# QUESTIONABLE PRACTICES IN THE ORGANIC LABORATORY: PART II

Joseph F. Solsky
Chemist
US Army Corps of Engineers (CENWO-HX-C), 12565 W Center Road, Omaha, NE 68144

## ABSTRACT

During recent environmental laboratory audits conducted by the USACE, certain ‘questionable practices’ have been observed, especially in the organic analysis areas.

Most people have a relatively good idea of what constitutes a fraudulent activity today. The concepts of ‘dry-labing,’ ‘peak shaving,’ ‘peak enhancing,’ or ‘time-traveling’ are well understood. These practices clearly involve the deliberate manipulation and/or alteration of data, often to achieve or meet method QC criteria. Unfortunately, these practices are still being observed today. In addition, there are a new group of ‘questionable practices’ now being observed that often involve the selective exclusion of data to achieve or meet method QC criteria.

Examples of some of these practices include the following: (1) Dropping points during initial calibration to meet method criteria. (2) Reporting very tight QC performance ranges when actual lab control charts show a significantly wider range. (3) Dropping points to achieve a lower Method Detection Limit (MDL). (4) Performing tunes by picking the scan or series of scans that will meet the desired criteria after the original tune had failed. (5) Performing initial calibration curves but never verifying that the peaks used for the calibration actually represented the target analyte.

These practices are often described as ‘the common approach used by everyone,’ yet when described to people within EPA (e.g., the MICE Hotline), the clear response is that these approaches were never intended within the context of SW-846, although not explicitly addressed nor prohibited.

## INTRODUCTION

The US Army Corps of Engineers (USACE) currently executes remedial and compliance activities under several environmental regulatory programs. The analytical testing of various environmental samples is often a significant part of these activities. The data must be produced by a process or system of known quality to withstand scientific and legal challenge relative to the use for which the data are obtained. To give the USACE programs the greatest flexibility in the execution of its projects, the SW-846 methods, as published by EPA, are generally the methods employed for the analytical testing of environmental samples. These methods are comprehensive and flexible and can be readily adapted to individual project-specific requirements. As stated in the Final Rule that incorporated the Third Edition of SW-846 (and its updates) into the RCRA regulations, this appendix is required to be used for certain activities in the RCRA program. In other situations, this EPA publication functions as a guidance document setting forth acceptable, although not required, methods to be implemented by the user, as appropriate, in satisfying RCRA-related sampling and analysis requirements.

During recent laboratory audits conducted by the USACE, certain 'questionable practices' have been observed, especially in the organic analysis areas. Prior to project execution, the USACE may conduct a review of the laboratory that was proposed for use on that specific project. This review typically consists of three phases: (1) documentation review; (2) analysis of Performance Evaluation (PE) samples; and (3) on-site laboratory audit. Additional follow-up audits can also be conducted. These 'questionable practices' have been noted during all phases of these laboratory reviews.

The concepts of 'dry labbing', 'peak shaving', 'peak enhancing', or 'time traveling' are well understood. These practices clearly involve the deliberate direct manipulation and/or alteration of data, often to achieve or meet method QC criteria. Laboratory professionals clearly recognize these practices as inappropriate since no professional reason exits to employ them other than to meet specific contractual requirements and avoid potential penalties. There is no technical basis that can justify the use of these practices. The impact on data usability must be determined on a project by project basis. Unfortunately, these practices are still being observed today. When fraud is detected in conjunction with USACE projects, the Corps is attempting to separate any criminal/civil charges from the actual impact of the fraud on data usability (e.g. to separate legal from technical issues).

As the nation moves away from the use of strict method protocols to a more performance based approach, the laboratories will have more discretion as to how methods are actually implemented. This will allow the laboratory community to take faster advantage of new technologies to cut costs and improve data quality. This move will place pressure on the laboratory community to employ knowledgeable experts to properly implement these newer technologies in a scientifically justifiable manner and to provide the enhanced documentation that will be needed. Current market over capacity has caused bidding wars and corner cutting. This move will place pressure on the regulator community to properly define what a performance based measurement system is and how its quality should be defined. This move will place pressure on the buyer of analytical services to better define the Data Quality Objectives (DQOs) such that the appropriate data can be obtained for any given project at a fair and appropriate cost. At the present time, issues exist in all these areas that can and are compromising data quality.

During this transition, USACE is observing a new group of 'questionable practices'. Many of these practices involve the selective exclusion of data to achieve or meet current method QC criteria rather than the direct manipulation of any single data point.

## QUESTIONABLE PRACTICES

The first example of such 'questionable practices' involves laboratory documentation, including Quality Control Plans and Standard Operating Procedures (SOPs), that do not accurately reflect what the laboratory actually does. Many of these plans contain statements that are misleading, in error, or simply incomplete. These laboratory documents are often directly incorporated into project specific Quality Assurance Project Plans (QAPPs) or Work Plans. Often, these laboratory documents are not carefully read or reviewed before incorporation. They should be. Do misleading, erroneous, or incomplete statements justify these practices?  Probably not.

The second example of such 'questionable practices' involves establishing initial calibration curves. Laboratories have been observed running six or more standards for methods that state 'a minimum of five points should be used to establish the initial calibration curve'. Points are then discarded, while maintaining a minimum of five calibration points, throughout the curve until the appropriate QC criteria can be met. No technical justification existed for the deletion of these points other than to meet the method QC criteria. This practice is often justified by using the rationalization that a 'better curve' is generated. Another reason heard is that 'everyone is doing it'. Points can only be rejected for inclusion in the curve if a known error was made or if a statistical evaluation indicates that the point can be discarded. When multiple target analytes are included in each calibration standard, it may become necessary to discard selected upper or lower points for individual target analytes. Points can be discarded at the upper end of the curve if the linear range of the detector has been exceeded. For these cases, the laboratory must dilute samples that exceed the highest point of the calibration curve. Points can be discarded at the lower end of the curve if the detector is not producing a response. For these cases, the laboratory quantitation limit must be adjusted accordingly. Under no other circumstances can points be discarded. If QC criteria cannot be met, the instrument system may be unstable or the calibration solutions may be incorrectly prepared. The 'best curve' is obtained when all valid points are included in the initial calibration curve.

The third example of such 'questionable practices' involves the verification of initial calibration curves through the use of continuing calibration verification (CCV) solutions. Laboratories have been observed averaging the % difference or % drift across all target analytes even when several of the target analytes exceed the criteria by a significant amount such that the average still meets the criteria as stated in the method. For example, when method 8021 is used, it is often difficult for laboratories to meet the CCV criteria for many of the gaseous target analytes. Method 8000B states the following: '…, if the average of the responses for all analytes is within 15%, then the calibration has been verified'. This language was chosen to make it easier for laboratories to implement this method when certain problem analytes, i.e. the gases in method 8021, marginally exceed the stated method criteria. It was never intended to allow the inclusion of obviously 'bad' data to make it 'acceptable'. Method 8000B goes on to say: '…, and the data user must be provided with the calibration verification data or a list of those analytes that exceeded the 15% limit'. If the QC criteria cannot be met, the instrument system may be unstable or the calibration verification solution may be incorrectly prepared.

The fourth example of such 'questionable practices' involves the reporting of acceptance ranges for laboratory control samples (to include surrogates). Laboratories have been observed reporting a very tight range for these QC samples on laboratory report sheets, indicating that they have good method control. However, an examination of actual control charts maintained by the laboratory shows a significantly wider range, if control charts are even available. This practice is often justified by using the rationalization that 'but the LCS was within the QC range, therefore, it must be okay'. Method 8000B stresses the importance of control charts to track laboratory performance. The ranges generated should then be compared to method established criteria. If a 'match' is not obtained, then the laboratory should consider modifying their method to improve its performance. Simply reporting data under this circumstance since the LCS 'met the method criteria' is unacceptable since it misleads the user of the data and misrepresents the laboratory's reported data quality. Control chart ranges must also be reasonable. The issue of control charts as related to what analytes need to be charted (all target analytes or just a subset), in what QC samples (LCSs, MSs, LCSs and MSs combined, etc.), at what spiking levels (action levels or mid-level), and appropriate recovery ranges (what would be considered a reasonable range for a given method) needs further clarification in the SW-846 methods. Many laboratories do not understand the significance of these charts and how to properly implement and use them.

The fifth example of such 'questionable practices' involves the reporting of wide matrix spike (MS) recovery ranges. This item is related to the fourth example as given above. Laboratories have been observed reporting very wide ranges for these QC samples on laboratory report sheets. However, an example of the actual ranges as derived in the laboratory shows a significantly narrower range. This practice is often justified by using the rationalization that 'by widening the ranges, less of our data is rejected'. No method is immune to all possible interferences and not all interferences can be predicted. Therefore, it is important to monitor for these effects. The purpose of the matrix spike (MS) is to see if a possible matrix effect is impacting the data quality. When the MS QC range is exceeded, clients would normally be contacted to see if data flagging is appropriate, sample(s) should be rerun, the method should be modified (i.e., add a clean-up step) to better deal with the interference, or a different method chosen that is not affected by the interference. Data users should not penalize a laboratory, or its data, due to the presence of reported potential matrix interferences. At the same time, laboratories should not flag all poor recoveries as possible matrix effects, especially in blanks, LCSs, etc. Good judgment should be used by all parties involved.

The sixth example of such 'questionable practices' involves the determination of the method detection limit (MDL). Laboratories have been observed running eight or more standards and then discarding points to achieve a lower MDL. No technical justification existed for the deletion of these points other than to achieve a lower MDL. This practice is often justified by using the rationalization that a 'better (lower) MDL' is generated. Points can only be rejected if a known error was made or if a statistical evaluation indicates that the point can be discarded. Under no other circumstances can points be discarded. The MDL study must be performed at the appropriate level with a reasonable recovery of the target analyte(s) noted. The 'best MDL' is obtained when all valid points are included. It appears that the industry is placing too much emphasis on this concept. Laboratories are being driven to report lower and lower levels of contaminants. Perhaps the industry would be best served by using the performance-based concept to demonstrate what a given method run by a given lab could actually 'see' (the concept of the MDL check sample). The issue of the 'not detected' target analyte has caused great confusion ('detection limit' versus 'quantitation limit' versus 'reporting limit').

The seventh example of such 'questionable practices' involves tuning a GC/MS detector. Laboratories have been observed performing tunes in an inconsistent manner, such as picking a single scan or a series of scans that meet the desired criteria. Single scans have been observed being used at various locations across the peak, including single points being used on the peak tail. The use of an average of two or more scans have been observed over various parts of the peak (front, tail, over apex), to even include more background scans than peak scans in the average. These various schemes would be used when the recommended approach (average of three scans over the peak apex minus a background scan) would fail the desired criteria. No technical justification existed for using these various approaches other than to meet the method QC criteria. This practice is often justified by using the rationalization that 'as long as a scan(s) can be found that passes, the instrument is in tune'. Different tune parameters may be needed to optimize instruments from a given manufacturer. However, a consistent approach must be used to evaluate whether the instrument is 'in tune'. A laboratory cannot simply pick and choose whatever scan(s) happens to meet criteria on any given day.

The eighth example of such 'questionable practices' involves the misidentification of GC/MS peaks during initial calibrations and during continuing calibration verifications. Laboratories have been observed performing these calibrations but never verifying the identity of the peaks observed. These systems can make errors in the identification of target analytes especially when more than one peak is present in the retention time window. As a consequence, laboratories can generate calibration curves for the wrong target analyte. This has been observed for certain Appendix IX compounds and for certain poor performing target analytes in methods 8260 and 8270. Instrument raw data must be reviewed by the analyst to ensure that all peaks have been correctly identified, all peaks are clearly visible and all peak shapes are appropriate for the target analyte being measured.

The ninth example of such 'questionable practices' involves performing continuing calibration verifications where the majority of the target analytes have missed their assigned retention time windows. Laboratories are performing unnecessary manual integrations to 'find' peaks that have missed these windows. This is very dangerous since peaks can be easily missed during the analysis of samples resulting in the reporting of false negative data. The SW-846 methods directly address retention time window criteria for the internal standards for the GC/MS methods but do not address any requirements for these windows for the target analytes. When such windows are missed, this should be a clear signal to the analyst that the system is out of control and corrective action is required. Such corrective action should include a system inspection along with repeating the initial calibration or updating the retention time windows for the target analytes.

Should the above 'questionable practices' be considered as examples of fraudulent activities? Some of the laboratories have described these practices as 'the common approach used by everyone', yet when described to people within EPA (e.g., the MICE Hotline), the clear response is that these approaches were never intended. Potential solutions might include the following: 1) More prescriptive methods (probably not) or more clearly written guidance? 2) Training for lab staff on GLP to include statistics? 3) Rethinking the way laboratory services are contracted for? 4) Collecting additional data from the laboratory for more detailed data validations? 5) The use a standardized data reporting format and better data validation software? 6) Etc.

## SUMMARY

The problems/issues noted above are very serious and directly impact on the usability of the data generated. Oftentimes the impact is equivalent to the impacts observed during past demonstrated cases of fraudulent data manipulation. How did we get to this point? There probably is no one single cause. One certain contributing factor is the price paid for these services. It is not uncommon to encounter projects that were bid low simply 'to get one’s foot in the Federal door'. Simply put, the price paid for these services was not sufficient to cover the costs of producing the product. The fault lies both with the laboratory community for bidding in this manner and the government for accepting bids based on low price only without considering the quality factor ('best value' procurement strategy). However, this is a free market economy. The age old adages 'let the buyer beware' or 'you get what you pay for' certainly apply here.

Another factor is the level of expertise that now exists at the laboratory level. Some laboratories have let go their most experienced staff since they could no longer 'afford' them. Many people feel that the computer attached to the instrument in use will give them the correct answer without additional thought. If anything, more expertise is needed to evaluate the larger magnitude of data moving through the laboratory and the complexity of today’s instrumentation. Laboratories should not be treated as black boxes. It is not uncommon for this author to visit a given laboratory and find that laboratory staff know very little about the fundamental chemistry of the method (or the software) in use. One common phrase heard often is 'but the method doesn't specify the approach to use'. This raises the question as to whether or not very prescriptive methods should be written. Yet each of the SW-846 methods typically state the following: 'This method is restricted to use by, or under the supervision of, analysts experienced in the use of gas chromatography/mass spectrometers, and skilled in the interpretation of mass spectra and their use as a qualitative tool'. Additional training of laboratory staff should be emphasized. Peer review of raw data should also be emphasized. Audit trails should be 'turned on' when available and reviewed on a regular basis.

More review of raw data would be encouraged. Most of the data generated within a laboratory is generated in an electronic format. Yet much of the data is still manually managed and reviewed. A greater emphasis should be placed on receiving data electronically and for the electronic screening/review of this data. To assist this process, standardized electronic data reporting formats should be used. Standard file formats have been developed for several of the instrumental methods that can transfer data electronically in a standard file format between an instrument, or its data station, and a laboratory LIMS system. However, this standard is not often used. No standard file format has been developed for the transfer of information from a laboratory, or its LIMS system, to the data user. The use of a common data dictionary along with a common file structure, such as that proposed by the Department of Energy Environmental Management Electronic Data Deliverable Master Specification (DEEMS) would be encouraged.

Certainly, other contributing factors are involved. The 'CLP' prescriptive mentality is still with us. Data validation is still often performed using a modified version of the National Functional Guidelines. This is not appropriate for the SW-846 methods and further emphasizes the prescriptive, rather than performance based, approach. The move to a performance based approach for the analysis of environmental samples is a welcome one. This move is also being greeted with uneasiness. The approach will place additional burdens on the laboratory community, regulator community, and the buyer of analytical services. Good communication will be very important to ensure that the needs of everyone involved have been met. The writers of the methods must work together with the users of the methods to minimize misunderstandings. It would be recommended that EPA revise Chapter One of SW-846 to more clearly describe this approach. The 'questionable practices' as described above are serious issues that must be resolved. Their timely resolution will give data users the confidence they need to make appropriate project decisions while at the same time using our tax dollars wisely.