Response to SC&A Comments on ORAUT-RPRT-0053

Rev. 0
August 21, 2013

Thomas LaBone, Nancy Chalmers, and Daniel Stancescu

Introduction

ORAUT-RPRT-0053, *Analysis of Stratified Coworker Datasets*, was issued in July 2012. In April 2013, SC&A issued comments and findings on this report [Chmelynski 2013], eight of which were entered into the Board Review Application. In this document, the authors of RPRT-0053 offer responses to the findings and comments made. The responses are given in three sections:

- Section 1 contains a high-level overview of what we feel are the major points where there is disagreement between SC&A and NIOSH concerning RPRT-0053.
- Section 2 contains responses to the eight findings on RPRT-0053 given in the SC&A report.
- Section 3 contains detailed responses to comments made in the SC&A report.

The technical detail increases as one goes from Section 1 to Section 3. Material taken from the SC&A document is presented in italics font in this report. SC&A comments related to the application of RPRT-0053 to data from the Savannah River Site (e.g., RPRT-0056) are not addressed here.

Section 1: Overview

Dr. Chris Chatfield gave what we consider to be a useful definition of statistics in his book *Problem Solving: A Statistician's Guide:*

> "Statistics is concerned with collecting, analyzing, and interpreting data in the best possible way, where the meaning of "best" depends on the particular circumstances of the practical situation."

The authors of RPRT-0053 have presented what we view to be the "best" approach to the problems at hand, based on our view of the situation and our personal experiences. We are confident that SC&A has done the same, but this is no guarantee that our "best" approach will be the same as their "best" approach. Below we have summarized what we consider to be the most important differences between our views and those of SC&A. A more detailed discussion of these issues is given in Section 3 of this report.

Scope and Purpose of RPRT-0053

In coworker modeling we have some dosimetry data from workers monitored in a given year and from that data develop a model to predict dosimetry data for other workers who

were not monitored. One wonders if we could derive more accurate predictions of dose to unmonitored workers if we take into account information like work location and job title, i.e., stratify the coworker model. The answer is "perhaps", depending on whether or not there is truly a difference in the doses to the workers in the different strata and whether or not we can accurately assign workers to the right stratum.

RPRT-0053 was written to provide statistical tests to help us decide whether or not two groups (i.e., strata) in a population of monitored workers have intakes or doses that are different enough to warrant a coworker model that explicitly incorporates this difference. If there is not enough of a difference to justify stratification, the standard unstratified model is used. SC&A seems to have reviewed RPRT-0053 with a different approach in mind, i.e., they propose the use of a stratified model by default and revert to the unstratified model only if the data are sufficient to justify such an action. Not unexpectedly, RPRT-0053 is not well suited to this application and the SC&A comments reflect this.

Power and Retrospective Data

The statistical procedures in RPRT-0053 are designed for the analysis of retrospective coworker data, i.e., data that are provided to us "as is." SC&A devotes a considerable portion of their report discussing topics like *a priori* power, gray regions, etc, which are only applicable during the design phase of an experiment, survey, or other data-generation effort. Because there is no design phase associated with coworker data, we feel that most of this discussion is not applicable to RPRT-0053.

OPOS

In our opinion, the most relevant issue discussed by SC&A is the use of the "one person - one sample" (OPOS) statistic. OPOS was designed to address the problems associated with individuals submitting more than one sample in a year. These problems are:

- Data dominance: a large fraction of the samples being submitted by small fraction of the individuals.
- Correlated data: multiple samples submitted by an individual can be correlated, which greatly complicates the use of statistical tests.

We consider these to be major issues, ones with which the use of the OPOS statistic effectively deals. While more rigorous solutions to these problems may be available, we do not think it is feasible to use them in our situation. SC&A did not comment on the problems of data dominance or correlated data or whether or not the use of OPOS statistics is useful for dealing with them. We would be interested in SC&A's comments

on these issues because the review appears to advocate the continued use of individual bioassay results over the use of OPOS statistics. We do not feel that this is a technically viable path forward if one wants to test for differences in strata.

Section 2: Responses to Findings

RPRT-0053-1

Due to the dependencies that exist in the ranked data, the R^2 for ROS does not have the usual interpretation. The recommendations in RPRT-0053 for using ROS do not address this concern.

R^2 is not mentioned anywhere in the text of RPRT-0053 as a goodness of fit criteria. However, the R^2 statistic appears in some ROS plots. We think this was done at the request of someone in ORAUT, perhaps to be consistent with previous practice (i.e., PROC-95). R^2 was not used by the statisticians to evaluate fits in ROS plots so we don't think this topic warrants a "finding" and does not need to be addressed in RPRT-0053.

The applicability of the R^2 statistic in the evaluation of cumulative probability plots was previously raised by SC&A in their reviews of PROC-0095 and OTIB-0019. All findings related to this issue were resolved and closed in 2007 under the OTIB-0019 review. The closure language can be found in the Board's review system.

(See Comments 4 and 5).

RPRT-0053-2

In the application of the procedures recommended in RPRT-0053, the issue of completeness of the available coworker data has not been addressed. If the unmonitored workers are from a different population, the applicability of a coworker model derived from monitored coworkers would be in question. The matter of the relative exposure potential of the monitored workers needs to be demonstrated rather than assumed. The methods proposed in RPRT-0053 for analyzing the coworker datasets require verification that (1) the available coworker data are representative of all groups of workers, and (2) the manner of use of the data is claimant favorable for the specific datasets to which the method is applied. A sound statistical methodology is subject to these two important caveats.

An implicit assumption in any statistical analysis, including those in RPRT-0053, is that the data being analyzed are representative of the population in question (e.g., are "complete"). In our opinion, the issue of data completeness is not within the scope of RPRT-0053 and should not be identified as a "finding."

We agree with the statement "*If the unmonitored workers are from a different population, the applicability of a coworker model derived from monitored coworkers would be in question."* This relates back to the issue of stating why unmonitored workers were not monitored. In the development of coworker models we assume that either:

- unmonitored individuals are members of the monitored population who were not monitored completely at random, or
- unmonitored individuals were unmonitored because they had no potential for exposure to radioactive materials.

In the first case we have the right model and in the second a conservative model. One can also theorize that these assumptions are wrong and that perhaps unmonitored workers were highly exposed and intentionally not monitored because of this. This fundamental and largely unstated difference in assumptions probably needs to be discussed and eventually resolved.

We disagree with the statement "The matter of the relative exposure potential of the monitored workers needs to be demonstrated rather than assumed." The reason why individuals were unmonitored will, in general, always be largely an assumption that cannot be "proved" to the satisfaction of everyone. The validity of this assumption is basically a function of the maturity of the radiation protection program in place at a given facility and the level of documentation available.

The concept of "claimant favorability" expressed in this comment is confusing. To illustrate the problem, assume Group A and Group B have very different dose distributions, with A higher than B, and we combine the doses from the two groups to form Group C. The dose distribution for C will overestimate the dose to B (be "claimant favorable") and underestimate the dose to A (not be "claimant favorable"). But, we typically assume that workers with higher doses are less likely to be unmonitored, so this is not considered a problem in practice. Now, if we stratify the model, the doses to both A and B will be more accurate. However, the dose to B will be lower and less "claimant favorable" and the dose to A will be higher and more "claimant favorable" than the combined dose distribution. So, in principle, no coworker model(s) can be "claimant favorable" to all strata in the model at the same time.

(See Comments 7, 8, and 26)

RPRT-0053-3

The OPOS statistic methodology summarizes a worker's exposure by averaging overall urine samples collected during the specified time period. The use of average values does not account for variability of the samples within the time period and the procedure will result in lower values of the GSD used in the coworker model.

In the presence of data dominance and dependent data (see Comment 9), the GM and GSD calculated with individual bioassay measurements do not have familiar statistical properties and are therefore not useful measures of central tendency and variance of the data. The OPOS statistic was adopted in an effort to deal with these major issues. We feel that the use of the OPOS statistic better achieves the goal of accurately estimating the intake rates and ultimately the dose to workers than does the use of individual bioassay results. Thus, it is not relevant whether or not the OPOS statistics have a higher or lower GSD than the individual data.

(See Comment 10)

RPRT-0053-4

The OPOS method must strictly be applied to comparisons where the sampling protocol was the same. Specifically, when there is evidence that the sampling protocol for one group of workers was different than the protocol used for the other group, the tests do not provide a valid comparison. For example, if the monitoring of one group of workers is incident-driven and the other is not, then the OPOS approach is not appropriate for comparing the two distributions.

We believe that there may be some confusion here concerning the use of the statistical term "sampling protocol." One definition of the term is[1]

> "The sampling protocol is the procedure used to select units from the study population to be measured. The goal of the sampling protocol is to select units that are representative of the study population with respect to the attribute(s) of

[1] http://sas.uwaterloo.ca/~rwoldfor/papers/sci-method/paperrev/node40.html

interest. The sampling protocol deals with how and when the units are selected and how many units are selected."

Thus, the sampling protocol tells how one might select individuals (i.e., a sample) from a population of people with the intent of inferring population parameters from the sample.

In this comment the statistical term "sampling protocol" is incorrectly used as being synonymous with the term "internal dosimetry monitoring program." There is no statistical requirement that all workers be on the same monitoring program in order to use the data to develop a coworker model, as long as the monitoring programs adequately characterize all significant intakes. Further, most sites had *graded* monitoring programs where the frequency and types of bioassay performed were based on the likelihood of the workers having a significant intake of radioactive material[2]. Even today this is standard radiation protection practice, so we would expect the bioassay (i.e., sampling) protocols to be different for different groups of workers.

Given all of this, we feel that it is appropriate (for example) to compare intakes calculated from "special" and "task-related" bioassay performed in one group to intakes calculated from "special", "task-related", and "confirmatory bioassay[3]" in another group.

(See Comment 12)

RPRT-0053-5

The methods in RPRT-0053 require a high level of confidence before deciding that the two worker groups are significantly different. The requirement for a high level of confidence in this decision is not claimant favorable when using a null hypothesis of "No Difference." The power of the tests to detect differences given the limited quantity of available data has not been established. The Data Quality Objectives (DQO) process should be used to balance Type 1 and Type 2 decision errors.

As discussed in the response RPRT-0053-2, one cannot be "claimant favorable" to both groups at the same time. Therefore, the statement that "... *requirement for a high level of confidence in this decision is not claimant favorable...."* is ambiguous. In addition, we consider the 95% confidence level to be consistent with standard statistical practice and

[2] Graded monitoring is also common in external dosimetry programs.
[3] The terms *special*, *task-related*, and *confirmatory* are defined in paragraphs 5.3 and 5.4 in ICRP 1997.

appropriate in this application. This and the issue of the statistical power of the test are mentioned in several comments and are discussed in the response to RPRT-0053-8.

All the statistical tests in RPRT-0053 for comparing two groups use the null hypothesis of no difference between the groups, i.e., they are two-sided tests. A one-sided test is used when the question at hand concerns one group tending to have higher values than the other group, and the reverse relationship is of no importance. This is not the case in RPRT-0053, where we are concerned whether the two groups are significantly different (direction is not specified). This is because, in general, the decision to stratify is based on non-directional differences, and RPRT-0053 is a generic procedure.

The DQO process is usually discussed in the context of situations where one can control the amount and overall quality of the data used to reach a conclusion (as discussed in Volume 1 of MARLAP for example). This is not the situation in a historic dose reconstruction, where we have little or no control over the quantity and quality of data, so it is unclear how one could use DQO to balance Type 1 and 2 errors. This issue, which is related to the power of the test, is discussed in more detail in the response to RPRT-0053-8.

(See Comments 20 and 26)

RPRT-0053-6

For many years, given the small number of CTW data points, the tests cannot reliably detect differences smaller than a factor of 4 to 10 in the CTW/non-CTW ratio of geometric means. Larger differences have a 95% or better chance of detection. Smaller differences would be in the "gray region" for the test, sometimes detected, sometimes not. Overall, SC&A concludes that the NIOSH method of concluding that there are no significant differences would often lead to very claimant-unfavorable results.

The retrospective data used to develop coworker models "are what they are" and we have no opportunity to change them. Failure to reject the null with retrospective datasets is inherently neither "claimant favorable" nor "claimant favorable" and is not indicative of "bad" data or inappropriate statistical methods. The small CTW datasets mentioned in this comment argue for the use of an unstratified coworker model, perhaps used in conjunction with the 95th percentile intake rates if there is evidence that a particular construction trade worker had potential for exposure on a par with the higher exposed

workgroups. The "gray region" is used in the prospective design of data collection and, once the data are collected, it is not relevant.

(See Comment 21).

RPRT-0053-7

The statistical tests for comparing two strata require that the samples in each group be independent. If a worker in one group is exposed to radionuclides with long retention in the body, then changes jobs and becomes part of the other group in the same year, the OPOS values are correlated for this worker. This correlation not only violates the assumptions of the tests, but also creates a bias toward a decision of "No Difference" between the two groups.

First and foremost, we consider the technical benefits realized by using the OPOS statistic to far outweigh relatively rare problems like the one mentioned in this comment. Second, to stratify coworker models one has to be able to assign individuals to specific and meaningful job titles (i.e., develop a *job exposure matrix*). The difficulty in determining an individual's job title, as postulated by SC&A in this comment, is a general problem associated with assembling a job exposure matrix and really has little to do with the use of the OPOS statistic. In fact, the problem raised by SC&A in this comment is an argument for not stratifying a dataset.

We assume that the "violation of assumptions for the tests" mentioned in the comment is the assumption of data independence. The main reason OPOS statistic was adopted was to achieve data independence, which can be grossly violated in the dataset of individual bioassay results.

(See Comment 24)

RPRT-0053-8

Although one example where a significant difference is found is presented in the report, NIOSH has not provided any measure of the power of the hypothesis test procedure to detect differences within the worker population. This deficiency should be corrected before the test is adopted as an appropriate procedure for coworker models. Conducting the tests at a 90% level of confidence would be claimant favorable.

The *a priori* power of a statistical test is usually considered during the design phase of the data collection procedure (e.g., the experiment or survey) so that the information collected is adequate to answer the questions being asked. In coworker modeling, we are presented with a predetermined dataset and cannot collect more, so it is not possible to perform an *a priori* power calculation. Even though an *a priori* power analysis could not be performed, we did make efforts to select the most powerful tests available. Based on our research the Peto-Prentice is the most powerful test available that can be used for comparing two groups with left-censored lognormal data, and while the power of this test will vary depending on the actual data used, there is no better statistical test that can be used for this purpose.

To perform an *a priori* power analysis, an acceptable level of power 1 - β has to be defined. To define β we must first define the *size of the effect*[4] that we want to detect, i.e., the size of the effect that is of *practical significance*. If we could define practical significance (we tried and were unsuccessful), we would perform an *equivalence test* [Streiner 2003], which tells us if the difference in the two groups is of practical significance, rather than a *null-hypothesis test*, which tells us if the difference in the two groups is of statistical significance.

In this comment SC&A may be referring to a *post-hoc* power analysis, which is the determination of power after the data are collected and the test performed. A *post-hoc* power analysis is an attempt to extract something useful from a null hypothesis test where we fail to reject the null hypothesis. Unfortunately, this analysis provides no additional information beyond that given in the confidence intervals of the estimated parameters and its use is generally discouraged (see [Ellis 2010, pg 58] and [Hoenig 2001]).

If conducting tests at an $\alpha = 0.1$ significance level (90% confidence level) would be "claimant favorable" as claimed in this comment, one might conclude that conducting the tests at a 50% confidence level would be even more "claimant favorable." Where does it end? The answer to that question is that the significance level chosen for a null hypothesis test is ultimately a judgment based primarily on the conventions established in a particular scientific field. More specifically, a significance level of $\alpha = 0.05$ (95% confidence level) appears to be the standard significance level used in the most areas of science[5].

[4] The magnitude of the difference between the two groups.
[5] http://www.jerrydallal.com/LHSP/p05.htm

SC&A has offered no justification for using 90% confidence level versus the standard 95% confidence level other than that it is more "claimant favorable." We do not feel that this is a reasonable justification, especially considering the ambiguity of "claimant favorable" in this context. In fact, by advocating a 90% level of confidence, using a one-sided null hypothesis test, and mis-specifying the null hypotheses in these tests, we feel that SC&A has created a situation which renders any conclusions that are drawn to be equivocal. For example, is a significant result of a statistical test oriented towards favoring significance really significant? For these reasons we feel that a 95% confidence level for the tests in RPRT-0053 is the most appropriate confidence level to use.

(See Comments 20 and 25)

Section 3: Detailed Comments

Comment 1

Referring to page 6 of the SC&A report:

RPRT-0053 reviews several statistical methods that are available for analyzing the coworker datasets. A range of methods is included for analysis of datasets with a varying proportion of nondetects, ranging from none to essentially all or most of the available data from monitored workers.

As stated on page 6 of RPRT-0053:

> "The purpose of this report is to detail statistical tests that can be used to decide if two strata from a given group of monitored workers are significantly different. Significantly different strata could warrant coworker models based on the strata rather than the entire population of monitored workers if the difference is of practical significance."

Thus, RPRT-0053 was written to provide generic guidance on how to decide whether or not two groups (i.e., strata) in a population of monitored workers have intakes or doses that are significantly different, where *significant* means statistically significant and not practically significant. RPRT-0053 was designed to test for non-directional differences. So, for example, a stratified coworker model would be considered when the dose to

Group A is significantly different that the dose to Group B, regardless of which is larger. This led us to use "two-sided" hypotheses tests in RPRT-0053. SC&A appears to have a different emphasis in using directional or "one-sided" tests. For example, the report advocates testing whether the dose to Group A is larger than the dose to Group B, without regard for any other relationships.

We feel that the decision to stratify a coworker model should be based on significant differences, not just the specific differences that are of interest to a given concern. This difference in philosophy is noted here because we feel that it is the basis for a number of the comments offered by SC&A. The ultimate resolution of this issue may fall under the category of a policy decision on when, why, and how coworker models should be stratified.

One possible response to a significant difference in strata would be to develop a fully stratified coworker model that incorporates all major job titles, facilities, and related information. Such a model requires the development of a job exposure matrix (JEM), which can often be difficult or impossible to accomplish with the retrospective data available in the project. Another possible response, one that SC&A hints at in their comments, would be to develop two separate coworker models for the two strata. It is unclear to us how such a model could be implemented in practice.

Comment 2

Referring to page 7 of the SC&A report:

The use of average values does not account for variability of the samples within the time period and the procedure will result in lower values of the GSD used in the coworker model. The OPOS approach represents a significant departure from the previous coworker model methodologies. This change may require re-evaluation of all previous cases with determinations that were based on coworker model estimates.

The issue of a coworker model based on OPOS statistics having a lower GSD is addressed in Comment 10. We agree that the OPOS approach represents a significant change in coworker model methodologies as does the use of stratification -- the two are intimately related. Coworker datasets that warrant stratification, and as a result have their coworker models updated to include stratification, will most likely require the re-evaluation of claims based on the unstratified model.

Comment 3

Referring to page 7 of the SC&A report:

The recommended hypothesis tests apply to only two groups, although several methods for multiple comparisons are discussed. Stratified models generally contain more than two strata with the objective of developing estimates for each strata.

In RPRT-0053, multiple comparisons are considered for two strata compared for multiple years, not multiple strata.

There is usually no a priori requirement that the strata be significantly different, although the resulting estimates may have sufficient precision to determine significant differences if the sample sizes are sufficiently large.

Stratifying a dataset when the strata don't really exist (i.e., there are no significant differences) will produce a coworker model with unnecessarily large uncertainties in the estimated parameters. So, to get the best coworker model we should be concerned with whether or not the proposed strata have a reason to be significantly different.

SC&A has shown in prior work that more than two strata are necessary in at least some cases, so as to ensure that coworker dose estimates are claimant favorable. In one example, we found that SRS construction workers need to be subdivided by job type and area of work (SC&A 2010a, SC&A 2010b).

One needs to be careful in the analysis of data to avoid what has been called "data dredging." One definition[6] of "data dredging" involves the practice of using the same set of data to both

- form a hypothesis to test, and
- subsequently test the hypothesis.

In other words, if we randomly sift through data looking for significant differences and then use the same data to test for those differences, we are bound to find significance more often than we should. Data dredging is strongly frowned on because the probability of falsely identifying significant differences is not readily quantified or controlled. Thus, any conclusions reached after data dredging are of dubious statistical or practical value.

[6] http://en.wikipedia.org/wiki/Data_dredging

In the context of the current discussion, the process of sorting through the construction coworker data to identify strata that look significantly different and then performing hypothesis tests on these strata is a form of data dredging. The preferable approach is to identify all strata to be tested prior to looking at the data and then test those strata for significant differences.

Comment 4

Referring to page 16 of the SC&A report:

A determination of the goodness-of-fit of the lognormal distribution is based on regression R^2 although, due to the dependencies that exist in the regression estimates derived from ranked data, the R^2 does not have the usual interpretation. RPRT-0053 states on page 8:

> *Operational bioassay programs can generate multiple results for an individual in a given period (e.g., a year), which creates a related problem if an individual is involved in an incident and has more (…) bioassay results than other workers. If these are not accounted for, the problems of correlated data and unequal number of samples per person can violate the assumptions on which the linear regression used to model the data and the statistical tests used to compare strata in the population are based (…).*

Although NIOSH has an apparent concern that the assumptions of linear regression apply, the data values in the ROS scatter plot are not independent observations.

The regression we refer to in this quote from page 8 of RPRT-0053 is not ROS but rather is the linear regression performed on the 50th and 84th percentile OPOS statistics in order to calculate the 50th and 84th percentile intake rates. Whether or not the assumptions of linear regression apply is a concern for this regression.

ROS is one of several methods that can be used to calculate the GM and GSD of the data. For example, one can also use maximum likelihood (ML) to do the calculation instead of ROS. If the data from which the GM and GSD are calculated are not *iid*[7] then the GM and GSD do not have familiar properties like having a 95% confidence interval of

[7] Independent and identically distributed.

$(GM*GSD^{-1.96}, GM*GSD^{1.96})$.

The fact that the order statistics of the data are not *iid* is not relevant to the problem at hand, which becomes obvious if one uses ML instead of ROS to calculate the GM and GSD (i.e., there is no need to calculate order statistics with ML).

Comment 5

Referring to page 17 of the SC&A report:

Finding No. 1: Due to the dependencies that exist in the ranked data, the R^2 for ROS does not have the usual interpretation. The recommendations in RPRT-0053 for using ROS do not address this concern.

R^2 is not mentioned anywhere in the text of RPRT-0053 as a goodness of fit criteria. However, the R^2 statistic appears in some ROS plots. We think this was done at the request of someone in ORAUT, perhaps to be consistent with previous practice (i.e., PROC-95). R^2 was not used by the statisticians to evaluate fits in ROS plots so we don't think this topic warrants a "Finding" and does not need to be addressed in RPRT-0053.

As mentioned in our response to finding 1, this issue was raised, addressed, and closed under SC&A's previous review of PROC-0095 and OTIB-019.

Comment 6

Referring to page 17 of the SC&A report:

Maximum likelihood techniques are used to estimate the parameters of the mixed model. If a dataset contains urine results for which most of the workers do not have analyte in their urine but a small fraction of the workers do, then the methods presented in RPRT-0053 are an improvement over the PROC-0095. However, NIOSH does not offer any consideration relating to the pattern or time distribution of the positive results. It is necessary to know if the positive results occur every year, and if those results are related to a particular procedure. For example, the positive results could be present x times per year, during defined periods of time, or during a specific campaign.

The most sophisticated coworker models being contemplated are those that stratify the workforce based on a detailed job exposure matrix that incorporates job title, work location, etc. It is unclear to us how one would go about incorporating patterns or time distributions other than the annual patterns currently used. For example, given monthly TLD external doses for a workforce, how would one implement this SC&A recommendation?

Comment 7

Referring to page 17 of the SC&A report:

In the application of the procedure recommended in RPRT-0053, the issue of completeness of the available coworker data has not been addressed. Some workers were not monitored; otherwise there would be no need for a coworker model. The underlying assumption appears to be that the workers with the most exposure potential were monitored, but we have seen in a number of cases that this was not necessarily true. If the unmonitored workers are from a different population, the applicability of a coworker model derived from monitored workers would be in question.

RPRT-0053 was written to provide generic guidance on ways to decide whether or not two groups in a population of monitored workers have intakes or doses that are significantly different. It is not a goal of RPRT-0053 to address the issue of *data completeness*. Nevertheless, this comment raises a critically important point that we would like to discuss further.

The worker monitoring data were collected in the past to demonstrate compliance with the applicable occupational dose limits that were in place at the time. We are provided these *retrospective* data and are asked to perform statistical analyses on the data to answer questions being asked in the EEOICPA program today. We did not have the opportunity to select the workers and monitoring programs needed to ensure that we could develop definitive answers to these questions.

In most cases, we believe that the radiation protection staff who worked in the facilities when these retrospective data were collected made a concerted effort to monitor all individuals who they felt had a likelihood of receiving significant intakes. The monitored workers were intended to be a *census* rather than a *sample* (random or otherwise).

The goal of the coworker model is to estimate intakes for workers who were inadvertently not monitored. The implicit assumptions of the current coworker models are that unmonitored workers

- are more likely to be the workers who had lower potential for significant intakes than they are to be workers who had higher potential for significant intakes, and
- are not monitored *completely at random*.

If valid, these two assumptions ensure that any intakes assigned to unmonitored workers are conservative. In cases where the two assumptions given above may not be valid, the current practice is to assign the 95th percentile intake to minimize the chances of underestimating the dose to the worker. Whether or not the monitoring data collected from a site meets these assumptions to a degree adequate to permit a useful coworker model to be developed is a decision that must be made before the methods of RPRT-0053 are applied and is therefore not discussed in RPRT-0053.

Comment 8

Referring to page 17 of the SC&A report:

Finding No. 2: In the application of the procedures recommended in RPRT-0053, the issue of completeness of the available coworker data has not been addressed. If the unmonitored workers are not from a population that had the highest exposure potential, the applicability of a coworker model derived from monitored coworkers would be in question. The matter of the relative exposure potential of the monitored workers needs to be demonstrated rather than assumed. The methods proposed in RPRT-0053 for analyzing the coworker datasets require verification that (i) the available coworker data are representative of all groups of workers, and (ii) the manner of use of the data is claimant favorable for the specific datasets to which the method is applied. A sound statistical methodology is subject to these two important caveats.

We suspect that the statement

If the <u>unmonitored</u> workers are not from a population that had the highest exposure potential, the applicability of a coworker model derived from monitored coworkers would be in question.

was meant to be (underlining is ours)

If the monitored workers are not from a population that had the highest exposure potential, the applicability of a coworker model derived from monitored coworkers would be in question.

We will respond to this comment (assuming our version is what SC&A meant) by saying that we agree. This relates back to the issue of stating why unmonitored workers were not monitored and monitored workers were monitored (see Comment 7). In the development of coworker models we assume that either

- unmonitored individuals are members of the monitored population who were not monitored completely at random, or
- unmonitored individuals were unmonitored because they had no potential for exposure to radioactive materials.

In the first case we have the right model and in the second a conservative model. One can also theorize that these assumptions are wrong and that perhaps unmonitored workers were highly exposed and intentionally not monitored because of this. This fundamental and largely unstated difference in assumptions probably needs to be discussed and eventually resolved.

We disagree with the statement "The matter of the relative exposure potential of the monitored workers needs to be demonstrated rather than assumed." The reason why individuals were unmonitored will, in general, always be largely an assumption that cannot be "proved" to the satisfaction of everyone. The validity of this assumption is basically a function of the maturity of the radiation protection program in place at a given facility and the level of documentation available.

The concept of "claimant favorability" expressed in this comment is confusing. To illustrate the problem, assume Group A and Group B have very different dose distributions, with A higher than B, and we combine the doses from the two groups to form Group C. The dose distribution for C will overestimate the dose to B (be "claimant favorable") and underestimate the dose to A (not be "claimant favorable"). But, we typically assume that workers with higher doses are less likely to be unmonitored, so this is not considered a problem in practice. Now, if we stratify the model, the doses to both A and B will be more accurate. However, the dose to B will be lower and less "claimant favorable" and the dose to A will be higher and more "claimant favorable" than the combined dose distribution. So, in principle, no coworker model(s) can be "claimant favorable" to all strata in the model at the same time. Additional discussion on this topic is given in Comment 26.

Comment 9

Referring to Section 3, starting on page 19 of the SC&A report.

One of the main concerns expressed by SC&A about the methodology given in RPRT-0053 appears to be centered around the use of the *one person, one sample* (OPOS) statistic in coworker modeling. For uncensored data, the OPOS statistic for an individual is simply the mean of his bioassay results for a given time period. The OPOS statistic is used in order to deal with two significant issues: dependent coworker data and coworker data dominated by a small number of individuals. Below, we discuss these two problems in more detail and how the OPOS statistic is used to solve these problems.

When multiple bioassays are performed on an individual, the results can be correlated if the individual has had an intake of radioactive material. For example, if an individual has detectable levels of Pu in one urine sample, the next urine sample is also likely to contain Pu. Datasets composed of such *dependent* data usually cannot be analyzed with standard statistical methods, which require *independence* of the data. This issue may have been of marginal importance when all we were interested in was estimating parameters (like intake rates). However, once we start asking if the intake rate of one part of a cohort is different than the intake rate of another part, the issue of data independence becomes critical.

Another problem associated with coworker modeling the bioassay data is that of "data dominance", where a small number of individuals (perhaps even one) submit a significant fraction of the total number of samples collected from the cohort. The resulting coworker model is not representative of the monitored population but instead is dominated by a small number of individuals.

The solution to both of these problems is to model the intakes (or intake rates) rather than the bioassay data. Intakes are independent and if we model the sum of intakes for each individual in a given time period, each person contributes equally to the coworker model. The problem is that it is, in general, not feasible to evaluate each individual's bioassay data in terms of intake.

Given that we cannot use intakes in a coworker model, we chose to use the OPOS statistic as a surrogate for the intake. As shown below, the intake is proportional to the mean of the bioassay data (the OPOS statistic), where the constant of proportionality is

the mean of the intake retention fractions[8]. Like an intake, an OPOS statistic is independent and gives each individual in a cohort equal weight in the final model.

While perhaps not the perfect solution to the problems discussed above, we feel that the OPOS statistic is the best available solution, and is undoubtedly better than just modeling the individual bioassay results -- which is the approach that SC&A appears to recommend.

Intakes and the OPOS Statistic

To perform an internal dose coworker model in the most technically correct fashion, we would model the *intakes* of the monitored workers for each year rather than their bioassay data. In IAEA Report 37 [IAEA 2004, pg 22], the equation for the weighted least squares estimate of an intake I is given as

$$I = \frac{\sum_{i=1}^{n} w_i M_i m_i}{\sum_{i=1}^{n} w_i \left[m_i \right]^2},$$

where

n = number of bioassay measurements,

M_i = bioassay measurements,

m_i = intake retention fractions, and

w_i = regression weighting factors.

This is a weighted regression through the origin of the bioassay measurements on the intake retention fractions, and the regression weighting factors are usually taken to be equal to the inverse of the variance of the measurements. If we assume that the variance is proportional to intake retention fraction [Skrable 1994, pg 442], the weighting factor w_i is given by

[8] This mean of the intake retention fractions is the part of the intake calculation that is not feasible to determine.

$$w_i = \frac{1}{\left(k \cdot \sqrt{m_i}\right)^2} \ ,$$

where *k* is an unknown constant of proportionality. Under these conditions, the weighted least squares estimate of the intake simplifies to [Skrable 1994, pg 442]:

$$I = \frac{\frac{1}{n}\sum_{i=1}^{n} M_i}{\frac{1}{n}\sum_{i=1}^{n} m_i} = \frac{\bar{M}}{\bar{m}} \ ,$$

which is usually referred to as the "ratio of the means" estimate of intake. Thus, the intake estimate is basically the average of the bioassay results divided by a constant, $\bar{m}$, that is determined by the choice of biokinetic models and exposure scenario for that intake. In the case where there are censored bioassay results, one could substitute the OPOS statistic for $\bar{M}$ to obtain an overestimate of the intake (i.e., the OPOS statistic equals $\bar{M}$ if all data are uncensored).

Comment 10

Referring to page 20 of the SC&A report:

The use of average values does not account for variability of the samples within the time period, and the procedure will result in lower values of the GSD used in the coworker model compared with previous procedures. A GSD must be assigned for the missing dose to a worker in each year, and that GSD should reflect the variability in that worker's exposure during the year. The OPOS GSD measures the variability of average annual dose across workers, and ignores variability for an individual worker within the year.

In the presence of data dominance and dependent data (see Comment 9), the GM and GSD calculated with individual bioassay measurements do not have familiar statistical properties and are therefore not useful measures of central tendency and variance of the data. The OPOS statistic was adopted in an effort to deal with these major issues (on which SC&A did not comment). We feel that the use of the OPOS statistic better achieves the goal of accurately estimating the intake rates and ultimately the dose to workers than does the use of individual bioassay results.

Finding No. 3: The OPOS statistic methodology summarizes a worker's exposure by averaging overall urine samples collected during the specified time period. The use of average values does not account for variability of the samples within the time period and the procedure will result in lower values of the GSD used in the coworker model.

One might infer from this finding that a higher GSD calculated incorrectly is preferable to a lower GSD calculated correctly, perhaps on the basis of "claimant favorability", i.e., higher GSD = higher POC. Below is an excerpt from 42CFR82 that discusses how dose reconstructions should be performed:

> "Several commenters requested HHS define what constitutes a ''reasonable estimate'' of the radiation doses incurred by an employee. EEOICPA requires the dose reconstruction program to arrive at ''reasonable estimates'' of these doses (42 U.S.C. 7384n(d)). HHS interprets this term to mean estimates calculated using a substantial basis of fact and the application of science-based, logical assumptions to supplement or interpret the factual basis. As discussed in the interim final rule, assumptions applied by NIOSH will give the benefit of the doubt to claimants in cases of scientific or factual uncertainty or unknowns."

Thus, if we are presented with multiple, equally valid solutions to a given problem during the process of developing coworker models, we should adopt the solution that gives the benefit of doubt to the claimant. This "claimant favorable" answer is usually taken to be the one that results in the highest dose. The concept of "claimant favorability" is not applicable in the case where there is a solution that is clearly better than the other solutions. More specifically, 42CFR82 does not guide us to adopt an inferior answer simply because it might result in a higher dose than the technically superior answer. Thus, the fact that the GSD will most likely be lower with the OPOS statistics than it is with the individual bioassay results is not relevant because the use of OPOS is a technically superior approach.

Comment 11

Referring to page 20 of the SC&A report:

The OPOS methodology does not examine the temporal pattern of individual exposures for longer than one time period.

We are unsure as to exactly what this comment means. In coworker models based on the individual bioassay results, the 50th and 84th percentile excretion rates (or retention) are estimated for each time period -- a year for example. With the use of the OPOS statistics the 50th and 84th percentile excretion rates are also estimated for a year. Any subsequent manipulation of the 50th and 84th percentiles is independent of how they were calculated. This is a property of coworker models in general and is the same whether one derives the percentiles with OPOS statistics or the individual bioassay results. So, if the OPOS methodology does not examine the "temporal pattern" of individual exposures for longer than a year then neither does the methodology that uses individual bioassay results.

Comment 12

Referring to page 20 of the SC&A report:

When comparing two populations using a statistical test for differences, it is important that the data are collected following the same protocol for both groups of workers. In the specific case of CTW versus non-CTW comparisons in RPRT-0056, NIOSH has said that sampling was incident-related for CTWs and routine for non-CTWs, so the OPOS method does not appear appropriate for comparing the two distributions

We believe that there may be some confusion here concerning the use of the statistical term "sampling protocol." One definition of the term is[9]

> "The sampling protocol is the procedure used to select units from the study population to be measured. The goal of the sampling protocol is to select units that are representative of the study population with respect to the attribute(s) of interest. The sampling protocol deals with how and when the units are selected and how many units are selected."

Thus, the sampling protocol tells how one might select individuals (i.e., a sample) from a population of people with the intent of inferring population parameters from the sample.

In this comment the statistical term "sampling protocol" is incorrectly used as being synonymous with the term "internal dosimetry monitoring program." There is no statistical requirement that all workers be on the same monitoring program in order to use the data to develop a coworker model as long as the monitoring programs adequately

[9] http://sas.uwaterloo.ca/~rwoldfor/papers/sci-method/paperrev/node40.html

characterize all significant intakes. Further, most sites had *graded* monitoring programs where the frequency and types of bioassay performed were based on the likelihood of the workers having a significant intake of radioactive material[10]. Even today this is standard radiation protection practice, so we would expect the bioassay (i.e., sampling) protocols to be different for different groups of workers.

Given all of this, we feel that it is appropriate (for example) to compare intakes calculated from "special" and "task-related" bioassay performed in one group to intakes calculated from "special", "task-related", and "confirmatory bioassay[11]" in another group. Therefore we disagree with Finding 4 given on page 22:

Finding No. 4: The OPOS method must strictly be applied to comparisons where the sampling protocol was the same. Specifically, when there is evidence that the sampling protocol for one group of workers was different than the protocol used for the other group, the tests do not provide a valid comparison. For example, if the monitoring of one group of workers is incident-driven and the other is not, then the OPOS approach is not appropriate for comparing the two distributions.

and the following recommendations offered by SC&A, also on page 22 of their report:

Given the problems introduced by the use of OPOS when there are different sampling protocols for each group, SC&A recommends that:

(1) OPOS values should not be combined into a single lognormal distribution when the sampling protocols for subsets of workers in the group differ

(2) Distributions of OPOS values can be compared only when the sampling protocols are the same for both groups.

Comment 13

Referring to page 21 of the SC&A report:

The answers to these questions are important because the use of OPOS values introduces complications in the subsequent coworker model analyses that rely on these values. OPOS values are not measurements, but are statistics derived from a set of

[10] Graded monitoring is also common in external dosimetry programs.
[11] The terms *special*, *task-related*, and *confirmatory* are defined in paragraphs 5.3 and 5.4 in ICRP 78.

measurements. The OPOS values are averages of a varying number of samples, with a different number for each worker. Since it is an average, each OPOS value has an uncertainty associated with the calculated value.

It is important to realize that all measurements have associated measurement uncertainties and that these uncertainties are not trivial to assess. These facts apply equally to individual bioassay results and to OPOS statistics derived from those individual results. Thus, we find it somewhat inconsistent to take issue with the OPOS statistic for reasons that exist in all measurement data.

Comment 14

Referring to page 21 of the SC&A report:

A difference in the number of samples available for the workers in each group implies a difference in the uncertainty for the OPOS values for each group. In general, more samples are available for the onsite workers who are part of an ongoing monitoring program. Due to the larger number of samples, the OPOS values for the onsite workers may be measured with greater precision than is available for other groups of workers.

One reason the OPOS statistic was adopted was to give all workers equal weight in the final coworker model - hence the "one person, one sample" moniker. This prevents workers with a larger number of samples per year (onsite workers perhaps) from dominating the coworker model. We find it interesting that SC&A did not offer any comments in this issue or whether or not the use of the OPOS statistic provides any advantages over the use of individual bioassay results.

Since there is uncertainty in the OPOS statistics, and this uncertainty varies from worker to worker and from one group of workers to another, all subsequent analyses based on OPOS values are conducted using heteroscedastic data.

There is measurement uncertainty in all personal dosimetry results (both internal and external -- see Comment 13) and all personal dosimetry results are heteroscedastic to some extent. Thus, the issue raised here is not specific to the use of the OPOS statistic.

Finding 1 in Section 2.1 indicates that the ROS method conducted on individual samples ignores the heteroscedastic nature of the order statistics derived from the sample values. If the order statistics are derived from OPOS values, this introduces a second problem

unique to the use of OPOS values in ROS: values that are being ranked may not come from the same distribution unless the monitoring protocol is the same for all members of the group.

The concern about ROS and order statistics were addressed in Comments 4 and 5, and the concern about monitoring protocols in Comment 12.

Comment 15

Referring to page 21 of the SC&A report:

The assumptions underlying the tests are violated if the nonparametric tests are applied using data with different variances in each group. The WRS test and the generalized WRS tests, including the Peto-Prentice test, are based on an assumption that the only difference between the two groups is a difference in the location of the distributions (Conover 1980, p. 217). This means that the shapes and variances of the two distributions should be approximately the same.

More than the usual amount of care needs to be exercised with regard to the discussion of the Peto-Prentice test, the WRS test, the generalized WRS test, and Peto-Peto test. As stated by [Leton 2005]:

> "We have seen in the literature and in the statistical software that sometimes the same tests receive different names and the same name is used for different tests."

Thus, there may be different tests that go by the name "Peto-Prentice." Being aware of this problem, we put Attachment B in RPRT-0053, which explains that the Peto-Prentice test used in RPRT-0053 compares the empirical cumulative distributions (ecd) of the two groups, looking for any difference (not just a shift in location). Further, the R function that implements what we are calling the Peto-Prentice test is called `cendiff` [Helsel 2012] and it has the following description:

> "Tests if there is a difference between two or more empirical cumulative distribution functions (ECDF) using the G-rho family of tests, or for a single curve against a known alternative."

Thus, contrary to what SC&A stated in this comment, the shapes and variances of the two distributions do not have to be the same for the Peto-Prentice test used in RRRT-0053.

Comment 16

Referring to pages 22-26 of the SC&A report:

In Section 3.2, SC&A uses simulations to show that, for a given dataset, a coworker model based on OPOS statistics can have a lower variance (GSD) than a coworker model based on the individual bioassay results and can also give a different estimate of the GM. For example:

The simulation analysis indicates that the OPOS approach results in underestimation of the range of variability across workers reflected in estimates of the GSD and 95th percentile, which are biased low relative to the original samples.

We agree with these conclusions, which are pretty much what one would have expected to conclude before doing the simulation. However, one might mistakenly infer from this discussion that the model derived from OPOS statistics is somehow "wrong" because it produces estimates of model parameters that are different than the estimates obtained with the individual bioassay results. We feel that we have provided ample technical justification for using the OPOS statistic rather than the individual bioassay results and that any such inference is incorrect.

Comment 17

Referring to page 27 of the SC&A report:

A hypothesis testing procedure is proposed for determining when there are "significantly different strata." The hypothesis test procedure compares the two strata using an MCPT and the nonparametric Peto-Prentice test. In the analysis of previously collected data, it is necessary to determine if the sample size was sufficient. NIOSH has made no effort to determine sample sizes that allow for sufficient power to detect differences.

SC&A is referring to a post-hoc power analysis. See Comment 20 for a discussion of why we feel that this is an inappropriate procedure.

More than two strata would be required to characterize properly the varied worker populations at many sites, including SRS. Multiple comparisons when there are more than two strata may be possible, but could be complex and suffer from limits imposed by

small sample sizes. The analysis may spiral into large numbers of comparisons with inconclusive results.

This comment brings us back to the fact that policy decisions on when and how to stratify data in coworker models is a separate issue. . We envision that stratification will be considered if there is a practically significant difference in any two preselected strata and there is sufficient information to construct a job-exposure matrix. There will be no need for multiple comparisons between a large number of strata as discussed in this comment. Note that RPRT-0053 was designed with this idea in mind, which is one reason why it only considers tests between two strata.

Comment 18

Referring to page 27 of the SC&A report:

An incorrect variation of the MCPT was described in ORAUT-RPRT-0049, Discussion of Tritium Coworker Models at the Savannah River Site – Part 1 (ORAUT 2010a). In that report, NIOSH compared the distribution of one group of workers to the entire population of workers to test for a significant difference, violating the independence of the two samples.

Referring to page 36 of the SC&A report:

One improvement that should be noted; the MCPT approach proposed in RPRT-0053 is based on samples from two mutually exclusive populations of workers. In RPRT-0049, the MCPT procedure compared coworker samples for one group of workers with samples drawn from the set of all workers. The current report properly compares the parameters of the lognormal distributions estimated separately for each group of workers.

RPRT-0049 and RPRT-0050 were designed to answer different questions. In RPRT-0049 we wanted to compare a coworker model based on construction trade workers (CTW) with the coworker model based on all workers (AW). The CTW is a subset of AW, and we were interested in this comparison because, in practice, if the CTW model is not applied to CTW then the AW model (which is the usual coworker model) would be applied.

In RPRT-0050, the CTW model is compared to the model based on all other workers (AOW). Note that if the CTW model is not used we would apply the AW model, not the

AOW model. The CTW/AOW comparison was examined because CTW and AOW are independent, which allowed us to use statistical tests (such as the Kuiper test) that require independence of the two strata. In RPRT-0053, the Peto-Prentice test requires independent samples so the CTW/AOW stratification was used.

Contrary to the statement offered by SC&A, the MCPT as used in RPRT-49 is correct because the simulation properly accounts for the dependence between the CTW and AW datasets. We mentioned this in RPRT-0050 on page 14:

> "The Monte Carlo permutation test used in Part 1 of this report (ORAUT 2010) is valid under the stated conditions, one of which is that independence of the stratum and complete sample is not required."

A simulation is given below to illustrate that the MCPT gives correct results when the two samples are dependent.

MCPT Simulation for Dependent Samples

Consider the simulation below, which is coded in R. In this simulation the mean of 500 numbers (Sample A) drawn from an iid normally distributed population ~N(100,10) is compared to the mean of a subset (Sample B) consisting of 50 numbers randomly drawn from Sample A. Thus, Sample A and Sample B are not independent because Sample A contains Sample B. The creation of Samples A and B are given in lines 11-14 of the R code. In lines 22-23, a t-test is used to compare the mean of A to the mean of B and a decision is made at the 95% confidence level as to whether or not A and B were drawn from the same the parent population -- which they were. A decision that A and B are different is a false positive and is recorded. Repeating this experiment 10,000 times we would like the false positive rate to be ~0.05, but it was in fact 0.03. The "coverage" of the t-test is not what is advertised for the test because of the dependence between the two samples. In other words, the t-test requires that the two samples be independent of each other and we violated that assumption.

On lines 17-19 a MCPT is performed in tandem with the t-test. After repeating the simulation 10,000 times at the 95% confidence level, the false positive rate of this test is indeed 0.05, i.e., it has proper coverage even with dependent samples. This is because the dependence between samples is accounted for in the simulation.

```
set.seed(123)
m <- 10000
n <- 5000
fp.b <- numeric(m)
fp.t <- numeric(m)
diff.b <- numeric(n)

for (i in 1:m) {
  x.all <- rnorm(500,100,10)
  mu <- mean(x.all)
  x.samp <- sample(x.all,50,replace=FALSE)
  diff <- mean(x.samp) - mu

  # MCPT
  for (k in 1:n) { diff.b[k] <- mean(sample(x.all,50)) - mu }
  lims <- quantile(diff.b,probs=c(0.025,0.975),type=9)
  fp.b[i] <- (diff > lims[2]) | (diff < lims[1])

  # t test
  lims <- t.test(x.samp,x.all)$conf.int
  fp.t[i] <- (0 > lims[2]) | (0 < lims[1])
}

sum(fp.b) / m
sum(fp.t) / m
```

The output of the simulation is:

```
> sum(fp.b) / m
[1] 0.0497
> sum(fp.t) / m
[1] 0.0304
```

In summary, the MCPT has the proper coverage under the null hypothesis (5% false positive rate) whereas the t-test does not (3% false positive rate) because the t-test assumes that the two samples are independent. In an analogous fashion, the MCPT used in RPRT-0049 properly accounts for the dependence between the samples and has proper coverage under the null hypothesis.

Comment 19

Referring to pages 22-26 of the SC&A report.

Every comment in this section that refers to *a priori* power in the context of MARISSIM, ProUCL, and DQO implicitly assumes that one can ask questions and then design a sampling program that is capable of answering these questions. This is not possible for coworker studies so we feel that the information in these documents is not relevant. See Comment 20 for a more detailed discussion of our views on this topic.

Referring to page 28 of the SC&A report:

The statistical methods developed for MARSSIM and Superfund are particularly useful in this discussion.

Many of the conventions established in MARSSIM are specific to its intended application and all of the associated regulations:

> "The MARSSIM's objective is to describe a consistent approach for planning, performing, and assessing building surface and surface soil final status surveys to meet established dose or risk based release criteria, while at the same time encouraging an effective use of resources."

Because of this we feel that MARSSIM is not generally applicable to the issues associated with coworker modeling.

Comment 20

Referring to page 27 of the SC&A report:

NIOSH proposes that strong evidence ($\alpha = 0.05$ or a 95% level of confidence) is necessary before any differences between groups of workers should be considered in the coworker model. In hypothesis testing, the demand for a high degree of confidence in a decision (α or Type 1 error) is usually balanced by a requirement for adequate power (β) to ensure the test has a capability of detecting differences thought to be of importance. Although there is a general discussion of power in the literature review included in Attachment B of RPRT-0053, NIOSH has not provided any measure of the power of the MCPT to detect differences given the sample sizes and variability encountered in the available datasets. One example where a significant difference was found is presented in the report. This deficiency should be corrected before the MCPT is adopted as an appropriate testing procedure.

Referring to page 29 of the SC&A report:

Finding No. 5: The methods in RPRT-0053 require a high level of confidence before deciding that the two worker groups are significantly different. The requirement for a high level of confidence in this decision is not claimant favorable when using a null hypothesis of "No Difference." The power of the tests to detect differences given the limited quantity of available data has not been established. The Data Quality Objectives (DQO) process should be used to balance Type 1 and Type 2 decision errors.

Null hypothesis testing (NHT) is used in RPRT-0053 to decide if the doses to two groups of workers are significantly different. To perform NHT one must define

- a null hypothesis, which is usually that there is no effect (e.g., the dose to the two groups are equal), and
- an alternate hypothesis, which is that there is an effect (e.g., the doses to the two groups are not equal)

The goal of a researcher is to design the experiment and specify the hypotheses so that the uninteresting hypothesis (the null) is rejected and the interesting hypothesis (the alternate) is accepted. Two important points be to made here are that

- the failure to reject the null hypothesis is not the same as proving that the null hypothesis is true[12], and
- in designed experiments, the failure to reject the null hypothesis is considered to be a "failure" of sorts because scientific journals tend to not publish uninteresting results, i.e., studies where the null is not rejected.

The statistical power of a hypothesis test is a measure of its ability to reject the null hypothesis if the null hypothesis is in fact false. Throughout their review of RPRT-0053 SC&A give a considerable amount of attention to the issue of statistical power and how the procedures in RPRT-0053 may be deficient with regard to power. We would like to address this important issue.

The *a priori* power of a statistical test is considered during the design phase of the data collection procedure (e.g., the experiment or survey) so that the information collected is adequate to answer the questions being asked. The worker monitoring data used for developing coworker models were collected in the past to demonstrate compliance with the applicable occupational dose limits that were in place at the time. We are provided

[12] Technically, the null hypothesis is never accepted in a NHT -- absence of proof is not proof of absence.

with these *retrospective* data and are asked to perform statistical analyses on the data to answer questions being asked in the EEOICPA program today. We did not have the opportunity to select the workers and monitoring programs needed to ensure that we could develop definitive answers to these contemporary questions. In summary, in coworker modeling we are presented with a predetermined dataset and cannot collect more data, so it is not useful to perform an *a priori* power calculation.

To perform an *a priori* power analysis, an acceptable level of power 1 - β has to be defined. To define β we must first define the *size of the effect*[13] that we want to detect, i.e., the size of the effect that is of *practical significance*. If we could define practical significance, we would perform an *equivalence test*, which tells us if the difference in the two groups is of practical significance, rather than a *null-hypothesis test*, which tells us if the difference in the two groups is of statistical significance (see Comment 29). The problem is that we don't know how to define practical significance in a way that would be acceptable to all stakeholders in the EEOICPA program.

Thus, we feel that the entire discussion of *a priori* power in the context of coworker models is basically irrelevant. The failure to reject the null is not a "failure" as in the case of a designed experiment. The data are what they are, and they will either lead you to reject the null hypothesis or not reject the null hypothesis. Rejection of the null hypothesis results in consideration of a stratified coworker model whereas failure to reject the null hypothesis results in the use of the standard coworker model. Neither course of action is inherently claimant favorable nor claimant unfavorable. Thus, we feel that the recommendations given by SC&A designed to increase the chances of rejecting the null hypothesis for a given set of retrospective data (e.g., the use of the 90% confidence level in a one-sided test) to be inappropriate.

In their comment SC&A may be referring to a *post-hoc* power analysis, which is the determination of power after the data are collected and the test performed. A *post-hoc* power analysis is an attempt to extract something useful from a null hypothesis test where we fail to reject the null hypothesis. Unfortunately, this analysis provides no additional information beyond that given in the confidence intervals of the estimated parameters and its use is generally discouraged. On the subject of *post-hoc* power, Ellis states [Ellis 2010]:

[13] The magnitude of the difference between the two groups.

> "The post-hoc analysis of nonsignificant results is sometimes painted as controversial, but it really isn't. It is just wrong."

Hoenig states [Hoenig 2001]:

> "It is well known that statistical power calculations can be valuable in planning an experiment. There is also a large literature advocating that power calculations be made whenever one performs a statistical test of a hypothesis and one obtains a statistically nonsignificant result. Advocates of such post-experiment power calculations claim the calculations should be used to aid in the interpretation of the experimental results. This approach, which appears in various forms, is fundamentally flawed."

There is considerable discussion in the SC&A comments related to EPA guidance on the determination of *a priori* power, e.g., from page 28 of the SC&A report:

In addition to the general advice of 10–15 samples, the Draft ProUCL Technical Guide contains further advice to use the DQO process. Appendix B, Section B1.3.2, of the same document (EPA 2010) contains detailed instructions for determining the required sample size based on data variability and DQO parameters. Instructions for 1-sided and 2-sided tests are provided. NIOSH has made no effort to determine sample sizes that allow for sufficient power to detect differences given the available sample sizes and variability.

The hypothesis testing framework recommended in the multi-agency document MARSSIM (EPA 2000) provides a basis for determining the necessary sample size for controlling decision errors of both types.

The DQO process is relevant only in the context of situations where one can control the amount and overall quality of the data used to reach a conclusion. As discussed above, this is not the situation with retrospective datasets, where we have no control over the quantity and quality of data. The EPA documents are a useful resource to us with regard to statistical methods but the guidance relative to *a prior power* is not relevant to the problem at hand.

Finally, given the same data and hypotheses, different statistical tests can have different *a priori* power. We made considerable effort to select the test that had the highest power for the questions we were asking in RPRT-0053 and selected the Peto-Prentice test. A detailed discussion of power for this test is given below.

Power of the Peto-Prentice Test

There is a detailed discussion in the Attachment B of the ORAUT-RPRT-0053 report, based on an extensive literature review, of the rationale for choosing the Peto-Prentice test, since it is known to be the most powerful test that can be used when comparing two groups with censored lognormal data. While the better known and easier to compute Gehan test can also be used for the same purpose, it was decided to use the Peto-Prentice test, due to its increased power when comparing two groups with left-censored lognormal data.

While there has been extensive research to compute the power of the Gehan and the Peto-Prentice tests in various situations, it is very difficult to summarize all the published results in the literature, since the power varies a lot for different scenarios, depending on the assumed distributions for the data, the sample sizes in the two groups, the censoring percentages in each group, and the relative positions of the cumulative distribution functions (CDF) in the two groups. Most of the power studies were performed in the context of the survival analysis field (see [Leton 2002], [Leton 2005], [Leton 2008]), and the results are usually summarized for four different scenarios: proportional hazards (PH), early hazard differences (EHD), late hazard differences (LHD), and middle hazard differences (MHD). The most common scenarios for the left-censored data correspond to CDF's that do not cross, or CDF's that intersect either at the lowest or at the highest values (PH and EHD scenarios). While the papers usually report various tables with the power for several tests (including Gehan and Peto-Prentice), they only apply to the specified distributions, samples sizes, and censoring percentages used in that particular simulation study. As an example of the PH and EHD scenarios, the power for the Peto-Prentice test can vary widely, from 61% to 97% [Leton 2002], from 43% to 80% [Leton 2005], and from 61% to 99% [Leton 2008]), depending on the distributions assumed for the data, the sample sizes in each group, and the censoring distributions and censoring percentages for each group.

Here are some concluding remarks from [Leton 2008], regarding the power of tests used in comparing two groups with censored data (including the Gehan, and Peto-Prentice tests):

- There is a similar power between score and weighted tests, although sometimes the power for the score tests is better than for the weighted tests.
- The differences in power are greater between score and weighted tests if the sample sizes are different.

- The greater sample sizes give greater power for the score and weighted tests.
- There is a great variability in the power for each test in different scenarios. The worse power is observed in the LHD scenario.
- In unbalanced groups, it is observed that power depends on the scenario, hazard of the groups and sample sizes.

In addition to the power studies performed in the context of survival analysis, there are also studies that provide estimates for the power of the Gehan and Peto-Prentice tests when comparing two groups with censored data, but with the more restrictive assumption that the distributions in the two groups differ from one another only by a location shift [Magel 1991], [Wamil 1997]. Magel [1991] provides tables with the power of the Gehan test when data follows either normal or exponential distributions in the two groups, and when the two groups differ only by a location shift. Similarly, Wamil 1[997] provides tables with the power of the Peto-Prentice test when data follows either uniform, Gamma, exponential, or normal distributions in the two groups, and when the two groups differ only by a location shift.

A simulation study, similar to those performed in Magel [1991] and Wamil [1997] was conducted, to develop estimates for the power of the Gehan and Peto-Prentice tests, for additional settings involving two shifted normal distributions, as well as for the case of two lognormal distributions, with the same GSD. The results, presented as power curves in Appendix A, show how the power of the two tests varies as a function of the shift in the means, for normal distributions, and as a function of geometric means for lognormal distributions. The normal samples from the two groups were generated as described in Magel [1991] and Wamil [1997], and were censored using a uniform distribution. The shifts in the means or geometric means were selected to be equally spaced, until the power reaches the 95% level. For each setting presented in Appendix A, 50000 samples were generated, and the power was computed based on the number of times the null hypothesis was rejected. Figures 1 through 4 show the power results for the four settings with shifted normal distributions presented in Wamil [1997, pg. 26-29], for the Peto-Prentice test, as well as the Gehan test; in all these four cases, when the two groups have the same samples sizes, and the same censoring percentages, the two tests have basically the same power. Figures 5 and 6 show the power results of two additional settings for the case of two shifted normal distributions, when either the samples sizes or the censoring percentages are different between the two groups; in these two cases, the Peto-Prentice test is slightly more powerful than the Gehan test, in detecting the differences in the means of the two shifted normal distributions. Figures 7 through 10 show the power results for four different settings involving two lognormal distributions with the same

GSD; in all these four cases, where the two groups have the same samples sizes, and the same censoring percentages, the two tests have basically the same power. Figures 11 and 12 show the power results of two settings involving two lognormal distributions with the same GSD, when either the samples sizes or the censoring percentages are different between the two groups; in both cases, the Peto-Prentice test is slightly more powerful than the Gehan test, in detecting the differences in the geometric means of the two lognormal distributions.

In all the 12 settings presented in Appendix A, the power increases when the difference between the means of the two distributions increases, and these differences in means are a function of the samples sizes, and the censoring percentages in the two groups. For this reason, it is very hard to generalize these results to all the possible situations encountered in the real-life examples, when the distributions, the samples sizes, and the censoring percentages in the two groups vary widely.

The fact that the Peto-Prentice test is the most powerful available test when comparing two groups with left-censored lognormal data, is also reiterated in Helsel [2012, pg. 178], in the summary section of tests used for comparing two groups: 'The Gehan and Peto-Prentice tests exhibited the most power when the underlying data were lognormal, the distribution most often used to model environmental data. The test with the overall best performance, including being able to accommodate unequal sample sizes and some measure of unequal censoring mechanisms, was the Peto-Prentice test using the asymptotic variance estimate. Environmental scientists would do well to look for software performing this version of a score test. When using other statistical software, look for the Peto-Prentice or Peto-Peto tests to achieve high power for multiply censored environmental data that are shaped close to a lognormal distribution.'

In conclusion, the Peto-Prentice is the most powerful test available that can be used for comparing two groups with left-censored lognormal data, and while the power of this test will vary depending on the actual data used, there is no better statistical test that can be used for this purpose.

Comment 21

Referring to page 30 of the SC&A report:

RPRT-0053 discusses hypothesis testing as though there were only two possible outcomes of the test. When the sample sizes are fixed by circumstance, there are, in fact, three possible outcomes, not two. The three outcomes are:

(1) Accept the Null hypothesis of No Difference
(2) Reject the Null hypothesis of No Difference
(3) No conclusion can be reached from these data

This 3-way list characterizes the "win," "lose," or "tie" nature of the decision-making under uncertainty. The process is best described in terms of the gray region for the test. The gray region is related to item #3 in the list above.

We are not familiar with the "win, lose, or tie nature of decision making" mentioned in this comment. As we discussed in Comment 20, there are only two outcomes for a null hypothesis test:

- reject the null, or
- fail to reject the null.

Failure to reject the null means that no conclusion can be reached from the data (#3 in the list above), but it is often referred to as "accepting the null" because the non-statistical life decision[14] is made to proceed with the null as if it was in fact true (#1 in the list above). Based on a statement made on page 31 of their report, we feel that SC&A understands this:

If a test concludes that there is no significant difference, this should not be taken as evidence that there is no difference, but rather that the data are insufficient to decide if there is a difference.

We are concerned that SC&A generates confusion in their comments by not keeping the concepts associated with the prospective design of an experiment separate from the concepts associated with the retrospective analysis of the data collected in an experiment. The "gray area" mentioned here and discussed in more detail in Section 4.1.1 is a good example of this, which is a concept associated with *a priori* power calculations in the

[14] Basically, before we perform the test we proclaim that if we fail to reject the null hypothesis we will take Action A whereas if we reject the null hypothesis we will take Action B. We are not trying to "prove" anything, we are making decisions [Casella 2002, pg 374.]. So, the whole issue of "failing to reject the null" versus "accepting the null" becomes somewhat moot.

design of an "experiment" (typically an environmental monitoring program) and not relevant to the retrospective analysis of datasets.

Comment 22

Referring to page 31 of the SC&A report:

In retrospective analysis of data, the gray region is also a useful tool for evaluating the performance of a test applied with sample sizes that are fixed and cannot be increased. When both the number of samples and the sample variability are fixed by circumstance, the width of the gray region is also fixed. In this case, it is necessary to determine if there is sufficient power in the available data to detect differences of the size of interest. One tool recommended in MARSSIM for analyzing the power of a hypothesis test is the test performance plot. This curve and its use in decision-making in retrospective analyses are discussed in the following section.

Referring to page 34 of the SC&A report:

In retrospective analysis of power, the sample sizes and the variability are known and the Type 1 error rate α is specified by selecting the confidence level used for the test. In this case, the power (100-β) and the width of the gray region are unknown.

In these comments SC&A is advocating the use of a *post-hoc* power analysis, which we discussed in Comment 20 and summarized as being ill-advised and of little practical value. In general, the discussion in MARSSIM relates to an *a priori* power analysis, not a *post-hoc* power analysis.

Comment 23

Referring to page 32 of the SC&A report:

Finding No. 6: For many years, given the small number of CTW data points, the tests cannot reliably detect differences smaller than a factor of 4 to 10 in the CTW/non-CTW ratio of geometric means. Larger differences have a 95% or better chance of detection. Smaller differences would be in the "gray region" for the test, sometimes detected, sometimes not. Overall, SC&A concludes that the NIOSH method of concluding that there are no significant differences would often lead to very claimant-unfavorable results.

Summarizing points made in previous comments:

- Failure to reject the null with retrospective datasets is inherently neither "claimant favorable" nor "claimant favorable" and is not indicative of "bad" data or inappropriate statistical methods.
- Once the data are collected the "gray region" of the test is not relevant.

The small CTW datasets argue for the use of an unstratified coworker model, perhaps used in conjunction with the 95th percentile intake rates if there is evidence that a particular construction trade worker had potential for exposure on a par with the higher exposed workgroups.

Comment 24

Referring to page 37 of the SC&A report:

A second concern with the hypothesis test strategy is that cases may arise when both groups contain the same worker. For example, in the derivative report RPRT-0056 (p. 12), NIOSH states the following [essentially the same passage appears in RPRT-0058 (page 12)]:

> *Because it was possible for a worker to change jobs during the course of a single evaluated period, it is possible that a worker would have some samples identified as nonCTW and others as CTW in the same period. Therefore, one person might have as many as four different OPOS results, one each for the AMW, CTW, nonCTW, and nonCTW+unk strata.*

When the radionuclide is long-lived, the OPOS values generated in each group for that worker will be strongly correlated.

So will the individual bioassay results. This is not a problem that is created by using the OPOS statistic and it can't be solved by using the individual bioassay results.

Referring to page 38 of the SC&A report:

Finding No. 7: The statistical tests for comparing two strata require that the samples in each group be independent. If a worker in one group is exposed to radionuclides with long retention in the body, then changes jobs and becomes part of the other group in the same year, the OPOS values are correlated for this worker. This correlation not only

violates the assumptions of the tests, but also creates a bias toward a decision of "No Difference" between the two groups.

The correlations in OPOS statistics caused by an individual changing jobs in any given year are considered to be a minor problem because it occurs relatively infrequently. We do not understand the interest SC&A has in this relatively minor contributor to correlations in the data, while at the same time ignoring the relatively significant correlations in individual bioassay results created by individuals submitting multiple samples per year -- a problem the OPOS statistic was meant to address.

Perhaps a more important issue here is that to stratify coworker models one has to be able to assign individuals to specific, unique, and meaningful job titles (i.e., develop a *job exposure matrix*) for all times of employment. The difficulty associated with doing this, as discussed by SC&A in this section is a general problem associated with assembling a job exposure matrix and really has little to do with the use of the OPOS statistic. In fact, the problem raised by SC&A in this section is an argument for not stratifying a dataset and using the standard coworker model.

Comment 25

Referring to page 40 of the SC&A report:

Claimant favorability is always of concern when setting the standards for the level of significance. NIOSH has proposed that strong evidence is necessary before any differences between groups of workers should be considered in the coworker model. In the examples in Sections 5.1 and 5.2 of RPRT-0053, and in subsequent applications to neptunium (ORAUT 2012b), mixed fission and activation products (ORAUT 2012c), and exotic trivalent radionuclides (ORAUT 2012a) at SRS, NIOSH conducts the statistical tests for a significant difference at the $\alpha = 0.05$ probability level requiring a 95% level of confidence. A higher level of confidence makes it more difficult to decide if there are differences between the two groups. A 90% level of confidence for the MCPT would be more claimant-favorable. The issue of confidence levels and power are further addressed in Sections 4.1 and 4.2.

Finding No. 8: Although one example where a significant difference is found is presented in the report, NIOSH has not provided any measure of the power of the hypothesis test procedure to detect differences within the worker population. This deficiency should be

corrected before the test is adopted as an appropriate procedure for coworker models. Conducting the tests at a 90% level of confidence would be claimant favorable.

If conducting tests at an $\alpha = 0.1$ significance level (90% confidence level) would be "claimant favorable" as claimed in this comment, one might conclude that conducting the tests at a 50% confidence level would be even more "claimant favorable." Where does it end? The answer to that question is that the significance level chosen for a null hypothesis test is ultimately a judgment based primarily on the conventions established in a particular scientific field. More specifically, a significance level of $\alpha = 0.05$ (95% confidence level) appears to be the standard significance level used in the most areas of science[15].

SC&A has offered no justification for using 90% confidence level versus the standard 95% confidence level other than that it is more "claimant favorable." We do not feel that this is a reasonable justification, especially considering the ambiguity of "claimant favorable" in this context. In fact, by advocating a 90% level of confidence, using a one-sided null hypothesis test, and mis-specifying the null hypotheses in these tests, we feel that SC&A has created a situation which renders any conclusions that are drawn to be equivocal. For example, is a significant result of a statistical test rigged to favor significance really significant? For these reasons we feel that a 95% confidence level for the tests in RPRT-0053 is the most appropriate confidence level to use.

Comment 26

Referring to Section 4.3.1, which starts on page 45 of the SC&A report:

The Peto-Prentice test is a generalization of the WRS test which is a test for the location of one distribution relative to the other.

As discussed in Comment 15, the Peto-Prentice test used in RPRT-0053 tests for different empirical cumulative distributions, denoted here by F(x), not different location of distributions.

Tests of location may be applied using three different forms of the hypothesis test, which differ in terms of the null hypothesis (H_0). Three hypothesis test forms may be tested using the Peto-Prentice statistic z:

[15] http://www.jerrydallal.com/LHSP/p05.htm

(A) H_0: The distribution of the bioassay data is the same for CTW and non-CTW vs. H_A: the distribution of data is not the same for CTW and non-CTW

Note that SC&A correctly specified "distribution" in the hypotheses in (A) as opposed to "mean" or some other measure of location. The null and alternate hypotheses in this case may be specified as:

$$H_0 : F_{ctw}(x) = F_{nctw}(x) \quad \textit{for all } x$$
$$H_A : F_{ctw}(x) \neq F_{nctw}(x) \quad \textit{for at least one } x$$

(B) H_0: The distribution of the bioassay data non-CTW is higher than for CTW vs. H_A: the distribution of data for CTW is higher than for non-CTW

The null and alternate hypotheses as stated by SC&A for (B) are incorrectly stated as

$$H_0 : F_{ctw}(x) < F_{nctw}(x)$$
$$H_A : F_{ctw}(x) > F_{nctw}(x)$$

The null and alternate hypotheses must be complementary and exhaustive. In other words, the null and alternate hypotheses must cover all possibilities and if one hypothesis is not true then the other must be. In addition, the null hypothesis must contain an equality (i.e., =, ≥, ≤) or else the null distribution can not be calculated. Case (B) can be correctly expressed as:

$$H_0 : F_{ctw}(x) \leq F_{nctw}(x) \quad \textit{for all } x$$
$$H_A : F_{ctw}(x) > F_{nctw}(x) \quad \textit{for at least one } x$$

(C) H_0: The distribution of the bioassay data for CTW is higher than for non-CTW vs. H_A: the distribution of data for non-CTW is higher than for CTW

The null and alternate hypotheses as stated by SC&A for (C) are incorrectly stated as

$$H_0 : F_{ctw}(x) > F_{nctw}(x)$$
$$H_A : F_{ctw}(x) < F_{nctw}(x)$$

As discussed above, the hypotheses should be stated like this:

$$H_0 : F_{ctw}(x) \geq F_{nctw}(x) \quad \textit{for all } x$$
$$H_A : F_{ctw}(x) < F_{nctw}(x) \quad \textit{for at least one } x$$

Although the three test forms may appear similar, in practice, there are large differences between the three test forms in terms of claimant favorability. The differences arise because the null hypothesis is assumed true until the data provide sufficient evidence to reject the null hypothesis.

For clarity, we would like to remind the reader that in these tests the null hypothesis is either rejected or not rejected. Note that not rejected is not the same as proven true. Basically, if we fail to reject the null hypothesis we have in fact proven nothing. This is a fundamental property of *null hypothesis testing* (NHT). In practice, phrases like the "null hypothesis is accepted" or "the null hypothesis is retained" are used when there is failure to reject the null, but we must remember that this is a life decision that is not supported by the data.

If the sample size for one or both groups is too small, the test would not have sufficient power to reject the null hypothesis. Test form A is a 2-sided test. With a 2-sided test, the null hypothesis of "No Difference" is rejected if the CTW data are either significantly higher or lower than the non-CTW data. If the sample size is too small, the test may have insufficient power to reject the null hypothesis of No Difference. Using this test form, the null hypothesis is accepted due only to a lack of evidence in the data that proves the CTW are different. This is not claimant favorable, as it places the burden of proof on the claimants despite the known lack of sufficient data to provide such proof.

Problems are associated with the concept of "claimant favorability" in this situation. This is discussed in Comment 8, but it boils down to the fact that a coworker model is a "zero sum game." If you take a coworker model with a certain GM and split it into two groups, the GM in one group will go up and the GM in the other group will go down (or they can both stay the same). It is not possible to have the GM go up in both groups, so stratification has to be unfavorable to one group or the other. We feel it is inequitable to declare a stratification to be "claimant favorable" if the dose to the group you are

interested in goes up and "claimant unfavorable" if the dose goes down. One of the reasons why we favor two-sided hypothesis tests over one-sided hypothesis tests is that we think stratification should be consider in the case of significant differences in doses to different strata regardless of the direction of the difference.

Test forms B and C are both 1-sided tests. In test form B, the null hypothesis is that the non-CTW data are higher than the CTW data.

As discussed above, this is an incorrect expression of the null hypothesis. The correct way to state this is

> "In test form B, the null hypothesis is that the non-CTW data are higher than or equal to the CTW data."

In this form of a 1-sided test, the null hypothesis is rejected if the CTW data are significantly higher than the non-CTW data. If the sample size is too small, the test may have insufficient power to reject the null hypothesis that the non-CTW distribution is at least as high as the CTW distribution. As with test form A, the null hypothesis may be accepted due only to a lack of evidence in the data to prove the CTWs are different from non-CTWs. Test form B is also not claimant favorable, as it places an unreasonable burden of proof on the claimant to show that the CTW data are higher than the non-CTW data despite the known lack of sufficient data.

The 1-sided test form B is more relevant than the 2-sided test form A. Unlike test form A, test form B at least provides a clear answer as to whether the CTW are higher than the non-CTW data, which is the issue in question.

Test form C is also a 1-sided test. In test form C, the null hypothesis is that the CTW data are higher than the non-CTW data.

As discussed above, this is an incorrect expression of the null hypothesis. The correct way to state this is

> "In test form C, the null hypothesis is that the CTW data are higher than or equal to the non-CTW data."

In this form of a 1-sided test, the null hypothesis is rejected if the non-CTW data are significantly higher than the CTW data. If the sample size is too small, the test may have

insufficient power to reject the null hypothesis that the CTW distribution is higher than the non-CTW distribution. Of the three test forms, only test form C is claimant favorable when the sample sizes are too small to provide clear evidence.

Note that accepting the properly stated null hypothesis includes the possibility that the distribution of the CTW bioassay is the same as the distribution of the non-CTW bioassay, i.e., there is no difference between the two. The SC&A goal of "claimant favorability" is only achieved with an improperly expressed null hypothesis.

Unless there is significant statistical evidence to the contrary, the null hypothesis that the CTW samples are higher than non-CTW should be accepted as claimant favorable.

A common theme throughout the SC&A report that is exemplified by this comment is that they feel that the doses to the CTW should be assumed to significantly greater than the doses to non-CTW until the data prove otherwise. This is the "reversing of the null hypothesis" mentioned in SC&A's first recommendation given on page 9 of their report:

NIOSH might consider reversing the null hypothesis for the Peto-Prentice test. NIOSH's implementation of the hypothesis tests to test for differences between CTWs and non-CTWs at SRS uses a null hypothesis that is not claimant favorable, as it places the burden of proof on the CTW claimants to prove a significant difference.

Adopting this approach means that we could go looking for a difference in strata, fail to detect that difference, and then develop a model that incorporates the difference anyway. We see fundamental difficulties associated with this approach. If the data are not adequate to demonstrate a significant difference between the strata then it is not clear to us how incorporating this difference into the coworker model will improve the estimates of the intake rates. This is critically important when there are in fact no significant differences between the strata. In this case stratification (the default action taken when we fail to reject the null) will always result in poorer estimates of the intake rates because it unnecessarily reduces the size of the sample used to estimate the intake rate.

In theory, a better approach to achieving what SC&A appears to be after here would be to define a difference *d* between the CTW and non-CTW empirical cumulative distributions that is considered to be of practical significance and rearrange the hypotheses to form an *equivalence test* [Streiner 2003, Wellek 2010]:

$$H_0 : \{F_{ctw}(x) - F_{nctw}(x)\} \geq d \quad for\ all\ x$$
$$H_A : \{F_{ctw}(x) - F_{nctw}(x)\} < d \quad for\ at\ least\ one\ x$$

Here, H_0 is that the distributions are not equivalent and H_A is that they are. Thus, the data can prove that the CTW and non-CTW are equivalent, i.e., that there is no practically significant difference between the two, by rejecting the null hypothesis. If the null is retained we can proceed under the assumption that the CTW distribution is significantly different than the non-CTW distribution.

The difficulty in implementing an equivalence test is that we have to define doses or bioassay results that are of practical significance to the compensation decision before looking at the data. During the development of RPRT-0053 we tried to define *d* and were unsuccessful, which is why we used the null hypothesis test of statistical significance rather than the equivalence test of practical significance. Note that if one decides that any difference is of practical significance (d = 0), the equivalence test reduces to test (C), which as we have seen does not answer the question SC&A is asking.

Comment 27

Referring to recommendation 5 on page 9 of the SC&A report:

In principle, multiple comparisons can be done for more refined groupings, like CTWs by job type with all non-CTWs. But this will run into difficulties in many cases, as we found in prior analyses even for a 10-sample threshold. It will be much more difficult to meet the 30-sample threshold needed for the tests recommended in RPRT-0053, but this is essential for a valid comparison. Moreover, a valid comparison requires that the 30 sample threshold be met for each of the two groups, not just one. RPRT-0053 is not explicit on this point, though it is implied in footnote 6 on page 9. The 30-sample threshold for each group should be made explicit.

We struggled quite a bit over establishing minimum strata sizes for the statistical tests, especially in the presence of censoring. Extremely small strata are undesirable and should be combined with other strata with similar key characteristics when possible. Small strata that represent small samples of some larger group can have large uncertainties in the estimated parameters. On the other hand, if a small stratum is basically a census of all workers who should have been monitored and all of the data are uncensored, the situation may not be as bad. In the end, we recommended a minimum of 30 OPOS statistics, i.e.

data from 30 individuals, in each stratum. However, because the procedures in RPRT-0053 were going to be used only by degreed statisticians, we gave them the latitude to exercise professional judgment on this subject.

Comment 28

Referring to page 22 of the SC&A report:

The literature search does not include reference to the Brunner-Munzel test (Brunner and Munzel 2000). This nonparametric test is another generalization of the WRS test designed for comparisons of populations with different variances.

The Brunner-Munzel test (Brunner 2000) is a modification of the Wilcoxon-Mann-Whitney test designed to handle ties and unequal variances. Instead of associating ranks with the sample observations, midranks are computed. Midranks are equal to ranks when there are no tied values. For tied values, the midranks are the average of their ranks. For example, the midranks of 2, 5, 5, 6, 9, 9, 9, 10 are 1, 2.5, 2.5, 4, 6, 6, 6, 8. If $\bar{M}_X$ and $\bar{M}_Y$ are the means of the midranks associated with the samples *X* and *Y*, when the data are pooled, then the Brunner-Munzel test statistic is computed by the following formula:

$$B = \frac{\bar{M}_Y - \bar{M}_X}{(m+n)\sqrt{\frac{SB_X^2}{mn^2} + \frac{SB_Y^2}{m^2 n}}}$$

where *m* and *n* are the number of observations for samples *X* and *Y*, SB_X^2 and SB_Y^2 are the variance estimates for the two samples (see Fagerland 2011b for exact formulas for SB_X^2 and SB_Y^2). The distribution of *B* can be approximated by a *t*-distribution with $f_B$degrees of freedom (see Fagerland 2011b for exact formula for f_B).

The Brunner-Munzel test is looking at the stochastic equality of two different populations. Stochastic equality is a measure of similarity between two populations, and is defined as $\Pr(X < Y) = \Pr(X > Y)$, which means than neither population has a much larger frequency of greater values than the other population. The hypothesis of the Brunner-Munzel test is expressed in term of the stochastic superiority, which is defined as:

$$P = \Pr(X < Y) + 0.5\Pr(X = Y)$$

Then, the null hypothesis used by the Brunner-Munzel test is:

$$H_0: P = 0.5,$$

meaning that neither group generally has larger values than the other,
versus the alternative:

$$H_a: P \neq 0.5,$$

meaning than in one of the groups greater values occur more frequently than in the other group. If the null hypothesis holds, then this implies the stochastic equality of the two populations. If the null hypothesis is rejected, the interpretation is not straightforward, but the conclusion that can be drawn is that the two populations differ in some way.
It is worth mentioning that stochastic equality of two populations does not imply equality of the means of the two populations (unless the two populations are symmetric), and equality of the means does not imply stochastic equality; the same is true with respect to the medians. So, in the general case, the concept of stochastic equality is different than the equality of means or medians.

While most of the practical applications of the Brunner-Munzel test are for ordered categorical data (e.g., pain scores example in Brunner 2000, pg. 22-23, examples in Fagerland 2011a, pg. 6-7), the test can be used for both continuous and discrete distributions (Delaney 2002, pg. 486). However, we couldn't find any application of the Brunner-Munzel test used to compare two groups with censored data. Since this test is using the midranks in order to rank the observations, is not clear how one will assign the midranks to censored data with multiple detection limits (e.g. <0.3, <1.5, <2.7). This is in contrast with the tests that are designed to compare two groups with censored data, like the Peto-Prentice test, that can handle data censored at multiple detection limits, using the information in detected values between detection limits in addition to the information in the proportion of values below each detection limit.

The Brunner-Munzel test is not implemented in standard statistical software (Skovlund 2010, pg. 595), and while there are some macros available in SAS and R, they seem to provide inconsistent results (for example, the brunner.munzel.test available in the R package 'lawstat', produce a p-value of 0.788961 for the pain scores example in Brunner 2000, while the reported p-value in the article is 0.792).

In conclusion, while the Brunner-Munzel test was suggested by SC&A as an alternative to the Peto-Prentice test, there is no available reference that shows how this test can be used in the comparison of two groups with censored data.

References

Casella 2002
George Casella and Roger Berger Statistical Inference (Pacific Grove: Duxbury) 2002.

Chmelynski 2013
H. Chmelynski, J. Lipsztein, S. Marschke, and J. Stiver Draft Review of ORAUT-RPRT-0053: Analysis of Stratified Coworker Datasets, Rev. 1, April 2013.

Delaney 2002
Delaney H. D., Vargha A., 2002, *Comparing Several Robust Tests of Stochastic Equality With Ordinally Scaled Variables and Small to Moderate Sized Samples*, Psychological Methods, vol. 7, no. 4, pp. 485-503.

Ellis 2010
Paul D. Ellis, The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results (Cambridge: Cambridge University Press) 2010.

Fagerland 2011a
Fagerland M. W., Sandvik L., Mowinckel P., 2011, *Parametric methods outperformed non-parametric methods in comparisons of discrete numerical variables*, BMC Medical Research Methodology, 11:44.

Fagerland 2011b
Fagerland M.W., Sandvik L., Mowinckel P., 2011, *Parametric methods outperformed non-parametric methods in comparisons of discrete numerical variables – Additional File 1 : Test statistics*, BMC Medical Research Methodology, 11:44.

Helsel 2012
Helsel H. Statistics for Censored Environmental Data Using Minitab and R, Wiley, Hoboken, New Jersey, 2012.

Hoenig 2001
J. M. Hoenig and D. M. Heisey *The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis* American Statistician (55), No. 1, pp. 19-24, 2001.

IAEA 2004
IAEA. Methods for Assessing Occupational Radiation Doses Due to Intakes of Radionuclides. Number 37 in Safety Report Series. International Atomic Energy Agency, 2004.

ICRP 1997
Individual Monitoring for Internal Exposure of Workers, ICRP Publication 78 Annals of the ICRP 27 (3-4), 1997.

Leton 2002
Leton E. and Zuluaga P. *Survival Tests for r Groups*, Biometrical Journal, vol. 44, no. 1, pp. 15-27, 2002.

Leton 2005
Leton E. and Zuluaga P. *Relationships Among Tests for Censored Data*, Biometrical Journal, vol. 3, pp. 377-387, 2005.

Leton 2008
Leton E. and Zuluaga P. *Unbalanced Groups in Nonparametric Survival Tests*, Statistics and Econometrics Series 15, Working Paper 08-52, 2008.

Magel 1991
Magel R. C. *Estimating the Power of the Gehan Test,* Biometrical Journal, vol. 88, no. 8, pp. 985-997, 1991.

ORAUT 2012
Analysis of Stratified Coworker Datasets, Rev 1. ORAUT-RPRT-0053, 2012.

Skovlund 2010
Skovlund E., 2010 *A nonparametric two-sample comparison for skewed data with unequal variances*, J. of Clinical Epidemiology vol. 63, pp. 594-595.

Skrable 1994
Skrable, K. W. et al., *Estimation of Intakes from Repetitive Bioassay Measurements*, Chapter 19 in Internal Radiation Dosimetry, Otto G. Raabe, editor. (Madison: Medical Physics Publisher), 1994.

Streiner 2003
David L Streiner *Unicorns Do Exist: A Tutorial on "Proving" the Null Hypothesis* Canadian Journal of Psychiatry, (48) No 11, December 2003.

Wamil 1997
Wamil J. K. *Estimating the Power of the Peto-Prentice Test*, MS Thesis, North Dakota State University.

Wellek 2010
Stefan Wellek, Testing Statistical Hypotheses of Equivalence and Noninferiorty (Boca Raton: CRC Press), 2010.

Appendix A : Power plots for the Peto-Prentice and Gehan tests

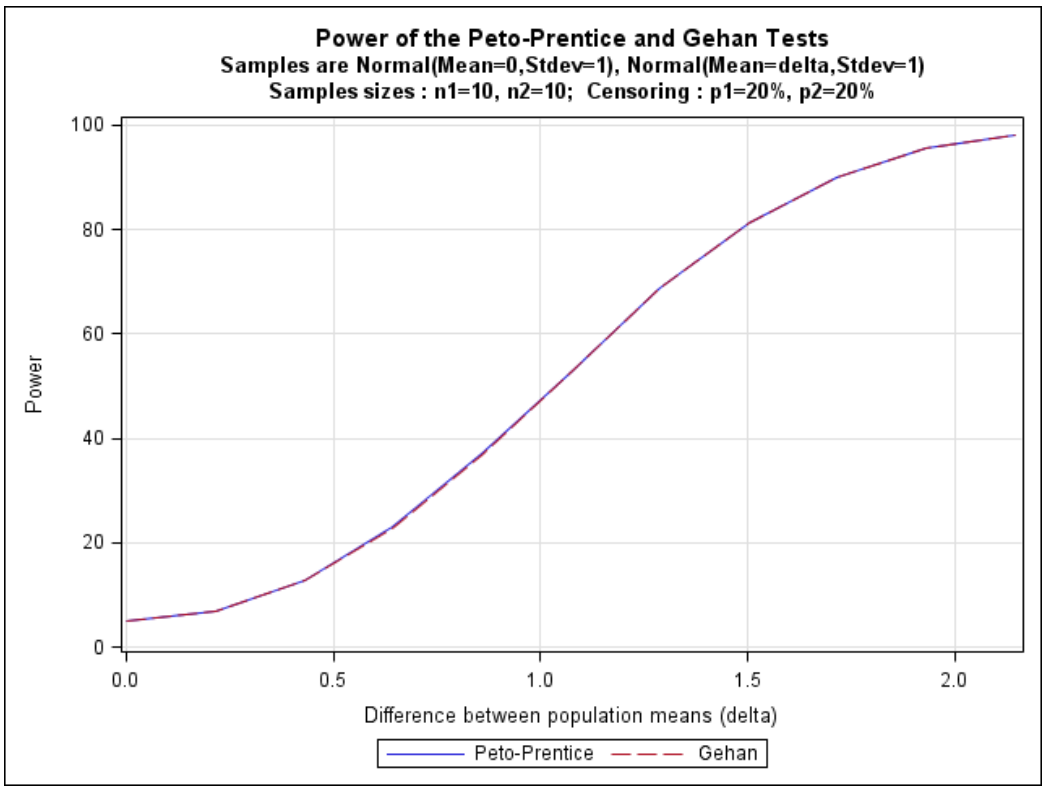

Figure 1

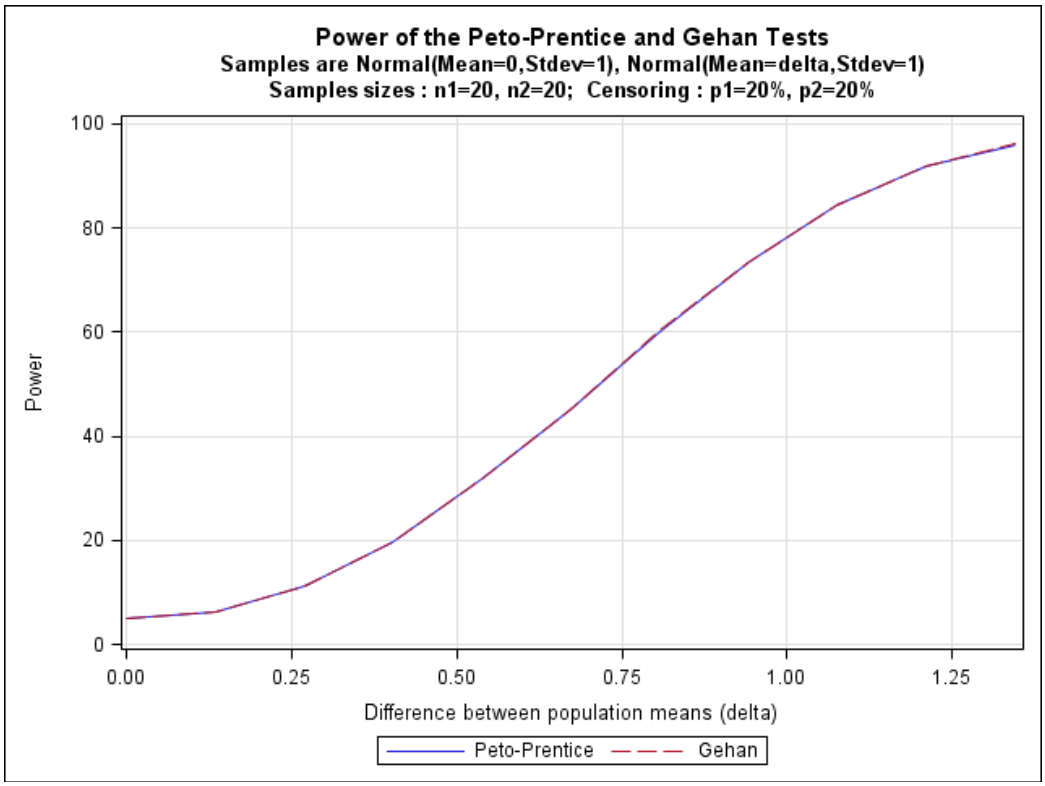

Figure 2

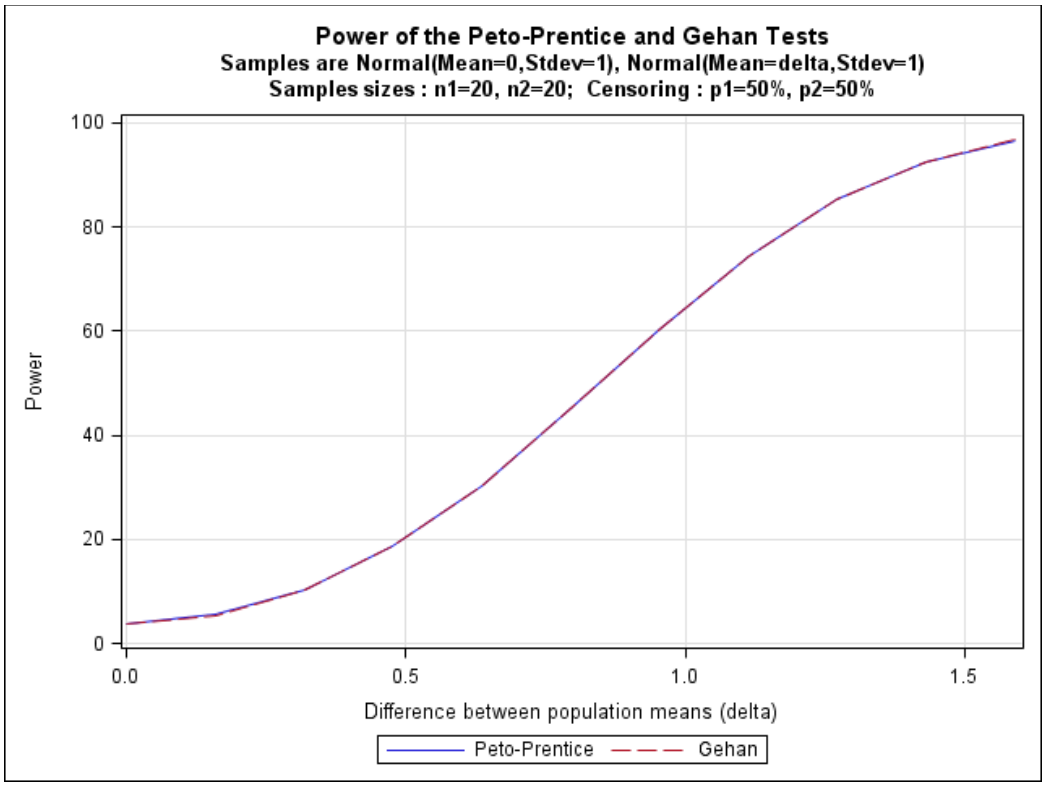

Figure 3

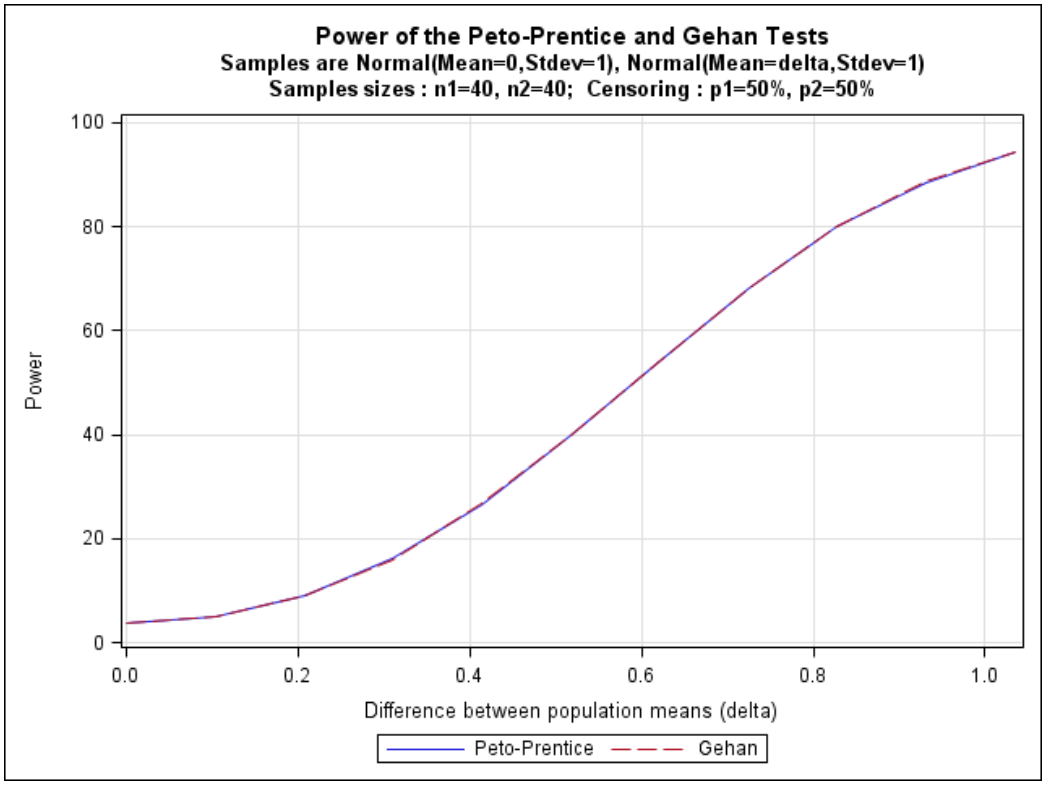

Figure 4

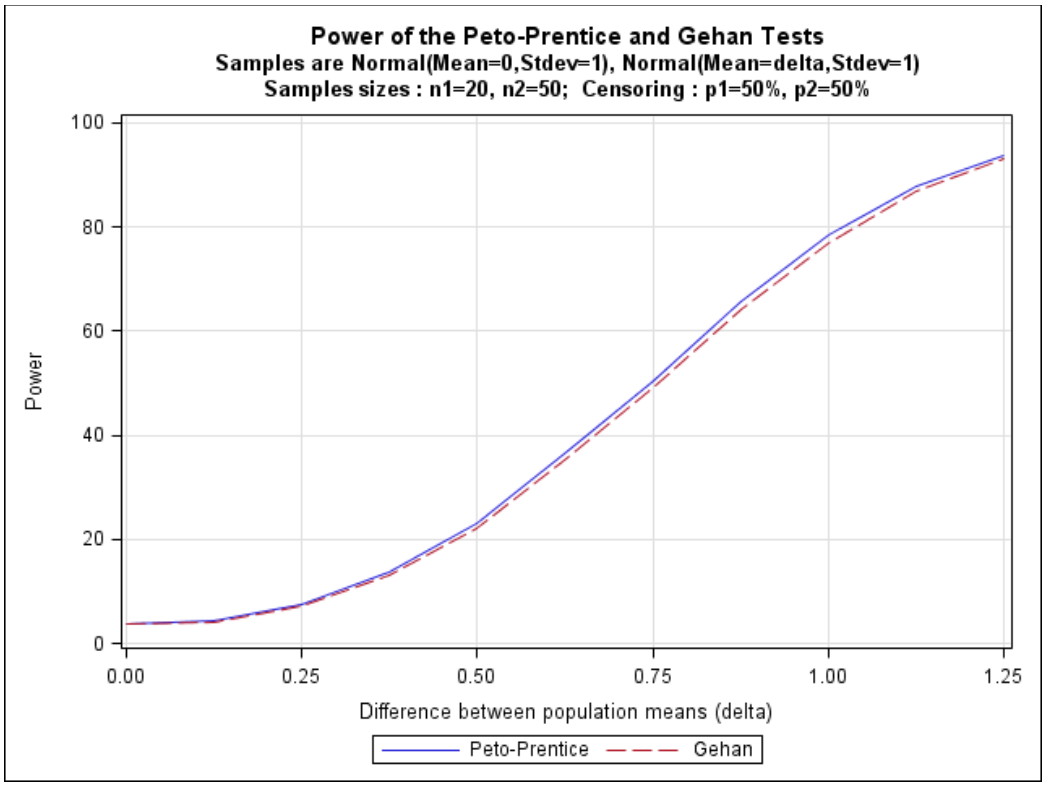

Figure 5

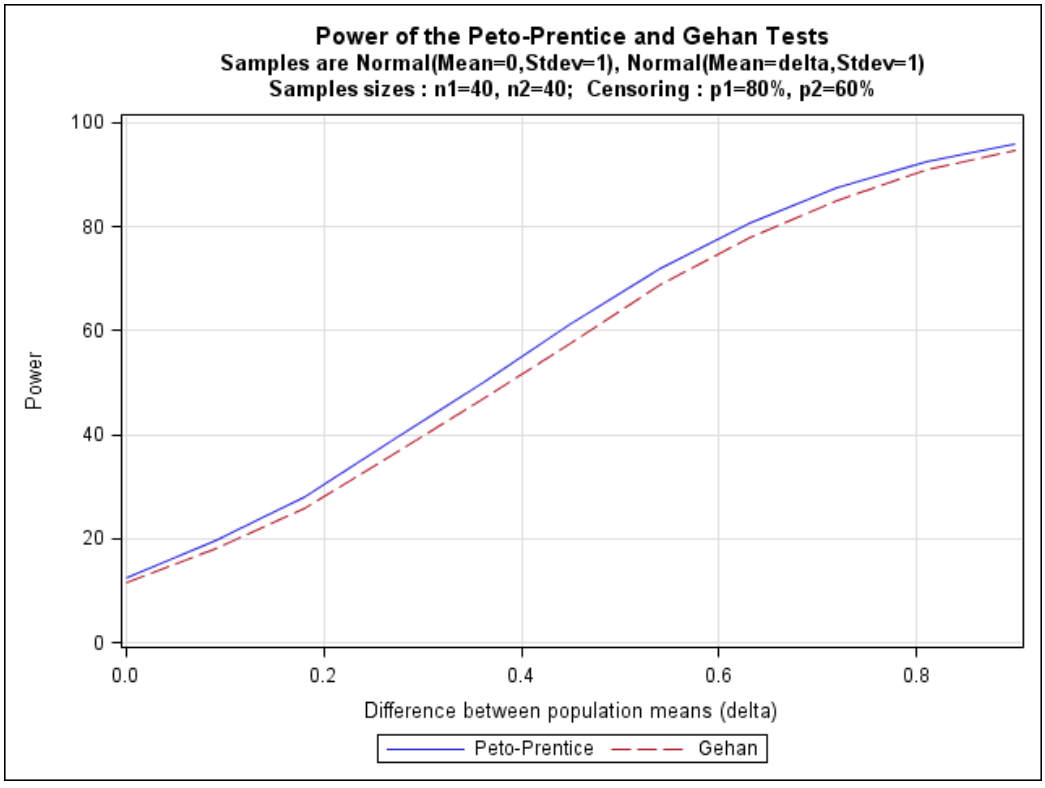

Figure 6

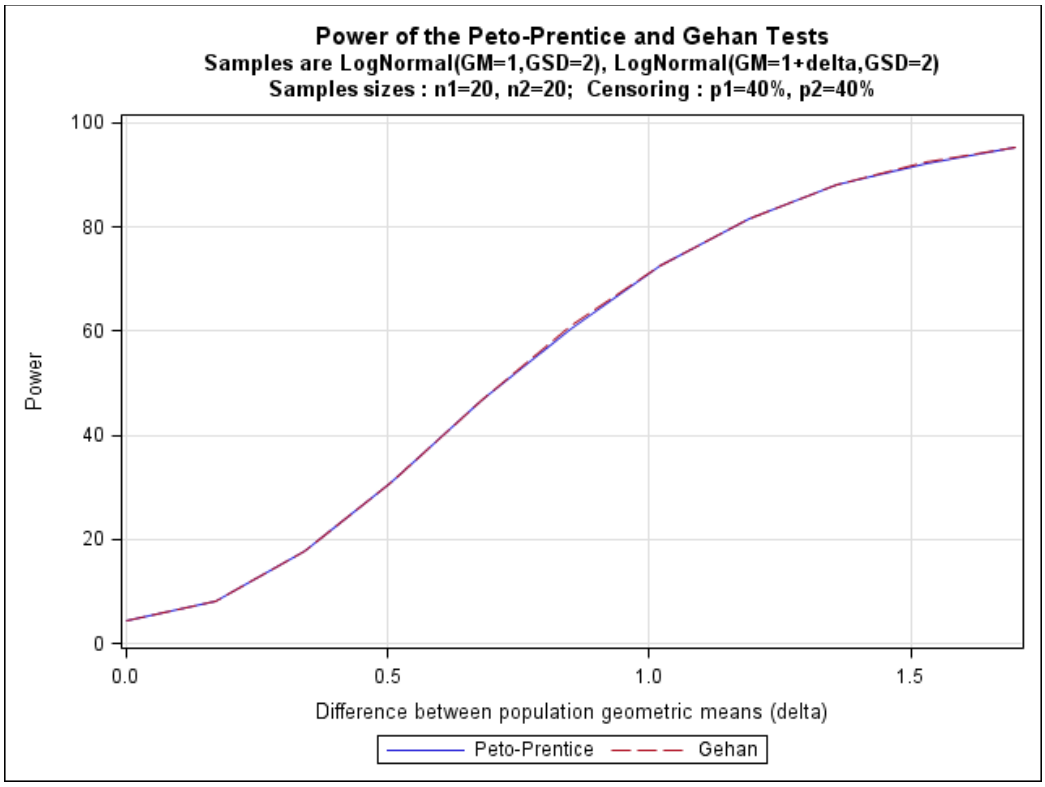

Figure 7

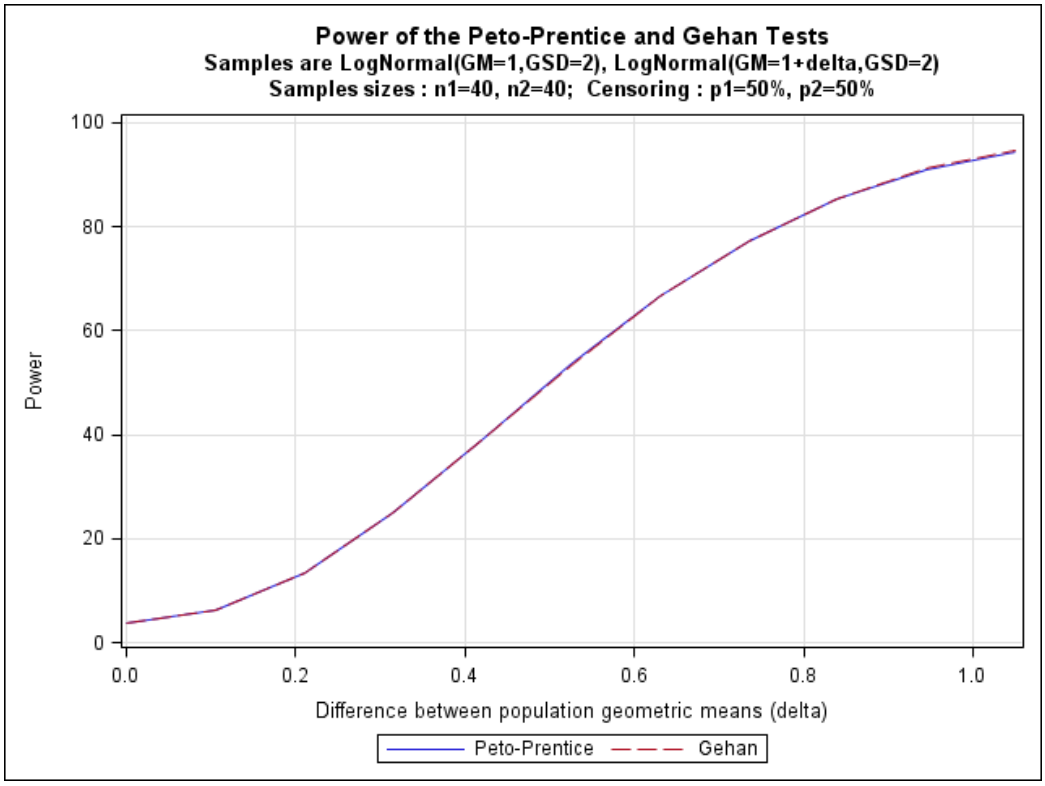

Figure 8

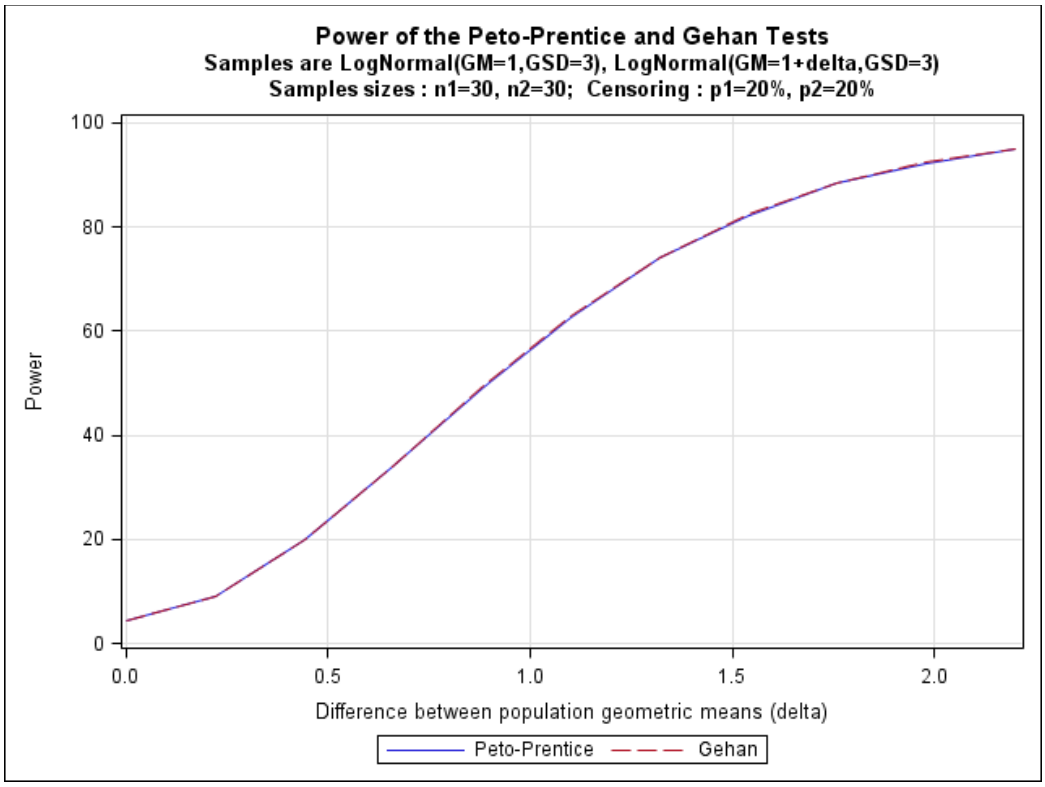

Figure 9

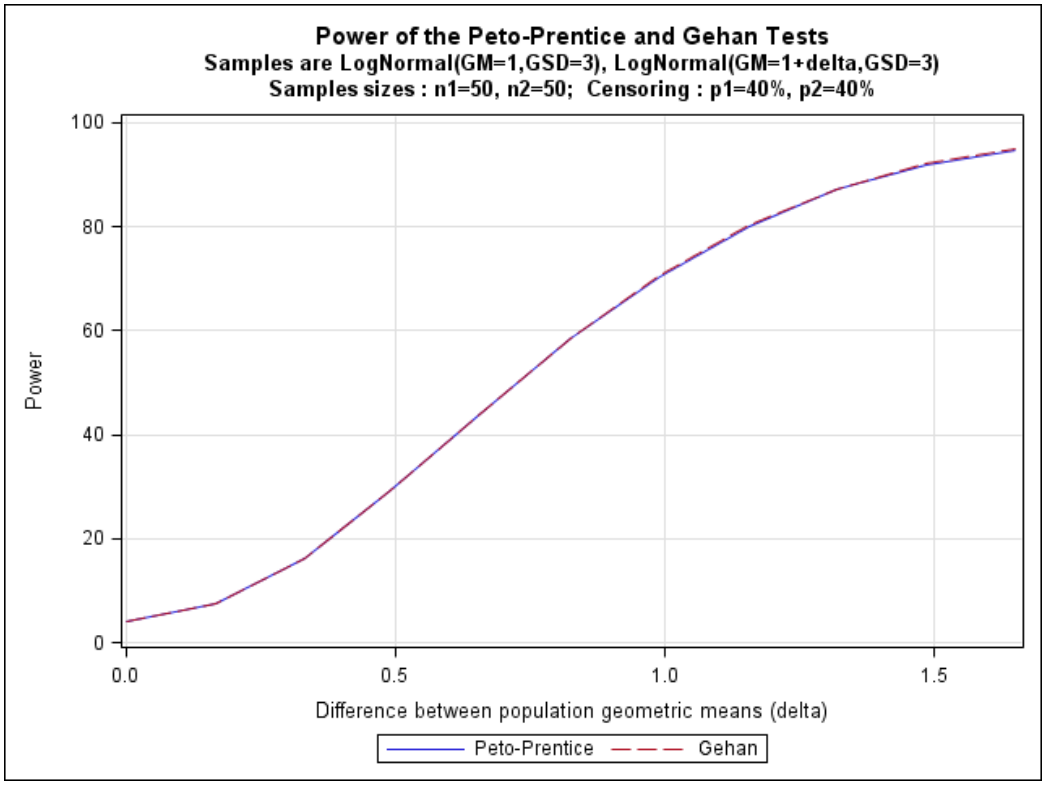

Figure 10

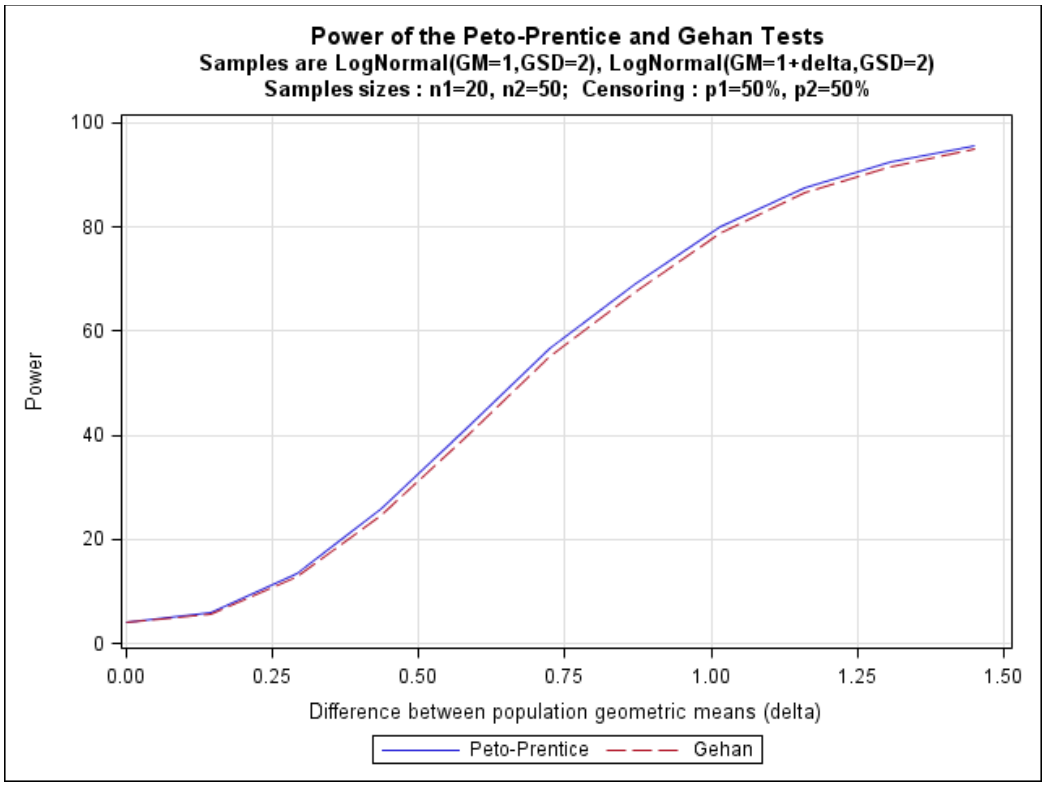

Figure 11

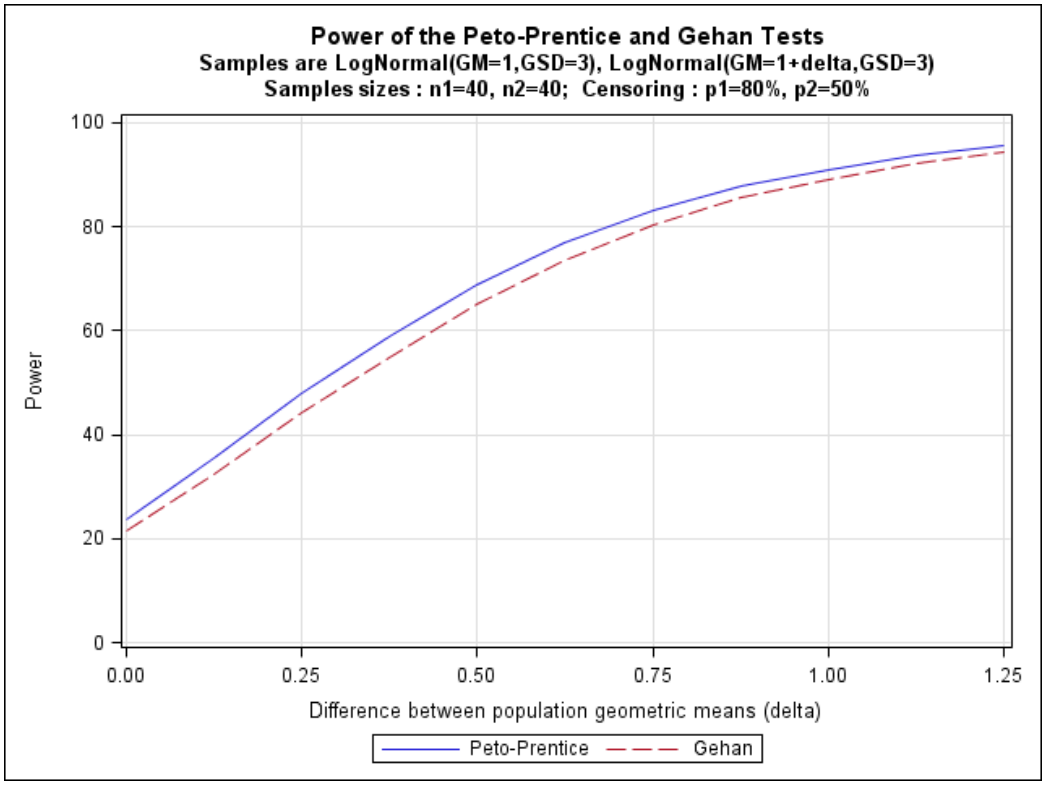

Figure 12