EPA Responsive Summary to Peer Review and Other Major Comments on the September 2004 Draft Unified Guidance

prepared by Mike Gansecki, Work Assignment Manager
July 30, 2009

The first part of this document addresses the main comments provided by Dr. Dennis Helsel, Dr. Jim Loftis, and Dr. William Huber in their technical peer review of the 2004 draft Unified Guidance (UG) document. The summarized points were identified by the EPA Work Assignment Manager (WAM). The pages in parentheses reflecting the source of the reviewer comments are based on a compilation of comments and potential responses developed by the WAM on October 4, 2005. The EPA responses and summarized comments for each peer reviewer can be found on the following pages of this document:

The second part of this document addresses the main comments provided by 19 Other State, EPA Regional, Industry, Statistician and Consultant reviewers of the 2004 draft UG document. A number of reviewers had also participated in the workgroup which helped to develop the guidance. The pages in parentheses reflecting the source of the reviewer comments are based on a second compilation of reviewer comments and potential responses developed by the WAM on April 5, 2005.

The EPA responses and summarized comments for each State or individual reviewer can be found on the following pages of this document:

PART I. PEER REVIEW COMMENTS AND EPA'S RESPONSE TO THOSE COMMENTS

Peer Reviewer — Dr. Dennis R. Helsel, PhD. U.S. Geologic Survey

- **<u>UG Structure and Organization</u> — Dr. Helsel recommends condensing and shortening the document. Equations should be correctly attributed to authors and explanatory footnotes/references should be provided in all UG tables. Consistent terminology should be used (p.2).**

EPA Response: We agree that the Unified Guidance should be shortened and condensed in places, especially where there was duplication of material or language, such as in the 2004 draft chapters 16 and 17. While duplicative material was removed, the overall size of the March 2009 guidance document is slightly larger than the reviewed version, since considerable new material was added. All equations were checked and references to authors provided. The tables in the 2009 Unified Guidance now contain reference footnotes for published tables with permissions secured and explanations of relevant variables. It is noted, however, that the vast majority of tables in the current draft were generated by the statistician Contractor and are not copies of existing published tables.

- **<u>Non-Detect Issues [Chapter 10]</u> — Look into Kaplan-Meier and Helsel-Cohn methods for dealing with multiple detection limits for non-detects (pp.6, 20). Evaluate the UG document and decide on consistent recommendations for direct substitution and other ND% limitations (p.25 and other comments). Consider use of Helsel's regression method for non-detects and censored probability plot tests (pp.12, 20). He questions discussions of MDL, QL, etc. and recommends a more consistent approach (p.16).**

EPA Response: We agree that the treatment of non-detects in the Unified Guidance needed to be revised. Certain methods in the 2004 draft guidance were dropped (e.g., the modified Aitchison), while others were added (e.g., Kaplan-Meier, Helsel-Cohn robust regression on order statistics (ROS), and parametric ROS). On the other hand, EPA disagrees with Helsel that all direct substitution methods are simply a matter of 'fabricating' data. While the 2009 Unified Guidance places greater emphasis on newer methods for handling and adjusting for non-detects, a distinction is made between non-detect treatments for parameter estimation vs. statistical test performance. The methods recommended by Helsel are generally superior to simple substitution for estimating the true underlying mean and standard deviation. However, at least two published studies suggest that simple substitution methods perform adequately in maintaining the statistical performance (Type I and II error rates) of prediction limits and possibly control charts.

Helsel's major point is the use of a variety of techniques to correctly handle non-detect values, especially when multiple detection limits are involved. Even using these techniques, potential problems can occur either because of the presence of trends over time (perhaps due to changing

analytical methods) or with certain facility data suggesting that both detects and non-detects vary with the nominal detection limit (even in cases of identical analytical technique).

On balance, Chapter 15 of the 2009 Unified Guidance describes a consistent framework for simple substitution and when it is appropriate, as well as other techniques that can be used in place of direct substitution. The Unified Guidance was also reorganized so that adjustments for non-detects occur in various chapters, immediately after methods recommended for cases with no non-detects. For example, the chapter on two-sample tests was revised so that the Tarone-Ware test, a variety of Gehan's test (itself an extension of the Wilcoxon rank-sum test for use with non-detects) is introduced after the t-test and Wilcoxon methods. Thus, the guidance provides a corrective process for non-detects after a set of guidance methods that describe the 'complete sample' case (i.e., no non-detects).

Proper statistical reporting of non-detects is a somewhat difficult area. EPA agrees that 'uncertain' measurements (e.g., estimated values below a quantitation or reporting limit) should not typically be used *by themselves* for assessing compliance in statistical testing. Use of estimated values should instead be limited to testing of larger data sets where they can usefully provide added information about the statistical properties of the data. The Unified Guidance has been reviewed to make sure that consistent recommendations are made regarding the reporting of non-detects. However, due to the historical changing nature of detection and quantitation limits, there may be sites where a statistical treatment must be made of a mixture of estimated values and 'less than' values. A general discussion of MDLs, PQLs and other uncertain data is found in Chapter 6 of the 2009 document (p 6-36f).

- **<u>Non-parametric tests for correlation and trend [Chapters 9 & 14]</u> — Helsel provides recommendations for other tests (27). The Kruskal-Wallis test should be added to ANOVAs (pp.3, 14, 15, 17).**

EPA Response: We agree that the Kruskal-Wallis test should be re-introduced into the Unified Guidance for identifying trends; it is located in Chapter 14 of the Unified Guidance as the Seasonal Mann-Kendall test. For identifying spatial variability, the Kruskal-Wallis test is mentioned as an alternative procedure on page 13-6 of Chapter 13. Finally, it is also formally recognized as an ANOVA detection monitoring test option in Chapter 17 (p. 17-9f), while the Mann-Kendall test is used to test for trends (p.17-30f) The Unified Guidance also provides a description and references to additional non-parametric tests for correlation (similar to Kendall's tau-b) and trend (Theil-Sen trend line) in Chapter 17.

- **<u>Power analysis</u> — Remove or put detailed non-central t equations and discussion in an appendix. Provide simpler analyses for the reader (pp.1, 5, 22).**

EPA Response: We agree that the discussion of the non-central t-distribution should be reorganized as supporting technical material and was placed in an appendix (Appendix C— Technical Appendix). Descriptions of power and its relevance to users — including easily understood examples — were added to the Unified Guidance. A new method using R-script

allows for direct calculation of κ-factors and reference power (Chapters 13, 19 and Appendix C of the 2009 guidance).

- **Special Problem Areas**

    - **Aitchison's method and simulations (pp.18, 21, 28-30);**

EPA Response: The simulation study regarding the modified Aitchison method was reviewed and dropped from the 2009 Unified Guidance. In the Chapter 15 discussions, however, it is suggested that the modified delta Aitchison method may be useful in the presence of certain bimodal distributions (p.15-5f).

    - **Limitations of Cohen's method (p.19);**

EPA Response: We agree that Cohen's original method based on a single detection limit was too limiting. It is still retained in Chapter 15, but greater emphasis is placed on the two methods that can incorporate multiple detection limits and are more robust with non-normal data.

    - **Use of Land confidence intervals for lognormal data (pp.4, 26);**

EPA Response: We agree that Land's upper confidence limit for lognormal data can often be unrealistically high. However, no comparable problem exists for the Land lower confidence limit, the kind recommended by the Unified Guidance for use in compliance/assessment monitoring. Land's upper limit would only be used during corrective action in assessing clean-up or remediation success. The 2009 Unified Guidance now references other methods for constructing upper confidence limits on skewed data, recommended by reviewer Dr. Anita Singh (Chapter 21, p.21-9).

    - **Spearman's and Mann-Kendall tests (pp.10, 27);**

EPA Response: Since we agree that the Spearman and Mann-Kendall trend tests were somewhat duplicative, the 2009 Unified Guidance dropped Spearman's test, especially since the Sen's slope method and the Theil-Sen trend line more closely correspond to the Mann-Kendall test.

    - **Some criticism of Chapter 6 normal/lognormal special study (pp.12, 14);**

EPA Response: The study of the performance of prediction limits when the underlying model (normal/lognormal) is mis-specified, was reviewed for accuracy and completeness. Most reviewers, however, agreed that the major conclusions of the study seem sound. The details of the study are now located in Appendix C, with a general summary in Chapter 10, p. 10-7. In general, the 2009 guidance recommends that explicit distributional testing be conducted for all but the smallest of data sets, and that no single distribution be treated as the default if proper goodness-of-fit testing can reasonably be applied.

- **Serial vs. Temporal Correlation (p.15);**

EPA Response: We agree that the 2004 draft guidance chapter on temporal correlation should be reorganized so that serial correlation and temporal differences are not intertwined in presentation, but are instead separated by topic. Further, we split Chapter 9 of the 2004 draft document into distinct chapters on spatial and temporal variability (Chapters 13 and 14 in the 2009 guidance). Chapter 14 identifies a number of distinct types of temporal variation (p. 14-1) and suggests potential adjustments for each.

- **Small Sample Study (pp.16-17)**

EPA Response: The small sample study from Chapter 10 of the draft 2004 Unified Guidance was reviewed and dropped from the 2009 guidance. Despite its omission, any conclusions from statistical testing are tempered by the fact that very small samples make testing in either detection monitoring or compliance/assessment monitoring rather difficult and often inconclusive. Other reviewers (notably Charles Davis) have suggested increasing sample sizes for testing by considering all available historical upgradient and downgradient data for use as potential background, a suggestion mentioned in the 2009 guidance. The use of the pooled variance approach (Chapter 13) can also provide larger background sample sizes.

- **Mention Available Software — (p.13)**

EPA Response: Review of statistical software programs is beyond the purview of the Unified Guidance, except perhaps for certain cases of very specialized software tailored specifically to implementing the guidance. In addition to the Optimal Rank Values Calculator prepared for this guidance, we have introduced the use of R-script for certain difficult or repetitive calculations.

Peer Reviewer — Dr. James C. Loftis, PhD. Colorado State University

- **Certain of the UG presentations [e.g., power, prediction limits] are complex information which might be better presented and simplified (pp.32, 43-44, 60).**

EPA Response: We agree that the highly technical discussions in the 2004 Unified Guidance text should be moved to one or more technical support appendices. Detailed power discussions are now located in Appendix C of the 2009 guidance (including R-script software). Further, power discussions were evaluated for clarity and simplified whenever possible in the main text.

- **Loftis was not certain how Chapter 5 (flowcharts) can best be presented and clarified (p.33). He also noted some mistakes, errors, and omissions in the charts (pp.63-64).**

EPA Response: The draft 2004 Chapter 5 flowcharts were reviewed for accuracy and omissions. While certain Unified Guidance reviewers and other users found the charts helpful in implementing the guidance, we decided to drop these detailed charts in favor of a more flexible data evaluation approach presented in the 2009 guidance.

- **Loftis recommends a table of critical skewness coefficient values (p.38).**

EPA Response: We disagree that a formal test using critical tabular values is necessary for the coefficient of skewness applications presented in the 2009 guidance (Chapter 10). This statistic is not intended to be used for formal distributional testing, but rather as an informal measure of data symmetry.

- **Loftis disagreed with the UG recommendation not to use non-parametric prediction limits for intrawell comparisons (pp.35, 48).**

EPA Response: We agree that the Unified Guidance should not rule out the use of non-parametric prediction limits for intrawell comparisons. Nevertheless, without substantial retesting and when the intrawell background sample size is small, it is difficult to achieve adequate statistical performance with non-parametric limits. For this reason, the 2009 Unified Guidance encourages users to carefully balance the tradeoffs incurred when using non-parametric limits under these conditions.

- **Loftis had some reservations about default normality [Chapter 6] (p.37).**

EPA Response: The 2009 Unified Guidance generally indicates that a default assumption of normality for rather small background sample sizes ($n < 8$) does not hurt the statistical power and performance of prediction limits, even if the underlying model has been mis-specified. This conclusion was based on the special study now presented in Appendix B and summarized in new Chapter 10 of the 2009 Unified Guidance. The guidance is careful not to overgeneralize these

results to other statistical methods, nor to downplay the importance of doing proper distributional testing whenever possible.

- **Loftis had serious reservations regarding pooling variances from different wells for intrawell tests, primarily due to differences in degrees of freedom of well means versus the pooled variance. He thought it affected prediction limit usage (pp.35, 39, 47-48).**

EPA Response: We agree that the formula for improving intrawell prediction limits through the use of ANOVA was incorrect in the 2004 draft guidance and has been revised in the 2009 guidance. Other reviewers (including Charles Davis) also noted this error. We believe, however, that the fundamental method is sound and can be applied in certain circumstances, using a corrected methodology. The final decision to include this method in the Unified Guidance (Chapter 13) was made after evaluation of William Huber's simulation study on this issue and his comments about the method's accuracy.

- **Loftis didn't think Aitchison's mixing model reflected typical environmental data and had suggestions for further evaluations (pp.35, 42, 57, 59).**

EPA Response: Various reviewers questioned whether a mixture model (non-detects representing one distribution, and detects representing another distribution) was valid for most environmental data. However, groundwater measurement, flow, and hydrogeology are complex enough that EPA believes there can be cases where such mixture models are appropriate. Further, the fitting of data to a distributional model is not always connectable to a well-understood physical process. At times, it is more of an empirical matter to see which provides the better fit. The 2009 Unified Guidance has eliminated a direct application of Aitchison's method. Chapter 15 primarily describes methods assuming a single population for both detect and non-detect data, but notes that the modified delta method of Aitchison's model for a mixture distribution may at times be appropriate.

- **UG presentations on temporal variation are somewhat confused, mixing ANOVA, seasonality and serial correlation. Loftis provided a number of recommendations (pp.40-41, 45).**

EPA Response: The 2009 Unified Guidance discussion of temporal variation has been expanded into a separate, full chapter. The discussion in Chapter 14 addresses several different types of temporal variation (e.g., seasonality, serial correlation), recommends techniques to identify each kind of problem, and then provides corrective measures.

- **Loftis thought that both Monte Carlo and formal power calculations would be beyond the capability of most UG users and thus, he would like to see simplifications (pp.43-44, 46-47).**

EPA Response: We agree that many modern statistical calculations either cannot be performed by hand or are very difficult to do without aid of a computer. Both Monte Carlo simulations and power calculations fall in this area. However, a number of statistical software packages now exist that can fairly easily perform these kinds of computations, including some specifically tailored to statistical analysis of groundwater monitoring data and others of low cost (including the free package R) or widespread availability (e.g., Excel). The 2009 Unified Guidance does discuss simplified approximations whenever appropriate, but EPA does not believe that ability to compute statistical tests by hand or only by spreadsheet should be a prerequisite for inclusion in the Unified Guidance. We have provided simplified instructions and scripts for certain prediction limit calculations in Appendix C of the guidance using R-script along with more detailed power discussions.

- **Loftis raised a number of issues with use of the Poisson distribution in Chapter 12 and believes they must be addressed to retain this test (p.45).**

EPA Response: Several reviewers in addition to Loftis objected to aspects of the Poisson prediction limits as presented in the draft 2004 Unified Guidance. We agree that the treatment of this method was subject to confusion and potential misuse by users (especially given its lack of scale invariance). In addition, we agree that the method may be of limited applicability compared to other methods (e.g., non-parametric prediction limits). Thus, the 2009 Unified Guidance no longer retains this method.

- **Loftis raised an issue of well/constituent correlation, which can affect the i.i.d. assumptions of many tests (p.45).**

EPA Response: We agree that correlations can exist among wells and/or constituents that would tend to violate the independence assumption of currently recommended tests. The simultaneous prediction limit formulations in the Unified Guidance account for the dependence between background data in interwell tests across multiple compliance wells. However, they do not account for correlated compliance data across those same wells, or between constituents collected from the same well. Accounting for this correlation structure in testing is difficult and beyond the scope of the Unified Guidance. To the extent that wells and/or constituents are *positively* correlated, an assumption of independence is somewhat conservative. In other words, it will generally be easier to detect statistically significant changes (either via simultaneous increases in a number of wells for one constituent, or in a number of correlated constituents in a single monitoring well). This may be less of a problem if the correlated monitoring parameters are hazardous constituents rather than naturally occurring indicators. Chapter 14 of the 2009 guidance includes a discussion of interwell correlation and correlation among constituents in a single well, as well as potential remedies.

- **Loftis commented on the power evaluation of lognormal data in Example 13-2. This discussion and example needs to be simplified while illustrating effects of lognormal data (p.47).**

EPA Response: We agree that this discussion might be confusing and should be simplified. While retained, this discussion was placed in Appendix C of the 2009 Unified Guidance and is not emphasized in the main text.

- **Loftis had trouble understanding the per-constituent comparison calculations in Chapter 13 (pp.47-48).**

EPA Response: We agree that the treatment of single-test, per-constituent, and site-wide false positive rates needed to be clarified, and the discussion has been expanded in the 2009 Unified Guidance. Chapter 19, Section 19.2 has been considerably revised to aid usability and understanding in making per-constituent comparison calculations.

- **Loftis recommends linear regression on ranks for Chapter 14 [trend analyses] (p.49).**

EPA Response: We considered this technique, but did not include it in the 2009 Unified Guidance. Instead, we have added other non-parametric techniques in Chapters 14 and 17 for evaluating trends.

- **Equations and presentations of the binomial distribution need clarification as well as references (p.49).**

EPA Response: We agree that the presentation of confidence limits based on the binomial distribution needed to be clarified. All equations in the 2009 Unified Guidance were checked for accuracy. A simplified discussion has been provided in Chapter 21 (pp. 21-14ff). We have also consolidated discussions of binomial median and upper proportion confidence intervals for easier understanding.

- **Loftis had numerous problems with the methods for varying false positives to meet power objectives in Chapters 16 & 17. He didn't think the equations were correct based on Zar. He also felt that the two options were unnecessary and only one should be presented; he did think the overall approach was reasonable. He also noted a mistake in Chapter 17 sample sizes example (pp.52-55).**

EPA Response: All equations for the two methods in 2004 draft Chapters 16 and 17 were checked for accuracy. The example mistake was corrected. We have retained both methods (a nomograph approach vs. tabled values) in the 2009 guidance (Chapter 22), because we believe the added flexibility in making different sample variability assumptions for each method is useful.

- **The guidance needs to include other options for when the GWPS is a background standard [t-tests, Mann-Whitney, etc.] (p.56).**

EPA Response: Other reviewers (notably the State of California) also commented on statistical options that are needed when a GWPS is not a fixed numerical limit, but is instead derived as a set of background measurements. In these situations, Chapter 7, Section 7.5 of the 2009 Unified Guidance now recommends possible techniques other than confidence interval tests. When the comparison is between two sets of distinct measurements as opposed to a one-sample comparison against a fixed numerical limit, various kinds of two- or multiple- sample tests described in Chapter 6 for detection monitoring (including t-tests) can be used.

- **Loftis also provided some thoughtful suggestions on improving UG tables [references, explanations, examples, definitions] (p.57).**

EPA Response: Other reviewers also commented on the presentation of the 2004 draft tables and their usability. All tables were reviewed for errors, references cited, terms and symbols defined, and further explanations were provided as necessary in the main text of the 2009 guidance.

- **Loftis recommended more examples in applying equations in the UG (p.56).**

EPA Response: We considered this suggestion, but generally retained the existing examples in the 2009 guidance with a few additions. The primary presentation structure followed in the guidance is to identify a procedural method followed by an example. Where further elaboration is needed to illustrate a point of difference, additional examples have been provided.

- **Finally, Loftis commented favorably on the non-parametric statistical software. He noticed that the UG tables and software power outputs didn't always agree on 'transition zones' among ratings. The software also needs updating to reflect the more recent UG rating system.**

EPA Response: The approximate power estimates have been improved and more closely match the 2009 guidance Appendix D tables in the transition zones. Categorical ratings in the revised Optimal Rank Calculator (2006 version) are based on the same 2009 guidance criteria.

Peer Reviewer — Dr. William A. Huber, PhD. Quantitative Decisions Inc.

- **The UG should indicate what alternative tests, procedures or methods might be applicable. Many State programs will only use what is in guidance (p.68).**

EPA Response: In the Unified Guidance, we have striven for a balance between technical accuracy and usability/simplicity. It is not designed to be an exhaustive compendium of possible statistical techniques. However, we have provided references to additional recommended methods for some discussions in the 2009 guidance.

- **The UG needs to address how statistical guidance can be translated into permit requirements (p.69). It should distinguish the statistical aspects of "permit development" from "routine application" (pp.80-81). Many of the diagnostic tests in Chapters 6-10 need not be run at every application instance; formal tests are in Chapters 11-17 (p.86). The guidance should specify how rigorous EPA intends the power evaluations to be (pp. 69, 116).**

EPA Response: We agree that the Unified Guidance should be reorganized to reflect the suggestions of the reviewer. The 2009 guidance contains separate Chapters 6 and 7 on statistical design issues, which would be used to initially develop a statistical program or write statistical portions of a permit, selecting from among statistical tests described in later chapters. The point regarding separating "development" from "routine application" is mentioned at the outset in Chapter 6. The new Chapter 5 on background data development and review also makes this point regarding the need for only periodic re-evaluation of data.

The 2009 Unified Guidance has also been clarified in its discussions of using power curves and other power evaluations as a tool for the selection of appropriate test methods, especially in Chapters 6 and 7. It should now be clear that power evaluations are primarily useful for permit writing and statistical design. We have also added language indicating that design approaches including assessment of statistical power are intended to allow considerable flexibility.

- **The UG should clearly indicate that "test shopping" or "post-hoc test selection" should not be allowed. The UG provides many options, but a single protocol must be followed once decided upon (pp. 69, 95, 106).**

EPA Response: We agree with the reviewer, that the guidance should not suggest that *post-hoc* test selection is allowable or desirable. This point is made on page 4-2 of Chapter 4, 2009 guidance describing the groundwater monitoring context. However, it should be clear as stated in the initial Disclaimer, that this is guidance and does not impose any regulatory requirements. Final RCRA permit requirements will determine the appropriate set of tests for implementation.

- **The guidance needs to consider how to incorporate retest samples into the overall record, especially for updating background (p.69).**

EPA Response: The 2009 Unified Guidance includes specific recommendations for how to incorporate resamples into the decision record and database for a given site in Chapter 5. In general, resamples should be identified separately from regularly scheduled monitoring samples (possibly by assigning special flags within the database). We also agree that treatment and tracking of resamples deserves special emphasis and discussion when it comes to updating of background in intrawell testing (discussed on pp. 5-13).

- **Huber thinks that upgradient-downgradient well comparisons are needed even if intrawell (downgradient) well testing is used (p.70).**

EPA Response: Formal upgradient-to-downgradient well comparisons — in situations where intrawell testing has been selected and justified at a site — would be potentially confusing and misleading to the regulator and the regulated party alike. However, we do agree that even if intrawell testing is conducted, wells upgradient to the site should continue to be monitored, especially as sentinels of either regional changes in background groundwater quality or migration from off-site sources of contamination. This point is made on page 5-2 of the 2009 guidance.

- **All formulas should have explicit references. All tables should identify explicitly how numbers were derived (p.70). Huber also suggests very specific criteria for all tables and figures [clear and consistent explanations, labeling, examples, definitions of all terms used in tables, etc.] (pp.79-80). Key terms like 'background" must be defined and used consistently (pp.76-78). Chapter 13 tables in particular need to have the exact algorithms spelled out (p.111).**

EPA Response: We agree that all equations, where appropriate, should be referenced; however, many are standard in statistical texts. The reference section in Appendix A covers many of the sources. The 2009 Unified Guidance provides greater detail about the derivation of statistical tables and algorithms, as well as references where appropriate. All figures and tables were reviewed for clarity in labeling, explanations, etc. Chapter 3 on basic statistical concepts provides much greater definition and explanation of key terms. A new Chapter 5 on background data has been added. Algorithms for parametric κ-tables found in Appendix D are provided using two R-scripts in Appendix C. We have also used terms consistently throughout the guidance.

- **Huber questioned the recommended sitewide false positive error rate approach for detection monitoring design. The guidance should indicate that there is flexibility in this choice. He also suggested that it not be too specific on some aspects of this presentation. Principles should be favored over specific levels (pp.72-73).**

EPA Response: The Unified Guidance is a guidance document without the force of regulation. This is emphasized in the Disclaimer at the beginning of the 2009 guidance. Thus, our recommendations are suggestions based on experience. We do agree with the reviewer (as well as others such as Charles Davis), however, that there was an imbalance in the treatment of false positives and false negatives in the 2004 draft guidance document. We also agree that the existing reference power curve approach could not be applied to control charts in the same way as to prediction limits. To balance the current recommendation of an *annual* site-wide false positive rate, the 2009 Unified Guidance recommends a cumulative *annual* reference power standard for detection monitoring design. This change allows each of the main Unified Guidance statistical techniques to be evaluated according to a common standard. It is noted in Chapter 6 that even the recommended criteria might be adjusted by States (p. 6-8).

- **Huber provided general guidance on UG language and style — consistent use of terms, terms clearly defined (probably in the Glossary), use of the active voice. He recommends a systematic editorial revision aimed at making the writing style, terminology and formatting consistent throughout the guidance (pp.75-82).**

EPA Response: The 2009 Unified Guidance has been carefully edited to maintain consistency in voice, terms, formatting, and writing style.

- **Huber favors a systematic approach to power analyses (sitewide) that takes into account the duration of monitoring (over years). The UG power analyses are based on single well-constituent evaluations at one point in time (pp. 72-73, 129-132).**

EPA Response: EPA agrees with the reviewer's suggestion (see response above). The 2009 Unified Guidance recommends a cumulative annual reference power standard, replacing the one-point-in-time reference curve.

- **Organization of the UG — Huber suggests placing theoretical discussions and historical notes in separate sections of the guidance (perhaps as appendices), as well as separating diagnostic development from implementation (pp.80-81). Table of Contents should show UG divisions. Also for each method, adopt a consistent background discussion/procedure/example approach. Keep recommendations out of procedures and put into background discussion (p.82).**

EPA Response: We generally agreed with these suggestions. More technical discussions, as well as historical notes, have been placed in separate Appendices B & C of the 2009 guidance. In addition, discussions of statistical design and diagnostic evaluation were distinguished from

routine implementation (see the response to a similar earlier comment by the reviewer). In addition, a consistent format has been adopted for the presentation of each method in the Unified Guidance, and motivating discussions and background material were mostly moved to chapters on statistical design and test selection. The Table of Contents has been expanded following a Part/Chapter/Section arrangement maintained throughout the document, with additional detailed Tables of Contents for the Appendices.

- **Huber raises an issue of correlated well/constituent variation [similar to Loftis' comment]. This can affect the i.i.d assumption. Need to put in early discussions (pp.83-84).**

EPA Response: We agree that wells and/or constituents at some sites may exhibit correlations not accounted for by the independence assumption inherent in the Unified Guidance test methods. However, a formal treatment of such dependence, seeing as it would likely vary from site to site, is beyond the scope of the Unified Guidance except for temporal variation discussions in Chapter 14 of the 2009 guidance. Further, EPA believes the 2009 guidance approach is conservative (see an earlier response to the well/constituent correlation issue raised by Loftis).

- **Huber recommends always graphing the data as a general UG principle (p.84).**

EPA Response: We agree with the recommendation, and have emphasized this principle within the 2009 final Unified Guidance. A separate diagnostic Chapter 9 covering graphical and other exploratory data methods has been added.

- **Huber recommends additional tests for equality of variance and outliers (pp.87, 128).**

EPA Response: In response to the reviewer's recommendations, we have added an additional Mean-Standard Deviation plot exploratory tool to the 2009 guidance for assessing equality of variance. In addition, Tukey's procedure using box plots to evaluate outliers, a quasi non-parametric technique, was included in the 2009 guidance.

- **Huber raises a question of what to do when all transformations fail (p.88).**

EPA Response: The 2009 Unified Guidance discussions in Chapter 10 and elsewhere rely primarily on non-parametric options, when no suitable normalizing transformation can be found. The guidance recognizes that other distributions like the gamma or Weibull may be appropriate fits, but are beyond the scope of the present guidance.

- **The UG should address how to treat non-detects in trend and control charts (p.88).**

EPA Response: We agree with these suggestions and modifications. The Mann-Kendall test and the Theil-Sen trend line to handle non-detects are presented in Chapter17 of the 2009 guidance

on trend tests. Chapter 20 on control charts provides explicit guidance on how to handle and incorporate non-detect measurements.

- **Huber recommends putting Chapter 6 Monte Carlo results of the normal/lognormal study into an appendix (p.88).**

EPA Response: We agree with the suggestion, and the details of the normal/lognormal simulation study are now found in Appendix B of the 2009 guidance. The overall conclusions are summarized in Chapter 10.

- **Allow for probability plotting of other than normal distributions (p.88).**

EPA Response: We partly agree with this comment. Other goodness-of-fit probability plots based on the gamma or Weibull parametric distributions are recognized as potential options, but are beyond the scope of this guidance. The 2009 guidance confines probability plotting to normality and other possible data transformations to normality (e.g. in the Chapter 9 presentation).

- **The guidance should explain why particular significance levels for diagnostic tests were chosen and potential options. Consider factors like sample size and number of tests (pp.89, 91, 93).**

EPA Response: We partly agree with this comment. While a desirable goal to identify optimal significance levels for diagnostic tests, the cumulative effect of many repetitions of the same test or of sequential testing was determined to be beyond the scope of the guidance. Unlike formal monitoring tests, the Type I and II risks for diagnostic testing vary with the purpose of the test. Significance level selections are occasionally discussed for specific methods in the 2009 guidance, but generally work from typical choices (e.g., 1, 5 or 10%). We do appreciate the reviewer's concern over a potentially important and complex topic, but it is one that will require extensive research. Sample size is clearly an important consideration. However, we are less convinced that explicit accounting should be made of the number of diagnostic tests to be performed. Given the generally small sample sizes available in groundwater testing, lowering the single-test false positive rate to account for multiple diagnostic procedures risks eliminating the power of these tests to make any diagnostic determinations.

- **Include negative correlation as a possibility for von Neumann test Chapter 9. There is a need to rewrite this section for language clarity as well (p.91).**

EPA Response: The discussion of the rank von Neumann ratio test in Chapter 14 of the 2009 guidance has been rewritten to indicate that positive or negative (auto)correlation is a possibility for dependent temporal data. A negative test statistic ν for this test will also trigger a significant outcome.

- **Huber posed questions on the method for removal of seasonal correlation (p.92).**

EPA Response: We agree that the 2004 draft guidance recommendations on correcting for seasonal trends were contradictory and needed to be clarified. Chapter 14 of the 2009 guidance recognizes seasonal correlation as one of a number of possible types of temporal variation, with specific remedies provided.

- **For Chapter 10, Huber recommends moving test-specific discussions to later chapters (p.94). He also recommends moving the Monte Carlo small sample study simulation to an appendix (p.95).**

EPA Response: We agree and have restructured the entire document to cover general and design elements in early chapters, and specific detailed tests in later ones. The small sample Monte Carlo study was dropped from the 2009 guidance.

- **Huber thinks that Monte Carlo or numerical integration power analyses may be beyond most users' capability. The UG should clarify when additional power evaluations are required beyond what is already available in guidance (pp.96, 136).**

EPA Response: Although this issue was also raised by other reviewers, technical evaluations of any proposed statistical program should include an assessment of statistical power. Although computer software of some sort will usually be required to perform a power evaluation, the draft 2009 Unified Guidance does provide indications of power associated with specific prediction limits. The guidance has also extended power assessment to control charts in Chapter 20, albeit using Monte Carlo simulation. We do not believe it too burdensome to recommend that either the tables within the Unified Guidance be utilized or a computer-generated simulation of power be created as necessary. R-script software provided in Appendix C is also straightforward to use.

- **Costs when contamination are present should not be highlighted at the expense of the more likely background condition when not present. Huber questions the software calculations for additional samples required. He raises a point that resamples may occur in different wells and somewhat add to the expense (pp.96-97).**

EPA Response: We did restructure discussions in the 2009 guidance to reflect that "background levels" (i.e. uncontaminated situations) are the basis for most detection monitoring cost considerations (e.g., Chapter 6 of the 2009 guidance). While the peer reviewers point is well taken that additional sampling for repeat testing may randomly occur at different wells, the average rates for a given well-constituent are dependent on the single error rate for a test. The RCRA regulations do not allow cost to be a determining factor in balancing environmental protection needs.

- **A median-of-three plan may be reinterpreted as a 2:3 prediction limit test. The UG needs to make the exact protocol clear to avoid 'test-shopping' (p.98).**

EPA Response: We agree with this comment, and note this particular overlap in Chapter 18 (p. 18-20). More generally, with all retesting plans, the 2009 Unified Guidance clearly lays out the protocol for running each kind of plan, including how and when resamples need to be collected and when a 'stopping point' has been reached.

- **In UG Chapter 11, the rationale for use of the t-test in detection monitoring is unclear (p.99).**

EPA Response: Discussion of the rationale for using the t-test has been reviewed and revised in Chapter 16 of the 2009 guidance to reflect that the site-wide false positive rate may not be achievable under this test selection. The guidance indicates more generally (e.g., Chapters 5 & 6), that the t-test has applicability when updating background or for other kinds of diagnostic testing.

- **Huber suggests the possibility of a sensitivity analysis of test results (p.100).**

EPA Response: We did not add a discussion of the value of sensitivity analyses, since it is beyond the scope of the guidance.

- **Huber recommends an easier method for calculating ties in the Wilcoxon test (p.101)**

EPA Response: While we agree with the reviewer's suggestion, this alternate method was not included in the 2009 guidance. The guidance continues to use the more standard statistical techniques for dealing with tied observations. The method cited by the peer reviewer is specific to the Wilcoxon test and cannot be directly extended to other test situations dealing with ties.

- **Chapter 12 (introduction to prediction limits) should have a very limited role. Make clear that it is only providing simple examples to illustrate procedures. Don't confuse this chapter with complicated examples involving multiple tests (p.103).**

EPA Response: We agree with the reviewer's suggestions, and have provided simple examples and computation of the limits themselves in the introductory Chapter 18 on prediction limits. Retesting details are covered in Chapter 19 of the 2009 guidance.

- **Define the meaning of a prediction limit (p. 102). Clearly identify a number of key terms — background, compliance data, updating, resampling, etc. (p.105).**

EPA Response: Chapter 6 (pp.6-42 to 6-44) and Chapter 18 (pp. 18-1 to 18-4) of the 2009 guidance now contain a clear definition of prediction limits and their relationship to statistical intervals. An Index has also been provided in Appendix A to locate key definitions.

- **Make sure rankings are consistently used (p.105).**

EPA Response: While we agree with the reviewer's suggestion to ensure internal consistency in mathematical notation for ranking data, the conditions cited as an example were correct (rankings from least to greatest as 1 to *n*). For practical reasons (the meaning of the *n*-th largest value), it was preferable to utilize an inverse ranking scheme for the Optimal Rank Calculator. Where there are ranking differences, the 2009 Unified Guidance makes this clear.

- **Huber suggests a reorganization of Chapter 13 (p.107).**

EPA Response: We agree with the reviewer's suggestion and have reorganized Chapter 13 (now Chapter 19 in the 2009 guidance). Material of a more general nature, especially those sections and topics touching upon statistical design, have been moved to Chapter 6 of the 2009 Unified Guidance.

- **The guidance should define the Bonferroni approximation correctly (p.108)**

EPA Response: Corrected discussions of the Bonferroni approximation have been provided in Chapters 6, 17 and 19 of the 2009 Unified Guidance.

- **Holding times are an issue on page 13-38 (p.109)**

EPA Response: Footnote 7 in the 2004 draft Unified Guidance has been eliminated to avoid suggesting that multiple samples should be collected at a single event for future evaluation, avoiding the holding time issue. More generally, the 2009 guidance considers the number of potential repeat samples within a given test (e.g., with prediction limits) as a factor to consider. Particularly where regulatory constraints may dictate the period between tests, a degree of statistical independence needs to be assured between samples. This issue is covered in numerous chapters of the 2009 guidance.

- **Section 13.7 comparing parametric/non-parametric tests needs a rewrite (p.109).**

EPA Response: The language in this section has been reviewed for clarity and modified as necessary. The points comparing parametric vs. non-parametric prediction limits are still valid, but have been better distinguished in Chapters 6, 18 and 19 of the 2009 guidance.

- **Need to discuss how to interpolate tables (part of table revisions) (p.111)**

EPA Response: We partly agree with the reviewer's suggestion, and have provided explicit guidance for bilinear interpolation of parametric prediction limit κ- tables (pp. 19-13ff). It was impractical to provide similar suggestions for the other numerous tables in Appendix D of the 2009 guidance. Statistical texts will need to be consulted for more sophisticated interpolation methods.

- **Add use of the normal approximation to the binomial test in Chapter 15 when N > 20 (p.112)**

EPA Response: We agree with the reviewer's suggestion, but inadvertently left out a discussion of the normal approximation for cases of larger sample sizes (i.e., n > 20) in the 2009 guidance. However, this approximation methodology is found in most standard statistical texts. The discussions in Chapter 21 (pp.21-14ff) note that the procedure for n < 20 provides exact binomial probability levels.

- **UG Chapter 15: Need to clarify one- and two-sided confidence intervals (p.112)**

EPA Response: The relationship between confidence levels and significance levels for one-sided tests vs. two-sided tests has been clarified. The reviewer also suggested providing explicit discussion of two-sided confidence intervals in the Unified Guidance. We disagree with this approach, since only lower limits are generally needed during compliance/assessment monitoring and only upper limits during corrective action. Chapter 7, Section 7.2 of the 2009 guidance contains a discussion distinguishing one- and two-sided confidence intervals.

- **Huber disagreed with conclusion of compliance test wording (p.113).**

EPA Response: This language in the 2009 Unified Guidance has been clarified. A valid exceedance of a lower confidence interval test does indicate the potential need for corrective action, but the guidance recognizes that other factors may need to be considered.

- **Huber questioned use of geometric versus arithmetic mean for compliance (p.114)**

EPA Response: The Unified Guidance is not a document about risk assessment. In that sense, it may not always be clear whether a test for compliance using the geometric mean as opposed to the arithmetic mean is the better statistical 'match' to a given compliance standard. However, there is some uncertainty regarding the appropriate choice of parameter to measure compliance and non-compliance (see discussion in Chapter 7 of the 2009 guidance, pp.7-6 to 7-8). Therefore, the guidance provides a number of both centrality and upper limit parameter tests for compliance and corrective action in Chapters 21 and 22.

- **Huber questioned use of an imputed value for a non-detect in Example 15-2. Generally, he recommends specifically citing other UG sections when making calculations (p.114).**

EPA Response: We agree that the treatment of non-detects in the 2004 draft Unified Guidance needed significant revision. We considered, but did not follow the reviewer's suggestion to describe a sensitivity analysis of varying the simple substitution value. Chapter 15 of the 2009 guidance contains an expanded range of options for managing non-detect data including the imputation of arbitrary values like ½ the detection limit.

- **Huber recommends simplifying confidence intervals around medians, Chapter 15 (p.115)**

EPA Response: We followed the peer reviewer's suggestion to simplify the presentation of confidence intervals around medians by limiting the discussion to one-sided limits, and have consolidated the discussion of medians and upper percentiles in Chapter 21 of the 2009 guidance.

- **The UG needs to maintain consistency in applying confidence intervals with non-parametric techniques in UG 15-21 and 15-32 (p.115)**

EPA Response: In the 2009 Unified Guidance, we have maintained a consistent approach in comparing the achievable non-parametric confidence levels to a pre-specified error value. The approximate rounding from .9894 to .99 in the first guidance example mentioned by the reviewer was more for convenience in comparing it to the pre-specified target level. The actual achievable level would be 98.94% and has been so modified. Where the difference between an achievable and target confidence level was more substantial (the second example cited), the achievable level was used.

- **Huber has some recommendations regarding skewness in Chapter 6. He suggests identifying different measures and notes that 1.0 is an <u>absolute</u> value criterion (p.116).**

EPA Response: We did not substantially change the discussion in Chapter 10 on the use of the skewness coefficient. The 2009 Unified Guidance indicates that along with the sample coefficient of variation, the skewness coefficient is still an indirect measure for assessing normality and better formal tests are presented. We did not discuss differences in how skewness measures are computed by different software packages. The discussion also clarifies that the informal cutoff point of 1.0 is an absolute value criterion.

- **Huber recommends different language to describe the CLT at UG 6-4 (p.116)**

EPA Response: We agree with the reviewer's recommended wording change and revised the discussion of the Central Limit Theorem in Chapter 3 of the 2009 Unified Guidance.

- **Huber suggests not using anomalous data in an illustrative example to better make the point regarding normality assumption, UG Chapter 8 p.8-13. He didn't think the graphical presentation was convincing. He has a number of suggestions for the illustration of Rosner's test for outliers in Chapter 8 (p.117)**

EPA Response: We checked the naphthalene example for clarity of presentation and believe it is sufficient for the presentation intended. Thus, it is retained in Chapter 12 of the 2009 Unified Guidance. We did not make use of Huber's suggestions for Rosner's test example, since it would add unnecessary complications to the discussion.

- **Huber raised an issue involving deletion of an observation, Section 8-10 (p.118). This is part of a broader issue discussion of how to evaluate historical data.**

EPA Response: We agree that deletion of identified outliers should be temporary and for the purposes of correct statistical evaluation. The decision record for the facility should always retain such observations as flagged values for auditing purposes. In Chapters 5 and 12, the 2009 Unified Guidance covers this subject in some detail. The guidance does make distinctions between historical data considered for background and other monitoring data sets in these chapters when considering potential outliers.

- **Huber has strong disagreement regarding ANOVA-based prediction limit calculations in UG Section 9-23. He offers a correction formula for pooled variance calculations but argues against its use based on his simulations (pp. 119-121)**

EPA Response: We agree that the formula provided in the current draft guidance needed correction, and have modified the basic equation in Chapter 13 of the 2009 guidance. Huber based his overall conclusions on a limited study of the pooled variance method in conjunction with an ANOVA diagnostic test for well spatial differences. The potential range of inputs even for such a simulation would need to be considerably broadened to justify his conclusions. Moreover, his simulation did not include the use of an ANOVA test for equal variances, a much more important diagnostic test. The effects of sequential tests (including diagnostic ones) is a potential research topic, but beyond the scope of the guidance. Simulations of the pooled variance method itself confirm that the basic equations are valid. We have included R-script software in the 2009 guidance (Appendix C) to make exact calculations for prediction limits using pooled background data from multiple wells.

- **Huber has disagreement with power results in Example 11-2. (p.121)**

EPA Response: The power computations in this example were re-checked and found to be correct. However, the example has been shifted to Appendix C and is less emphasized in the guidance document.

- **Huber strongly favors removal of the Poisson distribution in Chapter 12. He doesn't think it can be justified either theoretically or in practice, and he provides some simulations. He dismisses the UG attempt at rationalizing scaling (pp. 122-123).**

EPA Response: We agree with his and other reviewers' recommendations, and have eliminated the Poisson prediction limit method in the 2009 Unified Guidance.

- **Huber offered some comments on Sen's slope estimator Chapter 14-31ff, and corrections to Land H-factors in UG appendix Table 15-1. He recommends extending the table to N =101 (p. 124-125)**.

EPA Response: The tables of Land's H-factors were checked for accuracy and have been corrected as necessary in Appendix D of the 2009 guidance. We did not extend these tables to *n* = 101 as suggested, since larger sample size H-factors can be obtained from other statistical texts including (Gilbert, 1987).

The reviewer suggested a possible method for developing a confidence band around a non-parametric trend line. Another suggestion developed in consultation with peer reviewer Dennis Helsel, was to form such a confidence band by bootstrapping the Theil-Sen trend line. We incorporated the latter's suggestion and modified the language regarding Sen's slope estimator under the discussion of the Theil-Sen trend line in Chapter 17 of the 2009 guidance. This technique, though more computationally intensive, was added to the guidance as an alternative when a confidence band around a parametric linear regression cannot be justified. An R-script has also been provided in Appendix C to perform these intensive computations.

- **Huber had problems with some phraseology in Chapter 4 and the discussion of the benzene distributions, in defining null and alternative hypotheses (pp. 126-127).**

EPA Response: The language was clarified and corrected as necessary in light of the reviewer's suggestions. These discussions are now located in Chapter 3 of the 2009 guidance.

- **UG Chapter 16: Huber thinks there is a need for retesting in compliance monitoring (p.130); he thinks the ratio/confidence power criteria are too inflexible. He believes there should be the option to choose different criteria.**

EPA Response: We are not making the change, as suggested by the reviewer for confidence interval tests, because we believe that formal retesting during compliance/assessment monitoring would be difficult to implement (see discussions in Chapters 7, 21 and 22 of the 2009 guidance). We agree, however, that guidance users might consider sending samples to multiple labs for QA/QC evaluations, but the subject is beyond the scope of the guidance.

The ratio/confidence power criteria mentioned by the reviewer are not being required, but are only guidance suggestions. Other options can be selected under the framework established in the 2009 Unified Guidance.

- **Chapter 16: Huber is most concerned with statistical power for a facility 'moving into corrective action' and favors consideration of sequential effects. He also disagrees with the UG evaluation of cumulative false positive control for compliance monitoring (p. 132-133).**

EPA Response: The issue of cumulative false positive rates during compliance/assessment monitoring is complex and difficult to simplify within guidance. The number of exceedances of a GWPS will vary by facility and potentially within a facility over its regulatory lifetime. Further, the correlations among constituents with exceedances will also vary. It is difficult to envision how to provide general guidance on this point, especially any practical recommendations on what

sorts of false positive rates would be acceptable. Therefore, we did not materially change this point from the 2004 draft document to the 2009 version of the guidance.

- **UG Chapter 6: Huber takes issue with the default normality assumption (pp. 133-4).**

EPA Response: We agree that the discussion of the default normality assumption in the 2004 draft Unified Guidance needed revision to clarify when distributional testing should be conducted. The discussion in the 2009 guidance now indicates that such testing will only be done periodically (e.g., initially when the statistical program is put in place; thereafter each time background is updated). The Unified Guidance recognizes that a distributional assumption chosen early on in the life of a facility when there are fewer accumulated background data, may not be the best assumption at later periodic evaluations.

- **Huber questions use of normality distributions as basis for power evaluations. He also suggests other tests for assessing normality (pp. 134-135).**

EPA Response: While we agree that power evaluations would ideally be based on facility-specific data distributions, we believe it is infeasible to develop appropriate power reference standards in the guidance for a wide range of other distributions. Assuming a normal distribution for power evaluations provides a consistency for evaluating different statistical techniques (including non-parametric methods) and different groundwater network configurations. The 2009 guidance takes the position that a default normality assumption (for small data sets which cannot be formally tested for distributional assumptions), still provides superior power while maintaining reasonable error rates.

We recognize that other tests for normality exist. Nonetheless, the Unified Guidance was not designed to be an exhaustive compendium of statistical methods. Rather, a small number of techniques have been chosen for reasons of good statistical performance, applicability to groundwater, and usability.

- **Huber recommends other diagnostic normality tests to be included, but he also suggests that the number of diagnostic tests be limited in application (p.135).**

EPA Response: The Unified Guidance effort has striven to strike a balance between technical accuracy and completeness, against ease of use and simplicity. A number of diagnostic tests have been referenced within the Unified Guidance; however, a full description is limited to those selected tests considered of greatest utility and easiest to implement by the target audience.

- **Although Huber generally liked the statistical method summaries and flowcharts in Chapter 5, he had a number of criticisms and suggestions. These include clarifying terms, integration of flowcharts/summaries, assumptions, etc. (pp. 136-145).**

EPA Response: The flowcharts and method summaries in Chapter 5 of the 2004 draft guidance were carefully reviewed. Summaries were edited and corrected as necessary (Chapter 8 of the

2009 guidance). It was decided not to include the detailed flow charts, in favor of a more flexible data evaluation approach.

- **Huber was generally supportive of the non-parametric software, but had problems with dialog prompts, software language and the associated documentation. He provided a number of suggestions for improving it (pp.96-97, 145-148).**

EPA Response: The 2006 version of the Optimal Rank Calculator (which accompanies the 2009 guidance) has been modified to provide more accurate dialog prompts, error checking, and clearer language. In addition, a more accurate power approximation technique has been added. Documentation has also been updated and included.

PART II. OTHER REVIEW COMMENTS AND EPA'S RESPONSE TO THOSE COMMENTS

Reviewer — State of Florida

- **Florida reviewers suggested an organizational revision of the 2004 document, following a Part/Chapter/Section/Bullets approach (p.1).**

EPA Response: Based on these and other reviewer comments, we have revised the 2009 guidance document along these suggested lines. A brief description is now provided at the beginning of each major Part of the guidance. Other comments regarding improvements to the preliminary information and Table of Contents were also followed.

- **Reviewers also made a number of recommendations regarding appropriate references (p.2-3).**

EPA Response: We checked these comments and made the necessary revisions in the 2009 document.

Reviewer — State of Michigan

- **Michigan reviewers had concerns regarding the 2004 draft guidance wording regarding the use of intrawell testing in the presence of spatial variability not the result of contamination, which might not be consistent with their regulatory program (p.5).**

EPA Response: We have revised the discussion regarding interwell versus intrawell testing in the 2009 guidance document (Chapter 6). We have removed the language "thus not subject to RCRA purview" in this context since it was not the intention to suggest any effect on regulatory status.

- **Reviewers had concerns regarding the 2004 draft guidance language regarding the Central Limit Theorem (CLT) (p. 6).**

EPA Response: We have revised the language of the CLT discussions in the 2009 document (Chapter 3, Section 3.5.2), following these and other reviewer suggestions on this topic.

- **Reviewers had a number of comments regarding the 2004 draft guidance language when applying logarithmic distribution assumptions, including problems with backtransformation of an arithmetic mean (p. 6-7)**

EPA Response: We have revised the discussions for using the logarithmic distribution in Chapters 3, 7, 10, 16 & 21 of the 2009 guidance. Specifically, we have clarified the difference in applying the logarithmic assumption to future values of the data set (e.g., prediction limits) as

opposed to estimates of the untransformed arithmetic mean (mean-based prediction limits or mean confidence intervals). We do provide cautionary language for both backtransformation bias, as well as limitations of the upper confidence level (UCL) using logarithmic distribution assumptions.

- **Michigan reviewers believed that additional guidance should be provided on screening for statistical outliers (p. 8).**

EPA Response: We have revised the discussion for review of background data (Chapter 5) and provide a separate diagnostic Chapter 12 for outliers in the 2009 guidance. Outlier discussions are also provided in other chapters as pertinent to evaluating assumptions.

- **Michigan reviewers believed that Chapter 13 of the 2004 draft guidance was slanted towards industry, and that a more balanced discussion of false positives and negatives was needed (pp. 9-11).**

EPA Response: The effort to provide a systematic approach to evaluating false positive and negative errors is an attempt to address such a balance, both in the 2004 draft and the 2009 guidance. This is especially true regarding detection monitoring design, covered in Chapter 6 of the 2009 guidance. The 2009 guidance suggests a framework which can place all facilities on a common basis. We have also provided a discussion for considering false positive errors and power in compliance monitoring tests (Chapters 7, 21, and 22), and conclude that a different approach was needed.

- **Additional guidance should be provided on updating background data (p. 11).**

EPA Response: We agree with this comment and that of other reviewers. A new Chapter 5 discussing review and updating of background data has been added to the 2009 guidance.

- **Michigan disagreed with the 2004 approach of using the lower confidence interval of the mean for testing compliance. (pp. 12-15).**

EPA Response: We partially agree with this comment and have revised certain portions of the 2009 guidance. Chapter 7 contains the basic design discussions for compliance testing. We note on page 7-3 that not all regulatory programs are constructed alike and that the relevant null and alternative hypotheses may differ by program. In the discussion of selecting the appropriate statistical parameter for compliance/corrective action testing (Section 7.3), the guidance accepts that individual State programs will need to determine the appropriate parameter for compliance testing given the uncertainty regarding which statistic is most appropriate. The 2009 guidance provides a range of centrality and upper limit tests to address the likely parameter comparisons.

Reviewer — State of California, Dept. of Toxic Substances Control [DTSC]

- **California DTSC suggested describing the overall process for designing the statistical aspects of monitoring programs. A second suggestion was to include a discussion of the need for "common sense" in evaluating groundwater monitoring data A third suggestion was the need for, perhaps a separate document without the addition of historical information, to focus on methods and evaluation.(p.18)**

EPA Response: We agree with the first suggestion and have revised the 2009 document to contain statistical design of detection monitoring programs in Chapter 6, and compliance/corrective action monitoring programs in Chapter 7. We also agree with the second suggestion; the more practical aspects of reviewing background data are covered in a separate Chapter 5. Regarding the third suggestion, we have not developed a separate document. We have, however, reorganized the document so that the first Part contains introductory and general material, the second Part Diagnostic Methods, the third Part Detection Monitoring Tests, and a final Part covering details of Compliance/Corrective Action monitoring. Historical notes and more technical discussions have been moved to Appendices B and C.

- **DTSC noted that the 2004 draft guidance compliance/corrective action tests only considered fixed limits; California regulations also require testing against background data in some instances. (p.19)**

EPA Response: We agree with the comment and have revised the 2009 document to consider the use of background data limits in compliance/corrective action monitoring programs. The subject is covered in Chapter 7, Section 7.5.

- **DTSC commenters felt greater cross-referencing, e.g., between diagnostic tests, other discussions and formal monitoring tests should be improved. (p.19)**

EPA Response: While we generally agree with the comment, we were unable to provide fully updated cross-referencing because of the complexity of the task. We did revise the 2009 document to contain discussions of assumptions relevant to formal tests, as well as interactions among different diagnostic test assumptions. We have also provided a new Index in Appendix A for the 2009 guidance, which should assist in cross-referencing the test methods and subject matter.

- **Commenters indicated concerns regarding the use of a logarithmic transformation as a first alternative in Chapter 9 of the 2004 draft guidance. They preferred consideration of a range of ladder-of-power transformations. (p.23)**

EPA Response: We agree with the comment, and have indicated in Chapter 10 of the 2009 guidance regarding fitting distributions, that a range of ladder-of-powers transformations including the logarithmic can be considered. We only suggest the logarithmic distribution as a first alternate since many data follow this pattern.

- **DTSC suggested the use of time series and probability plotting as diagnostic measures when evaluating data. They also indicated experience with 'early data' not reflecting later patterns (p.23). They also suggest including autocorrelation evaluations (p.24).**

EPA Response: In response to these and other reviewer comments, we have added a new Chapter 9, covering common exploratory data tools, including time series and probability plots. The 2009 guidance also notes the pattern of potentially unrepresentative early data when discussing background data evaluation in Chapters 4 and 5, and a discussion of autocorrelation methods has been added in Chapter 14.

- **DTSC suggested including examples of larger data sets, which may have somewhat unique problems of multiple reporting limits, temporal variability, greater potential for selecting baseline data, etc. (p.24).**

EPA Response: Other commenters also expressed a desire for consideration of larger data sets and tabular information. We could not provide such full coverage under the scope of this guidance. However, many of the diagnostic tests in PART II of the 2009 guidance can be applied to larger data sets. It may be necessary to consult other literature or text sources for complete tabular information; we have occasionally provided such references in the text. Chapter 15 on managing non-detect data does provide newer methods for dealing with multiple detection limits. In Chapter 5 discussions of background data (or baseline data), we have adopted statistician Charles Davis's suggestion to include historical downgradient well data as potential inclusions which can enhance overall background sample size.

Reviewer — Industry Representative

- **The commenter felt that the guidance should clarify that a regulated site should be able to return to a detection monitoring programs, even if a non-hazardous constituent triggered evidence of statistically significant increase (and no other hazardous constituents were detected in groundwater). (p.25)**

EPA Response: We agree that indicator and other non-hazardous monitoring constituents can exhibit statistically significant increases (SSI) that are not necessarily indicative of a release. A specific example in Section 6.3.2 (Chapter 6 of the 2009 guidance) highlights this problem. In Chapter 4, factors are listed which can trigger statistically significant exceedances (and which are not necessarily indicative of a release). The RCRA regulations do provide the owner/operator with the ability to demonstrate that the SSI was not the result of a release. It might be necessary to work with the regulatory agency to adjust monitoring constituents which can avoid this problem in the future.

- **The Industry commenter provided an extensive set of recommended language for distinguishing method detection limits (MDLs), practical quantification limits (PQLs) and reporting limits (RLs). (pp. 26 to 30). He also raises an interesting point regarding changes to levels of quantification as newer methods are utilized (p.32).**

EPA Response: The overall subject of selecting the most appropriate censoring and quantification limits is beyond the scope of the Unified Guidance. On a more limited basis, we recognize that reviewers need to primarily deal with retrospective analyses, when evaluating historical groundwater monitoring data. For future data, the reviewer's recommendations might apply to prospective data; however, issues of data incompatibility can arise. The overall subject of detection limit appropriateness is under review by the EPA and the present guidance cannot resolve such issues. A limited discussion of how varying non-detect values in a data set can be managed is found in Chapter 15 of the 2009 guidance. The document also offers practical suggestions on how data of various qualifications (e.g., "J", "B" or "E") and the different forms in which they are reported might be ranked for use with certain non-parametric tests in Section 6.3.4.

The commenter's point regarding changes of analytical methods resulting in lower censoring limits is also a valid concern, but one mostly outside the scope of this guidance. Some of the diagnostic tests in PART II (and also recommended for updating background data in Chapter 5), might have application in evaluating historical and recent data sets. As noted above, incompatibility between different data sets might require newer background data acquisition.

- **The commenter raised an issue regarding what is meant by 'small sample sizes and felt the guidance should explain more clearly (p. 34).**

EPA Response: This comment has also been raised by other reviewers. Within the scope of RCRA data collection, a small sample size can be in the range of 4 to 20 values, and is assumed as such in the 2009 guidance. However, it is difficult to generalize regarding the appropriateness of sample sizes (e.g., with logarithmic data). Larger data bases when practical (especially for background) are certainly recommended in Chapters 5 and 6. We have adopted the suggestion to include historical downgradient well data in evaluating potential background or baseline data. We were unable to provide extensive coverage of larger data sets (e.g., greater than 50-100) in tables, etc., within the scope of this guidance preparation.

- **The commenter raised issues regarding use of interwell versus intrawell testing and preferred use of intrawell testing (p. 34).**

EPA Response: In the 2009 guidance, criteria under which interwell versus intrawell testing should be applicable have been clarified. The guidance strongly urges the use of prior ANOVA testing (Chapter 13) to evaluate evidence for spatial variation. For constituents that exhibit spatiality, the guidance recommends the use of intrawell tests. However, this recommendation is

specific to a given constituent. Some trace elements, for example, may exhibit no site-wide spatial variation. Interwell testing may then be appropriate, as discussed in Chapter 6.

- **The Industry commenter raised an issue that some State regulatory agencies may require more sensitive analytical methods, when censored limits (i.e., non-detects) are reported in certain groundwater studies (p. 34).**

EPA Response: RCRA regulatory decisions by an authorized State can and will occur which may differ from EPA's position. The issue is beyond the scope of the guidance. We do note in passing that deciding upon the merits of varying non-detect limits can be a complex issue. As described in Chapter 5 of the 2009 guidance, non-statistical decisions about certain potential background outliers (including non-detect values) may need to be made if the data are considered unreasonable. This would be worked out between the regulated party and regulatory agency.

- **The commenter suggested an alternate method for assessing statistical power of detection monitoring tests based on multiple constituents and wells, such as is found in a commercial statistical software application (p. 39 & 40).**

EPA Response: We will not comment directly on the use of proprietary software. Individual State programs may wish to apply such software, but the issue is generally beyond the scope of this guidance. However, in Chapter 6 of the 2009 guidance, we do recommend against the use of a composite measure of power, for the reasons cited on page 6-22. The guidance allows for effect size measures of statistical power in certain situations (also described in Chapter 6).

Reviewer — State of Colorado

- **Colorado commenters recommended a companion document providing a table, list or decision tree for quick reference to statistical methods and applicability (p.41).**

EPA Response: We have not provided such a companion document in this 2009 guidance version. The summary of methods and table in Chapter 8 are intended to address this issue. However, we have eliminated the extensive flow charts found in the 2004 draft guidance in favor of more flexible data evaluation. We have also revised the document to provide general design elements in the early portions of the document, with more detailed method presentations in later Parts and Chapters.

- **The Table of Contents (TOC) needs to also list Appendix information. A list of acronyms and symbols should also be provided (p.41).**

EPA Response: The main TOC in the 2009 guidance has been revised to include the major Appendix listings. We have also included more detailed TOCs in the Appendices themselves. Many important acronyms are found in the Glossary (Appendix A). We felt that a list of symbols (many of which are used repeatedly and differently in portions of the document) was impractical and potentially confusing.

- **Colorado commenters noted that the 2004 draft guidance was not entirely consistent in recommendations for different techniques based on percent non-detects (p.42).**

EPA Response: We have revised the 2009 version to maintain consistent recommendations, as found in Chapter 15.

- **Colorado raised the issue of methods for multiple non-detection limits and suggested their inclusion. (p.45).**

EPA Response: We agree with the comment and have added two methods for managing multiple non-detection limit data in Chapter 15 of the 2009 guidance.

- **Commenters raised the issue of the use of the H-statistic for sample sizes less than 30 (UCL of the arithmetic mean) and indicated that the guidance didn't discuss minimum sample sizes (p.45).**

EPA Response: We agree with the comment and have added cautionary language for appropriate use of Land's upper confidence interval of the mean for lognormal distributions. In general, we favor at least a minimum number of 8 samples for using logarithmic data, but were unable to provide absolute sample size minima for this test. The latter depends in part on the sample logarithmic variance and significance level. Based on other comments by Dr. Anita Singh, we have also included a reference to her alternative methods for calculating confidence intervals of this type using ProUCL® software developed for EPA (Chapter 21, p. 21-9).

Reviewer — State of South Dakota

- **South Dakota commenters raised the issue of the required use of unfiltered trace element sampling for Solid Waste facilities, and noted their situation of excessive well turbidity affecting data quality. They suggested that this requirement should be removed from regulation (p.46).**

EPA Response: This guidance will not comment on the appropriateness of the existing regulations; rather, it provides statistical solutions in response to the regulatory requirements. In the 2004 draft guidance, we did note that unfiltered trace element data can create interpretation and usage problems, particularly in geologic and hydrologic environments like South Dakota. The 2009 guidance briefly identifies these interpretation difficulties in the Chapter 5 discussion of background. There, we included summary information from EPA Region 8 on typical statistical patterns in background data. It was noted that local conditions and the type of sample collection could result in unfiltered samples containing higher background solids. These solids also contain naturally occurring trace elements, which can affect the overall background levels. Statistical testing can still be performed, but the average levels, detectability and variability of the data may increase for some constituents. We suggest that total suspended solids (TSS) also be

measured simultaneously with unfiltered trace element data; suspect data due to high natural trace element well solids concentrations can be better identified.

- **Commenters were concerned about the implications of using a 'moving window' for updating background data. While supportive of using such a technique to remove potentially unreliable early data, they felt it might impede identification of trends in the future (p. 46).**

EPA Response: The 2004 draft recommendations for updating background data have been revised in Chapter 5 of the 2009 guidance. We favor increasing background sample size to enhance statistical power and suggest tests which would be appropriate for periodic updating. We did not specifically include the 'moving window' concept in the 2009 guidance. It should be noted that the guidance identifies review periods (e.g., during permit development), when a more complete assessment of historical data is appropriate. Unreliable early data can often occur, and this would be a point when such problematic data might be removed from background. We do stress that there should be some reasonable basis for such removal; substitution of other higher quality data (e.g., from historical downgradient wells not considered contaminated) might compensate and allow enhanced data base sizes. While we cover formal outlier testing in Chapter 12, it is recognized that non-statistical decisions also play a role in determining the best quality background data.

- **Commenters questioned whether applying reference power criteria should be used, since the guidance indicated that effect size power can prove superior (p. 47).**

EPA Response: The discussion of the recommended relative reference power criteria, as well as the occasional use of effect size power evaluations, is similar to the 2004 draft guidance but has been redrafted in Chapter 6 of the 2009 guidance. We still suggest the reference power comparison as a means of systematically comparing detection monitoring statistical tests during the design period. The guidance offers examples of when effect size power evaluations might better substitute, but stress that the lack of absolute power criteria for most monitoring constituents would likely limit this option to only infrequent use.

- **South Dakota commenters felt that the Bonferroni adjustment (especially applying to prediction limit design) needed to be discussed in greater detail (p. 47).**

EPA Response: A more extensive discussion of the Bonferroni adjustment applying to prediction limit design is found both in Chapter 6 (Section 6.2.2) and in Chapter 19 (Section 19.2). In the 2009 guidance, the exact Binomial formula has been used instead of the Bonferroni approximation. However, guidance users will generally not need to make these calculations, since the Appendix D κ-factor tables use simpler inputs (background sample size, number of wells, constituents and frequencies of evaluation). R-script software is also available in Appendix C for calculating κ-factors for recommended tests.

- **Commenters raised a question of how to deal with non-detect values which may vary over orders of magnitude. Use of the ½ Detection Limit would result in very unreliable data (p. 48).**

EPA Response: Chapter 5 of the 2009 guidance covers the subject of background data evaluation. Some reported non-detect levels may be due to artificial reporting limits rather than analytical limitations. Occasionally, the earliest collected data may not be consistent with later samples. Judgment is needed during a data review, which might eliminate some of the more egregious large detection limit values considered as background. Chapter 15 of the guidance offers two methods for evaluating multiple detection limits, which can address reasonable analytical detection limit variability.

Reviewer — State of Ohio

- **Ohio commenters provided a number of suggestions for improving the guidance document—a thorough subject Index, a hyperlink text to the Glossary, listing of symbols used throughout the document, and a clarification of which statistical assumptions are absolutes versus interpretations (p. 49).**

EPA Response: We provide a subject Index in Appendix A of the 2009 guidance. While hyperlinks to the glossary is an excellent idea, it was beyond the capabilities of this guidance preparation effort, although we have somewhat expanded the Glossary itself. Also, given the extensive overlap and use of common symbols, we believe it impractical to bring together all of this information in a single location. A new Chapter 3 has been added, however, discussing basic statistical concepts (and including common statistical measures and symbols). There is also a discussion of the most important statistical assumptions in this chapter, as well as in more specific chapter discussions. Minimum sample sizes are discussed, but not indicated as absolutes.

- **Ohio EPA suggests the current discussion of developing/establishing background was inadequate. They also recommended additional guidance on the updating of background (pp. 50, 51, 53 & 64).**

EPA Response: We agree with these comments and thus, the 2009 Unified Guidance provides a separate Chapter 5 devoted to proper establishment of background and includes explicit guidance on when and how to update background. These commenters also favored the use of trend testing when updating background. While not the primary test recommendations, the 2009 guidance does recognize trend testing as a possibility (along with caveats discussed in Section 5.3.2).

- **Commenters were critical of guidance recommendations regarding pooling of background data from different geologic strata (p. 52).**

EPA Response: We emphasize in the 2009 guidance that such pooling is constituent-specific and likely to be limited to those constituents which exhibit site-wide uniformity (e.g., certain solubility-limited trace elements or never-detected constituents).

- **Ohio commenters felt that spatial variability could never be eliminated, even using intrawell methods (p. 52).**

EPA Response: We believe that this position is more of a semantic difference. Spatial variation because of well differences is eliminated. Observed patterns from a spatially varying well-field would continue within individual wells, but would be statistically treated as temporal variation.

- **Commenters favored listing methods no longer recommended by the Unified Guidance (p. 53).**

EPA Response: While not an exhaustive list of all previous test recommendations, most tests no longer recommended in the 2009 guidance are identified and discussed in Appendix B under "Historical Notes."

- **Commenters were critical of some of the Unified Guidance recommendations regarding when outlier testing should be performed (pp. 55-58).**

EPA Response: We agree that the guidelines on outlier testing should be clarified. This topic is covered in Chapter 5 of the 2009 guidance (Section 5.2.3) and Chapter 12. Outlier discussions are also found in other chapters as well. Some differences between background and other monitoring data outlier criteria are discussed. While some form of routine screening of potential background data during permit or other design development periods is encouraged, the guidance does not favor automated screening for reasons described in Chapter 5. Values flagged as 'outliers' might represent heretofore unseen measurements from the upper tail of the background distribution. Automatically excluding these values would then tend to drive an increase in the expected false positive rate at the site, assuming no impacts had yet occurred. Overall, the Unified Guidance continues to recommend a more cautious approach to outlier testing, so that each flagged outlier is examined as to whether it is or is not consistent with current conditions and laboratory procedures at the site.

- **Ohio commenters were concerned with the 2004 draft guidance's abandonment of Darcy's equation for estimating minimum time intervals between sampling (pp.59-61, 64).**

EPA Response: In the 2009 guidance, we have included a fairly extensive presentation of how Darcy's equation can be used for assessing sampling intervals (Chapter 14, Section 14.3.2). Certain caveats to its use, however, are discussed in Chapter 6, page 6-26.

- **Commenters raised considerable comments regarding the use of method (MDL), and quantification limits (QL) (pp.61-62 & 63).**

EPA Response: For the most part, this issue lies outside the scope of the guidance. Nationally, EPA is assessing the general subject of detection and quantification limits. Also see the response to a similar comment by an Industry representative above. 2009 guidance discussions are limited to potential ways of ranking variable reporting limits for certain non-parametric tests.

- **Ohio EPA was concerned with guidance recommendations to limit the number of indicator parameters (e.g., through leachate analyses). They felt that some situations don't allow for such evaluation, and it might conflict with Federal or State regulations (p.63).**

EPA Response: The 2009 guidance continues to recommend leachate analyses as one potential means of reducing the number of monitored constituents. The overall subject is discussed on pages 6-9 and 6-10 of Chapter 6. We have also suggested that some indicator parameters might better serve as indirect evidence of water quality releases and not necessarily be formally tested. The logic is that fewer monitored constituents can improve the statistical power of those remaining subject to detection monitoring. However, the guidance fully recognizes that the statistical aspects of monitoring design are only one part of the decisions to be made. If a regulatory program requires that more constituents be monitored, that is the appropriate response. We emphasize that this guidance attempts to provide judgments regarding better statistical decisions, but is not a complete response to the regulatory requirements.

- **Ohio criticized various facets of the Poisson prediction limit method, including its handling of non-detects and its comparison of the sum of measurements as opposed to individual values (pp.61-62, 64).**

EPA Response: We agree with these and other reviewer comments on the Poisson prediction limit. The 2009 guidance no longer contains this method.

- **Commenters questioned as to when trend tests might be preferable to prediction limits or control charts. They were concerned that trend testing would take longer to recognize a statistically significant difference (p. 65).**

EPA Response: The 2009 guidance indicates that trend testing may be appropriate when the stationarity assumption cannot be met for either prediction limits or control charts. There are a number of potentially complex situations which could occur; generally, the guidance limits discussion for a decision to use a trend test as ideally reached during the period of background data review (described in Chapter 5).

- **Ohio questioned as to whether applying mean-based compliance testing to a fixed MCL can be justified in terms of published information or references (p. 65).**

EPA Response: In Chapter 7, Section 7.3, the 2009 guidance indicates that the choice of statistical parameter versus MCLs or other fixed limits is a State program decision. We indicate that various regulatory programs have chosen mean and other upper percentile parameters for testing and that there is no national consensus. An earlier Federal Water Pollution Control Act reference to older RCRA fixed limits is discussed, but we could not locate more recent information in this regard for Safe Drinking Water Act MCLs. The guidance provides a range of compliance test options in Chapters 7, 21 & 22, once the parameter choice has been made.

Reviewer — State of Indiana

- **Indiana reviewers provided a number of language recommendations for the early regulatory discussions (pp. 68- 70).**

EPA Response: We have revised these chapters in the 2009 guidance, based on the suggested language.

- **Commenters preferred that the guidance address how to replace current practices under Part 265 groundwater monitoring requirements (p.70).**

EPA Response: Chapter 2, Section 2.3.1 of the 2009 guidance addresses this issue for interim status monitoring. We suggest the use of an existing provision for a groundwater quality assessment plan to allow for flexibility in applying guidance tests and recommendations. This might also help avoid the use of dependent aliquot replicate data, also discussed in Chapter 2.

- **Commenters were concerned that given the default normal distribution assumption, there appeared to be no need for distribution testing. (p.71.**

EPA Response: We make it clear in Chapter 10, Section 10.3 that the default normal distribution assumption applies only when there are too few sample data to perform distributional

testing described elsewhere in this chapter. Generally, with 8 or more samples in a data set, formal distributional testing is preferred.

- **Indiana questioned the 2004 draft guidance recommendation to never use field splits or duplicates in statistical testing. Further a structured test to account for multiple sub-samples shouldn't be referenced if never used (p. 71).**

EPA Response: This recommendation has been modified in the discussions of data mixtures in Chapter 6, pages 6-27 to 6-28 of the 2009 guidance. We make it clear that techniques for a structured test are available, but other options are easier and may be preferable.

- **Reviewers felt that the distinction of natural versus synthetic spatial variation was questionable (p. 71).**

EPA Response: We agree and have revised the language in the 2009 guidance to this effect. Whatever the source of spatial variation, the issue is similar from a statistical standpoint.

- **Indiana made the point that under the RCRA regulations, the owner/operator proposes the use of MDLs versus PQLs, subject to approval by the regulatory agency (p. 72).**

EPA Response: We agree with the comment, but presume that this regulatory approval process provides the agency with discretion to establish reasonable analytical method performance criteria.

Reviewer — State of Utah

- **The Utah reviewer questioned whether stationarity should also be included under checks of assumptions (p. 74).**

EPA Response: We agree with the comment, and have added stationarity to the discussion of key statistical assumptions in Chapter 3 of the 2009 guidance. It is also further discussed in other chapters on design and specific methods found in the guidance.

- **The reviewer found the 2004 discussions of tolerance intervals, prediction limits, etc. overlapping and confusing (p. 74).**

EPA Response: We agree with the comment, and have revised the 2009 guidance. In Chapter 6, individual detection monitoring tests involving statistical intervals are first identified in Section 6.4.4. Chapter 17 for formal detection monitoring tests separately addresses tolerance intervals and other regulatory ANOVA or trend tests. Confidence intervals are distinguished in Chapter 7 and addressed more fully in Chapters 21 and 22 dealing with compliance or corrective action tests against a fixed limit. Other historical uses of the tolerance interval are discussed in Appendix B.

- **The Utah reviewer questioned whether use of the logarithmic transformation (which can be back-transformed for prediction limits) also applied to other centrality tests (e.g., means or medians) (p. 75).**

EPA Response: We have clarified these discussions and identify the bias problem in back-transforming logarithmic or other data to a parameter like the mean. This would also apply to future mean prediction limits.

- **The reviewer noted that Chapter 3 of the 2004 draft guidance seemed to suggest that an absolute effect size was preferable to the reference power method (p. 75).**

EPA Response: The 2009 guidance clarifies that effect size power criteria differ between detection and compliance tests. Reference power criteria are recommended for detection monitoring test design in Chapter 6 as a means of systematically evaluating different tests. Effect size estimates of power may occasionally apply in certain situations, but no consistent absolute criteria are available for differences above some background level. In contrast, compliance testing (discussed in Chapters 7, 21, & 22) does use a form of effect size power evaluations when fixed limits are to be tested. However, these are structured so as to allow for greater generality, and may appear different than in other statistical texts.

- **The Utah commenter questioned the 2004 draft guidance suggestion that there was a mutual desire on the part of regulators and the regulated community to minimize the false positive error rate. He felt that minimizing the false negative error rate was of greater concern to regulators (p. 76).**

EPA Response: The 2009 guidance clarifies that balancing false positive and negative errors is more of a mutual concern with detection monitoring testing (see discussion in Chapter 6, Section 6.2.1). There is a greater asymmetry in risks pertinent to regulators and regulated entities when considering these errors under fixed limit compliance testing (Chapter 7, Section 7.4.1). Regulators will be more concerned with adequate power for compliance monitoring, and a sufficiently low false positive rate for corrective action. A regulated entity will likely view the relative risks for comparable tests in opposite fashion.

- **The reviewer felt that guidance discussions on spatial variability should also include Kriging. He also recommended time series statistics, including the autocorrelation coefficient for a stationary, univariate model (pp. 77 &78).**

EPA Response: Kriging is a very useful technique for defining spatial variability, but is beyond the scope of the guidance and is better addressed in specific geostatistical guidance and literature. In Chapter 14 of the 2009 guidance, we have added a basic discussion of autocorrelation, although some of the reviewer's suggestions are beyond the scope of this guidance (partial correlation, multi-variate time series, etc.). Again, other texts cover these subjects in much greater depth. (Note: The focus of this guidance is on directly applicable and formal tests for

initial investigation under RCRA regulations. Once remedial actions are contemplated, these additional techniques are likely to be applied).

- **The Utah commenter raised questions involving assessment of temporal correlation using a rank-based test when non-detects and ties are present (p. 78).**

EPA Response: The 2009 guidance addresses both issues, but it should be recognized that ranking ties and multiple non-detects is not completely resolved. Chapter 15 provides ranking methods when multiple detection limits are present. Also see the discussion on problems with tied observations using the rank-based vonNeumann ratio test on page 14-17.

- **The reviewer felt that Shewhart-CUSUM control charts could be used in the presence of a trend (p. 79).**

EPA Response: The 2009 guidance in Chapter 20 limits recommendations for using Shewhart-CUSUM control charts to stationary data. A problem with modifying control charts to remove a trend is that one could not be confident that such a trend will continue in the future. The guidance suggests the use of formal trend testing in Chapter 17 to evaluate this prospect.

- **The Utah commenter preferred Land's H-statistic as more exact over alternatives suggested by Dr. Anita Singh, (including use of Tchebyscheff's inequality) (p. 80).**

EPA Response: We agreed with the commenter, so long as a true logarithmic distribution is involved. The 2009 guidance continues to present the test method for Land's confidence interval of the arithmetic mean in Chapter 21 (including the upper confidence limit) as an option, but also provides a caution that it may be inappropriate in some instances. Data can appear logarithmic but actually be a more complex distribution, including temporal changes occurring with remedial action, aquifer attenuation, changes in the source extent and concentration, etc. We have also included a reference citation to Dr. Singh's work, particularly the ProUCL® program which can generate other robust mean confidence intervals. We also indicate that compliance testing of the lower mean confidence interval of a logarithmic mean will have fewer problems in this regard.

- **The Utah commenter questioned the risk ratio approach for compliance testing. He felt that the multiplicative risk factor assumption needed further explanation (p. 81).**

EPA Response: We disagree with the commenter on this point. While actual risk calculations involve multiple input factors, most are essentially constant compared to concentration. In effect, calculated risk is nearly proportional to concentration. This position is also taken in some CERCLA documents, including the use of a 2x factor (which is only suggested in this guidance). The 2009 guidance contains essentially the same positions as the 2004 draft guidance document. The R-ratio used is only an approximation, and is based on normal distribution assumptions. We do not believe a full derivation of Equation 22.2 is necessary, since it can easily be generated from the normal mean confidence interval t-test against a fixed standard. A more exact method

using the non-central t distribution is discussed in Section 22.1 for the alternate constant variance approach, and can be applied to Equation 22.2 as well.

Reviewer — EPA Region 3

- **Region 3 recommends that outreach and training be provided to Unified Guidance users (p. 83).**

EPA Response: Once the guidance is approved and finalized, EPA HQ will determine how best to conduct outreach and provide training.

Reviewer — EPA HQ Statistician

- **The statistician recommended using EPA's Quality Assurance Project Planning (QAPP) in developing groundwater monitoring data (p. 84).**

EPA Response: While we agree that the QAPP principles are worthy of consideration, the focus of the guidance relies upon the owner/operator to develop plans as part of permit activities. The QAPP guidance is more oriented towards sample and analysis planning under control of EPA.

Reviewer — State of Virginia

- **The Virginia reviewers questioned why the use of ½ practical quantification limit (PQL) should be used rather than ½ the detection limit (DL) in imputing values to reported non-detects (p. 85).**

EPA Response: This is a complex issue, and a full discussion is beyond the scope of this guidance. Nationally, EPA is assessing the overall subject of quantifying and reporting limits. Chapter 15 of the 2009 guidance provides newer methods for fitting values in data sets containing multiple non-detect limits, which can be superior to imputing arbitrary levels. The choice of using laboratory detection limits versus a practical quantification limits depends somewhat on the analytical methods involved. The most frequently used trace organic methods have established limits (usually PQLs) that represent good analytical performance in tested laboratories. For some trace element methods, laboratory-specific detection limits may be more realistic. These decisions would best be made at the time of permit development when assessing historical background data (Chapter 5 of the 2009 guidance).

- **The reviewers raised the question regarding the effects of removing an outlier on the performance of tests. They requested an example of how such tests might be affected (p. 85).**

EPA Response: This discussion has been revised in the 2009 guidance. Chapter 5 on background data evaluation provides an example of potential consequences from inadvertently removing real, but extreme data (Section 5.2.3). In the same discussion, it is indicated that for

certain tests involving a non-parametric prediction limit using a maximum value, the real power of the test (although not the reference power) can be decreased if the value used as the limit is considerably higher than is typical of background. It is also possible to avoid this issue using prediction limits with more retesting samples where the value used as the limit can be lower and more typical of the background distribution. Future value non-parametric prediction limits with retesting can also be assessed using the Optimal Rank Calculator.

- **Commenters raised concerns about the Poisson prediction limit (p. 86).**

EPA Response: We agree with these and other reviewer comments, and have dropped the Poisson prediction limit as a detection monitoring test in the 2009 guidance.

- **Virginia reviewers questioned how the single test error rate should be calculated based on an overall annual facility rate for detection monitoring design (p. 86).**

EPA Response: Chapters 6 (Section 6.2.2) and 19 (Section 19.2) of the 2009 guidance contains a more comprehensive explanation of how these individual false positive error rates are calculated. The Appendix D tables for prediction limit κ-factors do not require such calculations, although they are based on them. Simpler inputs (background sample size, number of monitoring wells and constituents, and the number of annual tests) are instead used.

Reviewer — Consultant

- **The commenter disagreed with the 2004 draft guidance downplaying of automated diagnostic testing. For large databases reviewed at a facility, some form of automated testing is necessary (p. 88).**

EPA Response: The 2004 draft and 2009 guidance discussions are not intended to discourage the use of statistical software. However, the subject is beyond the scope of the guidance. Our focus is on basic statistical concepts and relatively simple applications. We accept that consultants may need to use techniques beyond the present approaches.

Even using automated analysis, some caveats should be observed in understanding the implications. For example, the 2009 guidance on page 5-5 offers an example of potential problems with an uncritical rejection of outliers.

- **The reviewer felt that more emphasis should be placed on ladder-of-power transformations (p. 88).**

EPA Response: The 2009 guidance does specifically suggest considering ladder-of-power transformations in Section 10.2. The guidance places greater emphasis on the more commonly used normal and lognormal transformations (included in the larger suite of ladder-of-powers), since these have been found to fit many groundwater monitoring data.

- **The reviewer felt that indicator monitors should be more closely considered in terms of controlling the facility-wide false positive error rate (p. 89).**

EPA Response: The 2009 guidance suggests limiting the number of indicators and other parameters used as detection monitoring constituents, to control the false positive error rate in Chapter 5 (Section 5.2.1) and Chapter 6 (Section 6.2.2).

- **The consultant requested additional guidance on managing non-detect data and retest samples (p. 90).**

EPA Response: Chapter 15 of the 2009 guidance on managing non-detect data provides an expanded suite of methods, including for when multiple detection limits are present. The guidance also indicates a preference for using qualified data (e.g., "J" or "B") values versus simple assumptions of a non-detect level. Chapter 5, Section 5.3.3, covers situations for adding data to background from retest samples.

Reviewer — EPA Region 7

- **The Region recommended a simpler user's guide to avoid lengthy statistical background discussions. They felt that flowcharts were helpful (p. 91).**

EPA Response: While we have not created a separate user's guide, the 2009 guidance has been revised to make it more user-friendly. Historical notes and detailed power discussions have been place in Appendices B and C. The first Part of the document is structured to cover general regulatory and statistical concepts, detection and compliance/corrective action monitoring design, and a summary of recommended methods. The second Part covers diagnostic testing, and the last two Parts cover detailed detection and compliance monitoring methods. We chose to eliminate the detailed flowcharts in favor of a more flexible approach to data evaluation and permit/plan development.

- **The reviewer questioned whether the Disclaimer caveat that the guidance does not necessarily represent EPA policies is needed (p. 91).**

EPA Response: We do believe that this statement is necessary, since this is only guidance and not regulation.

- **The Region questioned the guidance statement regarding the use of an arithmetic mean versus a median in compliance/corrective action monitoring and asked under what conditions each might be appropriate (p. 95).**

EPA Response: The 2009 guidance indicates that there is no clear national consensus on which statistical parameter should be recommended for compliance/corrective action testing (e.g., when the standard is an MCL). In Chapter 7, Section 7.3, we review what is known about the bases for

various limits, but indicate that State or Regional RCRA programs will need to decide which parameter is the most appropriate. This could also include upper percentiles as well as the two centrality parameters mentioned. The guidance provides options for these different parameter choices in Chapters 21 and 22.

Reviewer — State of Wisconsin

- **The State wished clarification that the Unified Guidance is only advisory and will not affect their programs (p. 97).**

EPA Response: The Disclaimer at the beginning of the document indicates that the 2009 Unified Guidance positions are merely recommendations, and that other methods, approaches, etc. may be equally valid.

- **Wisconsin reviewers felt that addressing censored data was an important consideration (p. 97).**

EPA Response: Chapter 15 of the 2009 guidance provides greater discussion of issues and methods for managing non-detect data. In a similar response to other reviewers, we do note that the more general subject of appropriate detection or quantification limits is under study by EPA and is not dealt with directly in this guidance.

Reviewer — Dr. Robert Gibbons, PhD. University of Illinois at Chicago

- **Gibbons expresses concern that the draft Unified Guidance references ASTM standard D6312, but differs from its recommendations in several important ways (p.99)**

EPA Response: We believe the ASTM standard is still a valuable guidance for the statistical analysis of groundwater monitoring data, and reflects alternative strategies worthy of consideration. However, since adoption of the D6312, new research and evolving practice have led to changes in recommended guidance approaches. Intrawell tests are still featured prominently in the 2009 Unified Guidance, as well as giving further emphasis to using control charts. However, we are less convinced that accurate control charts can be constructed with 75% non-detects; it is also the case that Poisson prediction limits were dropped from the 2009 Unified Guidance. EPA is also not convinced that control charts are always superior to parametric prediction limits. There will continue to be differences between the ASTM standard and some of the approaches recommended within the Unified Guidance.

- **Gibbons suggests that the annual target SWFPR of 10% will be almost impossible to meet unless the ASTM approach is followed (pp.99-100).**

EPA Response: The Unified Guidance continues to describe a variety of methods for designing reasonable statistical programs. In particular, the available retesting strategies are not limited to 1-of-2 plans (i.e., a single resample). By implementing greater resampling, EPA believes the target SWFPR and recommended power of the 2009 guidance can be met at almost all facilities.

- **Gibbons is concerned that the Unified Guidance only recommends non-parametric prediction limits for interwell usage, not intrawell (p.100).**

EPA Response: We agree that the Unified Guidance should not preclude use of intrawell non-parametric prediction limits. The 2009 guidance has been revised to reflect this change.

- **Gibbons recommends that the Unified Guidance add gamma-based prediction limits to the normal-based limits currently included (p.101).**

EPA Response: Since sufficient field-testing and the evaluation of gamma-based limits have not yet occurred, the Unified Guidance, while it references Gibbon's gamma-based limits paper, does not provide full treatment as with the current normal-based procedure. Additional work has also been done on Weibull-based limits. These have also been referenced in the 2009 Unified Guidance.

- **Gibbons is surprised that the draft Unified Guidance provides so little guidance on control charts relative to the attention given to prediction limits (p.101).**

EPA Response: We agree, and the 2009 guidance has been substantially revised to address this concern. The guidance provides more equivalent treatment between these two methods in Chapter 20, but difficulties limited the ability to provide fully equivalent comparisons. It should also be noted that single well/constituent power is analyzed on a cumulative annual basis to complement false positive error rates (Chapter 6). We believe this approach allows for more realistic comparisons between control chart and prediction limit performance.

- **Gibbons is concerned about the relative lack of discussion concerning facilities that must use a mixture of statistical methods, rather than just a single type of test. He is also disagrees with the power standard suggested by the Unified Guidance for these situations (p.102).**

EPA Response: We agree that many if not most facilities will ultimately need to utilize a mixture of statistical methods. Discussion of this topic in the 2009 Unified Guidance has been expanded and clarified in Chapter 6. However, we disagree that minimum power should not be evaluated using the 'weakest link' principle. Since each well and constituent needs to be tested separately pursuant to the RCRA regulations, requiring even the least powerful of the methods selected by a

facility to achieve EPA's recommended power criterion ensures that each test will have an adequate chance of identifying contaminant releases. We do not believe that measuring power according to the 'weakest link' principle will be burdensome. Subsets of wells and/or constituents will have similar amounts of background data and will employ the same testing method. For each of these subsets, perhaps a single power curve might need to be generated to demonstrate comparability with the EPA reference power curve. This would only need to be done periodically, for instance in the initial statistical design and then again when background is periodically updated. We also believe that while this strategy should be considered in detection monitoring design, it may not necessarily be the most important constraint (see p. 6-6).

- **Gibbons welcomed removal of recommending ANOVA as a formal detection monitoring test and upper tolerance limits in compliance tests (p.103, 104).**

EPA Response: We believe it is important that the guidance cover all of those tests identified in the RCRA regulations. Therefore, we have restored a discussion of both ANOVAs and tolerance limits as potential monitoring tests in Chapter 17 of the 2009 guidance. However, our conclusions remain the same as in the 2004 draft guidance that ANOVA is generally not useful as a formal test in the presence of well spatiality. We have noted the changes regarding the upper tolerance limit test for compliance monitoring in Chapter 2, Section 2.3.3, and provide the upper and lower confidence intervals of an upper proportion in Chapters 7, 21, and 22 for compliance monitoring using the appropriate test hypotheses.

- **Dr. Gibbons raised the issue of 'cycling' between detection and assessment monitoring specifically with respect to Appendices I and II of the 40 CFR Part 258 regulations, and encouraged the guidance to discuss this more thoroughly (p. 104).**

EPA Response: The Appendix I list contains 38 hazardous constituents. Presumably, valid exceedances of these parameters would be sufficient to trigger compliance/assessment monitoring. However, Dr. Gibbons has alluded to use of an indicator as an example, which can be an alternative monitoring parameter allowed by regulation. The guidance does recognize this 'cycling' problem particularly with common indicators and other water quality parameters. See, for example, the discussion of interim status indicator monitoring in Section 2.3.1. We suggest that where indicator data are problematic, developing and maintaining a more realistic assessment monitoring scheme might be appropriate. Further in Chapter 6, we suggest that some indicator and water quality monitoring parameters might be co-sampled and analyzed, but not necessarily used as formal detection monitoring constituents. On a case-by-case basis, this issue should be worked out between the facility and regulatory agency. We also discuss some potential problems with total (or unfiltered) trace element data required under Part 258 (see Chapter 4, Section 3.4 and Chapter 5, page 5-11). Cycling might also occur with some of these constituents if turbidity levels are variable, due primarily to entry of additional levels of these constituents from background geologic materials.

- **Gibbons welcomed removing 'never-detected' constituents from computing power, but had differences regarding trace elements (pp. 104).**

EPA Response: This discussion has been amplified in Chapter 6 of the 2009 guidance. We still believe that certain trace element data can meet the same criteria as organics (e.g., mercury or beryllium). If newer analytical methods (e.g., ICP-MS) are used, the lower sensitivities might change the outcome. However, a sufficient background would still need to be established.

- **Gibbons favored multiple well/constituent power evaluations over the 'least powerful test' (p. 105).**

EPA Response: Use of multiple power analysis is discussed in Chapter 6, page 6-22 of the 2009 guidance. We believe that the least powerful single constituent test or 'effective power' is most appropriate under the RCRA regulations, but note possible limitations (p. 6-6).

- **Gibbons favors intra-well comparisons for most monitoring (p. 106).**

EPA Response: The 2009 guidance indicates that such decisions are constituent-specific. It is noted in Chapter 5 that most common ions and indicators would be best tested using intra-well methods.

- **Gibbons didn't think single censoring point data assumptions were realistic (p. 107).**

EPA Response: We have revised Chapter 15 of the 2009 guidance to include and generally recommend multiple detection limit techniques. However, we also include single-censoring methods, which may still be appropriate for certain data sets, and are easier to apply.

- **Gibbons suggested that Sen's test could be used to derive a confidence interval on the slope estimator (p. 107).**

EPA Response: We have added a comparable technique as the Theil-Sen confidence interval in Chapter 21 of the 2009 guidance. An R-script is also provided in Appendix C to conduct the calculations.

- **Gibbons agreed with the use of the default normality assumption when insufficient sample size precludes formal distributional testing (p. 108).**

EPA Response: The 2009 guidance, Chapter 10, follows the same logic as in the 2004 draft guidance.

- **Gibbons favored limiting monitoring constituents through evaluating leachate data and felt that 3-4 indicators might be reasonable (p. 109).**

EPA Response: We generally followed this logic in Chapters 5 and 6 of the 2009 guidance, and suggest limiting the number of indicators (especially in Section 6.2.3).

- **Gibbons disagreed with the small sample study in section 10 of the 2004 draft guidance and provided reasons. He also disagrees with the modified Aitchison's method (p. 110).**

EPA Response: We agree with Gibbon's arguments. The 2009 guidance no longer contains the small sample study. In addition, Chapter 15 no longer presents Aitchison's method, but the modified delta model is suggested as a possible option for multiple distributions involving non-detect data.

- **Gibbons felt that future mean prediction limits might be less sensitive to gradually increasing contamination at a monitoring well (p. 112).**

EPA Response: While we partly agree with Gibbon's arguments, the situation is too site- and well-specific to generalize. Shorter periods of sampling might avoid the problem raised.

- **Gibbons questioned why repeat or verification testing couldn't make use of regularly scheduled sampling events (p. 113).**

EPA Response: We believe that, especially for prediction limit applications based on exact error calculations, the use of overlapping samples (using next regular sample as a repeat sample for the previous initial sample) would induce changes in the error rate due to this dependency. For higher level tests (e.g., a 1:4 prediction limit), it could also take two years to confirm a release. Therefore, we did not include this suggestion in the 2009 guidance.

Reviewer — Dr. Anita Singh, Statistician

- **Dr. Singh was concerned with the default use of the lognormal distribution for groundwater data sets (p. 120).**

EPA Response: Both the 2004 draft and the 2009 guidance suggest the use of a default normal distribution assumption, primarily when data sets are too small for formal testing. We encourage the use of formal tests in other circumstances. The 2009 guidance does not recommend using a default lognormal distribution assumption. Simulations summarized in Chapter 10 with details in Appendix B support the superiority of the normal distribution, even when the data may be truly lognormal. The guidance also suggests the use of the ladder-of-powers transformation options in Section 10.2. While the guidance mentions the potential use of Gamma or Weibull distributions,

we conclude that their application is beyond the capabilities of most users. Interested parties can use the references provided or consult the wider statistical literature for these options, however.

We do believe that the lognormal distribution has applicability to groundwater data sets. In particular, background distributions for trace elements and other monitoring parameters may be well fitted by this distribution in two- or multiple-sample comparisons. The guidance does caution against the uncritical use of the upper confidence limit of the logarithmic mean test (e.g., pp. 10-3; 21-9); it is recognized that compliance data in particular may exhibit apparent lognormality, but in fact be the result of more complicated effects including data trends. This guidance does not address applications to soil contamination, which can exhibit much greater variability than most groundwater data sets.

- **Dr. Singh was concerned with basing a normal distribution determination on the sample coefficient of variation (p. 120).**

EPA Response: Both the 2004 draft and the 2009 guidance only suggest the use of the sample coefficient of variation (CV) as an informal means of checking potential normality. We have added Chapter 3 to the 2009 guidance discussing basic statistics, including the normal and lognormal coefficient of variation estimators. Chapter 10 of the 2009 guidance covers the CVs and coefficient of skewness (Section 10.2) as approximate measures, and also provides Example 10-1, where the logarithmic CV better approximates the true data variability. Chapter 10 of guidance stresses the use of formal normal distribution tests, when there are sufficient data.

- **Dr. Singh raised an issue of adequate sample sizes to apply the Central Limit Theorem (CLT) (p. 120).**

EPA Response: Although we agree with the sense of Dr. Singh's comments, the guidance does not directly apply the CLT. It is identified and discussed in general terms in Chapter 3, Section 3.5.2 of the 2009 guidance, but more as a background statistical assumption. She is concerned with the validity of a normality assumption in evaluating (untransformed) mean data. For most detection monitoring tests covered in Part III, the arithmetic mean need not be the only parameter tested. Even when compliance or corrective action monitoring is involved (Part IV), the guidance concludes that the parameter of choice is somewhat ambiguous from a RCRA regulatory standpoint (Section 7.3). Therefore, the guidance provides a number of centrality and upper limit tests, which can address a variety of parameter test choices.

- **Dr. Singh cautioned against the use of Welch's test and ANOVA on logarithmically transformed data as potentially inappropriate for testing mean differences (p. 120).**

EPA Response: We have revised the discussion of Welch's t-test to recognize that unequal logarithmic variances between data sets can create ambiguous results (Section 16.1.3). We do not believe that this problem is as serious for the pooled variance t-test or ANOVA, both of which depend upon an assumption of a common variance. In Chapter 17, the guidance does argue against the likelihood of ANOVA tests where well spatiality is present. The issue might

have some relevance for the use of the Student- t or ANOVA as diagnostic tests (Chapters 5, 6, 13 and 14). But it is again noted that the untransformed mean is not necessarily the only parameter which might be compared in detection monitoring or diagnostic tests.

A new Chapter 9 has been added to the 2009 guidance, which covers a number of graphical exploratory techniques used throughout the document. Whenever informal or exploratory methods are used (e.g., in diagnostic evaluations), the guidance is careful to include formal tests as well.

- **One of Dr. Singh's most important concerns was to caution against the uncritical use of the upper confidence limit (UCL) of the lognormal mean and to consider other methods for generating confidence intervals. She also was concerned with various test applications in the presence of censored data (p. 120).**

EPA Response: We agree with Dr. Singh's comments, and have added a reference (p.21-9) to the website containing the software program she developed for EPA [ProUCL®]. This program generates a variety of parametric and non-parametric or bootstrap confidence intervals around the untransformed mean. We conclude that the subject of applying various tests in the presence of non-detect data is important, but one requiring considerable further research and mostly outside the scope of this guidance. The guidance does provide more robust methods for fitting multiple limit non-detect data (Chapter 15), as well as the Tarone-Ware test for two-sample comparisons involving non-detect data and Theil-Sen trend confidence intervals.

- **Dr. Singh recommended additional measures and robust techniques for assessing outliers, beyond Dixon's and Rosner's tests based on normality (p. 120).**

EPA Response: We agree with Dr. Singh's comments, and have added Tukey's box plot method in Section 12.2. An additional reference to Barnett and Lewis (1994) is provided on page 12-11, where the 'swamping' effect is addressed. We have also added language at the end of Chapter 12, which suggests that the wider statistical literature may need to be consulted for other robust outlier evaluation techniques.

Reviewer — California State Water Board (CASWB)

- **CASWB felt that the 2004 draft guidance did not satisfactorily address the use of background limits as a major option under compliance and corrective action monitoring. The test hypotheses differ between fixed limit and background tests (pp. 152, 159).**

EPA Response: We agree with the comment and have revised the 2009 guidance accordingly. Section 7.5 covering compliance/assessment and corrective action monitoring design discusses options for standards based on background. The test hypothesis structure there is identical to that for detection monitoring (Chapter 6). Chapter 2 also provides direct regulatory language in selecting ground water protection standards for both RCRA Subtitle C and D programs. We also

note on page 7-3 that regulatory programs have the discretion to define test hypotheses in a different manner than suggested in the guidance, based on their perceived program needs.

- **CASWB reviewers found discussion of groundwater protection standards (GWPS) to be inadequate between Subtitle C and D programs (pp. 152-6).**

EPA Response: We agree with the comment and amplified the discussion of similarities and differences in the Subtitle C and D groundwater protection standards (GWPS) in Chapter 2, Section 2.2.5 of the 2009 guidance. We have continued to use the GWPS term in a generic fashion throughout this guidance, since it is widely recognized even if specific regulatory applications differ. We did not agree, however, with the position of the commenters that the exact language differences in the two programs needed to be as closely spelled out as suggested.

- **CASWB felt that the guidance did not recognize that State RCRA programs can be more stringent than the Federal rules, particularly in applying background level standards in compliance or corrective action testing (pp. 152-3, 161 & 162).**

EPA Response: We agree that the RCRA State programs do have this authority to be more stringent. The existing RCRA regulations clearly allow for use of background as a comparison standard (pages 7-19 & 7-20 of the 2009 guidance). CASWB comments suggested particular strategies for dealing with plume contamination. While specific to State needs, a full discussion is beyond the scope of this guidance. But nothing in this guidance is intended to interfere with or supplant current state programs (see initial Disclaimer). In Section 7.3 of the 2009 guidance, we also suggest that State or Regional programs will have to make decisions regarding the most appropriate parameter for testing, among other aspects of compliance design.

- **Reviewers believed that the guidance did not recognize the parallel use of the terms "compliance" and "assessment" monitoring between Subtitle C & D programs (p. 157).**

EPA Response: Footnote 1 on page 4-4 of the 2009 guidance makes this distinction clear.

- **CASWB requested that Appendix tables be clarified to identify symbols used and indicate if one- or two-tailed tests are assumed (p. 158-9).**

EPA Response: Appendix D tables in the 2009 guidance have been clarified to indicate authorship, use of symbols, etc, as suggested by the reviewers. In addition, where appropriate, either the table or the guidance text identifies whether tabular data is for a one- or two-tailed test.

- **Reviewers approved of the use of "never-detected" constituent removal from false positive error calculations (e.g., for prediction limits in detection monitoring), but raised a concern with occasional detects as a result of analytical limitations (pp. 164-165). They also proposed a "non-statistical" test approach in Appendix III to their comments.**

EPA Response: We have modified the approach suggesting removal of "never-detected" constituents from detection monitoring false positive error calculations in Chapter 6, pages 6-11 through 6-13 of the 2009 guidance. We suggest flexibility in application including a possible second resample, if there is doubt about a true release. We disagree that it is not a "statistical" test, but it is accepted that no exact false positive error rate can be assigned. We also further discussed limiting the number of formally tested constituents in Section 6.2.2. Although we did not make use of reviewers' Appendix III suggested language, the approach appears to be very similar to the guidance.

- **CASWB provided a description of additional well installations and more sophisticated plume tracking as part of compliance/corrective action monitoring in their State. They suggested adding such language to the Unified Guidance (pp. 165-171).**

EPA Response: Although we conclude that such measures are beyond the scope of the guidance which focuses on direct regulatory monitoring, we acknowledge that more sophisticated measures are appropriate at the program level. We briefly mention geostatistical analysis or fate-and-transport modeling as possibilities at the beginning of the document. Details of such methods can be found in many statistical literature sources. Discussions regarding the number of wells, etc., are more appropriately handled in other groundwater monitoring guidance.

Reviewer — Dr. Charles Davis, PhD. Envirostat, Ltd.

- **Davis believes the Unified Guidance should not be released until the power characteristics of control charts can be evaluated relative to the ERPC (p.208).**

EPA Response: We agree that the 2004 draft Unified Guidance was incomplete with respect to its treatment of control charts. A new comparative study of control charts and prediction limits was attempted, but could not be completed as part of the 2009 guidance due to limited funds, timing and technical problems. The guidance, however, does provide greater detail in Chapter 20 on choosing an appropriate control limit based on well network size/configuration, background data, and the retesting strategy. This will make implementation of control charts more similar to prediction limits; however, power assessment for a specific control chart configuration still needs to be done using Monte Carlo simulation.

- **Davis commented favorably on the small sample study in Section 10 involving non-detect data, but suggested more technical reports for review (p.209).**

EPA Response: Based on peer reviewer and other comments, this study was dropped from the 2009 guidance.

- **Davis has a question about facilities 'stuck between' detection monitoring and compliance monitoring (p.209).**

EPA Response: While a 'cycling' back and forth between detection monitoring and compliance monitoring could potentially occur at facilities where background has been exceeded yet no well exceeds an applicable GWPS, such facilities may not need to return to detection monitoring but rather could stay in compliance monitoring indefinitely. A more difficult circumstance is encountered when there have been previous (perhaps historical) impacts which cause the GWPS to be exceeded, yet more recent data indicate downward trends and corrective action has been ruled unnecessary. In this setting, a facility may be warranted in crafting a hybrid statistical

program containing elements of both detection monitoring and compliance monitoring. One example would be to track intrawell trends (detection monitoring), but to compare recent data for an upward trend relative to the historical well situation. Brief discussions of this topic are found on pages 6-31ff, 6-41ff and 8-3.

- **Davis was concerned about the problem of addressing the four successive sample requirement in the RCRA regulations and felt better distinctions between Subparts C and D needed to be provided (p.210). He later raised an issue of the need for full Appendix IX monitoring following indications of a release (p. 217).**

EPA Response: Modifications made to the Part 264 hazardous waste regulations in 2006 are discussed on page 2-5 of the 2009 guidance. One of the new provisions allows other sampling frequencies than the four successive sample requirement, a position followed in this guidance. Differences in Subparts C and D for sampling requirements are also more clearly discussed in Section 2.2.4, and for groundwater protection standards in Section 2.2.5. Other provisions allow for discretion in the number of Appendix IX constituents and wells which must be monitored for.

- **Davis disagreed with the 2004 draft guidance in not recommending non-parametric intrawell prediction limits. He also offered a recommendation to utilize all available historical site data in developing sufficient background sample sizes (p. 211).**

EPA Response: We have revised the discussion of non-parametric prediction limit applicability to intrawell monitoring (page 19-27). As a general principle, we have adopted Davis' suggestion to consider all available historical monitoring data in developing background (p. 5-9).

- **Davis questioned the guidance's lack of discussion regarding 40 CFR Part 265 interim status monitoring (p. 213). In a subsequent comment, he favored the use of potential detection monitoring parameters (under Part 264) rather than the existing Part 265 indicators and other limited constituents (p. 219).**

EPA Response: We agree with his comments, and have added suggestions for improving interim status monitoring and design in Section 2.3.1 of the 2009 guidance. These include the use of a relatively permanent groundwater assessment plan, which could be designed with Part 264 monitoring in mind. We also note that the RCRA permit provisions in §270.14(c) require a characterization of potential groundwater hazardous constituents. These can and have been added to interim status monitoring lists by individual state and EPA Regional programs.

- **Davis raised a concern about the meaning of the term 'exceedance', particularly in comparisons to a GWPS. His experience included determinations that any single exceedance might be considered significant (p. 213).**

EPA Response: Chapter 7 of the 2009 guidance covers design of a compliance or corrective action monitoring program. Included there are discussions regarding the choice of testing parameters (Section 7.3). Because of overall ambiguity on the matter, the guidance indicates

that individual State or EPA Regional RCRA programs will need to make such decisions. The guidance provides a statistical framework for a number of test parameter choices in Chapters 21 and 22.

- **Davis favored the 2004 draft guidance approach of not including ANOVA tests as part of formal detection monitoring and raised an issue of sufficient sample sizes to conduct an ANOVA, particularly if replicate, dependent samples were used (p. 216).**

EPA Response: We agree with his comments, but note that the 2009 guidance has been restructured to include all of the tests covered in the RCRA regulations. However, the guidance clearly indicates those typical conditions where ANOVA testing would not be appropriate (Section 6.4.2). Minimum sample sizes are noted as a potential problem. Moreover, we have stressed the need for independent data and the problems occurring with replicate (aliquot) samples in Chapter 2, page 2-11.

- **Davis raised additional factors to consider in distinguishing physical from statistical independence (p.220)**

EPA Response: Chapter 3 has been added in the 2009 guidance to discuss basic statistical concepts. Statistical independence is presented more fully in Section 3.2.1; it is discussed further in regard to detection monitoring design (p. 6-4) and as opposed to physical independence on page 14-2.

- **Davis has concerns regarding the reference level test for detection monitoring power criteria. In particular, he provided fractional power data for this test for different background sample sizes (p. 224-225).**

EPA Response: We have modified this prediction limit reference test in the 2009 guidance to provide annual power to identify a significant increase, based on otherwise similar test assumptions used in the 2004 draft guidance. Thus there are now three reference curves, one each for 1, 2 or 4 tests per year. This was done to make power evaluations comparable to the annual cumulative false positive error rate. Chapter 6, Section 6.2.3 describes this approach in more detail. We also believe this will allow for more consistent comparisons between control charts and prediction limits.

- **Although agreeing with the approach, Davis questioned whether use of the 'never-detected' constituent criteria to remove these constituents from testing would violate the RCRA regulatory testing requirements (p. 225).**

EPA Response: We believe that the detection monitoring regulations in Part 264 clearly allow discretion in selecting monitoring constituents and that the 'never-detected' approach discussed in Chapter 6 of the 2009 guidance is permissible. As a point of clarification, these 'never-detected' constituents **are** tested, using the Double Quantification Rule. We have also suggested in Chapter 6 that not every indicator parameter need be formally tested, although they could

continue to be regularly sampled and analyzed. These additional water quality data might be useful in assessing the likelihood of a release based on a statistically significant determination for a formal monitoring constituent.

- **Davis provided suggestions for an alternative approach to utilizing effect size power, based largely on the usefulness of potential monitoring constituents and their presence in leachate (p. 225-6).**

EPA Response: Although we did not directly use this approach in the 2009 guidance, we have expanded discussions regarding the careful selection of monitoring constituents (including using leachate data information) in Chapter 6, Section 6.2.2. Effect size and data-based power analyses are discussed in Section 6.2.4.

- **Davis raised a very interesting concern regarding use of ANOVA to evaluate well spatial differences. It is possible that a negative conclusion might be reached in applying ANOVA solely to upgradient wells, yet have significant downgradient well spatial differences (p. 229).**

EPA Response: We agree that this is a valid possibility, and there have been instances where, for example, upgradient wells at a facility are located in a lower concentration saline zone than downgradient wells. Although not directly addressed in the 2009 guidance, it is implied in guidance to consider and evaluate all historical data during permit or plan development. This would include conducting one-way ANOVA tests for upgradient and downgradient well spatial differences, e.g., in developing background data (Chapter 5). Some non-statistical judgement might be needed to eliminate likely historically contaminated downgradient or other wells from the diagnostic tests. Evaluating all historical well data would also be applicable in evaluating prospects for pooled variance described in Chapter 13, a technique developed and recommended by Dr. Davis.

- **Davis felt that the guidance should provide a means of evaluating that background data used for detection monitoring are at 'steady-state' (p.234).**

EPA Response: We agree with his and the Utah reviewer comments in this regard, and have added 'stationarity' as a fundamental assumption (discussed in Chapter 3 of the 2009 guidance).

- **Davis indicated that Poisson prediction limits did not appear to be valid tests (p.235, 240).**

EPA Response: We have removed the Poisson prediction limit test from the 2009 guidance, based on peer reviewer and other comments.

- **Davis advocated the use of the (Filliben) probability plot correlation coefficient test in assessing normality, but suggested that formal test outcomes using exact false positive error rates might be unnecessary for preliminary testing (p.235). He advocates a similar approach using Levene's test for equality of variance (p. 236).**

EPA Response: The 2009 guidance does present Filliben's test formally, and provides tabular test outcomes based on critical false positive error levels in Table 10-5 in Appendix D. However, his point is well taken that a rough measure of normality may be sufficient (e.g., an r > .9) for preliminary diagnostic tests. The same applies to the equality of variance test.

- **Davis offered suggestions for practical limitations in using the ladder-of-power transformations (p. 241).**

EPA Response: We have adopted these suggestions on page 10-4 of the 2009 guidance.

- **Davis suggests use of a 2-way ANOVA for adjusting degrees of freedom in intrawell tests (e.g., prediction limits; p.244)**

EPA Response: We agree that when both temporal variation and spatial variability appear to be present, the 2-way ANOVA approach developed by Davis is appropriate. It is presented in the 2009 Unified Guidance under Chapter 14, Section 14.2.2, although labeled as a One-Way ANOVA for Temporal Effects (including a spatial component).