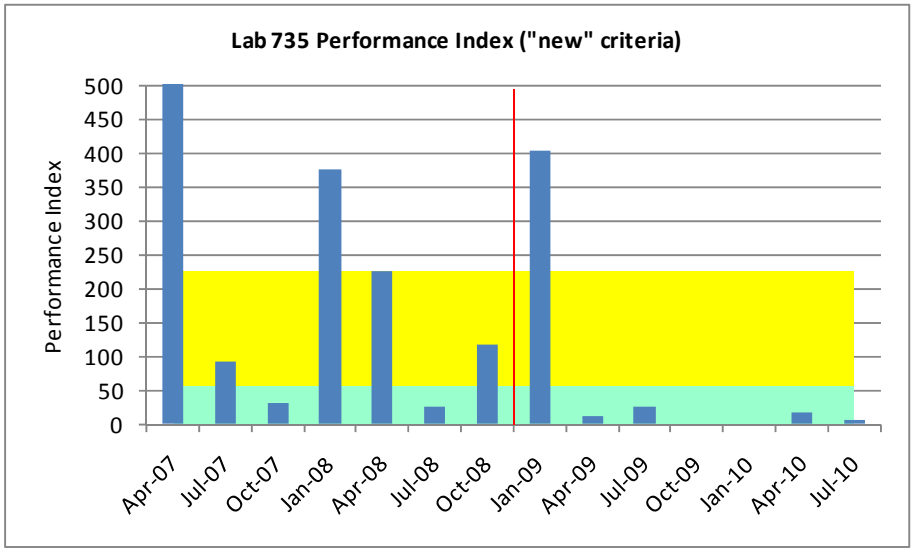
2. From adding the colour reagents separately, to adding them mixed.

It is possible that the first of these changes has helped improve the accuracy of Lab 1049's WASP analyses, by improving their calibrations.

3.5.1.4 Lab 735

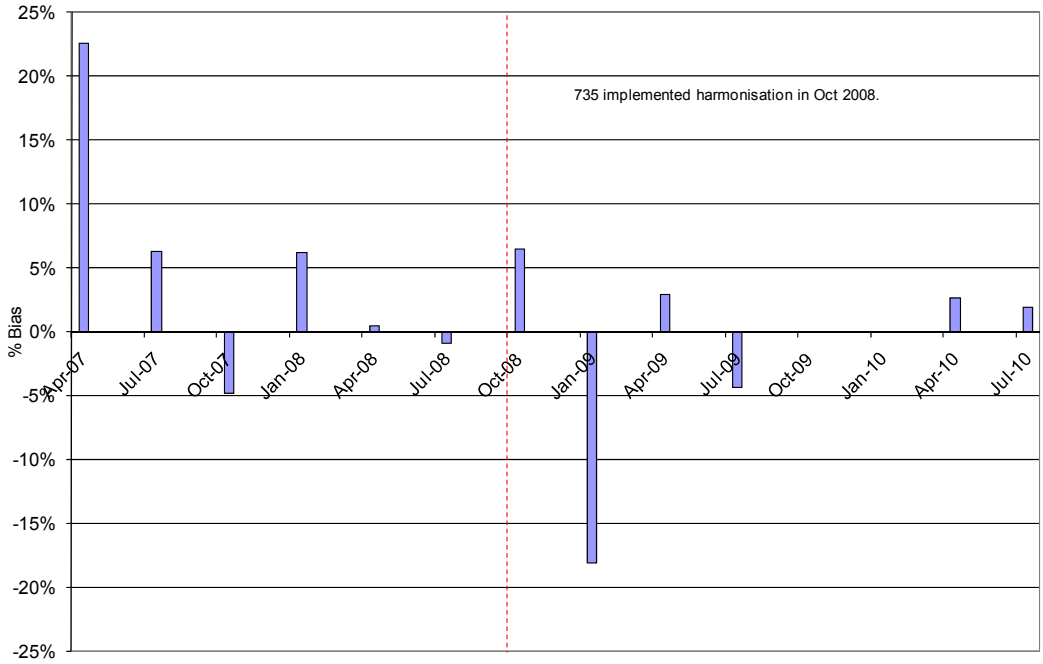
Lab 735's performance was variable prior to harmonisation (see Figure 3-15 below: the y-axis of this figure has been reduced for clarity – the very high PI score of 3100 that this lab obtained in April 2007 is not shown). Performance appears to have **improved** since harmonisation (although there were no results for the Oct 2010 or Jan 2010 rounds).

Figure 3-15 Lab 735 Performance Index (green band = "good", yellow band = "acceptable")



There appears to be no clear change in the typical bias of the tubes (Figure 3-16), but the results are typically within a few per cent of the spike value.

Figure 3-16 Lab 735 - Accuracy (% Bias) of WASP results relative to Spike value



No information is available on the exact changes made by Lab 735.

3.5.1.5 Summary on Laboratories with Improved Performance

Four participating laboratories showed clear improvement in WASP following harmonisation. In two cases (Lab 632 and Lab 174, both of which had previously been relatively poor performers compared with the other participants) this was due to the reduction of large negative bias. In the third case (Lab 1049) it was due to the reduction of large positive bias.

The fourth (Lab 735) had previously exhibited somewhat variable performance, which improved.

In all these cases, it should not be assumed that the improvement in performance is necessarily due to the changes implemented to harmonise procedures (although it may be). Other changes such as staff training, or upgrading laboratory equipment, may also be involved.

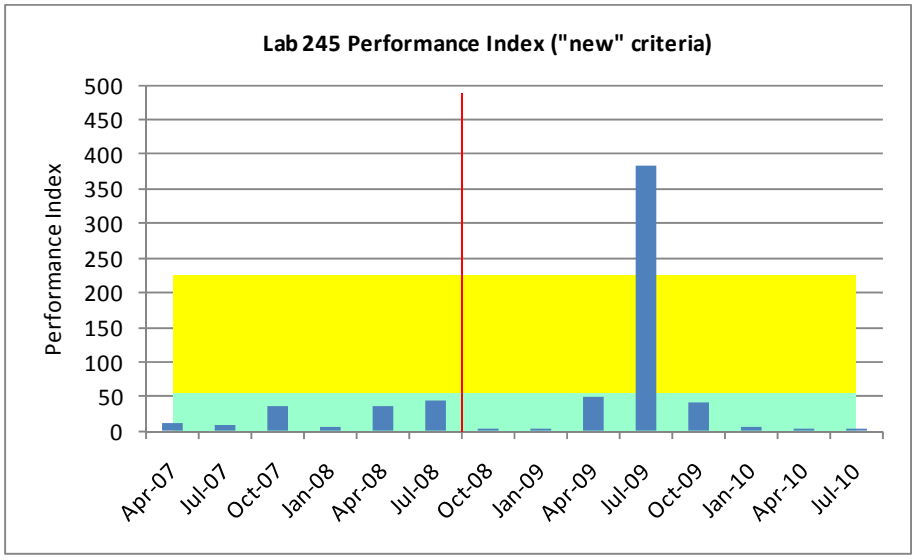
3.5.2 Laboratories That Performed Well Before and After Harmonisation

Some laboratories were performing relatively well in WASP prior to harmonisation, and continued to do so afterwards. These are discussed below. **Again, the Performance Index bands used here related to the proposed scoring system which was not adopted: the "good" and "acceptable" bands together would, under the current system, be together classified as "Satisfactory", which equates to a Z-score within ± 2.**

3.5.2.1 Lab 245

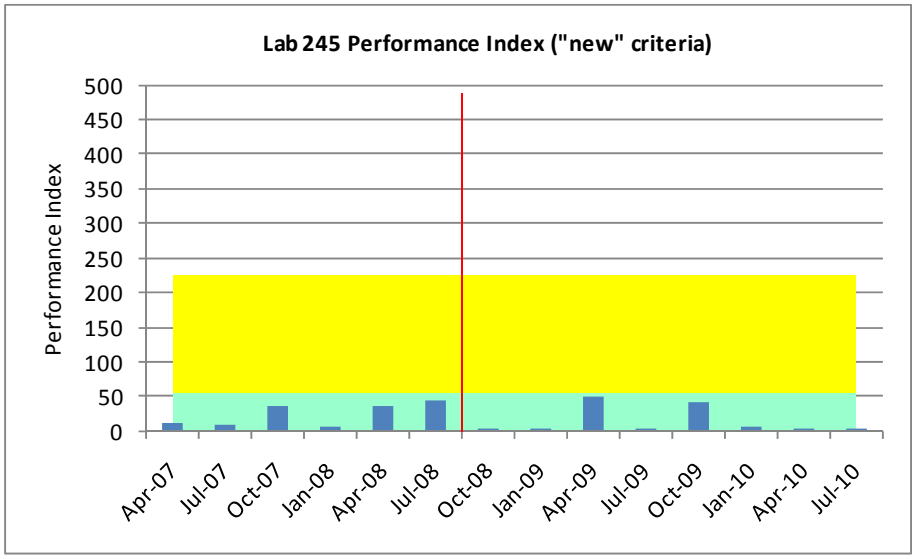
Lab 245 performed well prior to harmonisation (completed by Jan 2009). They are still doing so, despite one relatively poor round in July 2009. The PI scores are shown in Figure 3-17.

Figure 3-17 Lab 245 Performance Index (green band = "good", yellow band = "acceptable")



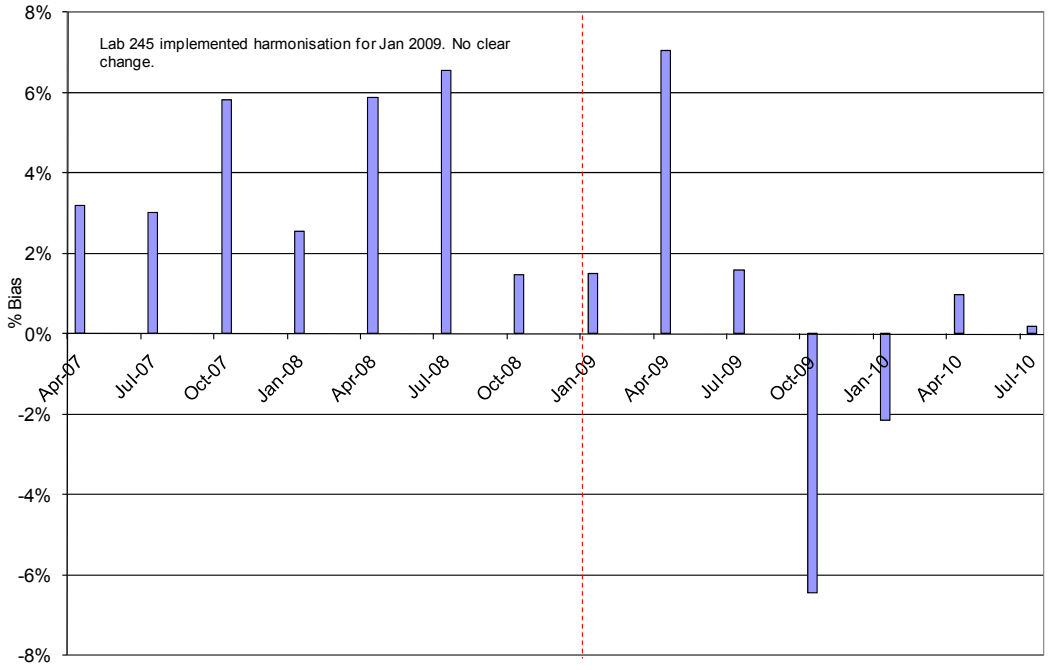
Two of the sample results in the July 2009 round appear to have been mixed up, giving rise to the poor July 2009 score. If this is corrected the "poor" result disappears, as shown in Figure 3-18.

Figure 3-18 Lab 245 Performance Index (green band = "good", yellow band = "acceptable") – Jul 2009 results for samples 2 and 3 interchanged



Examination of the bias of Lab 245's results shows that prior to harmonisation they typically over-estimated the spiking level of the WASP samples by a few per cent (Figure 3-19– the Jul 2009 result is based on data corrected as above).

Figure 3-19 Lab 245 - Accuracy (% Bias) of WASP results relative to Spike value



There is no clear change apparent as a result of the harmonisation. The positive bias prevalent before harmonisation may have ceased (the Oct 2009 results showed an uncharacteristic negative bias) but there appeared to be no change at the time of harmonisation. Harmonisation does not appear to have improved or impaired Lab 245's performance.

3.5.2.2 Lab 372

Lab 372 previously performed relatively well in WASP and continues to do so, with most PI scores in the "Good" range. This is illustrated by Figure 3-20. The April 2007 score of over 3000, and the Jan 2010 score of 879 (both of which are above the maximum shown below) were uncharacteristically poor.

Figure 3-20 Lab 372 Performance Index (green band = "good", yellow band = "acceptable")

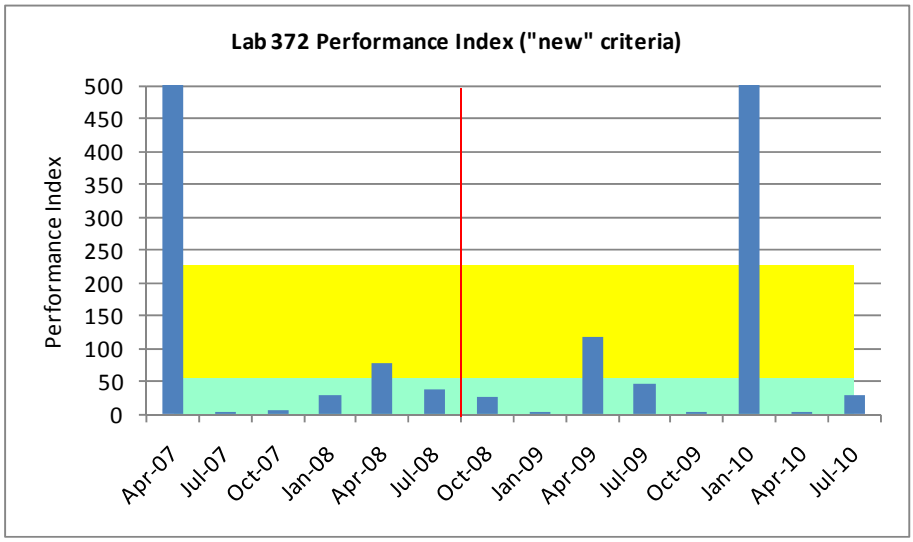
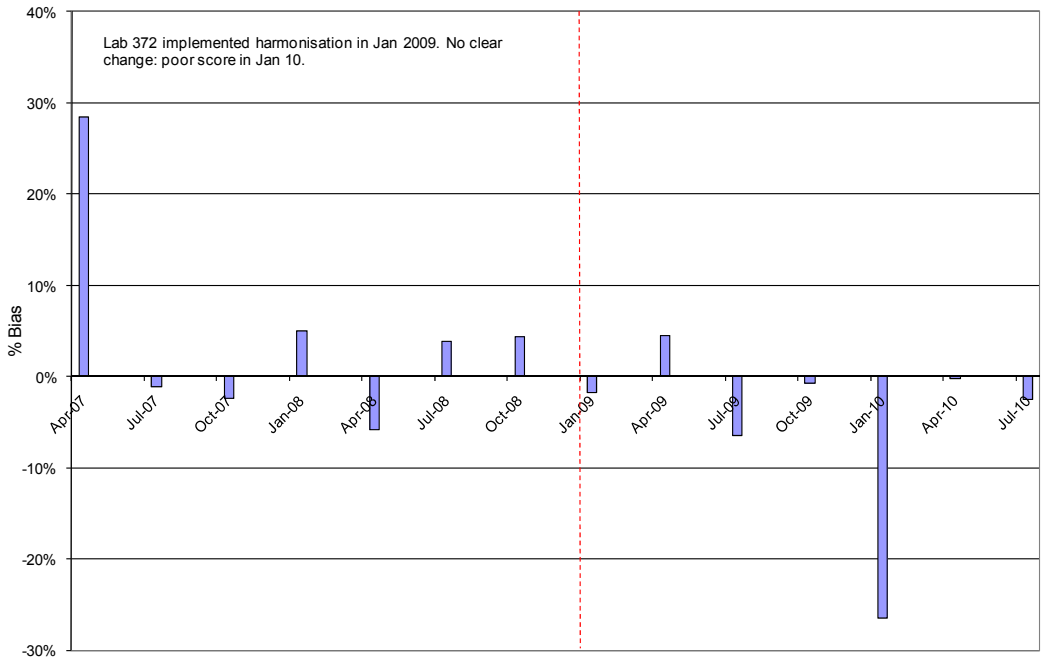


Figure 3-21 shows the bias of Lab 372's results: there does not appear to be any clear change as a result of the harmonisation.

Figure 3-21 Lab 372 - Accuracy (% Bias) of WASP results relative to Spike value



Overall, harmonisation does not appear to have improved or impaired Lab 372's performance.

3.5.2.3 Lab 935

Lab 935's performance in WASP was good before harmonisation, and has remained so. They claimed their original method differed little from the harmonised method, and therefore made only the following changes:

1. Changed extraction technique from a two-stage process, adding water first then reagents, to adding mixed reagent,
2. Changed from manual shaking to automated extraction.
3. Increased the frequency of calibration (now a full calibration with each batch).
4. More frequent calibration checks and blank checks.
5. Changed to using diffusion tubes for the standard solutions used in their calibrations.

Figure 3-22 shows their PI scores.

Figure 3-22 Lab 935 Performance Index (green band = "good", yellow band = "acceptable")

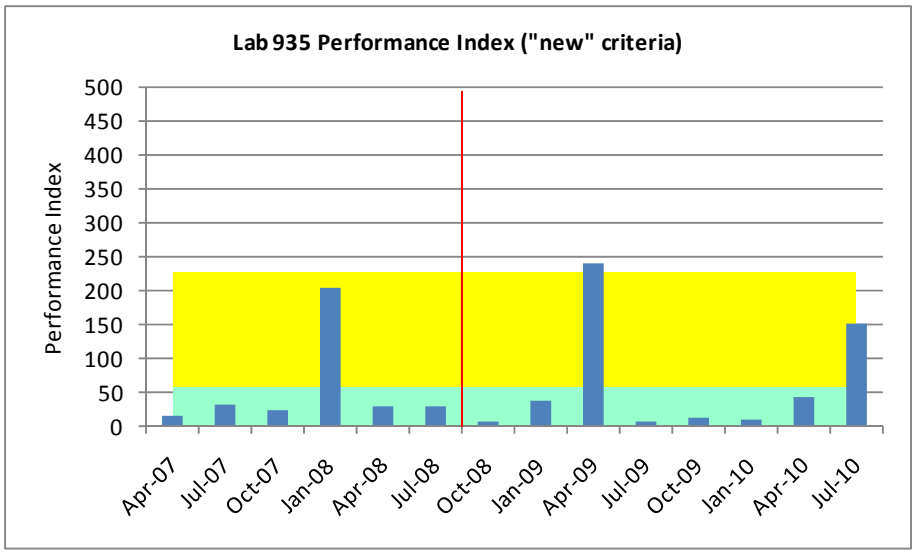
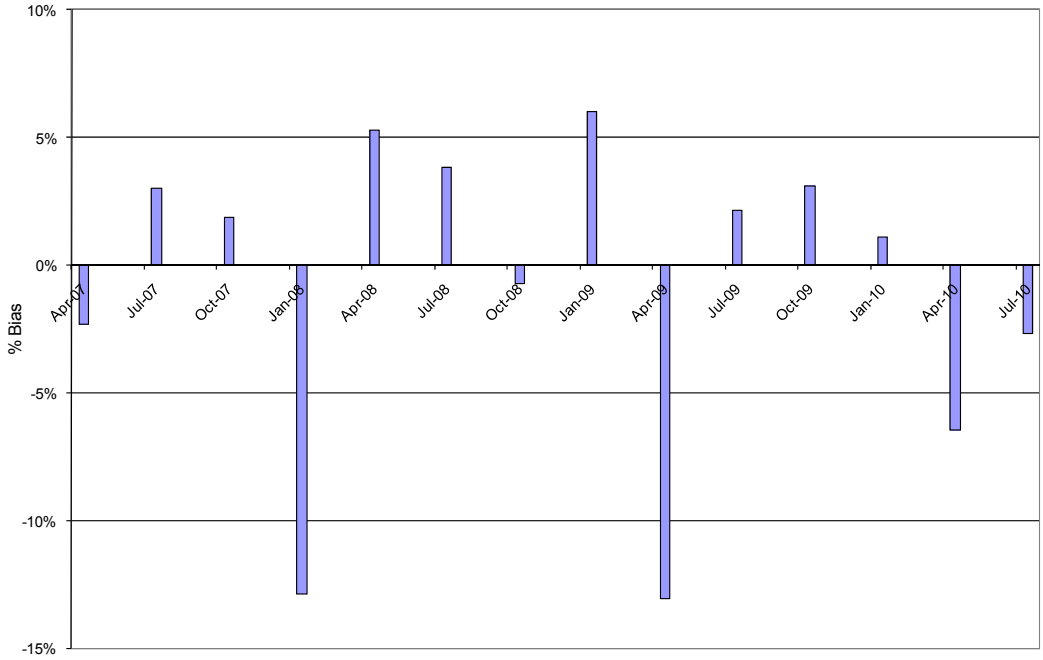


Figure 3-23 shows how Lab 935's bias in WASP has changed. There do not appear to be any clear changes as a result of the harmonisation.

Figure 3-23 Lab 935 - Accuracy (% Bias) of WASP results relative to Spike value

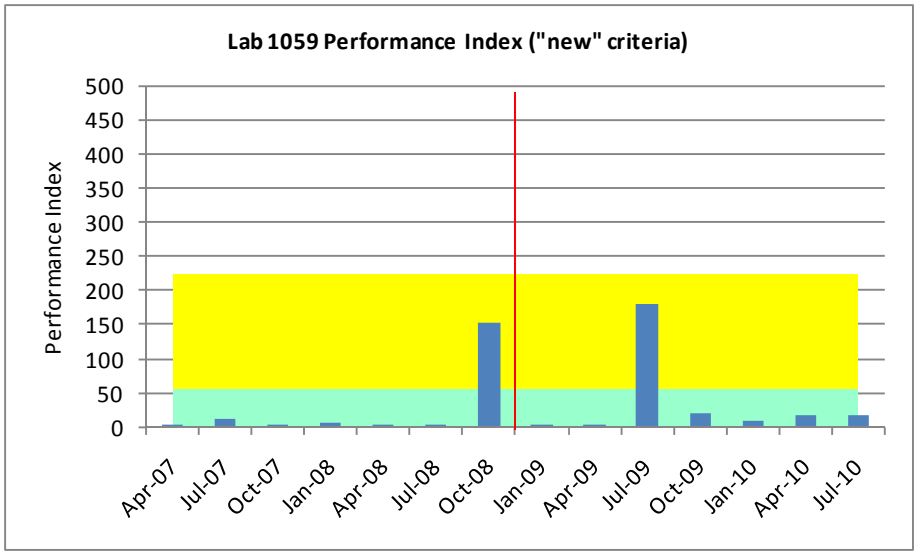


Lab 935's generally good performance in WASP does not appear to have changed as a result of harmonisation.

3.5.2.4 Lab 1059

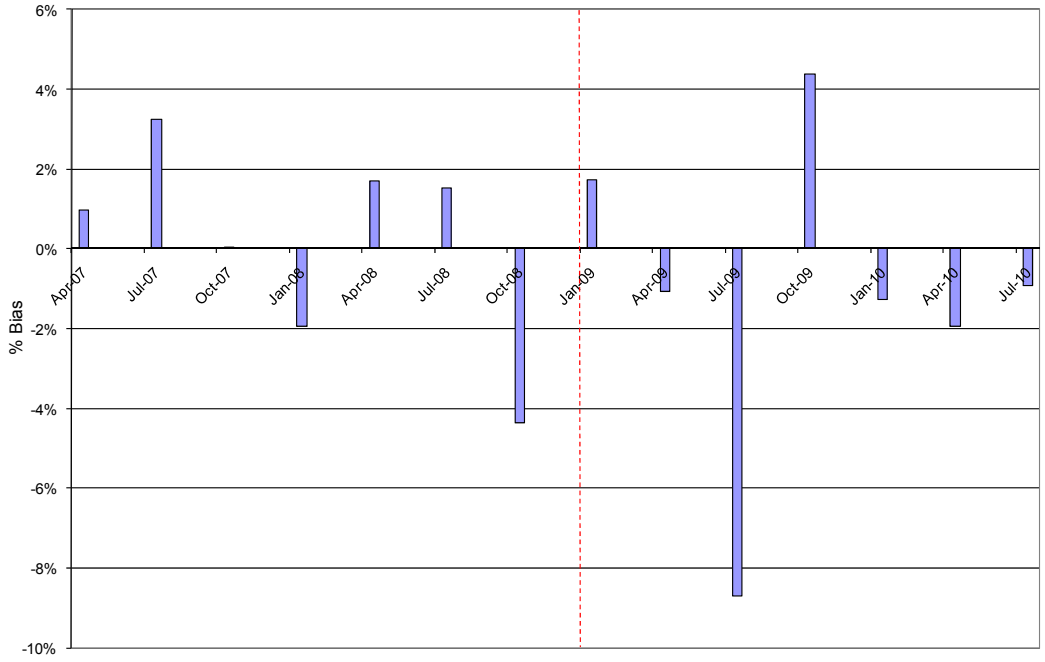
Lab 1059 performed well before and after harmonisation. Their PI scores are shown in Figure 3-24, and were usually well within the "good" band (occasionally in the "acceptable" band), both before and after harmonisation.

Figure 3-24 Lab 1059 Performance Index (green band = "good", yellow band = "acceptable")



The bias of Lab 1059's WASP results is shown in Figure 3-25. Harmonisation does not appear to have resulted in any clear changes.

Figure 3-25 Lab 1059 - Accuracy (% Bias) of WASP results relative to Spike value



Lab 1059's generally good performance in WASP does not appear to have changed as a result of harmonisation. The poor result in Jun 2009 resulted from underestimation of one tube result.

3.5.2.5 Summary on Laboratories that Performed Well Before and After Harmonisation

Four laboratories (Lab 245, Lab 372, Lab 935 and Lab 1059) were performing well prior to harmonisation, with PI scores typically in the "good" band. These laboratories continued to perform well once harmonisation had been completed: there were no obvious changes.

It is noted that even these "good" performers occasionally have a poor round, i.e. one with a PI score in the "unacceptable" range. When this happens, it is usually due to substantial under-read of one or more of the samples. It may be worth investigating the reasons for this.

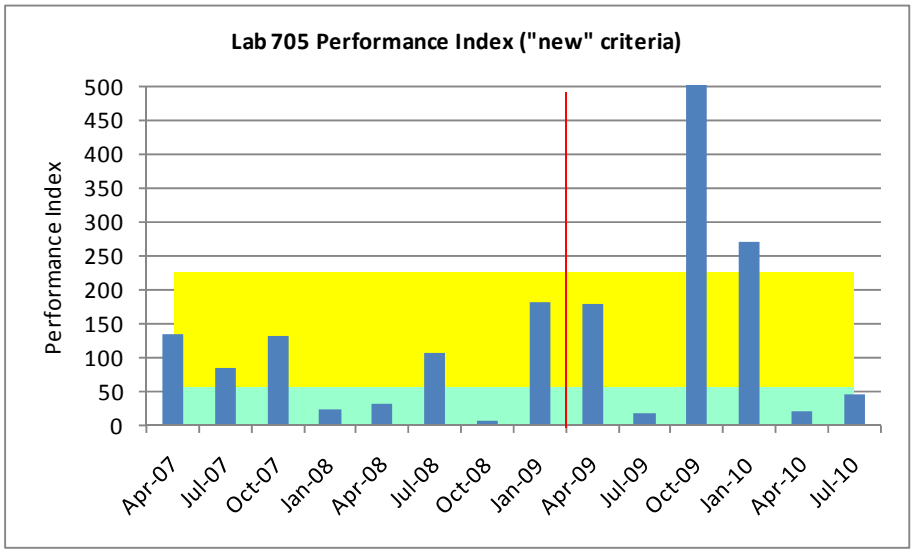
3.5.3 Laboratories with Variable Performance Before and After Harmonisation

A number of laboratories, (whose WASP performance before harmonisation was acceptable if rather variable), did not appear to show any clear improvement or deterioration. These are discussed below. **Again, the Performance Index bands used here related to the proposed scoring system which was not adopted: the "good" and "acceptable" bands together would, under the current system, be together classified as "Satisfactory", which equates to a Z-score within ± 2.**

3.5.3.1 Lab 705

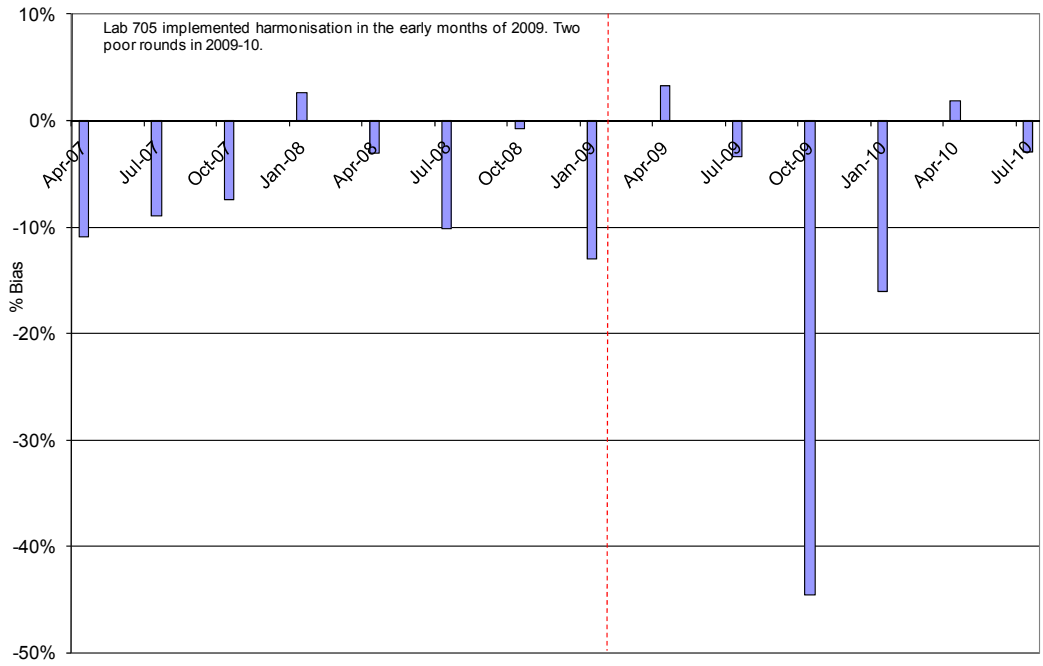
Lab 705 was slightly late in implementing the harmonisation. Lab 705's performance prior to harmonisation was variable but usually in the "acceptable" or "good" band. This has generally remained the case after harmonisation, although Lab 705 has had two consecutive poor rounds since harmonisation (Oct 2009 and Jan 2010). This is shown in Figure 3-26.

Figure 3-26 Lab 705 Performance Index (green band = "good", yellow band = "acceptable")



Looking at the "bias" of Lab 705's WASP results relative to the spike value, Figure 3-27 shows that prior to harmonisation this laboratory's WASP results typically underestimated relative to the spike value, frequently by around 10%. This appears to have been reduced subsequently, although the two poor scores obtained (in Oct 2009 and Jan 2010) involved substantial under-estimation.

Figure 3-27 Lab 705 - Accuracy (% Bias) of WASP results relative to Spike value



3.5.3.2 Lab 1054

Lab 1054 was performing inconsistently in WASP before harmonisation, as shown in Figure 3-28. Further investigation showed that their results often exhibited large positive bias, overestimating by more than 10% on occasions (Figure 3-29).

Figure 3-28 Lab 1054 Performance Index (green band = "good", yellow band = "acceptable")

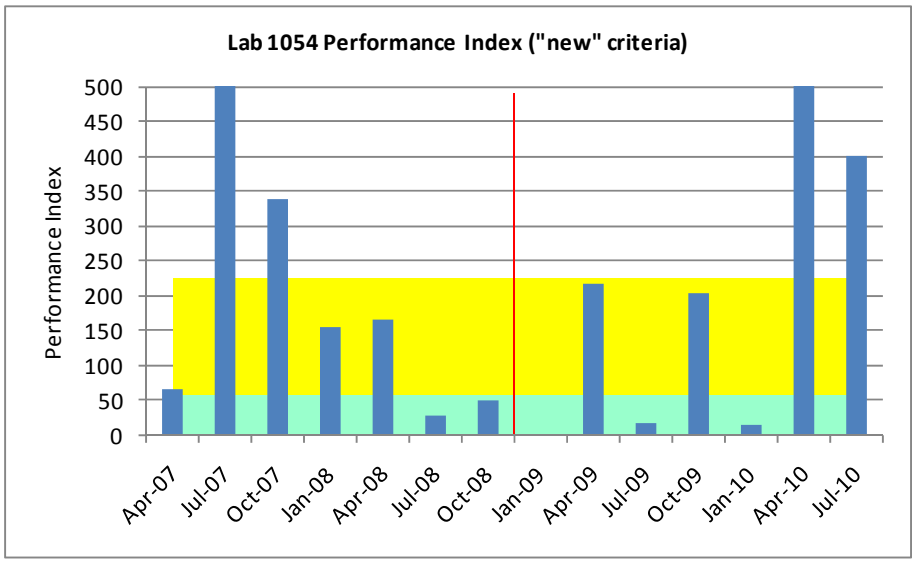
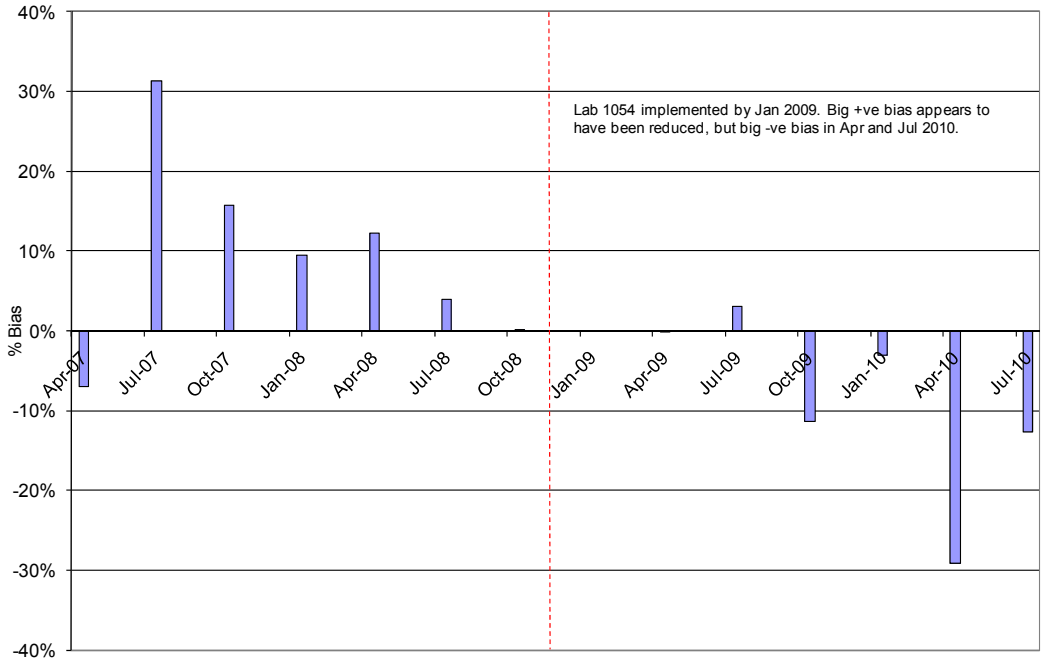


Figure 3-29 Lab 1054 - Accuracy (% Bias) of WASP results relative to Spike value



In the first few rounds after harmonisation, i.e. Apr 2009 - Jan 2010 (there were no data for Jan 2009) it appeared that the performance had improved. The positive bias frequently seen before harmonisation had apparently ceased. However, during 2010 negative bias has

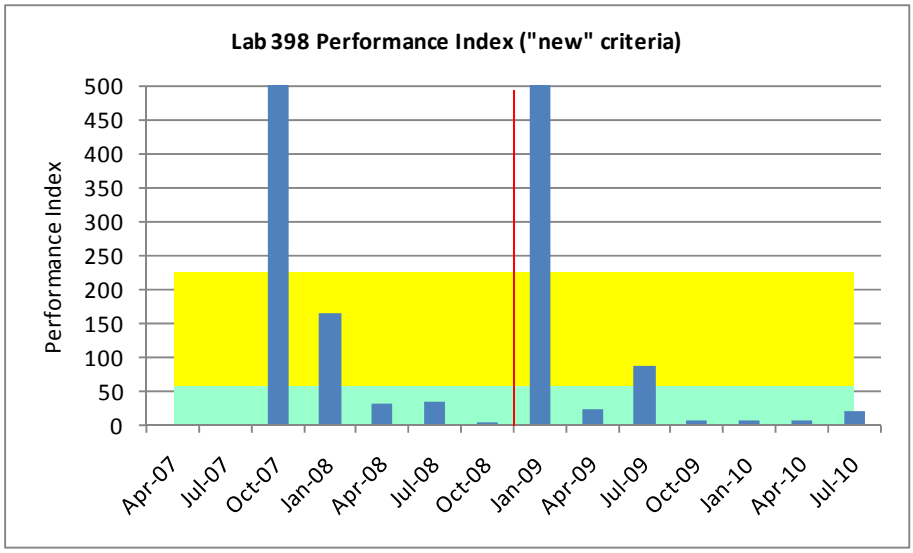
started to occur, and the Performance Index scores for the most recent two rounds (April and July 2010) were in the current “Unacceptable” band.

Therefore, despite an apparent improvement in the positive bias observed before harmonisation, Lab 1054’s performance is still variable. Moreover, two instances of large negative bias have occurred recently. A mixed outcome for this laboratory.

3.5.3.3 Lab 398

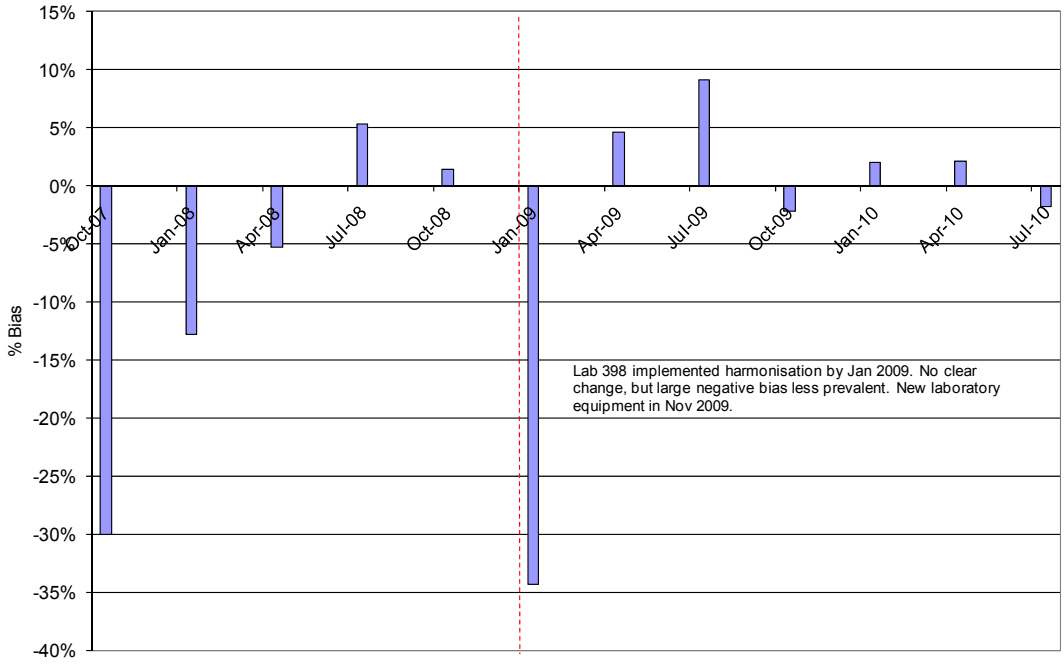
Lab 398’s performance was variable prior to harmonisation: predominantly in the “acceptable” and “good” bands but with an occasional poor score. This remains the case (Figure 3-30).

Figure 3-30 Lab 398 Performance Index (green band = "good", yellow band = "acceptable")



Lab 398’s WASP results appear to have improved in the four most recent rounds: they commented that they attribute the recent improvement in performance to new analytical equipment, installed in November 2009. A poor round in Jan 2009 (immediately after harmonisation) was due to substantial under-estimation of samples - Figure 3-31.

Figure 3-31 Lab 398 - Accuracy (% Bias) of WASP results relative to Spike value

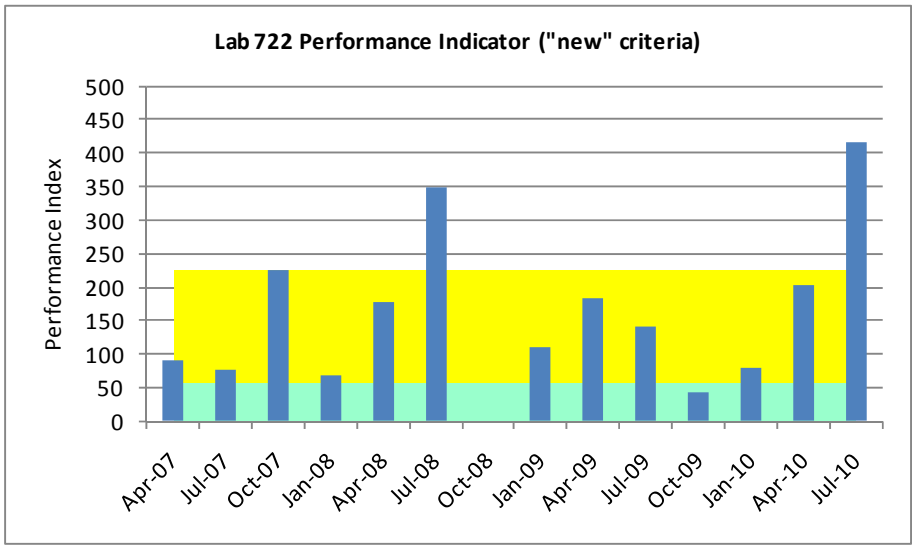


3.5.3.4 Lab 722

Lab 722 performed variably in WASP prior to harmonisation, with a PI usually in the "acceptable" band. This is shown in Figure 3-32. Laboratory moves, staff changes and other factors delayed harmonisation, and as of May 2009 they had not completed the changes. It is not known when harmonisation was completed - Lab 722 say that staff have changed and they are unable to provide details of what was changed and when. Therefore no red line (indicating completion of harmonisation) is shown in the figures below.

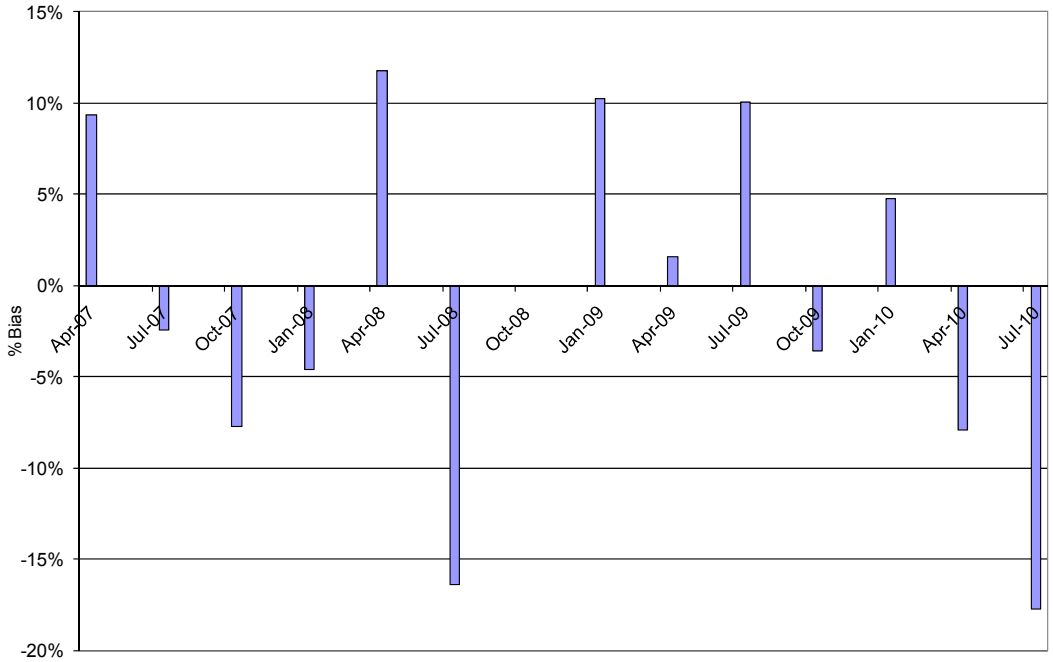
Lab 722's performance does not appear to have changed much since harmonisation, remaining still typically in the "acceptable" range.

Figure 3-32 Lab 722 Performance Index (green band = "good", yellow band = "acceptable")



There has been no clear change in the bias of Lab 722's tubes (Figure 3-33).

Figure 3-33 Lab 722 - Accuracy (% Bias) of WASP results relative to Spike value



3.5.3.5 Lab 775

The performance of Lab 775 in WASP was variable prior to harmonisation (sometimes good, sometimes poor). Since harmonisation, PI scores have mostly been in the "good" range, but one poor round occurred in Apr 2009 (Figure 3-34) which makes it impossible to conclude that there has been an improvement.

Figure 3-34 Lab 775 Performance Index (green band = "good", yellow band = "acceptable")

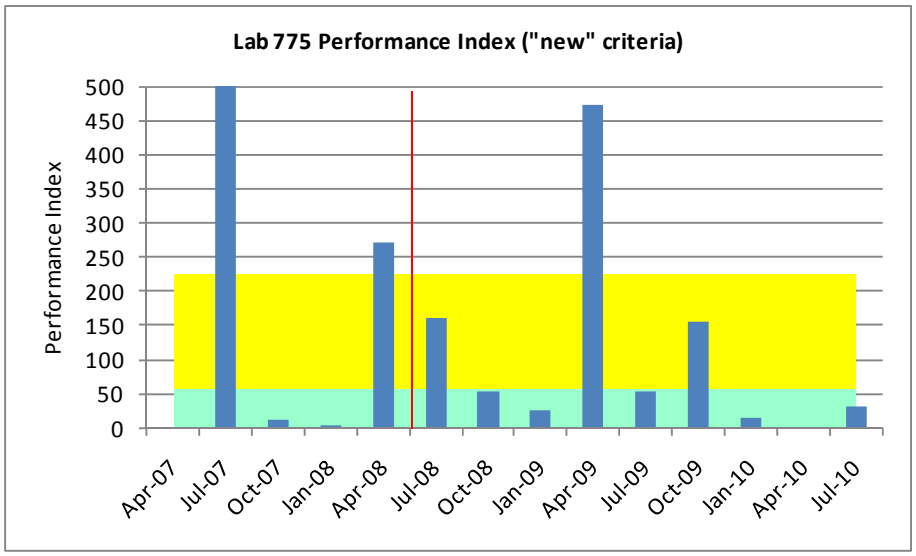
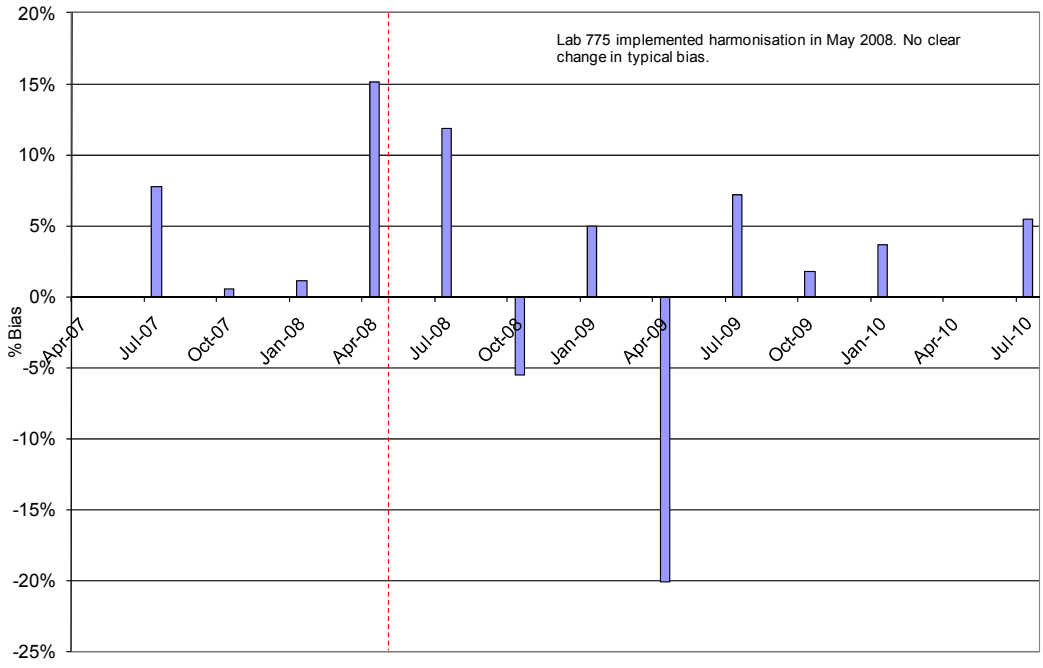


Figure 3-35 indicates that the typical bias exhibited by Lab 775's analysis of WASP samples has not changed markedly. It was more usually positive before and after harmonisation.

However, the poor round in Apr 2009 resulted from an uncharacteristic under-estimation of the spiking value.

Figure 3-35 Lab 775 - Accuracy (% Bias) of WASP results relative to Spike value

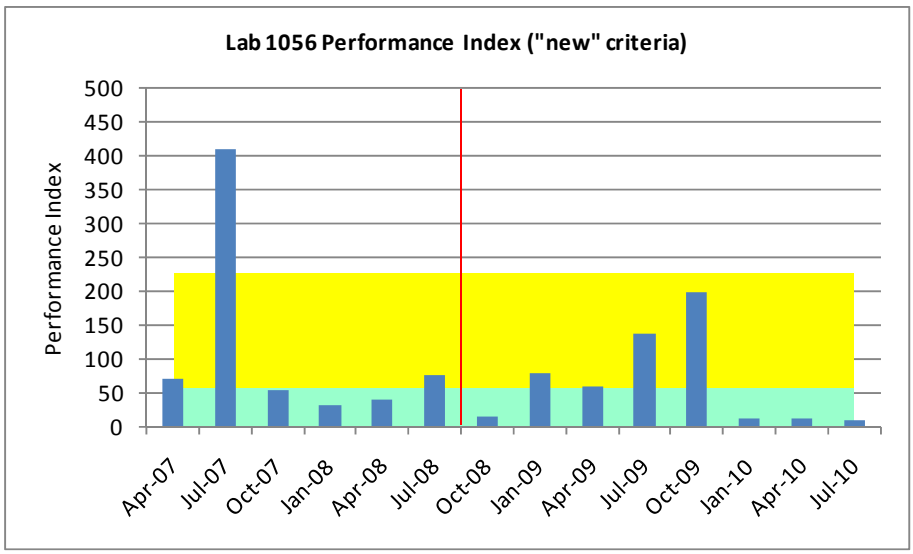


It does not appear that harmonisation has changed Lab 775's performance in WASP.

3.5.3.6 Lab 1056

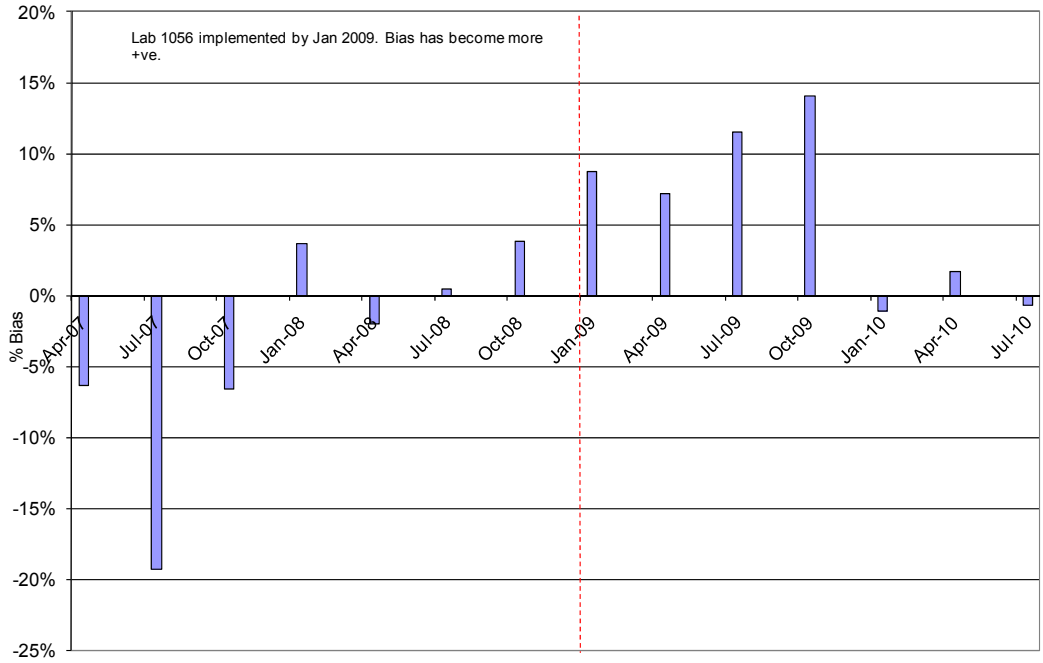
Lab 1056 showed no clear changes in their WASP PI scores (Figure 3-36) – although the most recent three rounds have been consistently good.

Figure 3-36 Lab 1056 Performance Index (green band = "good", yellow band = "acceptable")



It appears that the bias of their analyses of the WASP spiked samples became more positive in the four rounds immediately after harmonisation was implemented (Figure 3-37). However, this appears to have been reduced in the most recent three rounds.

Figure 3-37 Lab 1056 - Accuracy (% Bias) of WASP results relative to Spike value



3.5.3.7 Lab 1017

Lab 1017 performed reasonably well prior to harmonisation, despite an occasional poor score. This remains the case since harmonisation (Figure 3-38).

Figure 3-38 Lab 1017 Performance Index (green band = "good", yellow band = "acceptable")

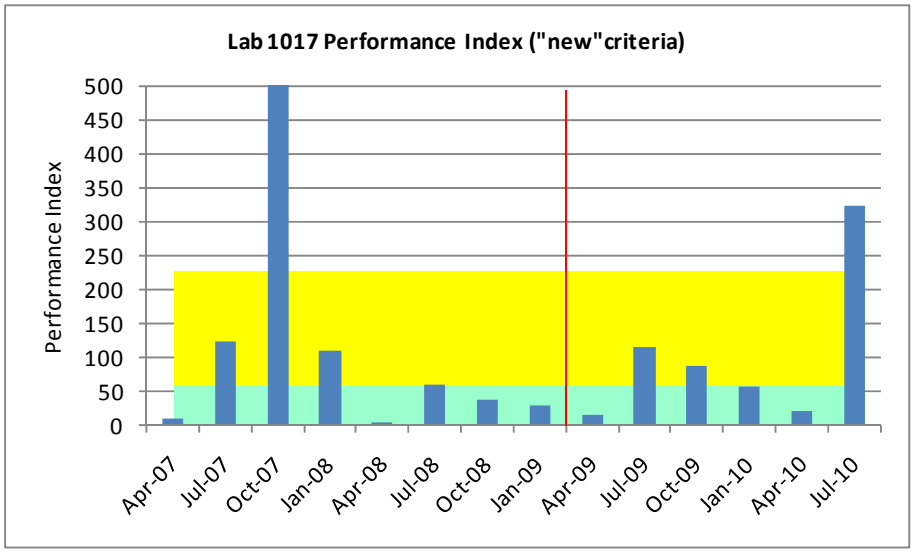
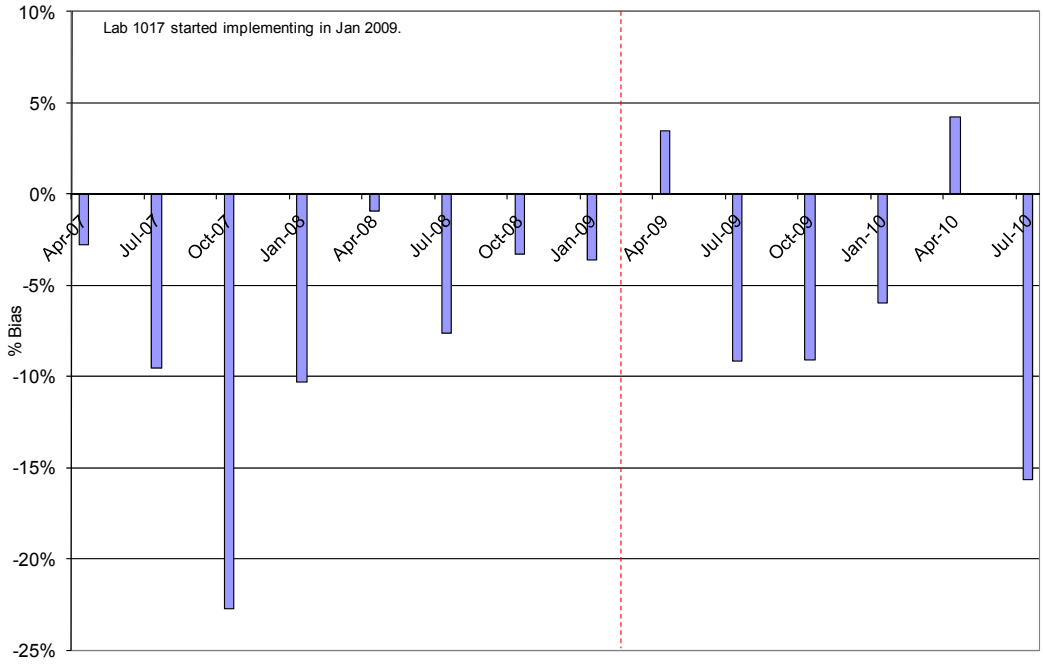


Figure shows that prior to harmonisation, Lab 1017's WASP analyses usually showed negative bias and that this has continued (Figure 3-39). The only "poor" score occurring since harmonisation was due to under-read.

Figure 3-39 Lab 1017 - Accuracy (% Bias) of WASP results relative to Spike value



Lab 1017 commented that the switch to using a vibrating tray for tube extraction caused them problems. They found that ***"although overall this improved the extraction from the metal discs it did not mix evenly throughout, we consistently got a high concentration in the bottom third of the tube. To overcome this we now invert the tray of tubes a number of times before reading to ensure a consistent mixture."***

3.5.3.8 Summary on Laboratories Whose Performance Remained Variable

Seven laboratories, whose performance was variable prior to harmonisation, showed no clear change. Their performance remained variable after harmonisation. These were; Lab 705, Lab 1054, Lab 398, Lab 722, Lab 775, Lab 1056,and Lab 1017.

Like the "good" performers discussed in section 3.3.2, these laboratories occasionally still have poor rounds, i.e. where the PI score is in the "unacceptable" range. When this happens, it is usually due to substantial under-read of one or more of the samples. It may be worth investigating the reasons for this.

3.5.4 Laboratories Whose Performance Has Worsened

Again, the Performance Index bands used here related to the proposed scoring system which was not adopted: the "good" and "acceptable" bands together would, under the current system, be together classified as "Satisfactory", which equates to a Z-score within ± 2.

3.5.4.1 Laboratory 1045

Lab 1045 was performing well in WASP prior to harmonisation (with the exception of a bad round in Oct 2008, giving a PI of almost 2000). This is shown in Figure 3-40 below: the y-axis has been reduced to a maximum of 500 for clarity. Since harmonisation (in Dec 2008) there have been appears to have been a **slight deterioration** in the WASP PI scores. Despite the improved consistency in the scores, and the fact that they remain mostly still in the "good" band, they are typically higher (i.e. worse).

Examining the bias of Lab 1045's WASP results (Figure 3-41) shows that a small negative bias appears to have been introduced around the time of the harmonisation changes.

Lab 1045's results have improved in one way: they appear to be more consistent. However, it appears that the changes they made were not entirely beneficial, as they are now typically slightly under-estimating the WASP samples. One possibility (although this has not been investigated in any way) is that the new extraction method is not as efficient at releasing the nitrite from the tubes as their previous procedures.

Figure 3-40 Lab 1045 Performance Index (green band = "good", yellow band = "acceptable")

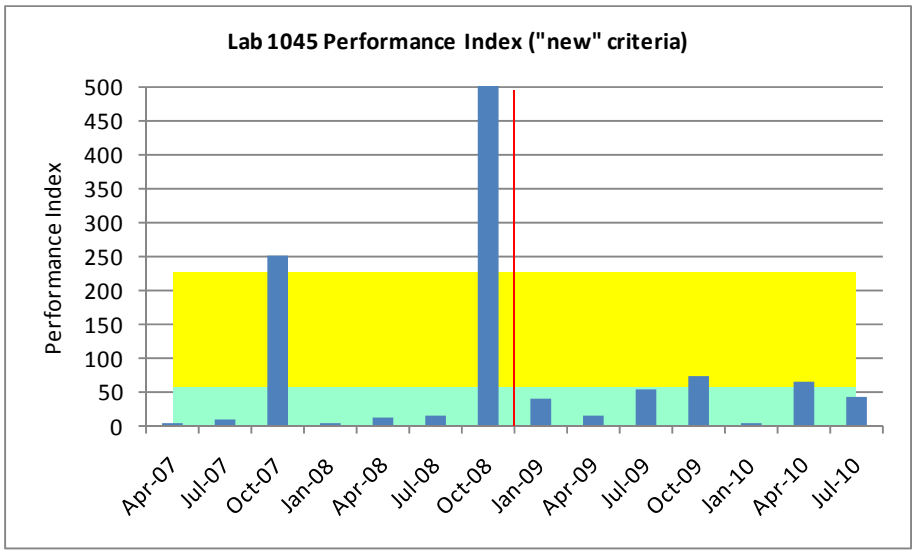
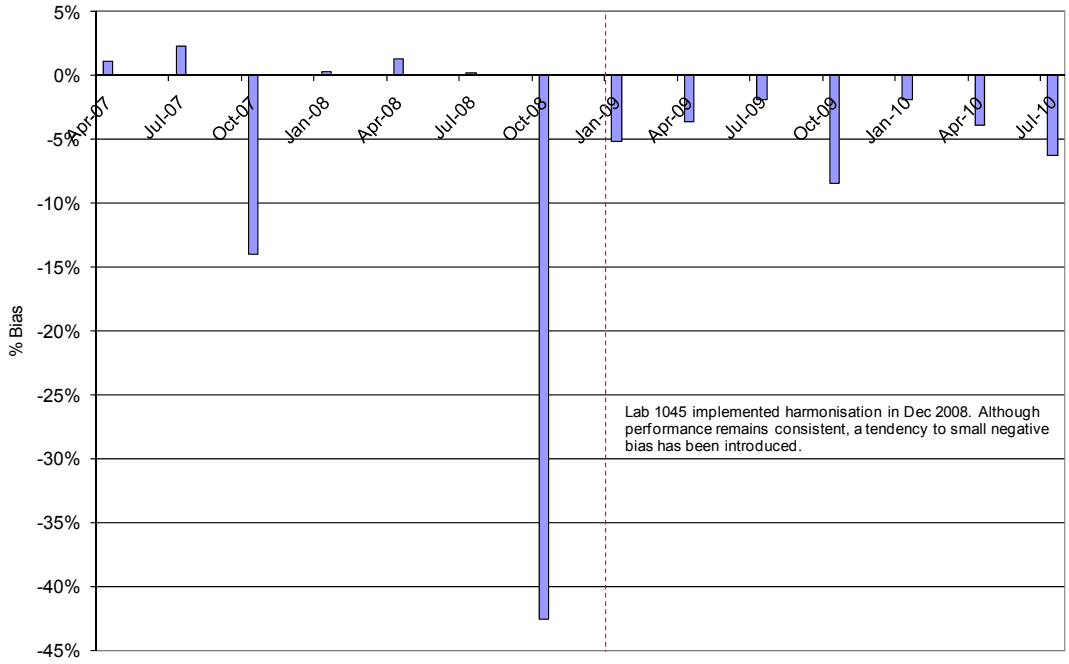


Figure 3-41 Lab 1045 - Accuracy (% Bias) of WASP results relative to Spike value



3.5.4.2 Laboratory 152

Lab 152 implemented harmonisation in July 2008. Their performance in WASP was good beforehand (apart from a poor result in April 2008). Although it is still predominantly in the good band (PI <56), the PI scores have become slightly ***worse*** since harmonisation –. This is shown in Figure 3-42.

More detailed examination of Lab 152's results shows that since harmonisation, they have tended to exhibit slight negative bias (i.e. underestimation of the "spiked" tube). This is shown in Figure 3-43. This is a very similar outcome to that observed for Lab 1045 above.

Figure 3-42 Lab 152 Performance Index (green band = "good", yellow band = "acceptable")

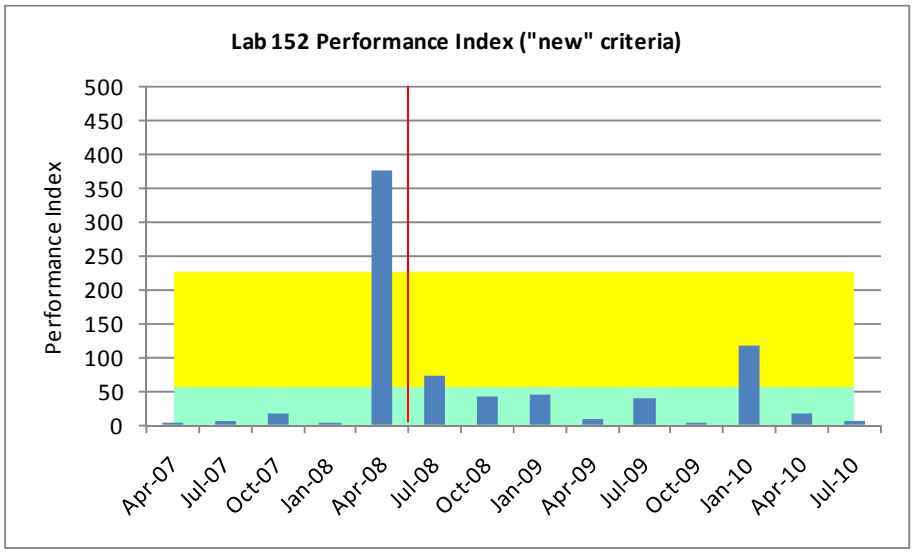
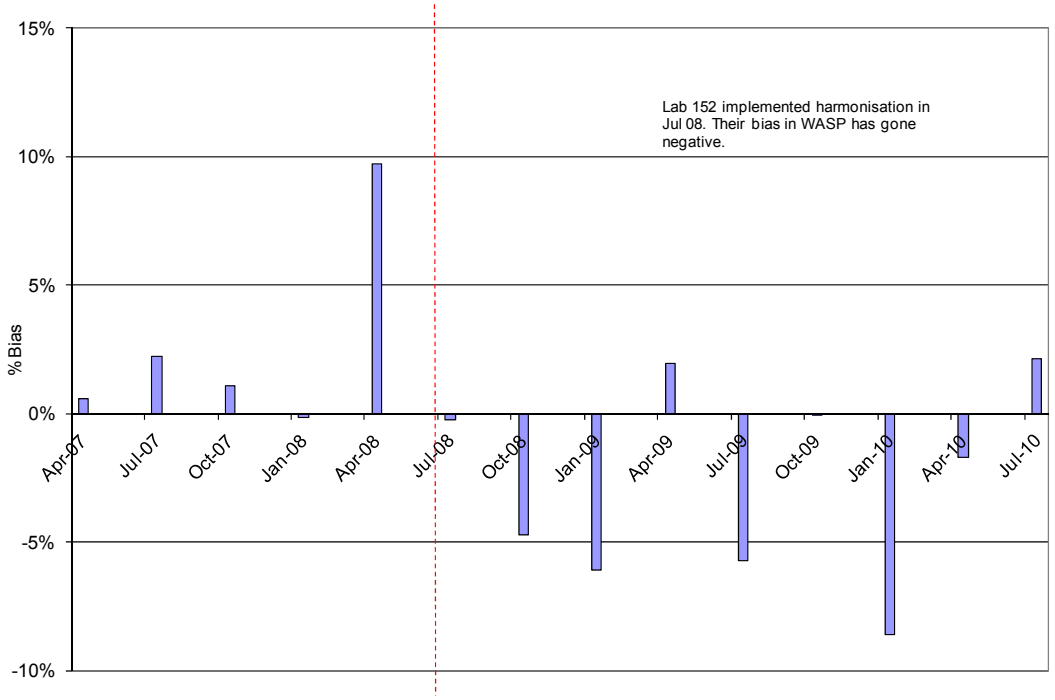


Figure 3-43 Lab 152 - Accuracy (% Bias) of WASP results relative to Spike value



3.5.4.3 Laboratory 922

Lab 922 performed consistently well before harmonisation, though with typically small negative bias (0 to -4%). They kept their original method of making up the calibration samples, with AEA's agreement.

Their performance, although still usually in the "Acceptable" range, got **worse** since harmonisation (Figure 3-44). Examination of the accuracy of their results shows that large negative bias became more prevalent (Figure 3-45).

The major changes that Lab 922 made were:

- more regular calibrations,
- more quality control samples,
- variation of reagent preparation procedure and
- mechanical shaking of samples.

The analyst commented that the samples were shaken rather than vibrated, on a tray shaker that shook them much more vigorously than the manual method previously used. However, the lab carried out an investigation into whether there was an effect from excessive shaking, and reported no effect. The analyst also commented that the reduction in recovery for the WASP samples in late 2009 and early 2010 was most likely due to the spiked tubes for these rounds having been stored for an extended period before analysis. They were also confident that none of the measures adopted in the course of harmonisation contributed to the problem.

Lab 922 has since become a different organisation – their last WASP round was Jan 2010.The new laboratory no longer uses the vibrating tray, but a combination of manual inversion and a vortex stirrer.

The new organisation now processes WASP samples as soon as possible after receipt. The new laboratory's WASP PI scores for Apr 2010 and Jul 2010 (not shown in the graph for Lab 922) were well within the "good" range.

Figure 3-44 Lab 922 Performance Index (green band = "good", yellow band = "acceptable")

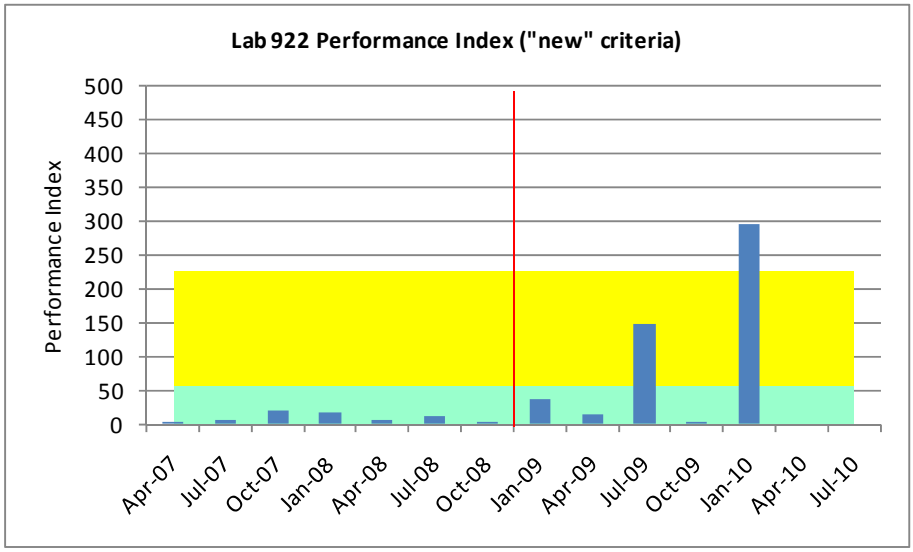
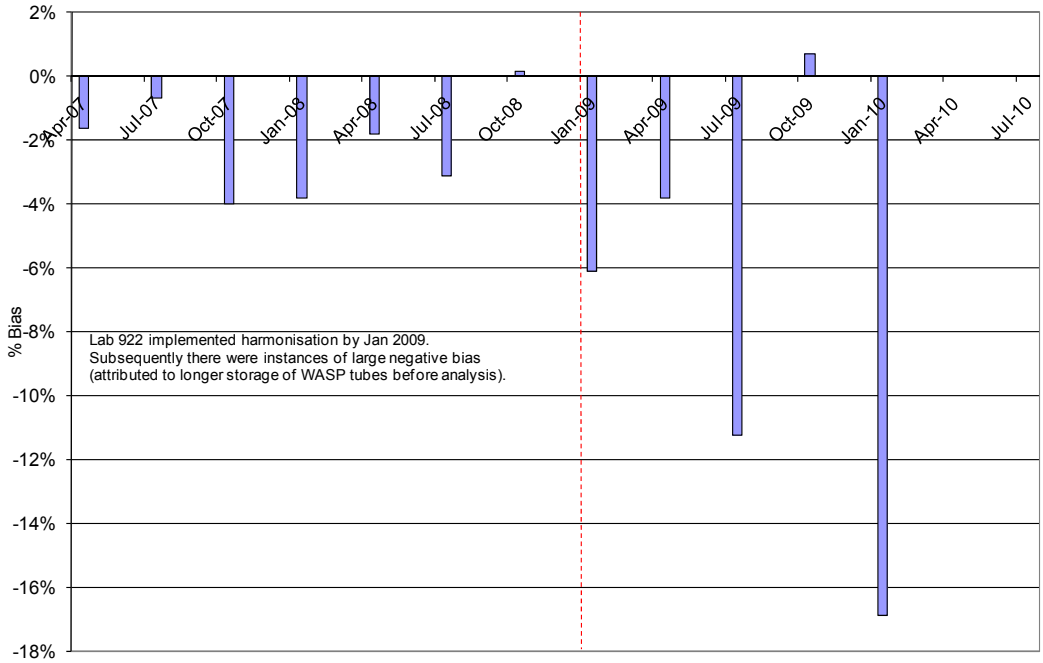


Figure 3-45 Lab 922 - Accuracy (% Bias) of WASP results relative to Spike value



3.5.4.4 Laboratory 346

Lab 346 was generally performing well in WASP before harmonisation in Oct 2008. Although their PI scores remain mostly in the “Good” and “Acceptable” bands, their scores are typically **worse** since harmonisation (Figure 3-46). Harmonisation appears to have introduced a tendency to negative bias (Figure 3-47).

Figure 3-46 Lab 346 Performance Index (green band = "good", yellow band = "acceptable")

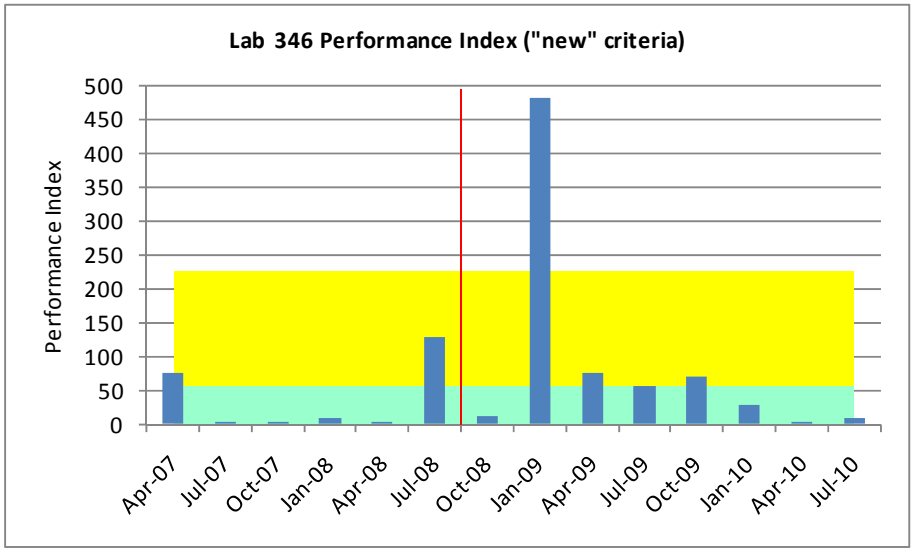
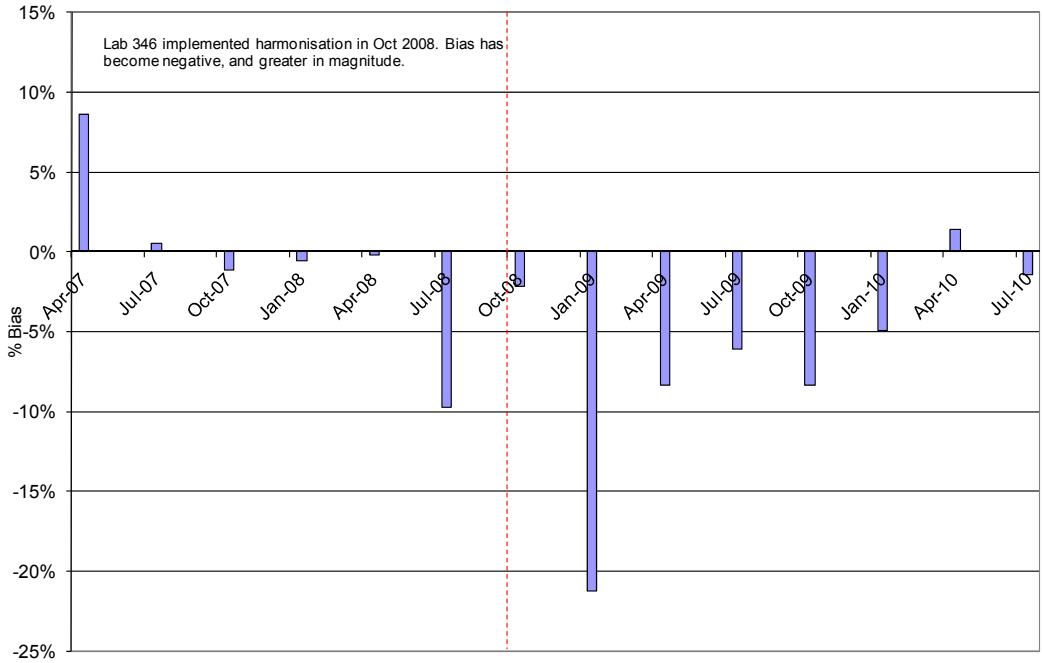


Figure 3-47 Lab 346 - Accuracy (% Bias) of WASP results relative to Spike value



Lab 346 told us that they needed to make few changes to implement the harmonised method. The changes were –

- Change to using mixed colour reagent rather than adding reagents separately
- The calibration standards were previously prepared using tubes with TEA-coated grids.
- Frequency of calibration (which was formerly every 3 months)

These changes would not be expected to be detrimental to their performance in WASP. It should also be noted that agitation was previously on a vibrating tray, but followed by manual shaking and inversion. It is possible that their new procedure (which does not involve any manual shaking or inversion) is resulting in less efficient extraction.

3.5.4.5 Laboratory 192

Lab 192 was performing well prior to harmonisation, with PI scores usually less than 50. Since harmonisation in October 2008, PI scores have been on average slightly higher (worse) although still well within the acceptable range (Figure 3-48).

Figure 3-48 Lab 192 Performance Index (green band = "good", yellow band = "acceptable")

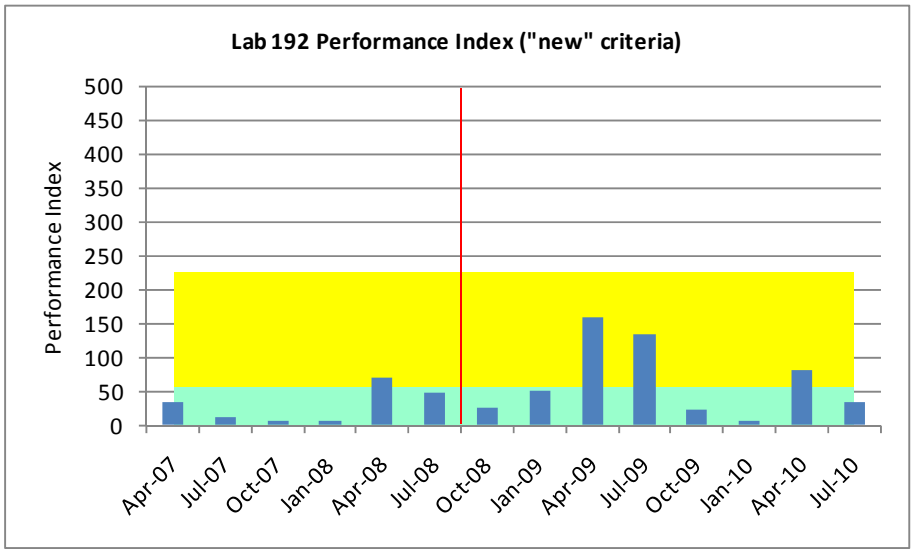
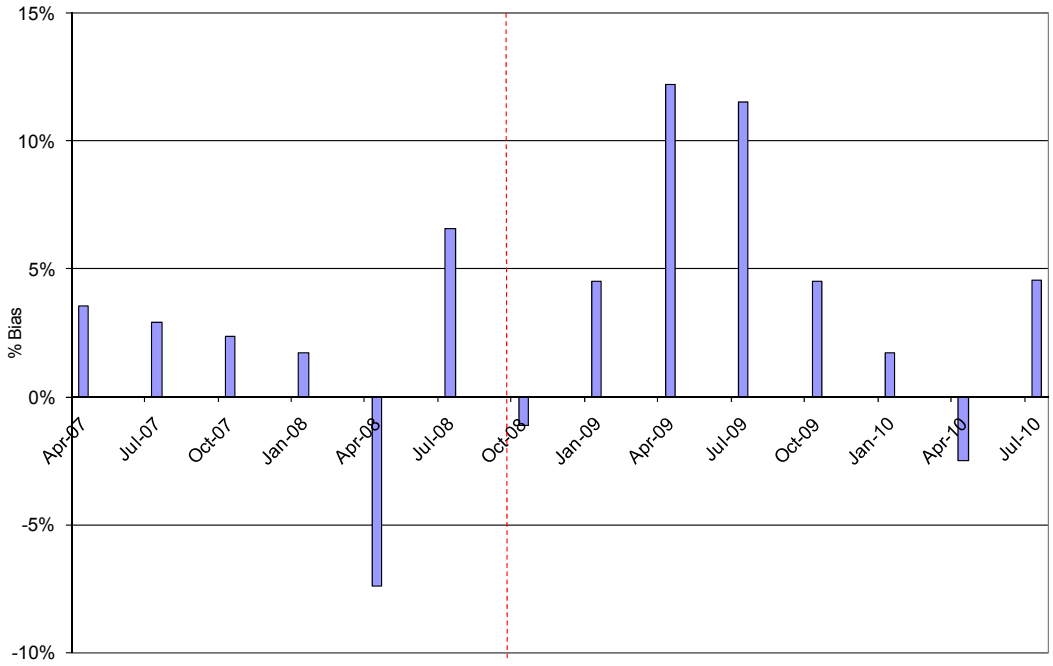


Figure 3-49 Lab 192 - Accuracy (% Bias) of WASP results relative to Spike value



This lab's analyses of the WASP tubes typically over-estimated: this appears slightly more pronounced after the harmonisation (Figure 3-49) compared to before. No information is available on the changes made by this laboratory in order to implement harmonisation.

3.5.4.6 Summary on Laboratories Whose Performance Worsened

Five laboratories have shown worse performance in WASP since harmonisation. These were laboratories 1045, 152, 922, 346 and 192. All were generally performing well (most PI scores in the "good" band) before harmonisation. Four of these (Lab 1045, Lab 152, Lab 922 and Lab 346) exhibited a similar pattern: harmonisation has appeared to introduce some negative bias, i.e. under-estimation of the spiking level on the WASP sample. PI scores remained mostly in the "good" and "acceptable" bands (**and therefore would be classified as "Satisfactory" under the current scoring system**) but were typically higher after harmonisation. One *possible* explanation is that the changes these laboratories have made to their methods of extraction *might* have resulted in slightly less effective release of the nitrite from the tube grids. However, this has not been investigated.

In the fourth case (Lab 192) the bias has become more variable, and often large and positive.

3.6 Information from the Field Intercomparison

The Field Intercomparison differs from WASP in that the diffusion tubes are prepared by the participating laboratories themselves, and are exposed at a real monitoring site rather than being artificially spiked. Examining results from the Field Intercomparison alongside those from WASP may reveal changes in diffusion tube performance resulting from preparation-related or exposure-related factors.

3.6.1 Changes in Accuracy Shown by the Field Intercomparison

The Field Intercomparison measures the accuracy of diffusion tubes relative to the reference method, i.e. the automatic analyser. The accuracy can be expressed in terms of the annual mean *standardised result*. This is the diffusion tube result divided by the reference measurement from the automatic analyser: the closer this value is to 1.00, the better the accuracy of the diffusion tubes.

However, diffusion tubes exposed outdoors are known to be affected by several sources of interference - in particular the following two:

- wind-induced reduction of the effective tube length
- in-tube photochemistry effects.

Both of these factors tend to cause over-read (positive bias). Therefore, in the case of the Field Intercomparison, the results would be expected to show an average positive bias. (As mentioned in Section 2, for consistency with typical practice in the UK, the diffusion tubes in this intercomparison are not exposed in any kind of wind shelters).

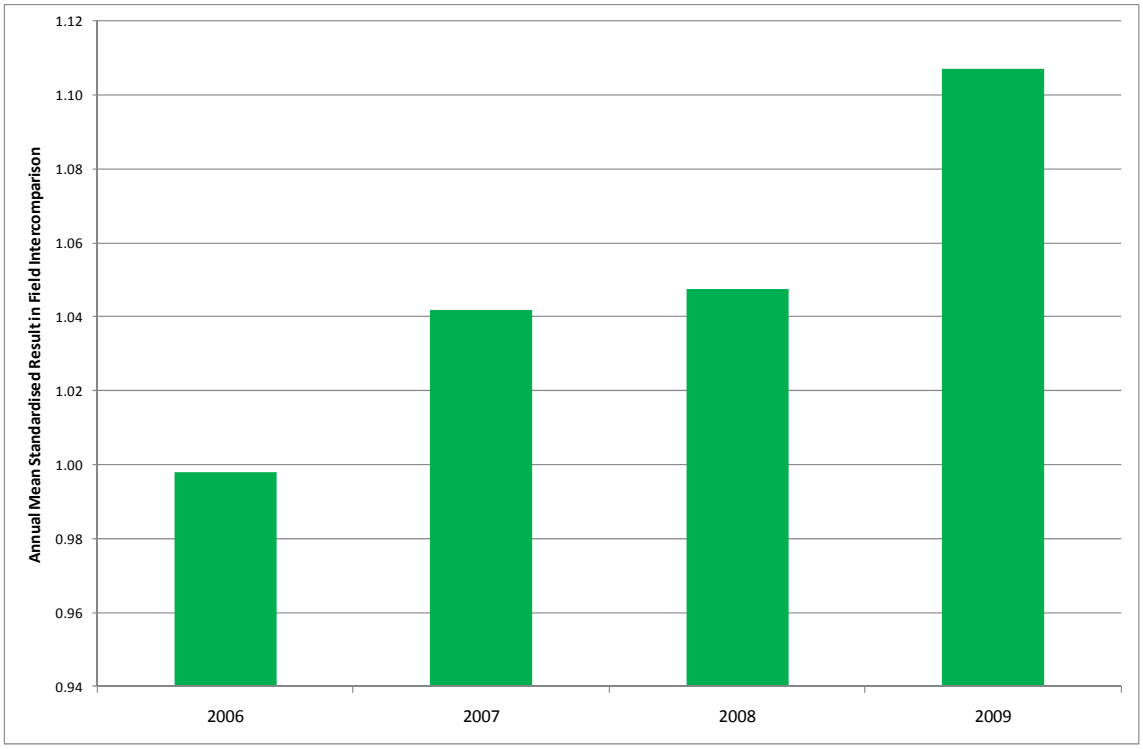
Also, harmonisation required some laboratories to make changes to how the ambient concentration was calculated from the mass of nitrite collected by the tube. These changes would have increased the reported ambient concentration by a small percentage – so a small increase in positive bias might be expected as a result of harmonisation.

Figure 3-50 shows how the annual mean standardised result in the Field Intercomparison (averaged over all participants) has changed between 2006 and 2009 (the most recent complete year's data available at the time of writing).

Because diffusion tube accuracy is affected by meteorological conditions (such as wind speed, UV light and temperature), bias varies considerably from month to month: for this reason, the annual mean is shown rather than individual monthly means. However, Figure

3-50 clearly shows that the bias has increased and become more positive year on year. The standardised result has increased from 0.985 in 2006, to 1.11 in 2009.

Figure 3-50 Mean Standardised Result from Field Intercomparison



The increase in positive bias over the years might at first be viewed as an indicator of worse performance. However, as explained above, positive bias is consistent with the expected behaviour of diffusion tubes exposed outdoors. This finding is consistent with that of a short-term field investigation of the effects of harmonisation, based on tubes from seven laboratories, and carried out in 2009[2].

It is possible that the increase in positive bias between 2008 and 2009 is partly due to the changes involved in harmonisation: however, the average bias shown by the tubes in the Field Intercomparison has been increasing each year since 2006 it is not possible to attribute the increase to this with any certainty.

It should also be noted that no such increase in bias is seen in the WASP results. This indicates that the increase is due to factors *other than* analysis.

3.6.2 Changes in Precision Shown by the Field Intercomparison

As highlighted above, improvement of precision was not a stated aim of harmonisation. However, data from the Field Intercomparison have been investigated here, to see whether any changes have occurred.

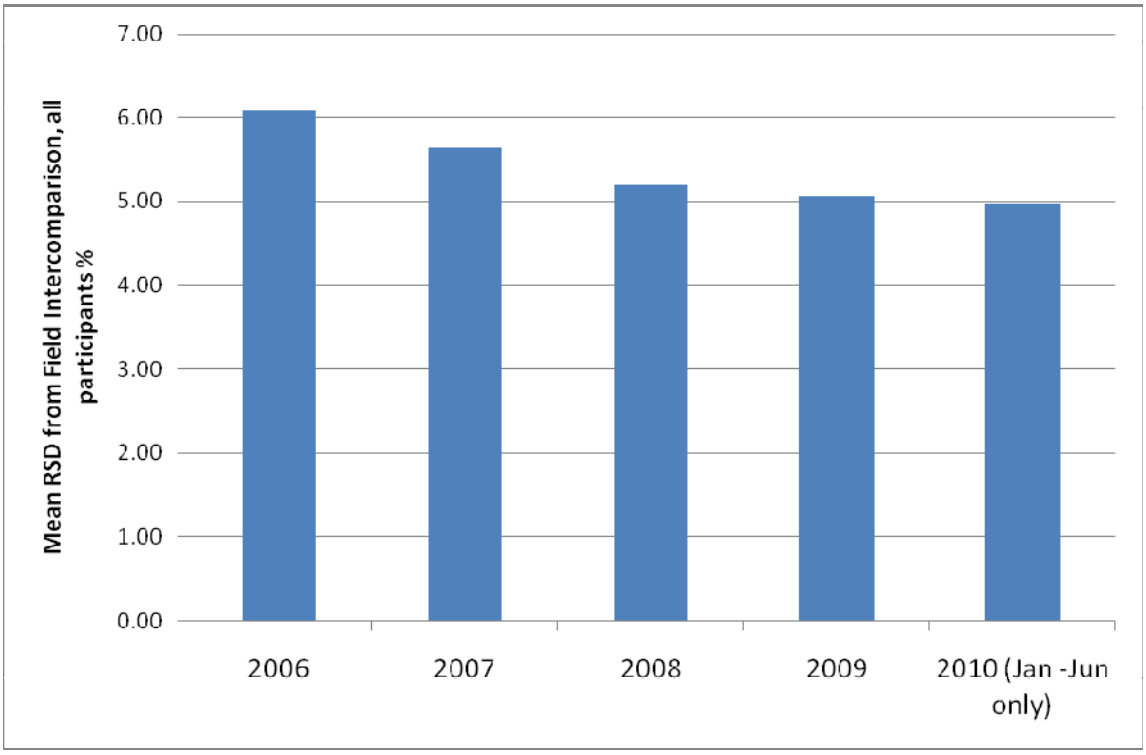
Diffusion tubes are exposed in triplicate in the Field Intercomparison, and the Relative Standard Deviation of a triplet of results (i.e. the standard deviation of the triplet, expressed as a percentage of the mean) is used as an indicator of precision. Figure 3-51 shows the annual mean RSD for all tube triplets in the Field Intercomparison, averaged over all participants, for years 2006 to 2009, and 2010 (January to June only : diffusion tube results from the Intercomparison were only available for the first half of the year, at the time of writing).

The mean RSD has improved slightly, from 6 % in 2006 to 5 % in 2009 (and 5% also in the first six months of 2010). The Field Intercomparison data indicate a modest improvement in

diffusion tube precision between 2006 and 2010: however, this improvement was happening prior to the implementation of harmonisation, and so cannot be attributed to harmonisation.

This improvement in precision is consistent with the findings of the earlier short-term field investigation of the effects of harmonisation on diffusion tube results from seven laboratories, referred to above[2].

Figure 3-51 Annual mean Precision of Triplicate Tubes in Field Intercomparison



No such improvement in precision is evident in the WASP results. This indicates that the improvement is not related to analytical procedures. One possible explanation (which has not been investigated) is that tube preparation has become more consistent.

3.6.3 Changes in Spread of Field Intercomparison Results

Table 3.1 shows the mean standardised result from the Field Intercomparison, for each participant, for years 2006 to 2009, and the first 6 months (Jan – Jun) of 2010 (at the time of writing, the full year's Intercomparison results for 2010 were not yet available). Also shown is the standard deviation, maximum, minimum, median, 90th percentile and 10th percentile for this parameter, together with the difference between the maximum and minimum, and the 90th and 10th percentiles (both of which are useful indicators of the spread of the results). All these parameters were calculated using the standard formulae in an Excel spreadsheet. Grubb's Test identified one outlier (the lowest value in the 2008 dataset, 0.65) but this was not removed. The 90th percentile, median, 10th percentile, and the difference between the 90th and 10th percentiles, are shown in Figure 3-52.

There is no indication that the spread of results has been decreased by the harmonisation: on the contrary, the spread of results in 2009 was greater than in either 2008 or 2007. However, it is noted that the gap between the median and 90th percentile (representing the spread of higher than average results) has decreased, while the gap between the median and 10th percentile result (representing the spread of lower than average results) has increased.

Table 3.1 Summary of Spread of Standardised Results in Field Intercomparison

	2006	2007	2008	2009	2010 (Jan-Jun only
Arith. Mean Standardised Result (SR) - all participants	1.00	1.04	1.05	1.11	1.16
Sample size n	23	23	23	22	22
Std. Deviation of SR's	0.14	0.11	0.15	0.18	0.13
Max SR	1.21	1.22	1.22	1.38	1.35
Min SR	0.67	0.86	0.65	0.66	0.85
Max - Min	0.54	0.36	0.57	0.71	0.50
Median SR	1.04	1.03	1.09	1.18	1.19
90^{th} %ile SR	1.15	1.16	1.20	1.28	1.28
10^{th} %ile SR	0.87	0.93	0.89	0.94	0.96
90^{th} %ile – 10^{th} %ile SR	0.28	0.23	0.31	0.34	0.31

Note: the values in the tables above are rounded to 2 decimal places, while Figure 3-50 shows the values exactly. This has resulted in some apparent discrepancy.

Figure 3-52 Median, 90^{th} and 10^{th} percentile of mean standardised result. Also difference between 90^{th} and 10^{th} percentile (shaded area).

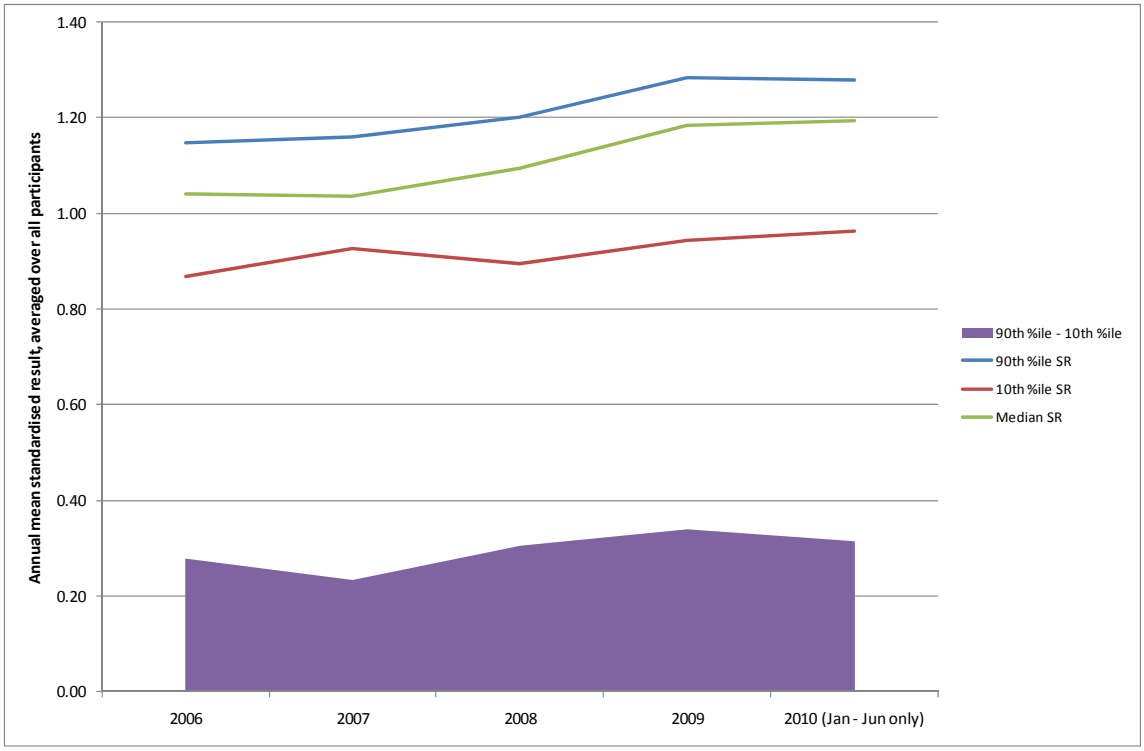
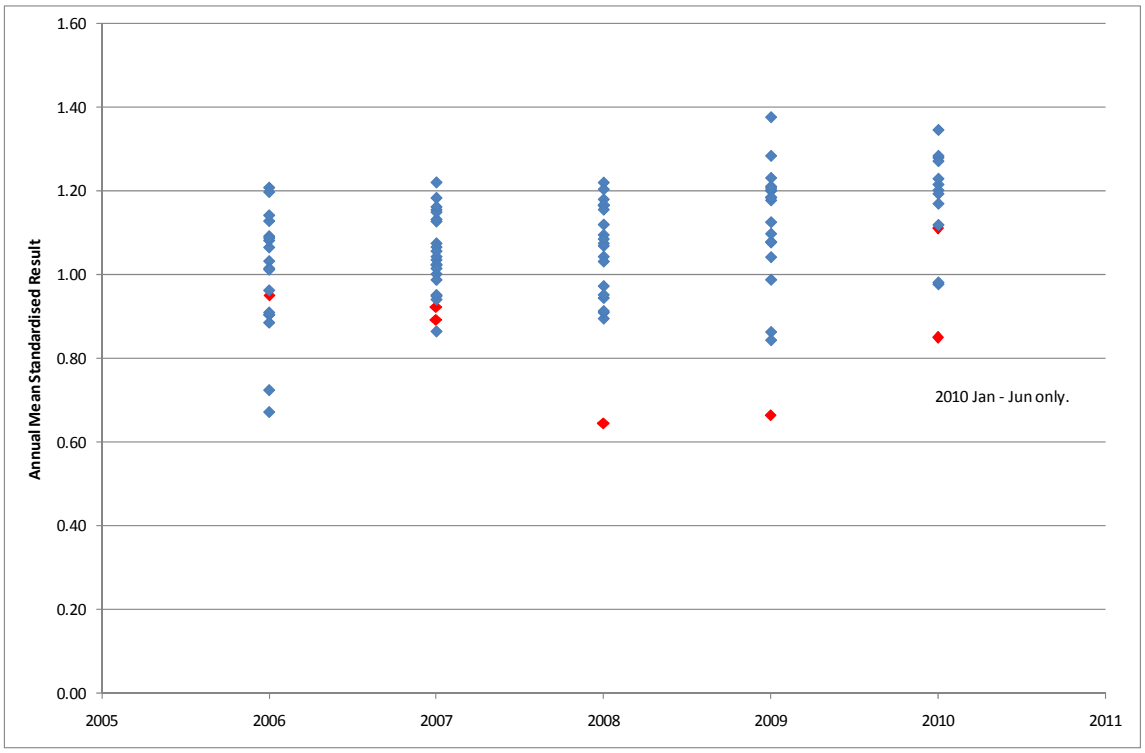


Figure 3-53 shows the annual mean standardised result from the Field Intercomparison, for each participating laboratory. A minimum data capture of 9 months has been applied (or 5 out of the 6 available months for 2010). Red markers indicate annual results for which average precision was poor. The criterion used here for "poor precision" is that the annual mean RSD of the triplicate tube results was greater than 10%.

In the most recent years, there appears to be a link between large negative bias and poor precision. One laboratory had unusually low Field Intercomparison results – and poor precision - in both 2008 and 2009. This was not reflected in its WASP results, indicating that in this particular case the problem was unlikely to be an analytical issue. At the upper end of the range, the laboratory with the highest standardised result in 2009 is Lab 1056 - one of the laboratories whose WASP results in 2009 showed a tendency to positive bias. This appears to have been resolved in 2010.

Figure 3-53 Spread of Annual Mean Standardised Results in Field Intercomparison



3.7 Data from the Combined Bias Adjustment Factor Database

As explained in section 2.3, many Local Authorities carry out co-location studies at their own monitoring sites. A national database of the results of Local Authority co-location studies is compiled and updated by Air Quality Consultants, and is available via the internet, and can be downloaded from http://laqm.defra.gov.uk/bias-adjustment-factors/national-bias.html .

This database contains not only the bias adjustment factors calculated from the individual co-location studies, but "combined" bias adjustment factors for all studies using the same type of diffusion tubes, calculated by orthogonal regression.

The database also contains information on the precision of the diffusion tubes used in each study. This is classified as "Good" (i.e. the RSD of the tube triplets was less than 20% in at least eight out of twelve monthly exposure periods, and the annual mean RSD was less than 10%) or "Poor" (i.e. the above criteria are not met). It should be noted that these precision

criteria are not particularly stringent, given the mean precision of around 5% found by the Field Intercomparison.

The data held in this database were used to investigate whether the spread of combined bias adjustment factors had changed, between 2006 and 2009 (the most recent year for which data were available).

Figure 3-54 shows the spread of combined bias adjustment factors contained in this database, for each year 2006-2009 (2010 data were not available at the time of writing). The bias adjustment factor (BAF) is the reciprocal of the standardised result, so in Figure 3-54 the points towards the top of the graph indicate instances of under-read, and those towards the bottom indicate instances of over-read. In some cases, *all* the co-location studies on which the BAF was based were indicated on the website as having had "poor" precision (as defined above). These are shown by red markers. In a further set of cases, at least half of the co-location studies on which the combined BAF was based had "poor" precision: these are shown by amber markers. The "red" and "amber" points are considered to be less reliable because of the poor precision of the bulk of the studies on which they are based. However it should be noted that poor precision is not necessarily due to poor laboratory performance – siting or exposure-related factors can contribute.

Table 3.2 shows the mean BAF for each year, and the range from minimum to maximum; with and without the poor-precision "red" and "amber" BAFs.

(Grubbs' test was applied to the maximum and minimum values for each year: none were identified as genuine outliers.)

Figure 3-54 Spread of Combined Bias Adjustment Factors

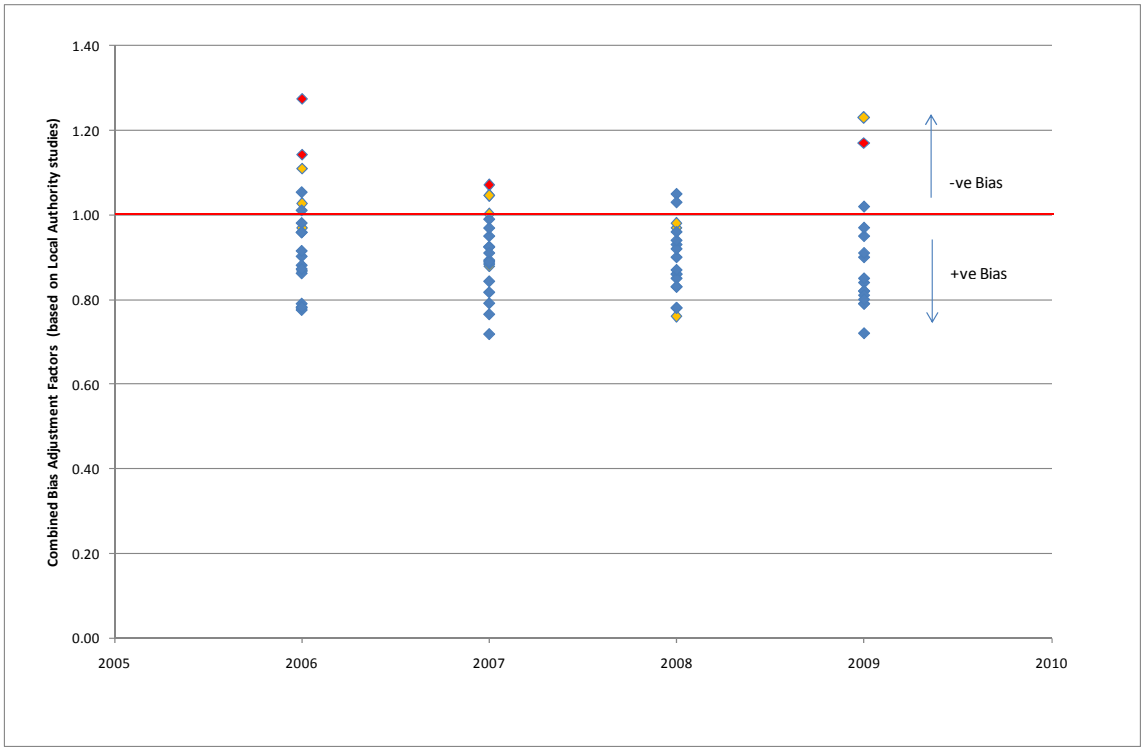


Table 3.2 Summary of Spread of Combined BAF's

	2006	2007	2008	2009
Arith. Mean combined BAF - all participants	0.95	0.91	0.89	0.90
Sample size n	21	19	18	16
Mean combined BAF – excluding those based mainly or entirely on "poor" precision trials	0.89	0.88	0.89	0.86
Max combined BAF	1.28	1.07	1.05	1.23
Min combined BAF	0.78	0.72	0.76	0.72
Max – Min	0.50	0.35	0.29	0.51
Max – min (excluding excluding those based mainly or entirely on "poor" precision trials)	0.28	0.27	0.27	0.30

Overall, the spread of combined BAFs in this database for 2009 is greater than that for the previous two years, and is similar to the range for 2006. This is the case whether or not the values shown as "amber" and "red" in Figure 3-54 are included. Therefore, these data do not appear to indicate that the harmonisation has reduced inter-laboratory variation in bias.

However, like the results from the Field Intercomparison, these data do show an apparent association between poor precision and diffusion tube under-estimation (here shown by high bias adjustment factors), in that most of the highest combined BAFs (except in 2007) were the "red" and "amber" points.

4 Conclusions and Recommendations

4.1 Conclusions from WASP

Please note that the "Good", "Acceptable" and "Poor" ranges referred to below relate to the proposed scoring system based on the Performance Index, which HSL were planning to adopt at the time the report was prepared. Under the current Z-score based scoring system, the "Good" and "Acceptable" categories would together be classed as "Satisfactory".

1. The quarterly Performance Index (PI) score, averaged over all participants in the WASP scheme who were involved in the harmonisation exercise, does not appear to show any improvement specifically resulting from the harmonisation (implemented by Jan 2009).

2. The average Rolling Performance Index (which is based on the best four of the last five rounds) has improved since the early WASP rounds, and this improvement appears to be continuing. This indicates that typical performance is improving. However, there is no specific change associated with the timing of the harmonisation.

3. There is no indication that harmonisation has improved precision in WASP. However, this was not a specific aim of harmonisation.

4. Apparently random outliers are still occurring in the WASP data. When very poor scores are obtained in WASP they are now usually low outliers rather than high outliers. Low outliers appear to be randomly distributed among the participants, and not confined to any particular laboratories.

5. The spread of WASP results appears to have decreased slightly over the period being investigated - indicating improved agreement between participants - but the improvement appears to be gradual, rather than specifically linked with the timing of the harmonisation.

6. Harmonisation appeared to *improve* the performance of the worst performers in WASP. Four participating laboratories that had previously been performing relatively poorly compared to most other participants, showed clear improvement in WASP following harmonisation. In two cases this was due to the reduction of large negative bias. In the third case, it was due to the reduction of large positive bias. The fourth had previously exhibited somewhat variable performance, which improved. (It is not assumed that the improvement in performance is necessarily due to harmonisation, as other changes such as staff training, or upgrading laboratory equipment, may also be involved).

7. Four laboratories that were performing relatively well prior to harmonisation, with PI scores typically in the "good" band (equating to Z-scores within ± 1) continued to perform well once harmonisation had been completed: there were no obvious changes in their performance.

8. Seven laboratories, whose performance was variable prior to harmonisation, showed no clear change. Their performance remained variable after harmonisation. Like the "good" performers mentioned above, these laboratories continued to have occasional "poor" rounds, i.e. PI scores in the "unacceptable" range: these equate to Z-scores outside the range ± 2, which would be classed as "unsatisfactory" under the current WASP scoring scheme. When this happens, it is usually due to substantial under-read of one or more of the samples, and the reasons should be investigated.

9. Five laboratories have shown ***worse*** performance in WASP since harmonisation. All were generally performing well (most PI scores in the "good" band) before harmonisation: subsequently, PI scores remained mostly in the "good" and "acceptable" bands, but were typically slightly higher (worse). Four of these exhibited a similar pattern: harmonisation has appeared to introduce some negative bias, i.e. under-estimation of the spiking level on the WASP sample. In one case, this was attributed to the WASP tubes being stored for a longer period before analysis, and procedures were changed to remedy this. Two of the four laboratories reported issues or concerns regarding extraction: it is possible that the changes these laboratories have made to their methods of extraction have resulted in slightly less effective release of the nitrite from the tube grids – this should be investigated.

10. It was noted that even the consistently good performers in WASP occasionally had a poor round, i.e. one with a PI score in the "unacceptable" range. When this happens, it was usually due to substantial under-estimation of one or more of the samples, i.e. a low outlier. It may be worth investigating the reasons for this.

4.2 Conclusions from the Field Intercomparison

1. The bias (relative to the reference chemiluminescent analyser) of the diffusion tubes exposed in the Field Intercomparison has been increasing year-on-year since 2006. The tubes in this study now typically show positive bias (i.e. over-read). This is consistent with the known sources of interference affecting diffusion tubes exposed in ambient air. This increase in bias is thought to be related to factors other than analysis, because no corresponding increase has been observed in the WASP results.

2. The precision of the triplicate tubes in the Field Intercomparison has improved by around 1% between 2006 and 2009. As the WASP data show no improvement in precision, this is not thought to be analysis-related, but linked to some other factor. One possibility is that consistency in the preparation of diffusion tubes has improved.

3. As of July 2010, the Field Intercomparison has not shown any improvement in inter-laboratory agreement: on the contrary, the spread of results in 2009 was greater than in 2007 or 2008.

4. Data from the Field Intercomparison appear to show an association between large negative bias, and poor precision. This may be due to the presence of low outliers in some datasets, and the reasons for these should be investigated.

4.3 Conclusions from Analysis of Combined Bias Adjustment Factors

5. Data from Air Quality Consultants' database of combined bias adjustment factors up to 2009 (based on collation of UK Local Authority co-location studies) has not shown

improvement in inter-laboratory agreement (on the basis of the range of combined bias adjustment factors).

6. Data from the above database do appear to show an association between large negative bias, and poor precision, as was the case for the Field Intercomparison data.

4.4 Recommendations

- The four laboratories whose WASP results began to show negative bias after harmonisation are a cause for concern and further investigation is recommended. One aspect of diffusion tube analysis which may not have been satisfactorily optimised in all cases is extraction (including ensuring that the extracted sample is homogeneous). This should be investigated further.

- Occasional, apparently random instances of poor performance in WASP continue to occur. When these happen, they are more commonly due to low (rather than high) outlying values. The reasons for these should be investigated where appropriate.

- Data from the Field Intercomparison, and from the database of nationwide combined bias adjustment factors managed by Air Quality Consultants, also appear to show an association between under-estimation and poor precision, in diffusion tube results. Field Intercomparison data indicate that this may be due to occasional low outliers. The reasons for these should be investigated.

- QA/QC of diffusion tubes used in LAQM should continue, with advice being provided to laboratories where appropriate.

5 Acknowledgement

The authors would like to thank the participating laboratories for their positive approach and co-operation throughout the harmonisation process, and for the feedback they provided for this report.

6 References

1. Defra Working Group on Harmonisation of Diffusion Tubes “Diffusion Tubes for Ambient NO_2 Monitoring: Practical Guidance for Laboratories and Users”. Report to Defra and the Devolved Administrations, AEA report number AEA/ENV/R/2504, Feb 2008.

2. A Loader, J Targa, T Clark, P Willis, S Winne “Field Investigation of Inter-Laboratory Variation in NO_2 Diffusion Tube Measurements 2009”. Report to Defra and the Devolved Administrations, AEA report number AEA/ENV/R/2936, May 2010.

The Gemini Building
Fermi Avenue
Harwell
Didcot
Oxfordshire
OX11 0QR

Tel: 0870 190 1900
Fax: 0870 190 6318

www.aeat.co.uk