Draft

ADVISORY BOARD ON
RADIATION AND WORKER HEALTH

National Institute for Occupational Safety and Health

SC&A'S EVALUATION OF ORAUT-RPRT-0078, REVISION 00, "TECHNICAL BASIS FOR SAMPLING PLAN"

Contract No. 211-2014-58081
SCA-TR-2017-PR009, Revision 0

Prepared by

Ron Buchanan, PhD, CHP
Harry Chmelynski, PhD

SC&A, Inc.
2200 Wilson Boulevard, Suite 300
Arlington, Virginia, 22201

Saliant, Inc.
5579 Catholic Church Road
Jefferson, Maryland 21755

October 2017

SC&A, INC.: ***Technical Support for the Advisory Board on Radiation and Worker Health Review of NIOSH Dose Reconstruction Program***

DOCUMENT TITLE:	SC&A's Evaluation of ORAUT-RPRT-0078, Revision 00, "Technical Basis for Sampling Plan"
DOCUMENT NUMBER/ DESCRIPTION:	SCA-TR-2017-PR009
REVISION NO.:	0 (Draft)
SUPERSEDES:	N/A
EFFECTIVE DATE:	October 3, 2017
TASK MANAGER:	Kathy Behling [signature on file]
PROJECT MANAGER:	John Stiver, MS, CHP [signature on file]
DOCUMENT REVIEWER(S):	Kathy Behling [signature on file] John Stiver, MS, CHP [signature on file]

Record of Revisions

Revision Number	Effective Date	Description of Revision
0 (Draft)	10/3/2017	Initial issue

TABLE OF CONTENTS

ABBREVIATIONS AND ACRONYMS

ABRWH	Advisory Board on Radiation and Worker Health
AQL	acceptable quality level
α	alpha
β	beta
γ	gamma
LTPD	lot tolerance percent defective
CI	confidence interval
dpm/l	disintegration per minute per liter
$f(m)$	probability of observing m typos in a sample
m	observed number of typos in a sample of data
M	true number of typos in a total data population
n	number of fields in a sample of data
N	number of fields in a total data population
NIOSH	National Institute for Occupational Safety and Health
OC	operational characteristic
ORAUT	Oak Ridge Associated Universities Team
RPRT	report
θ	theta

1.0 INTRODUCTION AND BACKGROUND

In June 2017, the Advisory Board on Radiation and Worker Health tasked SC&A with a technical review of ORAUT-RPRT-0078, *Technical Basis for Sampling Plan*, Revision 00, issued June 17, 2016 (NIOSH 2016, referred to as "RPRT-0078"). In RPRT-0078, the National Institute for Occupational Safety and Health (NIOSH) presents a method to select a sample of data stored in an electronic database and compare it to the corresponding hardcopy (or other primary forms of data) to estimate the number of typos present in the electronic database. NIOSH then provides a means for determining if the observed number of typos in the sample is acceptable, or if the sample should be rejected. This process is useful when a selection of data is to be used from a large population of data stored in an electronic database that has been populated from hardcopy records, such as for creating coworker dose or intake rates.

This report presents SC&A's evaluation of the technical approach, statistical methods, and documentation used by NIOSH in RPRT-0078.

2.0 OVERVIEW OF ORAUT-RPRT-0078

RPRT-0078 is a detailed document and, for evaluation purposes, it is advantageous to provide a brief outline, as follows:

- **Purpose** – Section 1.0 (page 5). The purpose of the document is to provide a statistical sampling technique in which a comparison of the data in the electronic dataset to the original data is performed (after the transcription from the original to the electronic database is complete) to estimate the typo rate. This information is used to determine if the specified typo rate has, or has not, been exceeded. One or more entry errors in a field in the electronic database is considered one typo. Additional errors in the same field are not consider additional typos.

  The main items addressed are:

  - The sample size that is needed (i.e., how many fields need to be compared) given preselected acceptance parameters
  - Determining the acceptable typo rate in a given sample of the total population

  The process in RPRT-0078 only applies to the selection of the appropriate sample and typo rate analysis. It provides no assurance that the hardcopy database is accurate or complete, or that the transcription of the data from the hardcopy to the electronic database is complete.

- **Terms** – Section 2.0 (pages 5–6). A field is an entry into the database (e.g., "0.12 dpm/l"). There are three types of fields in a hardcopy or electronic database: critical, noncritical, and irrelevant, as described below:

  - **Critical Field –** Data in which any error makes the data unusable.
  - **Noncritical Field –** Useful data, but not critical.

- **All Fields –** Consist of both critical and noncritical fields.
- **Irrelevant Field-** Data are not relevant and can be discarded.
- **Lot** – All the usable records in an electronic database.
- **Sample** – The set of entries (consisting of either critical or all fields) selected for analysis. There will be separate sets of entries (samples) for critical and for all fields from the electronic database, which will then be compared to the hardcopy.

The critical fields and all fields are analyzed for typos independently. An excess of typos in either type of field disqualifies the database for use as representative data.

- **Basic Equation** – Section 3.0 (page 6). The basic function used in RPRT-0078 is presented on page 6 as Equation 3-1. This is the probability mass function for the hypergeometric distribution that provides the probability $f(m)$ of observing m typos in a sample of size n number of fields. This is not a simple arithmetic equation, but a probability function that is analyzed using a computer program. It is illustrated in the example on page 7, where the probability of observing 30 typos in a sample of 3,000 fields, if the original transcription typo rate was 1% (50 out of 5,000), is 0.115 (11.5%). The probability of observing other numbers of typos (e.g., 0 to 50) under the same conditions is presented in Figure 3-1 on page 7.

- **Acceptable Error Rate** – Section 3.1 (pages 8–9). The acceptable quality level (AQL) is defined as the typo rate that is acceptable. As stated on page 8, NIOSH has defined the AQL for critical fields to be 0.005 (0.5%) and for all fields 0.025 (2.5%).

- **Unacceptable Error Rate** – Section 3.1 (pages 8–9). The lot tolerance percent defective (LTPD) is defined as the typo rate that is unacceptable. As stated on page 6, NIOSH has defined the LTPD for critical fields to be 0.01 (1%) and for all fields 0.05 (5%).

- **Consumer's Accept Number** – Section 3.1 (pages 8–9). The user of the data (the consumer) sets the limit at which the typo rate (e.g., 1%) can be exceeded and the data can still be used (in RPRT-0078, NIOSH used a limit of no more than 2.5%; i.e., 97.5% of the time the typo rate will be less than 1%). Under these parameters, the number of typos observed in a given sample that is acceptable is determined, and this value is termed the consumer's "accept number." For example, as illustrated in Figure 3-1, the accept number is 22 typos out of a sample of 3,000 from a population of 5,000. Using an accept number of 22 typos in this case means that there is less than a 2.5% chance of there being more than 50 typos in the total population when 22 (or fewer) typos are observed in a sample of 3,000 out of a population of 5,000.

- **Producer's Accept Number** – Section 3.1 (pages 8–9). The risk of rejecting a sample that has less than the acceptable typo rate is termed the producer risk and is set at 0.025 (2.5%) in the examples in RPRT-0078. Figure 3-2 illustrates the selection of the producer's accept number in a process that is similar to the selection of the consumer's accept number above, only with reverse emphasis. In this case, the producer's accept number is 20, meaning that observing more than 20 typos in a sample of 3,000 (out of a

population of 5,000 with 50 typos) would result in incorrect rejection of a valid sample 2.5% of the time.

- **Convergence of Accept Numbers –** Section 3.1 (pages 8–9). The goal is to adjust the sample size so that the accept number for the distribution in Figure 3-1 is the same as the accept number for the distribution in Figure 3-2. To accomplish this, using the parameters recommended in RPRT-0078, the sample size is adjusted until an accept number is obtained that leads to the acceptance of a good sample at least 97.5% of the time, while at the same time accepting a bad sample at most 2.5% of the time. As shown in Figure 3-3, this occurs at n = 2,435 where the acceptance number is 17. This means a random sample of n = 2,435 critical fields (out of a total population of 5,000) containing 17 or fewer observed typos is consistent with a population typo rate of 1% or less. Observing 18 or more typos means the population typo rate could be greater than 1%. Having determined the required sample size, n, and the final accept number, either of the two following methods can be used to accept or reject the sample data. NIOSH recommends Method B in Step 6 on page 16 and in the example on page 17.

- **Decision Method A - Use of Accept Number –** Section 3.2 (page 10). The number of observed typos (m) in a given sample (consisting of n fields) can be compared to the final accept number (which balances the consumer's and producer's risks, as illustrated above). If the observed number of typos is equal to or less than the final accept number, then the sample data are accepted. If the observed number of typos is greater than the final accept number, then the sample data are rejected as having an unacceptable typo rate.

- **Decision Method B – Use of Confidence Intervals –** Section 3.3. Pages 10–12 of RPRT-0078 provide a method to derive the 95% confidence interval (CI) from the observed number of typos (m) in a sample of size n out of a total population of N fields in the electronic database. The lower bound of the CI is obtained by running the program associated with the function in Equation 3-6. The upper bound of the CI is obtained by running the program associated with the function in Equation 3-7. Using a 95% CI provides for a producer risk of 2.5% and a consumer risk of 2.5%, as demonstrated on page 11.

- **Large Populations –** Section 5.0 (pages 12–14) describes a sampling plan for circumstances where the total number of fields (N) is much greater than the number of fields in the sample (n). In this case, the hypergeometric distribution can be replaced with a binomial distribution, as given in Equation 5-1 on page 12. For a hypergeometric distribution, the sample data are not replaced after they are used. For a binomial distribution, because N is much greater than n, the sample data are replaced after they are used, which simplifies the probability function.

- **Discussion of Issues –** Section 6.0 (pages 14–16) contains some points of discussion concerning the sampling plan.

    – **Size of Database –** It may appear that the number of fields required in a sample would increase in proportion to the number of fields in the original database.

However, as illustrated in Figure 6-1 (page 15), the sample size for the hypergeometric distribution (dotted line) approaches the binomial distribution (solid line) as the number of total fields increase. As the number of fields in the total population becomes infinite, the required number of fields in the sample increases to n = 4,511 for the set of parameters used to construct the curves in Figure 6-1 (i.e., risk = 2.5%, 1% maximum typo rate, etc.).

- **Fields in a Database Are Independent –** The results of one type of field in a database do not influence the results of another type of field in a database. For example, if the analysis from a sample for the critical fields passes the typo test, but the analysis of a sample for all fields fails the typo test, then that database fails.

- **Sampling Frame –** Generally, it is more expedient to select the field values from the electronic database and compare them to the corresponding fields in the hardcopy, instead of the other way around.

- **Example –** Section 8.0 (pages 17–18) provides an example of sampling a database. The following parameters were used in this example:

  - N = 157,336 critical fields, and 314,672 noncritical fields (472,008 fields total)
  - An acceptable error rate (the AQL) of $\gamma = 0.005$ for critical, and 0.025 for all fields
  - An unacceptable error rate (the LTPD) of $\theta = 0.01$ for critical, and 0.05 for all fields
  - A producer's risk of $\alpha = 0.025$
  - A consumer's risk of $\beta = 0.025$

  Running the appropriate program using these parameters results in the recommendation to sample 4,511 critical fields and 874 all fields.

  The critical-field sample data were drawn from the electronic database and compared to the corresponding hardcopy data; a total of 4 typos were found. Using the appropriate program, as illustrated on page 17, the 95% CI was determined to be 2.42E-4 to 2.27E-3, which was below the LTPD level of 0.01. Therefore, the critical-field sample passed the typo test.

  The all-field sample data were drawn from the electronic database and compared to the corresponding hardcopy data; a total of 33 typos were found. Using the appropriate program, as illustrated on page 18, the 95% CI was determined to be 0.0261 to 0.0526, which contains the LTPD level of 0.05. Therefore, the all-field sample failed the typo test. Therefore, the electronic database in its current condition is not useful under these testing parameters.

3.0 SC&A'S EVALUATION OF ORAUT-RPRT-0078

The following is a summary of SC&A's evaluation of the approach, statistical analysis, and documentation used by NIOSH in developing RPRT-0078.

3.1 EVALUATION OF NIOSH'S APPROACH TO SAMPLING PLAN

SC&A did not identify any issues with the general approach used in RPRT-0078 to develop a sampling plan.

3.2 EVALUATION OF NIOSH'S STATISTICAL METHODS

SC&A evaluated the statistical methods employed by NIOSH in RPRT-0078 for developing a sampling plan and found that it is an application of hypothesis testing with alpha (α) and beta (β) to determine whether the percentage of defects (i.e., entries with one or more typos) is within acceptable levels. The use of the binomial approximation for large populations, confidence intervals for the number of defectives, operating characteristic (OC) curves, and the example provide useful insight about how and why the plan works.

3.2.1 Parameters

The assumptions underlying the sampling plan are clearly stated in RPRT-0078 but scattered throughout the text. Therefore, while SC&A concurs with NIOSH's sampling plan, there are several important parameters (some are fixed and some are variables) that could affect the derived values. These parameters and their impacts need to be considered when being used for dose reconstruction purposes. The following is an outline of these items, which apply to both the critical fields and all fields.

- **Fixed parameters:**
  - Total population (N)
  - Total number of typos in population (M)
- **Variable parameters:** Each of these values will be selected independently for the critical fields and all fields by the user. Because the values selected may affect the results of the analyses, they are listed here as variables:
  - Variable 1: Producer's risk α (NIOSH recommends 0.025, i.e., 2.5% for both critical and all fields)
  - Variable 2: Consumer's risk β (NIOSH recommends 0.025, i.e., 2.5% for both critical and all fields)
  - Variable 3: Acceptable error rate (AQL) γ (NIOSH recommends 0.005 [i.e., 0.5%] for critical fields, and 0.025 [i.e., 2.5%] for all fields)
  - Variable 4: Unacceptable error rate (LTPD) θ (NIOSH recommends 0.01 [i.e., 1%] for critical fields, and 0.05 [i.e., 5%] for all fields)

- **Derived or observed values:** Each of these values will be determined independently for the critical fields and all fields:
  - The value of the number of fields to be sampled (n) under a given set of fixed and variable parameters
  - The value of the accept number (i.e., the number of typos in a sample of n fields) that balances the producer's and consumer's risk
  - The number of typos observed (m) in a sample of n fields
  - The OC curve
  - The CI

3.2.2 Changes in Parameters

Consumer's risk β: Increasing the consumer's risk β (in the examples in RPRT-0078, $\beta = 0.025$) increases the probability that a given sample would be accepted, but it also increases the risk of accepting a sample where the typo rate in the population exceeds the desired upper limit of typos (in this case, 1% for critical fields and 5% for all fields).

Producer's risk α: Likewise, decreasing the producer's risk α (in the examples in RPRT-0078, $\alpha = 0.025$) increases the probability that a given sample would be accepted, but it also increases the risk of accepting a sample where the typo rate in the population exceeds the desired upper limit of typos (in this case, 1% for critical fields and 5% for all fields).

Acceptable error rate (AQL) γ: The probability of accepting only a good sample is increased as the value of AQL (γ) is decreased (in the examples in RPRT-0078, the value of γ was 0.005 for critical fields and 0.025 for all fields). However, decreasing the AQL value also increases the probability of discarding a good sample.

Unacceptable error rate (LTPD) θ: The probability of accepting a bad sample is increased as the value of LTPD (θ) is increased (in the examples in RPRT-0078, the value of θ was 0.01 for critical fields and 0.05 for all fields). However, increasing the LTPD value also increases the probability of accepting a good sample.

3.3 EVALUATION OF DOCUMENTATION IN ORAUT-RPRT-0078

SC&A's review of RPRT-0078 did not identify any notable documentation or clarification issues.

4.0 SUMMARY AND CONCLUSIONS

SC&A found the approach used to develop a sampling plan to be reasonable and technically correct.

SC&A found the statistical methods used in the sampling plan to be acceptable. In Section 3.2, SC&A provided some expanded discussion concerning the effects that changes in variable parameters could have on the results.

SC&A did not identify any documentation issues that would affect the readability or application of the sampling plan.

5.0 REFERENCE

NIOSH 2016. *Technical Basis for Sampling Plan*, ORAUT-RPRT-0078, Revision 00, National Institute for Occupational Safety and Health, Cincinnati, Ohio. June 17, 2016.