Assessment of Alternatives to Correct Inventory Difference Statistical Treatment Deficiencies

Prepared by K. R. Byers, J. W. Johnston, C. A. Bennett, R. J. Brouns, M. F. Mullen, F. P. Roberts

Pacific Northwest Laboratory
Operated by
Battelle Memorial Institute

Prepared for
U.S. Nuclear Regulatory Commission
Notice

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, or any of their employees, makes any warranty, expressed or implied, or assumes any legal liability of responsibility for any third party's use, or the results of such use, of any information, apparatus, product or process disclosed in this report, or represents that its use by such third party would not infringe privately owned rights.

Availability of Reference Materials Cited in NRC Publications

Most documents cited in NRC publications will be available from one of the following sources:

1. The NRC Public Document Room, 1717 H Street, N.W.
   Washington, DC 20555
2. The NRC/GPO Sales Program, U.S. Nuclear Regulatory Commission,
   Washington, DC 20555
3. The National Technical Information Service, Springfield, VA 22161

Although the listing that follows represents the majority of documents cited in NRC publications, it is not intended to be exhaustive.

Referenced documents available for inspection and copying for a fee from the NRC Public Document Room include NRC correspondence and internal NRC memoranda; NRC Office of Inspection and Enforcement bulletins, circulars, information notices, inspection and investigation notices; Licensee Event Reports; vendor reports and correspondence; Commission papers; and applicant and licensee documents and correspondence.

The following documents in the NUREG series are available for purchase from the NRC/GPO Sales Program: formal NRC staff and contractor reports, NRC-sponsored conference proceedings, and NRC booklets and brochures. Also available are Regulatory Guides, NRC regulations in the Code of Federal Regulations, and Nuclear Regulatory Commission issuances.

Documents available from the National Technical Information Service include NUREG series reports and technical reports prepared by other federal agencies and reports prepared by the Atomic Energy Commission, forerunner agency to the Nuclear Regulatory Commission.

Documents available from public and special technical libraries include all open literature items, such as books, journal and periodical articles, and transactions, Federal Register notices, federal and state legislation, and congressional reports can usually be obtained from these libraries.

Documents such as theses, dissertations, foreign reports and translations, and non-NRC conference proceedings are available for purchase from the organization sponsoring the publication cited.

Single copies of NRC draft reports are available free upon written request to the Division of Technical Information and Document Control, U.S. Nuclear Regulatory Commission, Washington, DC 20555.

Copies of industry codes and standards used in a substantive manner in the NRC regulatory process are maintained at the NRC Library, 7920 Norfolk Avenue, Bethesda, Maryland, and are available there for reference use by the public. Codes and standards are usually copyrighted and may be purchased from the originating organization or, if they are American National Standards, from the American National Standards Institute, 1430 Broadway, New York, NY 10018.
Assessment of Alternatives to Correct Inventory Difference Statistical Treatment Deficiencies

Manuscript Completed: September 1983
Date Published: November 1983

Prepared by
K. R. Byers, J. W. Johnston, C. A. Bennett,
R. J. Brauns, M. F. Mullen, F. P. Roberts

Pacific Northwest Laboratory
Richland, WA 99352

Prepared for
Division of Safeguards
Office of Nuclear Material Safety and Safeguards
U.S. Nuclear Regulatory Commission
Washington, D.C. 20555
NRC FIN B2209
ABSTRACT

This document presents an analysis of alternatives to correct deficiencies in the statistical treatment of inventory differences in the NRC guidance documents and licensee practice. Pacific Northwest Laboratory's objective for this study was to assess alternatives developed by the NRC and a panel of safeguards statistical experts. Criteria were developed for the evaluation and the assessment was made considering the criteria. The results of this assessment are PNL recommendations, which are intended to provide NRC decision makers with a logical and statistically sound basis for correcting the deficiencies.
ACKNOWLEDGMENTS

The authors wish to acknowledge their indebtedness and gratitude to the members of the panel, who assisted in the development of the alternatives assessed in this study, provided an initial assessment of them, and also served as peer reviewers of the report:

Dr. Robert Easterling
Ms. Yvonne Ferris
Mr. John Jaech
Dr. Ralph Lumb
Dr. Richard Mensing
Dr. Richard Post
Dr. Jim Shipley
Dr. Fred Tingey.
SUMMARY

In November of 1980, a paper was prepared by several Nuclear Regulatory Commission (NRC) staff members on the validity of the statistical basis for treating inventory differences (ID) in licensee facilities (SECY-80-514, NRC 1980). The paper described 13 statistical deficiencies in current NRC guidance and in licensee practices in the areas of ID evaluation and modeling, determination of ID variability, and statistical terminology. The paper also described possible changes that could be made to remedy the deficiencies. The purpose of this study is to develop additional alternatives and perform an assessment of the safeguards statistical implications of the alternatives and the costs to licensees in the event of their implementation. This assessment was performed for eleven of the deficiencies. The other two are being studied by other NRC projects.

The assessment was performed in the following steps:

- A 10-member panel of experts in the development and application of statistical methods to the problems of nuclear material safeguards was convened to help develop the alternatives for correcting each statistical deficiency. Two statistical experts in safeguards from Pacific Northwest Laboratory (PNL) were members of the panel.

- A report of the results of the panel discussion (Appendix A) was prepared by PNL staff members.

- An assessment of the proposed alternatives for each deficiency was made by each panel member with respect to five criteria related to the objectives of nuclear safeguards.

- The panel assessment, combined with the notes from the panel discussions, was used to assist the PNL staff in assessing the alternatives.

- Finally, implementation and operating costs to a typical licensee for each alternative were estimated by PNL staff members.

Remediing the statistical deficiencies identified in the NRC staff paper will require coordinated actions on three fronts. PNL's conclusions and recommendations are structured in terms of these three fundamental areas for improvement, listed below in order of importance:

1. Proper modeling and analysis of inventory differences

2. Effective implementation of statistically sound methods

3. Selection and use of appropriate terminology.
These areas for improvement are closely related. Steps taken to improve modeling and analysis of inventory differences have immediate implications for the area of effective implementation. Similarly, the selection of appropriate terminology is influenced by the approach taken in the modeling and analysis.

**PROPER MODELING AND ANALYSIS**

The proper approach for statistical modeling and analysis of inventory differences depends on the role that inventory differences are expected to play. This role has evolved over the years, from the early emphasis on accounting for valuable materials and control of plant processes and operations to a more recent emphasis on detection of loss or diversion or assurance after the fact that diversion or loss has not occurred. The role of inventory differences continues to evolve.

Any inventory difference can be explained in terms of three possibilities:

1. A loss or diversion has occurred.
2. The inventory difference is a result of chance, i.e., random variability of the data used in calculating the inventory difference.
3. Other causes - for example, unmeasured process holdup, bookkeeping errors, and unrecognized biases.

When inventory differences are to be used as indicators of loss or diversion, it is necessary first to rule out, in some way, the possibility that the inventory difference was due to "other causes." Then only the first two possibilities remain, and if the inventory difference is too large to be explained by chance alone, the only remaining possibility is that a loss or diversion has occurred. (Diversion is an intentional removal of the material for unlawful uses. It includes theft. A loss is unintentional. Either lost or diverted material may still be within the plant.)

There are several complementary ways of dealing with the possibility that other causes may account for the inventory difference. One approach is to control directly the occurrence of such causes. It may be possible to accomplish this through formal control programs, such as measurement control, process control, and administrative control. A second approach is to investigate the possible causes of an ID after the fact and determine what effect each might have had on the ID. A third possibility is to incorporate the "other causes" explicitly into the statistical modeling and analysis on a routine basis. If all other causes can be properly accounted for in the statistical modeling and analysis, then all inventory differences will be explainable in terms of either 1) loss or diversion or 2) chance.

Several of the most important statistical deficiencies identified in the NRC staff paper are concerned with this issue of statistical modeling and analysis. These deficiencies are the following:
Deficiency C3. The regulations and guides do not explicitly state the error terms in the ID model and the model does not label random and fixed errors.

Deficiency B3. Not all sources of error are reflected in the calculation for LEID, nor are they otherwise substantially addressed in any guidelines.

Deficiency C4. As defined, ID may reflect impacts of other than current-period operations. In particular, the regulations do not require the adjustment of ID to eliminate prior-period contributions caused by shipper-receiver differences, waste and scrap operations, and possibly other operational factors.

Deficiency A1. The procedure for setting the current ID action limits reflects some nonstatistical considerations and is not based on the framework of the statistical test of hypotheses. As a result, this procedure does not control the false alarm rate nor the probability of alarming and initiating an investigation when a specified amount of material is lost or diverted. In addition, the current procedure does not address evaluation of multiple-period inventory differences.

To remedy these deficiencies, PNL makes the following recommendations:

- All potentially significant sources of error in the estimator of ID should be accounted for in the conceptual model; i.e., \( E(ID) = 0 \). This implies that terms for unmeasured inventory (e.g., process holdup), nonmeasurement errors (e.g., bookkeeping errors), and all possible sources of bias should be included in the model. Prior-period adjustments must also be handled correctly. (Procedures for making prior-period adjustments are discussed in Section 3.2.2.2). However, at the measurement system level of modeling where one specifies by statistical models the way the error terms arise, those terms for which no evaluation method is available at the time (e.g., certain nonmeasurement errors) are omitted from the variance model. Then the \( E(ID) = D \neq 0 \), where \( D \) is some constant error that does not affect \([SD(ID)]\). It is then recognized that those sources of error must be accounted for in any subsequent investigation of an excessive ID before one can conclude that the ID was due to an actual loss of material.

- An attempt should be made to estimate and correct for all potentially significant biases. To the extent that this is accomplished, the estimator of ID will be free of "fixed" errors, or biases. If all potential fixed errors are estimated without bias, any observed inventory difference will then be explainable in terms of either 1) loss or diversion, or 2) random variability.
A spectrum of tests and accompanying actions should be used in evaluating inventory differences. The allowable minimum power of such tests should be specified by balancing the power against the licensee's risk of a false alarm. Evaluation procedures should explicitly recognize that even with improved modeling, causes other than loss or diversion and random variability will sometimes be responsible for inventory differences. One of the functions of material accounting is to monitor the state of statistical control over these causes and to identify out-of-control situations. This viewpoint suggests the need for a variety of statistical tests to be used both for monitoring the system and for investigating the reasons for any loss of control that occurs. Examples of tests that could be performed as part of the evaluation procedures include tests of measurement control (e.g., tests aimed at detecting bias or a change in measurement error variance), tests of cumulative inventory differences, tests of accounting procedures to ensure that clerical errors are kept under control, and tests of material control. Such a broad approach to evaluation is more complex than a simple test based on estimated ID, but it is better suited to the inherent complexity of the system. The "spectrum of tests" recommended for the remedy of this deficiency is a licensee responsibility, and whether or not it should be a regulatory requirement is a matter for policy decision by the NRC (see Deficiency A2).

It should be noted that the simple model

\[ ID = \text{Beginning Inventory} + \text{Receipts} - \text{Removals} - \text{Ending Inventory} \]

masks much complexity. Each term may be a sum of quantities for many different material types. Each material type may have its own set of measurement systems for determining mass, element, and isotopic composition. The quantity of special nuclear material (SNM) in an item or batch is usually the result of weighing, sampling and analytical determination. Each of these measurement operations is subject to random errors and potential biases. It would take pages of algebraic formulas to specify the sums of products (mass times element or isotopic concentration) that are required to define ID in terms of the measurements involved. And this would just get us down to the level where the measurements could be statistically modeled, i.e., where the errors associated with the measurements can be specified. The major task of combining these errors to define an estimator for \([SD(ID)]\) requires carrying these errors back up the levels of modeling from item quantities to material types to ID components and finally ID itself. To further complicate the problem, the estimate, \(sd(ID)\), usually depends on estimates of the measurement error standard deviations and biases based on specially collected measurement control data, not the actual measurements used to determine the quantities entered on the inventory documents used to calculate ID. It was necessary to keep in mind that the modeling level under discussion may change; whether we are working down the ID model or up, the \(SD(ID)\) has an impact. As we progress down from ID to measurement data we may refer to a potential constant bias. But in working back up
with the measurement errors, that same bias will have been estimated and so will have a standard deviation associated with it, called the standard error of the bias estimate.

EFFECTIVE IMPLEMENTATION

Once statistically sound models and analysis methods have been developed, effective implementation of them must be ensured. Two of the statistical deficiencies identified in the NRC staff paper concern implementation, specifically:

Deficiency A2. Current alarm criteria are based on nonstatistical considerations and are set independently of the variability of ID.

Deficiency B5. Licensee ID variability models potentially contain weaknesses with respect to model structure and parameter estimation.

These deficiencies are closely tied to deficiencies of modeling and analysis. Satisfactory resolution of the modeling and analysis issues addresses a part of deficiencies A2 and B5. To more fully remedy these deficiencies, PNL further recommends the following actions:

• In considering remedies for Deficiency A2, it should be recognized that the establishment of alarm criteria is not a purely statistical question, although it has many statistical aspects. Policy judgments are necessary. PNL therefore recommends the development and implementation, within an appropriate statistical framework, of administrative limits. Administrative limits could be controls based on other than statistical considerations; e.g., controls on target amounts of loss that must be detected with high probability, and acceptable control limits for the variance of ID. The preferred statistical approach is to use hypothesis tests to determine whether the administrative limits are being achieved and whether the estimated ID and its standard error are in statistical control. The statistical approach for testing IDs is a change from the current approach, but monitoring of the standard error of ID is already required by regulations, 10 CFR 70.58(f).

• A procedure for independent external review of facility-specific ID variability models should be developed and implemented. This will help to ensure that the weaknesses referred to in Deficiency B5 will be identified and eliminated. The external review should be performed by a competent statistician, either from the NRC staff or an outside contractor.
APPROPRIATE TERMINOLOGY

Having modeled ID with an expected value equal to zero, some definitions of terms are more logical than others. In addition, the use of conventional statistical terminology will facilitate understanding.

Five of the statistical deficiencies identified in the NRC staff paper (SECY-80-514) are concerned with the use of inappropriate or imprecise statistical terminology. In some cases such terminology has led to misinterpretations or confusion in safeguards applications. These terminology deficiencies are the following:

Deficiency C1: Inventory difference is referenced in official regulations as MUF. Lack of uniformity in terminology is a potential source of misunderstanding.

Deficiency C2: Regulations and guides fail to distinguish between true ID and observed ID. This distinction is essential for complete understanding and modeling, and for rigorous statistical treatment of ID.

Deficiency B2: The regulations and guides fail to distinguish between the true value of a measure of variability, the formula for computing it (the estimator) and the numerical value (the estimate) obtained by substituting observed data into the estimator.

Deficiency B4: Concepts and definitions of error which are in wide use in the nuclear industry are imprecise and are not always consistent with definitions in other disciplines; this could be a source of confusion and could lead to inaccurate evaluation of ID variability.

Deficiency B1: The term "limit of error" has several inconsistent definitions.

The remedies for these deficiencies will probably lead to long-term benefits by promoting a better understanding of the concepts involved and smoother communications among those concerned with the use of statistics in safeguards.

To remedy these deficiencies, PNL makes the following recommendations:

- The term "MUF" (material unaccounted for) should be replaced by "ID" in the official documents (regulations and regulatory guides) of the NRC when the next document revision occurs. The term "ID" has already been adopted by DOE, NRC, and the nuclear industry in correspondence and publications. Furthermore, the term "ID" may be more descriptive: MUF is more suggestive of a loss or diversion, whereas ID is more consistent with the idea discussed previously under the heading of proper modeling and analysis, i.e., that an inventory may fail to balance exactly for any number of reasons.
(e.g., random variability, loss or diversion, or such other causes as unmeasured process holdups, unrecognized biases, or bookkeeping errors). However, the principal benefit resulting from this change is improved consistency of terminology.

• Understanding and proper usage of the terms "true value," "estimator," "estimate," "expected value," and "observed value" should be promoted by the NRC. This is probably most effectively accomplished through the publication of a regulatory guide or the revision of Regulatory Guide 5.3. The latter should be updated and expanded to include discussions of these statistical concepts, as well as describe recommended notation that would help reinforce the meanings of the concepts. In addition, the NRC should provide guidance on statistical concepts and notation to NRC personnel responsible for standards development, license reviews and evaluation, and inspection and enforcement in the area of material accounting.

Briefly, the interpretations of the terms in the previous paragraph are as follows: A "true value" is "the actual quantity of the property being measured in the object being measured" (Eisenhart 1969). The "true value" is a concept necessary to give us something to attain. Measurements are made in the hope of coming as close as possible to this "true value." The value of the statistical approach is that a probability can be assigned to the agreement of the estimate with a "true value" that can never be known with absolute certitude. Statistical "estimators" (functions) are used to provide "estimates" (numerical values) for population parameters. An "expected value" can be interpreted as the limiting mean of an estimator, when the number of observations used in calculating the estimator becomes very large. The expected value of a parameter for a statistical population, which is generated by theoretically applying a measurement system to an item, is not the conceptual "true value." All measurement systems have some bias, and measurement control procedures are used to estimate the bias or to provide a quantitative assessment that the bias is negligible. (These terms are defined and discussed in more detail in Section 3.2.1.2. Suitable notation is described in Section 4.0).

• The term "systematic error," used in safeguards applications, should be eliminated, and the terms "bias" and "random effect" should be adopted. The term "random effect" should be used whenever the error in question is regarded as a random sample drawn from a population of errors with mean zero and finite variance; random effects give rise to variance components. The term "bias" should be used whenever the error in question is regarded as fixed or constant; fixed effects give rise to estimated differences, and estimated differences also give rise to variance components.

Whenever the term "random effect" is used, the source of the effect should be clearly identified. For example, one should specify "the random effect due to difference between analytical laboratories," or "the random effect due to fluctuations over time in the bias of scale number three," or "the random effect due to differences between individual analysts."
The NRC should promote this change in terminology through an appropriate publication and, subsequently, consistent usage in NRC documents. An appropriate opportunity could be the publication of the planned reference manual on statistical methods for nuclear material accountability which is currently being prepared with NRC sponsorship. An alternative that could be more timely is publication of a technical report specifically devoted to terminology for modeling ID or ID variability.

- The terminology to describe the measure of ID variability, "limit of error," should be eliminated and "estimated standard error of the estimate of ID" should follow as a logical consequence of recognizing the functional dependence of the random variables involved. If a confidence limit is needed in a particular evaluative problem, the words "confidence limits," specifying the confidence level used and the usual caveats about interpreting confidence limits would be used. Consequently, Deficiency 81 would be remedied by discontinuing the use of the imprecise and misleading term "limit of error." An alternative terminology is "estimated standard deviation of the estimated ID," which has the same meaning as "estimated standard error of the estimated ID." The "standard error" terminology is more appropriate in reference to an ID because in statistical usage it normally refers to a function of random variables such as a mean value. "Standard deviation" is the usual and proper term for the variability of individual measurement results. Neither term should be used without identifying specifically, or in the context, what estimator is being referenced.

The NRC should promote this change in terminology for the measure of the variability of ID and other estimators by appropriate changes in regulations and regulatory guides at the next opportunity for revisions of the documents that now define or refer to "limit of error."

IMPACT ASSESSMENT

Choosing the PNL-recommended alternatives for the eleven deficiencies is estimated to result in a $51K implementation cost and an annual operating cost of $18K for each licensee. Resolution of the ID modeling deficiency, C3, accounts for more than half of the total implementation costs, since a facility-specific system analysis will be required that will include adapting statistical techniques to incorporate sources of ID variability not already included in the statistical model. Resolution of Deficiency B5, which is concerned with the evaluation of ID variability models used by licensees, is the other major implementation cost contributor. Working with the assumption that the licensee will bear the total cost of evaluating their ID model structure and use, PNL estimates that third-party evaluation will cost each licensee approximately $15K.
The additional annual operating cost is considered by PNL to be a small expense necessary to collect additional data and perform additional calculations and data evaluation associated with new components of the ID variability model. The impact assessment and the cost relationships among alternatives are summarized in Table 1.

### TABLE 1. Impact Assessment of PNL-Recommended Resolutions for the Statistical Treatment Deficiencies

<table>
<thead>
<tr>
<th>Deficiency</th>
<th>Implementation Cost ($K)</th>
<th>Additional Annual Operating Cost ($K)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Modeling and Analysis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>B3</td>
<td>*</td>
<td>7</td>
</tr>
<tr>
<td>C4</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>A1</td>
<td>*</td>
<td>6</td>
</tr>
<tr>
<td><strong>Effective Implementation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>*</td>
<td>0</td>
</tr>
<tr>
<td>B5</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td><strong>Appropriate Terminology</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>B4</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Cost included in the resolution of Deficiency C3.*
# CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>v</td>
</tr>
<tr>
<td>SUMMARY</td>
<td>vii</td>
</tr>
<tr>
<td>1.0 INTRODUCTION</td>
<td>vi</td>
</tr>
<tr>
<td>2.0 TECHNICAL APPROACH</td>
<td>v</td>
</tr>
<tr>
<td>2.1 DEFICIENCIES AND ALTERNATIVES</td>
<td>v</td>
</tr>
<tr>
<td>2.1.1 ID Evaluation</td>
<td>v</td>
</tr>
<tr>
<td>2.1.2 ID Variability</td>
<td>v</td>
</tr>
<tr>
<td>2.1.3 ID Definition and Modeling</td>
<td>v</td>
</tr>
<tr>
<td>2.2 DECISION ANALYSIS METHOD</td>
<td>v</td>
</tr>
<tr>
<td>2.3 COST ESTIMATION</td>
<td>v</td>
</tr>
<tr>
<td>3.0 ASSESSMENT</td>
<td>v</td>
</tr>
<tr>
<td>3.1 LOGICAL DEPENDENCE OF DEFICIENCIES</td>
<td>v</td>
</tr>
<tr>
<td>3.2 ASSESSMENT OF ALTERNATIVES</td>
<td>v</td>
</tr>
<tr>
<td>3.2.1 Definitional Deficiencies</td>
<td>v</td>
</tr>
<tr>
<td>3.2.1.1 Deficiency C1</td>
<td>v</td>
</tr>
<tr>
<td>3.2.1.2 Deficiencies C2 and B2</td>
<td>v</td>
</tr>
<tr>
<td>3.2.1.3 Deficiency B4</td>
<td>v</td>
</tr>
<tr>
<td>3.2.1.4 Deficiency B1</td>
<td>v</td>
</tr>
<tr>
<td>3.2.2 Modeling and Analysis Deficiencies</td>
<td>v</td>
</tr>
<tr>
<td>3.2.2.1 Deficiency C3</td>
<td>v</td>
</tr>
<tr>
<td>3.2.2.2 Deficiencies B3 and C4</td>
<td>v</td>
</tr>
<tr>
<td>3.2.2.3 Deficiency A1</td>
<td>v</td>
</tr>
</tbody>
</table>
3.2.3 Implementation Deficiencies (A2, B5)..........................3.29
  3.2.3.1 Deficiency A2........................................3.29
  3.2.3.2 Deficiency B5........................................3.30
4.0 RECOMMENDED NOTATION........................................4.1
5.0 REFERENCES.....................................................5.1
APPENDIX - ALTERNATIVES TO CORRECT INVENTORY DIFFERENCE STATISTICAL TREATMENT DEFICIENCIES - A PANEL DISCUSSION.............A.1
FIGURES

3.1 Logical Dependence of Deficiencies........................................3.3

TABLES

1 Impact Assessment of PNL-Recommended Resolutions for the Statistical Treatment Deficiencies...........................................xv
3.1 Definitions Relative to Deficiencies B2 and C2.........................3.6
3.2 Alternative Terminology for Error Concepts (Deficiency B4)..........3.12
3.3 Definitions Relative to Deficiency B4..................................3.14
3.4 Emphasis Dichotomies for Alternatives to Deficiency C3..............3.19
3.5 Hypothesis-Testing Decision Table........................................3.25
4.1 Notation for Some Statistical Concepts....................................4.2
1.0 INTRODUCTION

In March 1980, the Nuclear Regulatory Commission (NRC) requested that the Office of Management and Program Analysis and the Office of Nuclear Material Safety and Safeguards (NMSS) jointly prepare a paper identifying statistical deficiencies in the current treatment of inventory differences (ID), with the aim of placing this treatment on a valid statistical basis with respect to terminology, theoretical principles, and operating procedures. The paper, "Report on the Statistical Treatment of Inventory Differences," SECY-80-514 (NRC 1980), described 13 deficiencies in the areas of ID evaluation, determination of ID variability, and ID definitions and modeling. These deficiencies involve the ability to assure a high probability of alarming and of initiating an investigation when a specified significant amount of material is lost or diverted, as well as the problem of excessive alarms associated with ID evaluation. Although SECY-80-514 identifies possible changes that could be made, these modifications have not been evaluated for the benefits and costs associated with eliminating the deficiencies. In May 1982, NMSS requested the services of Pacific Northwest Laboratory (PNL) to identify and evaluate alternatives to correct these deficiencies.

On July 8 and 9, 1982, a workshop meeting was held at the Battelle Human Affairs Research Center facilities in Seattle, Washington. This workshop panel consisted of a group of people expert in the development and application of statistical methods to the problems of nuclear material safeguards. A copy of the report summarizing the meeting results is provided in Appendix A.

The purpose of the present work is to perform an assessment of the alternatives that were developed with the aid of the panel of experts for each deficiency, in an effort to identify safeguards value and cost to licensees. The technical approach is described in Chapter 2, with the deficiencies and alternatives to be considered described in Section 2.1. The decision analysis method is described in Section 2.2, which discusses the value attributes used in the analysis and the panel's assessment of their relative importance. Cost assumptions are discussed in Section 2.3.

Chapter 3 is a presentation of the assessment summarized by deficiency type. Each deficiency, the alternatives, the logic underlying the alternatives, and the assessment are discussed in a format to assist the NRC decision makers in selecting alternatives. Recommended notation for statistical concepts is presented in Chapter 4.

(a) PNL is operated for the U.S. Department of Energy by Battelle Memorial Institute.
2.0 TECHNICAL APPROACH

This chapter discusses the technical approach used to assess alternatives to correct each ID statistical treatment deficiency. Section 2.1 identifies and discusses the deficiencies and alternatives. The decision analysis method used for this assessment is described in Section 2.2, and cost estimation assumptions are provided in Section 2.3.

2.1 DEFICIENCIES AND ALTERNATIVES

Eleven of the thirteen ID statistical treatment deficiencies were evaluated by this study. The other two deficiencies are currently being studied by other NRC projects and are not included in this report. The deficiencies were classified in the NRC staff paper (SECY-80-514, NRC 1980) into three groups: ID evaluation, ID variability, and ID definition and modeling.

The alternatives for each deficiency presented are a combination of NRC alternatives and alternatives developed by the panel of statisticians. The NRC alternatives are indicated by the staff paper reference (SECY-80-514).

2.1.1 ID Evaluation

Deficiency A1

The procedure for setting the current ID action limits reflects some non-statistical considerations and is not based on the framework of the statistical test of hypotheses. As a result, this procedure does not control the false-alarm rate nor the probability of alarming and initiating an investigation when a specified amount of material is lost or diverted. In addition, the current procedure does not address evaluation of multiple-period inventory differences.

Alternatives

1a. For ID evaluation, adopt the framework and terminology of the statistical test of hypotheses (SECY-80-514).

1b. Shift emphasis in the regulations and guides pertaining to statistical tests from false-alarm rate to the probability of alarming and initiating an investigation when a specified amount of material is lost or diverted (SECY-80-514).

2. Use statistical decision theory.

3. Develop a spectrum of tests and accompanying actions that correspond to evaluations of measurement error, measurement error plus process variability, and ID.

4. Control the Type 2 error rather than the Type 1 error.

5. Control both types of error according to a defined tradeoff strategy.
Alternatives 1 (a and b), 4 and 5 emphasize hypothesis testing while Alternatives 2 and 3 allow for statistical control aspects. If Alternative 1a, which would adopt the framework of hypothesis testing for ID, is accepted, 1b follows logically from a statistical standpoint. No hypothesis testing should be applied without considering the probability of detecting a specified loss or diversion of material.

Statistical decision theory (Alternative 2) is a general approach which allows the use of costs of data collection and monetary or other quantifiable risks in designing optimal (cost-wise or risk-benefit) decision rules. Alternative 3 recognizes that setting ID action limits is necessary but not sufficient for a statistically sound application of the accounting data. It suggests the use of hypothesis testing for indices of statistical control as well as ID.

Alternative 4 is essentially the same as Alternative 1b but with more emphasis on the power of the test, 1-β, without regard to the type 1 error rate. Alternative 5 is in the mid-position, suggesting a tradeoff strategy (perhaps developed between the licensee and NRC) for setting α and β levels that will be acceptable.

Deficiency A2

Current alarm criteria are based on nonstatistical considerations and are set independently of the variability of ID.

Alternatives

1. Apply statistical principles to the construction of the test of hypotheses. In particular, establish alarm thresholds (or ID action limits) based on ID variability (SECY-80-514).

2. Develop and apply, within an appropriate statistical framework, administrative limits.

Currently, three action limits are specified by NRC for ID. The second and third action limits are based on specific percentages of the process throughput of the facility during the inventory period. Thus, these action limits are independent of the variability of ID, and neither the false-alarm rate nor the power of the test of ID are controlled. On the other hand, the first action limit for ID, which is essentially just a warning level, is based on the limit of error of ID (LEID), and LEID is controlled by the requirement that it not exceed 0.5% of throughput.

2.1.2 ID Variability

Deficiency B1

The term "limit of error" [used widely by NRC and the nuclear industry] has several inconsistent definitions.
Alternatives

1. Adopt the standard deviation as the measure of ID variability. Eliminate the use of the term "limit of error" (SECY-80-514).

2. Use standard error of the estimate as a measure of ID variability. Eliminate the term "limit of error." Find different words for instances where "limit of error" is used in the context of evaluation.

3. Use estimated mean square error as a measure of ID variability. Eliminate the term "limit of error." Find different words for the instances where "limit of error" is used in the context of evaluation.

4. Use standard error of the estimate or estimated mean square error as a measure of ID variability. Eliminate the use of the term "limit of error" as indicating a measure of ID variability, but retain the term where it is a defined interval for a specific evaluation problem.

The term "limit of error" does not have a precise or consistent meaning in the statistical literature, but it is given a specific meaning in safeguards. However, the definitions in the ANSI standard (ANSI N15.16-1974) and the Code of Federal Regulations [10 CFR 70.51(a)] are different. Many safeguards personnel in the nuclear industry favor the ANSI definition. It is usually impractical to calculate the limit of error of ID in the way intended by the Code because the degrees of freedom of the ID estimate are not known. Many statisticians believe that the use of a special term, "limit of error," for the uncertainty component at the 95 percent confidence interval of an estimate is unnecessary and can be misunderstood, and consequently, that it should be discontinued.

Deficiency B2

The regulations and guides fail to distinguish between the true value of a measure of variability, the formula for computing it (the estimator) and the numerical value obtained by substituting observed data into the estimator (the estimate).

Alternatives

1. Modify the regulations and guides to define and distinguish among the true standard deviation, its estimator, and its estimate (SECY-80-514).

This deficiency addresses the same concepts that were addressed by Deficiency C2. The panel did not suggest any alternatives to the recommendation given in the NRC staff paper.
Deficiency B3

Not all sources of error are reflected in the calculation for LEID, nor otherwise substantially addressed in any guidelines.

Alternatives

1. Improve guidelines to better model ID variability and to identify and characterize, to the extent possible, on a facility-specific basis, all contributors to the variability of inventory differences, including non-measurement errors (SECY-80-514).

2. The variance calculation that accompanies the modeling of ID such that it has expectation zero resolves B3. Specifically require the ID model to have a term for "process" variability where this term is carefully defined.

Deficiency B4

Concepts and definitions of error which are in wide use in the nuclear industry may be overly restrictive and are not always consistent with definitions in other disciplines; this could be a source of confusion and could lead to inaccurate evaluation of ID variability.

Alternatives

1. In the development of future guidance for licensees, eliminate concepts and definitions of error that are overly restrictive and confusing. In particular, investigate various approaches to the treatment of bias and its impact on ID variability and clarify the notions of random and systematic error (SECY-80-514).

2. Use the terms "random effects", "residual effects" and "bias" in place of current terminology. Clearly specify sources of error.

3. Replace the terms "random" and "systematic" with "uncorrelated" and "correlated" errors and "bias."

4. Do not change the terminology, but make sure that current terms are understood.

Deficiency B5

Licensee ID variability models potentially contain weaknesses with respect to model structure and parameter estimation.

(a) The authors of the original paper on the statistical deficiencies suggested changing this sentence to the following (see Appendix 4, SECY-80-514): "In the development of future guidance for licensees, eliminate concepts and definitions of error that are confusing or that do not conform to standard statistical practice."
Alternatives

1a. Assure that the model assumptions are appropriately tested and that sensitivity analyses to assess the consequences of wrong assumptions in the modeling of the variability of random errors are performed (SECY-80-514).

1b. Provide guidelines to control the magnitude and effect of modeling errors, including those stemming from nonmeasurement contributors (SECY-80-514).

2. Utilize third-party opinion to evaluate facility-specific variability models and their mode of application.

2.1.3 ID Definition and Modeling

Deficiency C1

Inventory difference [currently used by NRC and the nuclear industry] is referenced in official regulations as MUF. Lack of uniformity in terminology is a potential source of misunderstanding.

Alternatives

1. Replace the term "MUF" with "ID" in all regulations and official documents (SECY-80-514).

2. Replace "ID" with "MUF" to be consistent with world-wide usage.

3. Do nothing, but explain terms carefully.

This deficiency refers to the use of the term "ID" by the licensees and the NRC in recent and current literature, correspondence and reports, while the term defined and used in the Code of Federal Regulations (10 CFR 70) and the regulatory guides is "material unaccounted for" (MUF). This lack of uniformity in terminology is a potential source of misunderstanding.

Deficiency C2

Regulations and guides fail to distinguish between true ID and observed ID. This distinction is essential for complete understanding and modeling, and for rigorous statistical treatment of ID.

Alternatives

The following alternatives for alleviating the deficiency include the one recommended in the NRC staff paper and those suggested by the panel. These are not really alternatives in the sense that one would be chosen over the others; they support the same position in a variety of ways.
1. Modify regulations and guides to be consistent with the ANSI Standard N15.16 definitions of parameter, estimator and estimate for true ID, ID estimator and ID estimate, respectively (SECY-80-514).

2. Communicate the distinctions among "true value," "estimator" and "estimate" from fundamental statistical principles.

3. Reinforce the distinctions among "true value," "estimator" and "estimate" through suitable notation.

4. Develop the distinctions among "true value," "estimator" and "estimate" through education.

The NRC regulations and guides discuss only observed or measured values of ID and the flow and inventory quantities from which ID is calculated. For example, all of the terms on the right side of the equation

\[ ID = BI + A - EI - R \]  [see 10 CFR 70.51(a)(6)]

are functions of observed, or measured values. The regulations and guides do not explicitly consider the "true value" of ID. Because the "true value" can never be known, statisticians usually try to avoid the concept and talk of "population parameters" instead. In safeguards, statistical populations are defined in terms of measurement operations, and the parameters of interest are the mean and variance of the population of potential observations on an item. If the measurement operation were free of bias, then the limiting mean of the estimator of the mean (that is, in most cases, the expected value of the arithmetic averages) is the same as the "true value" sought. In modeling a measurement system, a statistician will use the sum of two terms for the limiting mean, a true value and a bias, allowing for the fact that the true value sought may not be the same as the population mean produced by the measurement system. Validation of the model requires assessing the significance of the bias. The conceptual "true value" is just a concept. The statistical "true value," if the term must be used, is operationally defined by the measurement procedures and methods of estimation used. Thus, the conceptual "true value" of ID differs from the statistical "true value" defined by the statistical model for ID. One necessary distinction is between the conceptual "true value" and the operationally defined population parameter. The conceptual "true value" of ID would be an exact measure of SNM loss. The statistical "true value" of ID is the expected value of the estimator of ID. Since all of the components of estimated ID are measured, or could be measured, and are subject to measurement errors, observed ID is always an estimated value with a variance that is also estimated.

---

(a) ANSI N15.16-1974 does not contain the required definitions, whereas ANSI N15.5-1982 (Draft) does.
Communication of the distinctions among "true value," "population parameter," "estimator" and "estimate" to the statistical layperson is a problem that statisticians in all areas of application have had to address. It is important that in communications the distinctions are consistently reinforced. The distinction must become specific, and one must deal with models and components of ID, since the statistical true value of ID is defined by the estimates (from data) used in the estimator (function) of ID specified by the model for ID. To this extent, deficiencies C2, C3, and C4 are related.

Deficiency C3

The regulations and guides do not explicitly state the error terms in the ID model, and the model does not label random and fixed errors.

Alternatives

1. Improve ID modeling by more explicit consideration of true ID and identification of random and fixed error components, and modify regulations and guides accordingly (SECY-80-514).

2. Model ID so that \( E(\text{ID}) = 0 \), and describe errors in terms of the effects they have on the observations (i.e., errors described in the context of the analysis of variance).

3. Model ID so that \( E(\text{ID}) = 0 \) and the definition of errors (long-term, short-term systematic, random) is relative to a particular data set.

4. Limit the terms in the ID models to those for which measurements are available, and describe errors in terms of the effects they have on the observations (i.e., errors described in the context of the analysis of variance).

5. Limit the terms of the ID models to those for which measurements are available, and so that the definition of errors (long term, short term systematic, random) is relative to a particular data set.

There were two fundamentally different notions discussed by the panel with respect to modeling ID. The first is to require that the estimator of ID be defined such that its expected value is zero (in the absence of any losses due to diversion). The second is to limit the terms in the ID model to those for which direct measurements are available. A requirement that ID be modeled so that it has expectation zero would seem to require the treatment of all true effects, regardless of replication, as random. Thus process deviations, measurement error--in fact, any deviation from the current material balance model--are all considered chance variables. This leads to a simple alternative hypothesis (to the null hypothesis that true ID = 0), but it also leads to the most complex estimation problem. If the components in the ID model are limited to those for which direct measurements are available, then the estimation problem is straightforward; but the alternatives to the null hypothesis are many, only one of which is that a material loss has occurred. Many of the current evaluation problems derive from the attempt to ignore the multiplicity of
alternatives to the null hypothesis that true ID is zero. Whatever route is chosen, it is mandatory that individual components in the ID model be described in terms of the sources of error that affect them and ways in which the effects are to be propagated in computing the error in the estimated ID.

Deficiency C4

As defined, ID may reflect impacts of other than current-period operations. In particular, the regulations do not require the adjustment of ID to eliminate prior-period contributions caused by shipper-receiver differences, waste and scrap operations, and possibly other operational factors.

Alternatives

1. Modify regulations and guides to incorporate statistically acceptable procedures to correct for bias, shipper-receiver differences, and waste and scrap operations (SECY-80-514).

2. From an accounting viewpoint, the important thing is that the books balance; from a statistical viewpoint the important thing is that the accounts reflect what actually happened. The regulations should recognize the distinction. A requirement that ID be estimated so that $E(ID)=0$ resolves this deficiency.

The important point here is that regardless of the overall requirements on modeling ID, i.e., defining true ID, any prior-period adjustments should be done in a way that reflects the actual occurrence. This process may, in fact, result in two sets of books: one that would be closed at the end of an accounting period where accounting variances are handled according to accounting principles, and another that would remain available for making prior-period adjustments so that valid statistical analyses can be performed.

2.2 DECISION ANALYSIS METHOD

The method begins with a clear, concise statement of the objective of the alternatives to be considered. For all deficiencies, the objective was stated as follows:

- Develop a regulatory framework which assures that IDs are treated on a valid statistical basis with respect to terminology, theoretical principles, and operational practices.

For this objective and the alternatives for each deficiency, five assessment attributes were developed. These five attributes were presented to the panel members for evaluation to assist PNL in estimating the safeguards and statistical value of the alternatives:

1. Performance - The degree to which the alternative meets the following goals of SNM accounting:

   a. quality control of SNM accounting effectiveness
b. detection of significant losses or diversions

c. provision of SNM accountability assurance for the public and the NRC.

Consideration should be given to the technical effectiveness in minimizing false-alarm rates, enhancing the power of the tests, and increasing the flexibility of alarm responses. Performance could involve any one, or all three, of these distinct goals of safeguards.

This attribute was considered as the most important assessment attribute by 70 percent of the panel members. When considering any alternative, the quality of its statistical performance is felt to be the most significant.

2. Feasibility - The likelihood that the alternative can be successfully implemented, i.e., put into operation fulfilling its assigned function. Assume state-of-the-art or near-term available technology.

This attribute was considered equally as important as the performance attribute by 40 percent of the panel members. Many others considered it almost as important. If the concept statistically performs well but cannot be implemented as designed, then it has failed.

3. Inspection and Enforcement - The extent to which the alternative may improve SNM control and accounting systems inspection and enforcement efforts.

The panel considered this attribute as having moderate importance. This rating would indicate that if the concept is statistically valid, it has value even if I&E efforts are not eased by its implementation.

4. Acceptance - The degree to which the MC&A staff of all affected licensees are motivated to accept the intent of the alternative. Assume, when making this judgement, that appropriate communication with the licensees regarding the pros and cons of each alternative has been made.

This attribute was considered by the panel to have moderate weight. The weighting would indicate that licensees should have some impact on how the regulatory framework for ID should be developed. Even if the alternative performs well, some degree of acceptance by the licensees is still indicated in order to achieve maximum safeguards effectiveness.

5. Public Awareness - The likelihood that the alternative will communicate the right concept of inventory difference to the public and generate a more favorable perception of MC&A safeguards.
This attribute was considered to be the least important. The panel felt that while public awareness is important, it is not as important as the other attributes. The argument is that the public is not currently aware of the right concept of ID, and any changes statistically would probably not significantly change the awareness. The panel members were allowed to interpret "right concept" according to their individual experiences.

2.3 COST ESTIMATION

The cost estimates in Chapter 3 are approximations derived using as a reference facility a high-enriched fuel fabrication facility for manufacturing high-temperature gas reactor fuel elements. The reference facility is similar to that described by Heaberlin and Byers (1980), and used the following assumptions:

- The facility is a high-enriched uranium fabrication plant for hexagonal-block, bonded-bed, High Temperature Gas-Cooled Reactor (HTGR) fuel elements.
- The design capacity of the facility is 4500 kg U/year and operates two shifts per day, 5 days per week.
- Uranium is enriched to 93% $^{235}$U and is received at the facility as cylinders of UF$_6$.
- Pyrocarbon-coated thorium particles and the graphite fuel blocks are obtained external to the facility.
- The facility is located in the United States and meets NRC licensing requirements using the status quo MC&A capabilities.
- The status quo facility is estimated to employ 358 people of which 30 are considered MC&A staff. There is allowance for overlap of MC&A and process personnel, including all who are assigned to MC&A or perform MC&A activities. The MC&A staff ratio is 2 professional people for each technical person.
- The process steps, equipment, and MC&A methodologies are based on established technology.
- The facility has a collocated scrap recovery operation which processes solid and liquid scrap and recycles the recovered uranium into the head-end of the fabrication process.

The cost estimates, which should be used for cost comparison of the alternatives, are not sensitive to differences in plant size or throughput, but may be influenced somewhat by the kind of facility because of possible differences in the complexities of the nuclear material accounting system. However, it is not anticipated that those differences will affect the ratio of costs among alternatives.
The cost estimates were derived for each alternative independently; in reality, however, efforts expended by the licensee to correct one deficiency may serve, in part or entirely, to correct several. For example, in correcting Deficiency A1, little or no additional work may be needed to correct deficiencies A2, B3 and possibly others. Some of the alternatives may possibly result in substantial savings to the licensee by reducing the frequency of investigations of inventory discrepancies and by aiding those investigations. No attempt was made to include such effects on cost to the licensee, since that was considered to be beyond the scope of the task.

The costs presented are the licensee's estimated costs for implementing the alternatives. They do not include costs that would be incurred by the NRC.
3.0 **ASSESSMENT**

This chapter contains assessments of the alternatives developed for each deficiency. The assessments were performed by PNL by considering the information and opinions expressed by the panelists (a summary of the discussions of the panel is attached as Appendix A), a review of the NRC staff paper and its attachments provided as SECY-80-514, and an assessment of each alternative with respect to the five attributes described in Section 2.2. Where the panel reached a consensus position on the resolution of a deficiency, that position is identified in this assessment.

The panel also considered the significance to the licensees and to safeguards effectiveness of alleviating each deficiency. The panel pointed out that the deficiencies are not independent and that there is a natural order in which the deficiencies would be addressed irrespective of individual impacts on licensees and safeguards. Further, the means chosen by the NRC to alleviate a deficiency is a crucial and, at this point, unspecified element in any evaluation of the significance of removing a deficiency. The alternative chosen to alleviate some of the deficiencies, e.g., C3, will alter the nature of other deficiencies. With the conclusion that it is not possible to assess the significance of individual deficiencies in isolation, the panel grouped and ordered the deficiencies as a way of aiding the PNL evaluation. This grouping into definitional, modeling and analysis, and implementation is discussed in Section 3.1.

The correction of the definitional deficiencies would not affect the remaining deficiencies. However, the analysis and implementation deficiencies could, perhaps, be more consistently dealt with if the definitional deficiencies were handled first. On the other hand, the decisions on modeling methodology would make one resolution of definitional deficiencies more logical than another.

3.1 **LOGICAL DEPENDENCE OF DEFICIENCIES**

In July 1982, PNL convened a panel of experts in the development of statistical methods and their application to the problems of nuclear materials safeguards, in order to develop alternatives to correct each ID statistical treatment deficiency. The results of that panel discussion were used to develop alternatives for this assessment. PNL reorganized the deficiencies into groups reflecting the statistical logic applied to the problem by the panel members, because definitions of terms are important for communication, while the central problem is the statistical modeling of ID. Once the model is specified, statistical procedures for estimating parameters of the model can be determined and decision rules for parameters of interest developed. These decision rules can be used in setting action limits. An oversight function is required to assure that the whole process is done properly. The grouping of deficiencies based on this logic will be used in this report.
Figure 3.1 shows the logical dependence of the deficiencies. The definitional deficiencies (C1, B2 and C2, B4, B1) impact on all the other deficiencies, but only to the extent that the definitions provide the basis for communication. The ID model (C3) is the most important aspect for resolving the other deficiencies. This is because ID is operationally defined by the error sources included in the model. The model specifies what is included in the conceptual ID. Deficiency B3 is subsumed to C3 because it specifies that "all" sources of error should be included in the ID model. In the view of the panel and PNL, "all" includes more than the "random and fixed errors" appropriate for historically used, measurement-control-based ID models. It includes, for example, the topic of Deficiency C4, prior-period adjustments such as those due to correcting for shipper-receiver differences or bias after the close of the material balance period. Some panel members also recommended inclusion of non-measurement contributors and process holdup in the ID model.

Figure 3.1 does not distinguish between an "ID model" (as used in C3) and an "ID variability model" (as used in B5 and inferred from the SECY-80-514 alternative for B3). This is because complete specification of the ID model should contain a description of how the parameters in the simple ID model are defined. These algebraic definitions require further definitions that generate submodels. The submodels for the data used to estimate the parameters, in turn, specify the random and fixed errors to be associated with the data. The estimate of the uncertainty to be associated with ID is then based on the combination of submodels for the data in the overall ID model. The ID modeling process steps down through more and more detailed specification of the true values to be estimated. Then the ID estimation process goes back up the steps, using estimates of the quantities and variances in the data to calculate the variance to be associated with the estimated ID.

When estimates of the ID and its standard deviation are made (PNL-recommended terminology is "estimated standard error of the estimated ID") they can be used in statistical decision rules (A1). A test of a null hypothesis, e.g., that population ID is zero, with a specified action to be taken based on what the data say about the truth or falseness of the hypothesis is a statistical decision rule. The action to be taken usually depends on the possible alternatives to the null hypothesis. For example, if the hypothesis that the population ID is zero is not supported by the data, the superficial alternative is that the ID is greater than (or less than) zero. But this alternative is consistent with a multitude of practical alternatives, only one of which is that material is missing. The possible alternatives depend on how ID was modeled and how the data were collected. See Section 3.2.2.3 for further discussion of this point.

Figure 3.1 has Deficiencies A2 and B5 under "Implementation." These deficiencies have aspects of material and modeling control which are NRC functions. The ID modeling provides a statistical basis for setting alarm criteria, but there is room for other criteria based on administrative limits (e.g., requirement of specified actions at certain fixed ID levels) or alternative models (e.g., comparing historical throughputs and IDs). The oversight of model development and use is viewed as an evaluation and inspection function for which NRC is responsible. The dashed "oversight" line indicates that NRC's procedures for setting alarm criteria should also be evaluated under this function.
3.2 ASSESSMENT OF ALTERNATIVES

The panel generally agreed on the causes of the problems pointed out in SECY-80-514. Alternative solutions to these problems frequently reflected the emphasis put on different aspects of the problem by different panel participants, but the panel was able to recognize and understand the points emphasized by the individual members. Choice of a "best" alternative was not a task at the panel workshop. A subsequent value assessment of deficiency alternatives by the panel provided PNL with input helpful in making a logical selection of the "best" alternative. The alternatives selected in Section 3 reflect the views and areas of emphasis of PNL staff. The following subsections include descriptions of how the alternatives arose and how each could impact the deficiency.
One aspect that pervades all of these considerations of alternatives is the role of statistics as viewed in nuclear materials accounting. To many statisticians, use of LEID to provide a critical value for testing the null hypothesis that true ID is zero is a questionable statistical practice. This is pointed out in SECY-80-514 on pp. 8 and 21 of Appendix 4 and pp. 25-29 of Enclosure 1. Two viewpoints of the role of statistics that sometimes conflict are the hypothesis testing view and the statistical control view. Bennett (1982) presented a paper on the control, evaluation, and decision aspects of statistics in safeguards, detailing the history and misuse of statistics designed to flag a failure of statistical control as a basis for a decision regarding diversion or loss. When ID is significant, even when and if it is more fully modeled, the statistical conclusion should be to look for the cause of the significant result. It is conjectured that in most cases the cause is really a failure to adequately model ID. It is recognized that establishing "statistical control" must precede using ID as a basis for decision. Those who recognize the complexity of the model underlying the estimation of ID question whether any decision should be made based on estimated ID and its estimated standard error of the estimate other than to take action to determine the cause of a "significant" ID. It should also be remembered that excluding a potential source of error in the model does not remove the effect of that source from the data. Note that the lack of statistical control is not an indicator of "out-of-control" facility operations (SECY-80-514, p. 5) but means that the statistics being generated do not follow the ID model in use. It requires looking at the model, its adequacy and the resulting estimated parameters to determine what caused the "out-of-control" conclusion.

3.2.1 Definitional Deficiencies (C1, B2 and C2, B4, B1)

Definitions of statistical terms have provided the opportunity for hours of interesting discussion and many papers in the safeguards literature. There are several concepts in safeguards which statisticians and users of statistics understand. A problem arises when people from within or between either group try to put these concepts into words. The set of words used by any individual to describe the concepts rarely agrees in total with that used by another individual. There is a need to recognize that conceptual definitions cannot cover all operational contingencies, and that operational definitions can become too complicated for concise communication. What any individual means by a particular statistical term depends upon how the terms are operationally related to the statistical model and/or data under consideration. Conceptual definitions are useful for focusing on desired operational definitions.

3.2.1.1 Deficiency C1

The deficiency addressed by C1 is the lack of consistency between NRC regulations and usage of the terms "ID" and "MUF." This is really not a problem of definition. At the conceptual level, the problem is merely one of uniformity in the acronym assigned to the concept for "the difference between the amount of material supposed to be on hand according to accounting records and the amount of material determined to be on hand according to an inventory" (ANSI N.15.5, 1982). The choices are ID or MUF. Most panel members agreed
with SECY-80-514 to use ID uniformly. Some recognized a problem with NRC’s use of ID while the IAEA continues to use MUF. Others suggested that the problem did not require any action since safeguards practitioners had no problem reading "ID" when "MUF" was printed.

**Recommendation.** The name itself, whether it is MUF or ID, has no statistical implications; thus, the choice between the alternatives must be supported by nonstatistical reasons. The public perception of the meanings, or connotations, of the terms is a consideration. PNL supports the use of "ID" on the basis that it will create consistency between the official documents of NRC and DOE and between NRC documents and actual current practices in the industry. The disadvantage of using ID while the world-wide IAEA practice is to use MUF is considered to be of lesser importance. The duality of terminology does not appear to cause significant difficulty or result in misunderstandings among those familiar with the history and practice of nuclear material safeguards. The PNL recommendation is that NRC change the term "material unaccounted for" to "inventory difference" in the Code of Federal Regulations, and in each regulatory guide that is affected, when the next amendment or revision of the document is made for any purpose.

**Impact Assessment.** The cost to the licensee for implementation of either Alternative 1 or 2 would be minor because only terminology changes in MC&A documents and reports are involved. If changes in the documentations or report forms are required, those changes would likely be made when more substantial amendments are made.

### 3.2.1.2 Deficiencies C2 and B2

Deficiencies C2 and B2 are combined because they require the same solution. They are concerned with making clear distinctions between the concepts of a "true value" and an "estimated value" of a parameter. The "alternatives" suggested by the Panel and recommended by PNL are actually complementary ways of communicating the distinctions among "true value," "estimator" and "estimate" of a population parameter. For C2, the "population parameter" is the ID; and for B2, it is the "standard deviation." The following section is intended to develop the required distinctions.

Alternative 1 (from SECY-80-514) of Deficiency C2 references ANSI N15.16 (presumably, 1974). This ANSI standard contains a definition of "estimator," but does not define "estimate" (as either a noun or verb) and does not mention "parameter" or "true value." Appendix 3 of SECY-80-514 remedies these deficiencies and the 1982 Draft of ANSI N15.5 contains the definitions required, except for "true value." Statisticians usually avoid the term "true value," except to point out that it is not necessarily the same as the population parameter nor the expected value of a statistic. The definitions given in ANSI N15.5, 1982 are recommended for use. Since that is a draft document, the definitions needed to make the required distinction are given here in Table 3.1.
TABLE 3.1. Definitions Relative to Deficiencies B2 and C2

<table>
<thead>
<tr>
<th>Reference</th>
<th>ANSI</th>
<th>N5.16</th>
<th>SECY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section</td>
<td>Section</td>
<td>80-514</td>
<td>19</td>
</tr>
<tr>
<td>1. Population: a collection of objects or events</td>
<td>2</td>
<td>3.1</td>
<td>2.3.1.1</td>
</tr>
<tr>
<td>2. Random Variable: a variable which takes on any one of the values in its range according to a specific but usually unknown probability distribution.</td>
<td>3.1</td>
<td>2.3.1.5</td>
<td></td>
</tr>
<tr>
<td>3. Parameter: An unknown constant which is associated with or is used to characterize a distribution or density function.</td>
<td></td>
<td>2.3.1.10</td>
<td>19</td>
</tr>
<tr>
<td>4. Estimator: A function of a sample ( (X_1, X_2, \ldots, X_n) ) used to estimate a population parameter.</td>
<td>3.1</td>
<td>2.3.6.1</td>
<td>15, 19</td>
</tr>
<tr>
<td>5. Expected Value: the value obtained by multiplying the value of a random variable by the value of the density function of the random variable and summing or integrating the product over the range of the random variable.</td>
<td></td>
<td>2.3.1.12</td>
<td></td>
</tr>
<tr>
<td>6. Sample: (a) A collection of elements selected from a population according to some sampling plan. (b) To select or choose a collection of elements from a population.</td>
<td></td>
<td>2.3.3.2</td>
<td></td>
</tr>
<tr>
<td>7. Estimate: (a) a particular value, or values realized by applying an estimator to a particular realization of a sample, i.e., to a particular set of sample values ( (x_1, x_2, \ldots, x_n) ). (b) to use an estimator.</td>
<td></td>
<td>2.3.6.2</td>
<td>19</td>
</tr>
<tr>
<td>8. Statistic: (a) a function of a sample ( (X_1, X_2, \ldots, X_n) ). (b) a value calculated from a set of sample values, usually (but not necessarily) resulting in an estimate of some population parameter.</td>
<td>2.3.6.3</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>9. Error: a deviation from correctness</td>
<td></td>
<td>2.3.13.3</td>
<td></td>
</tr>
<tr>
<td>10. Correctness: the property of conforming to a conventional or approved standard</td>
<td></td>
<td>2.3.13.3</td>
<td>11</td>
</tr>
</tbody>
</table>

"True value" is not included in the list of statistical terms. Eisenhart, in NBS Special Publication 300, Vol. 1 (1969, pp. 28-30), discusses "true value." (In reading that paper it should be remembered that an estimated ID is based on measurement procedures grossly different from an NBS estimate of the length of a gauge block. The gauge block is available for repeated measurements of its length; ID is not available for repeated determinations of its magnitude,.) The implicit definition of "true value" given by Eisenhart is "the true magnitude of the quantity measured," (p. 29). He goes on to say:
The step from quantitative inferential statements about the limiting mean associated with the measurement of a given quantity by a particular measurement process, to quantitative statements about the true magnitudes of the quantity concerned, may be based on subject matter, knowledge and skill, general information and intuition — but not on statistical methodology (Eisenhart 1969).

(Eisenhart defines "limiting mean" on p. 28. For purposes here, it can be equated to the population mean, the population parameter that distinguishes between the central location of different distributions.) Eisenhart and most statisticians clearly put such a "true value" outside of the realm of statistical methodology, even for such a simple problem as determining the length of a stable and accurately measurable gauge block. Our concern here is with the difference between the sum of amounts previously measured and recorded on paper and the sum of amounts currently measured. This is quite different from conceptualizing how much longer one gauge block is than another. When it comes to the magnitude of ID, it is difficult to even conceptualize what might be meant by a "true ID." A "working definition" attempted in ANSI N15.5 2.3.14.1(a) (1982 draft) is as follows:

A "true" ID (§1D) is the linear combination using the "true" amounts (constants) of material in the quantities EI, R, A and BI.

The key word "true" is left in quotes. This definition is not clarified by following the internal reference to Section 2.3.1.3.3. It uses the undefined concept "true" and so cannot convey more than an undefined concept despite its use of the algebraic term "linear combination." PNL recommends avoiding the use of "true ID" by defining the population ID as follows:

The population ID is the expected value of the estimator of ID, that is, of the random variable defined by the linear combination of the functions of the measurement random variables used to estimate ID.

This definition requires that all quantities used in the estimator of ID be based on some kind of measurement that can be characterized as a random variable. Unless the conceptual "true ID" is operationally defined as a population ID it cannot be made tractable statistically. A fuzzy conceptual model (see Section 3.2.2.1) is possible, but the operational model is inherently linked with a specific accountability system and the data it generates.

The application of statistics begins with modeling measurement systems, as far as ID is concerned. Measurement systems provide our connection with the real world of material accounting. The only statements we can make about this real world depend on the statistical populations (Table 3.1, Definition 1) defined by the application of the measurement systems. The events of concern are the application of the measurement systems to process items or measurement control standards. Each such application generates a random variable (Definition 2). Each random variable can be characterized by a distribution function with certain population parameters (Definition 3). Depending on the reasons for collecting data, estimators (Definition 4) are selected to help answer the questions of interest. The expected values (Definition 5) of the estimators can be determined from general statistical theory. The estimators
of major interest in safeguards are the means and variances of the random variables. The expected values of these estimators can be theoretically determined to show that they are "good" estimators of the population parameters of interest. All of this can be done without looking at any data. The distributional assumptions involved can be checked after the data have been collected, or past experience can be used to provide reasonable starting assumptions.

The actual work of determining numerical values for the estimators is done by collecting samples (Definition 6) of observations from the populations of possible measurements on process items and standards. These sample realizations of values for the random variables are our only connection with the "true" quantities sought. (At this basic measurement level, most people can conceptualize a "true" weight for the UO₂ in a can or the fraction of uranium in the UO₂.) For process items we frequently only have a sample of size one from the population of possible measurements, like the weight of the UO₂ in a can of UO₂ product. For a measurement control standard weight, we may have 15 or 20 measurements over an accounting period. Each one of these 15 or 20 observations constitutes a sample of size one from potentially different random variables. Over time, we take measurements to determine whether or not the weighing system is generating random variables from the same distribution; that is, whether or not each observation comes from a distribution with the same mean and the standard deviation characteristic of the weighing system. We can use an estimator of the mean and standard deviation to get the sample estimates (Definition 7a). These numerical estimates can then be compared with the standard value or historical sample standard deviation to determine whether or not the weighing system is in statistical control. (Control charts are commonly used to check the state of statistical control as the data are generated.) Definition 8, "statistic," does not seem to be practically different from the noun "estimator." Another term used in SECY-80-514 (p. 11 and p. 15) is "observed" (ID or LEID), which would be equivalent to estimated (ID or LEID) in the sense of Definition 7b.

It should be recognized that in going from the basic measured values to the estimated ID (or its standard error), an important statistical step has been taken. The estimator of ID is a function of the basic measurement random variables and so is itself a random variable. (See ANSI N15.5, 1982 Draft, Sections 2.3.1.3 and 2.3.1.5). The estimator of the standard error of the estimated ID is also a random variable derived from the basic random variables through their algebraic relationships specified in the operational definition of the ID estimator. Although the statistical distributions associated with the basic measurement random variables are usually well characterized, analytical characterization of the distribution of ID has proved to be intractable without numerous simplifying assumptions and recourse to statistical theory of sometimes questionable applicability.

The rather loose conceptual definition of error from ANSI N15.5, 1982 Draft is included to provide a connection for "correct quantity" as used on p. 11 of SECY-80-514 in the last paragraph before subsection (b). Correctness is conceptually defined as "conforming to a conventional or approved standard." In 2.3.13.3 of ANSI N15.5, 1982 Draft, "correctness" is conceptually expanded to include both "the expected value of the random variable" and "the
'true' value being sought" as well as "the accepted reference value." "Correct quantity" or "correctness" is thus a fuzzy concept, useless for precise discourse. The definition of "correct" depends on the operations used (take the expected value of a random variable, apply a measurement system to a standard, or philosophize about "true" values).

In order to avoid changing the words in SECY-80-514 and in the alternatives generated by the panel, true, as used in "true ID," "true value," and "true standard deviation" must be interpreted in the sense of the expected value (limiting mean) of a random variable. In this usage it must be remembered that the "true value" of a measurement random variable may include the bias of the measurement system. "Observed value" is identical to "estimated value."

Recommendation. The panel and PNL support the implementation of all four alternatives (i.e., use ANSI N15.5 definitions; develop, communicate, and reinforce the distinction among "true value," "parameter," "estimator," and "estimate") as a remedy for both deficiencies. Their implementation will aid in clarifying and reinforcing these important distinctions for the statistical layperson. The benefit of alleviating these deficiencies is clarification for nuclear material managers of the meaning of the statistical concepts and error models used in support of nuclear material accounting. The resulting improvement in communication between statisticians and nuclear material accountability personnel can be expected to have long-range benefits to safeguards.

Notation used in written material should indicate and distinguish among "true value," "estimator" and "estimate." Recommended notation that is consistent with ANSI N15.5 1982 Draft is given in Table 4.1 of Chapter 4.0.

To achieve the desired clarification of the above concepts, PNL recommends that the NRC incorporate explanations of the concepts and suggested notation for the true values, observed values, estimates and estimators of ID and of the measures of the variability of ID into future revisions of regulatory guides that pertain to statistical concepts, terminology, modeling and evaluation of ID and the variability of ID.

Impact Assessment. No increased costs to the licensee are anticipated as a result of implementing any or all of the alternatives associated with these deficiencies.

3.2.1.3 Deficiency B4

The problem here is with the terminology used to conceptualize and define the various kinds of errors encountered in nuclear materials accounting. Much of the cause of this problem is attributed to the usage of the terms "long-term systematic error" and "short-term systematic error" in J. Jaech's book, Statistical Methods in Nuclear Material Control (TID-26298, 1973). Two critical comments (see page 12 of SECY-80-514):

1. "The terms bias and systematic error are generally considered synonymous in the statistical literature and industrial practice but not in Jaech's book."
2. "Jaech treats bias only as a random variable."

PNL suggests that a close study of Jaech reveals the following:

1. Jaech (p. 81) clearly states: "... no distinction is made between long-term systematic error and bias."

2. Jaech treats biases as constants, but estimates of biases as random variables. Since biases may change over time, they can be modeled as being drawn from a distribution with a mean and a variance.

Jaech's book provides a bridge between the complexity of strict statistical development and the need for a practical calculational algorithm. The book defines, in words and copious examples, his concepts of long- and short-term systematic error. The definitions on p. 81 of the book provide only the conceptual definitions. The remainder of the book presents operations based on these definitions. Conceptual definitions are basically didactic tools. The error categories used by Jaech have produced considerable confusion, not so much because the words fail to cover the required concepts, but primarily because the role of a conceptual definition has been overburdened by the readers. The most critical deficiency in the application of these definitions has been the failure to understand the crucial role of the "data set" in these definitions. Jaech's algorithm for calculating the variance of ID is based on decomposing the data into subsets that are statistically homogeneous, i.e., subsets that have effects characterized by the same population means and variances. One must consider the random variables involved in obtaining the data. There may be a constant (fixed) bias in the basic measurement model used. Then one subset of the data in hand may be used to estimate the magnitude of the bias, and another subset may be corrected for the estimated bias. At this level, as noted before, we are dealing with measurement random variables. The problem comes when we move to models that aggregate the data into components of ID.

The panel members understood the concepts involved, but seemed of the mind that Jaech's terminology is not a viable didactic tool. (John Jaech himself did not object to this idea.) PNL doubts that any standard statistical terminology will be agreed upon that adequately conveys the concepts involved. The alternatives developed by the panel were based on the need to distinguish three concepts associated with the uncertainty in ID, the same three concepts as intended by Jaech's formulation. "Uncertainty" was defined as a deviation from the true value, where "true value" was understood in the sense of the expected value (limiting mean) of a measurement random variable. (However, in the generation of alternatives, the distinction between the uncertainty in the basic random variables and in ID, at a higher level of modeling, was lost.)

When considering the basic measurement random variable level of modeling there appears to be general consensus that only two kinds of errors were to be modeled, bias and random error. It is understood that bias refers to "the deviation of the expected value of a random variable from a corresponding correct value," (ANSI N15.5, 1982, Draft 2.3.13.8) and that a random error is "the deviation of a random variable from its expected value," (ANSI N15.5,
Conceptually, random errors attenuate over repeated applications of a measurement system to a specific item to be measured, but biases do not. Depending on the way the measurement system is operated, error sources may contribute to the bias and the random error terms. For example, if the measurement system is used to determine the uranium concentration in a can of UO$_2$ product by fluorimetry, several sources of bias exist: sampling, sample storage method, reagent purity, equipment contamination and operator skill. Each of these is also a source of random errors. An operator who is careless about avoiding parallax error may consistently read too high in measuring dilutions, causing a biased low result in the intermediate gram per liter result. Upon repeated analyses, this parallax error is variable, inducing random errors about the biased readings. The end result is that the statistically expected value of the determination is not the "true value" sought. The expected value of the measurement system (its limiting mean) still has the bias term. It is the attempt to estimate such bias terms that has led to the creation of various terminologies to account for the uncertainty that bias estimation contributes to the measurement results.

To further complicate the problem, some measurement practitioners advocate accounting for nonmeasurement errors in the total uncertainty of the final result. Two types of commonly recognized nonmeasurement errors are those whose magnitude is "estimated to be within some reasonable bound |ΔX|, perhaps by experience or judgment" (Ku 1969), and those "due to ... mistakes occurring in processing the measurement," (ANSI N15.5, 1982 Draft 2.3.13.13, Note). In the first type, the errors are "nonmeasurement" because they are not based on random variables generated by a measurement system, and so are not estimated statistically. In the second type, the errors are due to mistakes in handling the numbers produced by the measurement system. A third type of nonmeasurement error is modeling error, the failure to include important sources of variability in the model.

Ignoring, for simplicity, the nonmeasurement errors, there are then three concepts with which to deal: bias, random error, and the uncertainty induced by the elimination of bias (calibration). When a chemical analysis system is calibrated, or a weighing system adjusted for bias, or an estimator for a bias correction applied to the random variables produced by a basic measurement system, the result is a random variable that is a function of random variables. The uncertainty in such functions of random variables needs to be characterized in terms of both the original random variable and the new "adjusted" random variable. The critical question: How long does the specific adjusting function persist? Equivalently: To what subset of basic measurement random variables does the adjustment function apply? In terms of the basic measurement random variables, bias corrections and calibrations affect the population mean. In terms of the random variable defined through the adjusting function, bias corrections and calibrations affect the population variance, since the function defines a new statistical population.

At the level of ID conceptualization, the estimator of ID is a function of random variables, and so itself is a random variable. Operationally, its definition requires specifying a period of time, the inventory period. Over this time period a number of calibrations or adjustments for bias may be used.
The variance of the random errors associated with these different adjustment functions has an effect on the data (realized random variables) collected during specific time periods. The durations of these time periods may vary from the time it takes to do a single analysis to the entire inventory period.

The alternatives generated by the panel arose from emphasizing various aspects of this complicated process. Some terminology is more applicable to one level of modeling than another (from high-level modeling of ID to the low-level modeling of basic measurement random variables). In some cases, distinctions between the statistical populations being considered were not clear. The alternatives can be organized to distinguish among them, as in Table 3.2.

### TABLE 3.2. Alternative Terminology for Error Concepts (Deficiency B4)

<table>
<thead>
<tr>
<th>Alternative</th>
<th>Measurement Error Has an Effect on</th>
<th>Third Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1] (SECY-80-514)</td>
<td>Bias Random Error</td>
<td>(None)</td>
</tr>
<tr>
<td>[3]</td>
<td>Bias Uncorrelated Errors</td>
<td>Correlated Errors</td>
</tr>
<tr>
<td>[2]</td>
<td>Bias Random Effects</td>
<td>Residual Effects</td>
</tr>
</tbody>
</table>

In all of the alternatives, "bias" is the appropriate term for an error that makes the population mean different from the "true value" sought, regardless of statistical population. For errors that affect the population standard deviation, random error is offered implicitly in both Alternatives 1 and 4. Assuming that "current terms" means Jaech's terminology, his "random error" would be most appropriate to the statistical population defined by the function (estimator) for ID. This estimator of ID is free from recognized bias. The estimator of the ID population standard deviation has components for random errors that do not affect the mean of the basic random variables, and for fixed errors that do affect the mean of the basic measurement random variables. However, these fixed errors are estimated, and the result is used to correct the measurements for the fixed errors, which in the process affects the standard deviation associated with specific subsets of the ID data. These are the long- and short-term systematic errors. Operationally, "long-term" refers to an error component which applies to all the data generated by specific measurement systems during an accountability period. "Short-term" applies when more than one calibration or bias estimation was used for a measurement system during the accountability period.

Alternative A1 does not specify terminology for the third concept, requiring that the treatment and terminology be clarified. Words in SECY-80-514 indicate the distinction should be made between "random" and "fixed" effects (in the theory of components of variance sense) and that the term "systematic error" should be eliminated (p. 23 in "Additional Staff Views").
One explanation of Alternative A3 is that if random variables are modeled as

\[ Y_i = \mu_i + \varepsilon_i \]

i.e., simply, a true value and an error term, then inclusion of the potential for bias expands \( \varepsilon_i \) to \( (\varepsilon_i + \beta_i) \), and the errors in \( Y_i \) are correlated because each \( \varepsilon_i \) has a common error component. Another way is that if \( Y_i \) is based on a calibration from a regression fit; then, for simple linear regression

\[ Y_i = \phi_1 + \phi_2 x_i + \varepsilon_i \]

and the error in the \( y_i|x_i \) (the estimate of \( Y_i \) given \( x_i \)) includes a component due to estimating the regression parameters \( \phi_1 \) and \( \phi_2 \). Again, the errors associated with the observed \( y_i \) are related. The panel thought that the concept of correlated errors could be more difficult to explain than the current terminology or that suggested in Alternative 2. The populations envisioned by the proponents of this alternative would seem to be at some intermediate level of functions of random variables sufficient to define the error dependencies in terms of a time sequence.

The panel accepted the term "random effects," but rejected the term "fixed effects" in favor of "residual effects" for Alternative A2. This was done with some trepidation since "residual effects" has connotations of the errors left after fitting a regression model. It is true that calibrations often use regression models generating errors that affect estimates of both \( \mu \) and \( \sigma \), but other estimation procedures used in the industry also do this. It was thought that this regression specificity could be overcome and "residual effects" used to conceptualize any component of uncertainty that was neither a recognized bias nor random error. Thus, "residual effects" does not connote the same error concept as "long-, short-term systematic error" or "correlated errors." "Residual effects" merely connotes the errors, random or fixed, that are remaining after all modeled sources have been considered. The conceptual population level here is for any population defined in the ID estimator hierarchy. At any level we have a random variable that may have a biased population mean and that definitely has a well-defined population standard deviation (which may be difficult to estimate). At the highest level of the ID model, the "random effects" terminology implies that we can think of the errors as being a random selection from a population of errors due to some measurement error source. We can thus speak of the "random effect" due to scale bias, due to operator differences, due to sampling—due to whatever source we want to specify. This terminology eliminates the need for general terminology for the third concept, replacing it with a specific description of the kind of random effect involved.

**Recommendation.** PNL, after reviewing the four alternatives, recommends use of the terms "bias" and "random effect," specifying the source of the random effect. Once we take the first step from the basic observed random variables produced by a measurement system, we are dealing with functions of
random variables. Random variables are adjusted, in some sense, by other random variables. The result of such adjustments is a random variable based on some estimation procedure (usually for bias or calibration). Both the mean and variance of the resulting random variable are affected by the estimation procedure. Using "random effect due to" reminds us how the quantities of interest were defined. Additional definitions are recommended in Table 3.3.

### TABLE 3.3. Definitions Relative to Deficiency B4.

1. **Uncertainty:** A concept employed to describe the inability of a measurement process to measure the correct values exactly. (ANSI N15.5 Section 2.3.13.12)

2. **Bias:**
   - (a) The deviation of the expected value of a random variable from a corresponding correct value. (b) A fixed error which remains constant over replicate measurements (synonyms: deterministic error, fixed error, systematic error). (ANSI N15.5 Section 2.3.13.8)

3. **Random Error:**
   - (a) The deviation of a random variable from its expected value. (b) A deviation which occurs "randomly" (according to a probability distribution) over replicate measurements. (ANSI N15.5 Section 2.3.13.7)

4. **Function:**
   - A rule or procedure by which exactly one element of one set (called the "range" of the function) is associated with each element of another set (called the "domain" of the function).
   - A function of a random variable is itself a random variable. (ANSI N15.5 Section 2.3.1.5)

5. **Random Effect:**
   - The additive component for a specific source of variation in a linear model, assumed to be drawn from a population (distribution) with a mean of zero and a finite variance. (PNL)
3.2.1.4 Deficiency B1

This deficiency addresses the use of the term "limit of error" to designate the uncertainty component used in constructing a 95 percent confidence interval for the random variable, IO. Use of the term "limit of error" presents two problems: that it has different definitions in official documents [10 CFR 70.51(a) and ANSI N15.16, 1974], and that these definitions encourage its use to provide a critical value for a test of the hypothesis that the true ID is zero (SECY-80-514, p. 25, pp. 33-38). The panel agreed with SECY-80-514 that "limit of error," as used in LEID, should be dropped because its use leads to erroneous inferences (not supported by the ID data and underlying statistical model). If its use is eliminated, it will not have to be defined, and the deficiency is solved. However, there is a need for appropriate terminology to specify the concept of ID uncertainty. It would be best if standard statistical terminology could be used in the definition of the measure of dispersion to be associated with ID.

ANSI N15.5, 1982, Draft discusses measures of dispersion in Section 2.3.5. It should be noted that each accountability period generates a single estimated ID, and that we are concerned with a measure of variability for that ID. Since it takes a minimum of two observations to get an estimate of any measure of dispersion, we must have recourse to the underlying model of ID to "propagate" (ANSI N15.5, 1982, Draft 2.3.12.2) the variability associated with the basic random variables into an estimate for the variability of the estimated ID. It is true that ID is a function of random variables and so is itself a random variable with an associated probability distribution which has a definable population mean and variance. ANSI N15.5, 1982, Draft 2.3.5.1 defines "variance" as follows:

\[ \text{Variance. (a: population)} \quad \text{The expected value of the square of the difference between a random variable and its own expected value, i.e., the second moment about the mean}. \quad \text{(b: sample)} \quad \text{The sum of the squared deviations from the sample mean divided by one less than the number of values involved.} \]

The standard deviation (for either the population or the sample) is then defined as "the positive square root of the variance."

The estimator of the standard deviation, the sample standard deviation, is defined in terms of a sample (necessarily of size greater than one). Use of the term "standard deviation," unqualified by any adjectives (as recommended in Alternative 1), would imply an impossible calculation. We only have a sample of size one (i.e., the ID value). A sample mean is one of the most common estimators; but even its standard deviation is accorded a special name in standard statistical textbooks: "standard error of the mean" (Ostle 1963; Steel and Torrie 1960; Mendenhall and Ott 1980). In fact, "Standard Error
Alternative 2 recommends "standard error of the estimate" as the terminology for the measure of ID variability. This terminology is applicable to components of ID as well as ID itself. Its value lies in leading users to recognize that they are not dealing with a simple sample standard deviation, but with the standard deviation to be associated with a complicated estimator, and in fact the specific estimator of ID obtained for the accounting period of interest. The specific estimator for the standard error of the estimate of ID also requires the inclusion of calibration errors and bias estimation, concepts difficult to bring into the definition of a sample standard deviation.

Alternative 3 recommends the use of "estimated mean square error." ANSI N15.5, 1982 Draft, 2.3.6.5 defines the term as follows:

**Mean Square Error (of an estimator).** The expected value of the squared deviation of an estimator from the value of the population parameter being estimated, i.e., the second moment of an estimator taken about the value of the population parameter being estimated.  
**NOTE:** The mean square error is the sum of the variance and the square of the bias of an estimator.

This definition is concerned with the population mean square error. It has to do with assessing the magnitude of the bias of a statistically biased estimator (as defined in 2.3.6.4). For example, taking the expected value of "the sum of the squared deviations from the sample mean" (of a sample size n) and dividing by n, results in a biased estimator of the population variance, that is:

\[ E\left[ \frac{1}{n} \sum (\bar{x} - \bar{x})^2 \right] = \frac{n-1}{n} \sigma^2 \]

However,

\[ E\left[ \frac{1}{n-1} \sum (\bar{x} - \bar{x})^2 \right] = \sigma^2 \]

So that dividing by \( n-1 \) provides an unbiased estimator of \( \sigma^2 \). The bias of the biased estimator is

\[ \frac{n-1}{n} \sigma^2 - \sigma^2 = \sigma^2 \left( \frac{n-1}{n} \right) \]

so that the magnitude of the bias decreases as \( n \) increases. If we were forced to choose among biased estimators, the one having minimum mean square error might be preferred. Then the mean square error would have some theoretical interest, but as defined, it has little practical (i.e., calculation from data) interest.

3.16
Of practical interest is a measure of uncertainty which combines the random error and the other errors of estimation associated with the estimate of ID. In this context, an "estimated mean square error", defined as the sum of the estimated variance of an estimate and the square of the associated estimated bias, has been considered. If the estimate is ID, then this would imply adding up all the bias estimates, squaring them, and adding this result to the propagated variance of the random errors. Another approach would be to square the individual biases and add their sum to the propagated variance.

The choice between squaring and summing or summing and then squaring biases can be avoided if bias corrections are made whenever the estimated bias is greater than the estimated standard deviation of the bias, and if we use the rationale of treating errors based on how they relate to the data and the full ID model (as described in Sections 3.1 and 3.2.2 of this report). Under this approach, the terminology "standard error of the estimated ID" is most appropriate. We have to get back to the basic measurement random variables to operationally define how the associated measurement errors should be used in the estimate of ID variability.

Alternative 4 is noncommittal about the candidates "standard error of the estimate" or "estimated mean square error," and so does not suggest an alternative. Alternative 4 does allow for use of the words "limit of error" where it indicates "a defined interval for a specific evaluation problem." None of the other alternatives allow "limit of error." There does not seem to be any good reason for retaining "limit of error" in the safeguards glossary. If the evaluation problem has to do with comparing the uncertainty associated with estimated IDs at different times for the same facility, or under different ID models, or for different facilities, then the estimated standard error of the estimated ID (normalized as required for comparisons), is adequate. However, due to the complexity of the estimator, an appropriate procedure for establishing confidence limits will be hard to determine. In particular, assigning a property equivalent to degrees of freedom will be a problem. If the evaluation task is to define decision rules, then the required multiples of the standard error can be used. (Note that we have consciously shortened the recommended terminology since the context should make it clear that we mean "estimator of the standard error of the estimated ID."

Recommendation. PNL concludes that the use of the term "limit of error" in connection with ID should be discontinued. Instead, "the standard error of the estimate" or "the standard deviation of the estimate" should be used to define the variability of ID. "Standard error" is the preferred term because it is generally understood to mean the standard deviation of a complicated estimator, whereas the term "standard deviation" alone is generally understood to refer to a simple sample standard deviation. If the term "standard deviation" is applied to an estimator, such as ID, the estimator must be stated, i.e., "...the estimated standard deviation of the estimated ID..." or the estimator being referenced must be entirely obvious from the context.

The term "limit of error" could be applied to the designation of a defined interval for a specific measurement result, but this application is not recommended. The proper terms for a specific measurement result are "standard deviation" and "variance."
Impact Assessment. All of the alternatives proposed to correct this deficiency would involve small, one-time costs for implementation, but operating costs would not be affected. The implementation costs involve documents, report forms and computer programs to accommodate the new terminology and changed computation. Our analysis indicates that the total cost would be the same for all the alternatives; it is estimated to be 2 person-weeks or $2K per licensee.

3.2.2 Modeling and Analysis Deficiencies (C3, B3 and C4, A1)

3.2.2.1 Deficiency C3

This deficiency was concerned with the error terms in the ID model and with its lack of labeling random and fixed errors. In our strict definition of a statistical model (in Section 3.1), the error terms are a third-level specification. First came the definition of ID in terms of the parameters to be estimated from the data. The parameters to be estimated are such things as Ending Inventory, which could include a component for process hold-up that is closer to reality than is currently modeled, perhaps using a by-difference method with recovery data from past process cleanouts. At the second modeling level, the first-level parameters are modeled in more detail in terms of the measurement involved. At the third modeling level, the statistical models for each measurement made specify how the errors to be associated with the data arise. Error terms become explicit, and methods for estimating error components from data can be redefined. At this level, potential biases are included in the model, and methods for collecting data to estimate the magnitude of the biases are defined. Measurement methods requiring calibration are also specified, and methods of data collection to assess the uncertainty associated with calibration are defined. Methods to check that calibrations remain valid and that biases are constant or zero are also defined. All of these models of data collection generate data, each with its own random errors. If the model at this level contains a bias term, an attempt is made to estimate the bias and its associated standard error of estimation. Calibrations also have their associated errors of estimation. At this stage, all errors are treated as the result of random processes. That is, they are, statistically, random variables. When processes, as opposed to measurement control data, are generated, the effect of the bias estimation and the calibration random errors becomes fixed for some subsets of the process data. We can now talk of random and fixed errors in the second-level models. It should be remembered, however, that all estimates of standard deviations were based on how the random errors of estimation were modeled. An estimate of a bias is a random variable even when the population bias is modeled as a fixed quantity.

The above description shows Deficiency C3 as interpreted by the panel and PNL. Although SECY-80-514 Deficiency C3 focuses on error terms, the overall ID model and the lower-level models must be defined before we can specify how the error terms arise. Then, we develop a valid way for associating these errors with the components of ID, and ultimately with ID itself. Basically, the door was opened to full conceptual modeling of ID. In this context, the alternatives posed by the panel were ways to improve ID modeling by more explicit consideration of true ID.
There were two emphasis dichotomies operating in the panel deliberations and consequently in the alternatives generated. The first dichotomy was concerned with the overall model: whether ID should be modeled to have expectation zero or the parameters in the model should be restricted to those that can be estimated by direct measurement. The second emphasis dichotomy was concerned with whether the statistical errors are viewed in terms of their effects on the observations or relative to a particular data set. These dichotomies generated the four panel alternatives, as shown in Table 3.4.

**TABLE 3.4. Emphasis Dichotomies for Alternatives to Deficiency C3**

<table>
<thead>
<tr>
<th>Error Description</th>
<th>Model</th>
<th>Direct Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect on Observations</td>
<td>E(ID) = 0</td>
<td>Alternative 2</td>
</tr>
<tr>
<td>Relative to a Data Set</td>
<td></td>
<td>Alternative 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alternative 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alternative 5</td>
</tr>
</tbody>
</table>

The E(ID) = 0 emphasis arose from the perceived need to include all terms that could make ID nonzero in the model. It gives real meaning to the assumption that, in the absence of diversion or loss, ID should be zero. If all possible sources of error are included in ID, then a significant ID does imply diversion or loss. The problem here is that "all possible sources" includes such things as unaccounted-for process hold-up and operator and clerical "mistakes." Including such terms in the model does not guarantee that another kind of error, model error, is excluded. Furthermore, development is required of statistical methods for estimating these additional parameters and their uncertainties. On the other side of this dichotomy were those who emphasized the practical need to obtain estimates of the parameters specified in the model. If data cannot be collected to estimate the parameters with an associated uncertainty, then its use in the model is questioned. The more terms left off the model, the more alternatives there are to investigate when the estimated ID is significantly larger than its estimated standard error of estimation.

The line between this dichotomy becomes blurred when we realize that investigation of the cause for a significant ID under the direct measurement definition requires some kind of measurement for unaccounted-for process hold-up and perhaps even the frequency of operator and clerical "mistakes." It becomes a question of whether to include these "new" sources of errors in the model or recognize that they have always been implicit in the old models. In either case, measurement data of some sort is required, as well as a broad definition of "measurement."

The second emphasis dichotomy was between treating random and fixed errors in terms of the effect they have on the observations or in terms of the data set to which they apply. The brief words used to describe this dichotomy can
be misleading. The distinction is really between the effect of the errors on the observed (estimated) ID and their effect on a specific subset of the data. This distinction has impacts on the extent to which a model is developed, the meaning of random and fixed errors, the alternatives to the hypothesis that ID equals zero, and the time span of the data on the measurement systems allowed to be included in the estimate of ID and its estimated standard error.

With regard to the analysis of variance (AOV) analogy, it was conceded by the panel that the concepts of random and fixed effects, in the sense used in the AOV, are not analogous to the current distinction between random and systematic errors in safeguards usage. In the AOV the random effects are assumed to be sampled from an infinite statistical population, but for fixed effects, the effects constitute the entire population. Random effects are constrained by the statistical model to have an expectation of zero and finite population variance, but fixed effects are only constrained in the AOV model to sum to zero for convenience, and are constant for each experimental unit subject to a particular effect. In common safeguards practice, a random error might correspond to a random effect, but systematic errors imply too many different statistical entities to be analogous to fixed effects. In particular, the models for estimation of systematic errors (e.g., due to bias and calibration) include constraints about zero expectations and finite variances.

Considering the actual words used in the statements of the alternatives, the phrases "describe errors in terms of the effect they have on the observations" and "definition of errors should be relative to a particular data set" leave little room for a distinction. The only obvious way in which errors can be related to a particular data set is through the effect they have on the observations. Use of the specific terminology "long-term systematic," "short-term systematic," and "random" was not endorsed by the panel, nor is it recommended by PNL, so the distinction is not one of terminology. The conclusion is that the concepts of random and fixed errors, under whatever names, should be used according to their definition and treatment by the models. Since all errors are treated as random, or estimable from random variables, at the lowest level of modeling, we are only concerned with "fixed errors" at the first and second levels of modeling. The fixedness is in terms of the data subsets affected, or equivalently, the effects they have on the observations. If an error has the same effect on more than one observation, then the error is "fixed" for that set of observations. Valid use of the ID model requires that the data subset to which a "fixed error" is applied be kept in view so that the errors can be correctly combined to get the estimated standard error of the estimate for ID.

This logical analysis of alternatives has shown that based on the evaluation of the panel, the alternatives are not independent enough to provide operational direction.

Recommendation. The key issue is whether all chance (as opposed to deliberate) deviations of the observed ID from the true ID are to be treated as random effects, or whether only those effects estimable from measurement data are to be included in the error propagation, with the remaining effects
contributing jointly to a (possibly unknown) non-zero expectation for the ID. Those who favored Alternative 1 may be reflecting the general recommendation for explicit consideration and identification of random and fixed error components. This approach carries over to the choice of alternatives in Deficiency B3 as that one which calls for "better modeling of ID variability" and identification and characterization of all contributors to the variability of inventory differences.

PNL endorses the general approach of Alternative 2, which calls for modeling ID such that \( E(ID) = 0 \) and for the errors to be described in terms of the effect they have on the observations. However, there are two cautions:

- The distinction between the conceptual model of the total uncertainty in an ID and the model pertinent to the analysis of a particular data set must be recognized.

- A solution must be agreed upon to the semantic problem considered in Deficiency B4, i.e., calling a random contribution to the total uncertainty in the ID a "systematic effect." For example, an error term for human errors (mistakes) would be in the conceptual model, but it would be omitted from the data set model until a way is found to determine the distribution function and estimate the parameters of such errors (see Section 3.2.2.2). The principal error terms necessary for a model of the variability of ID is discussed in Section 3.2.2.2, and the recommended nomenclature is discussed in Section 3.2.1.3.

In Section 4.3, PNL suggests an integrated approach using a model with the expected value of ID equal to zero.

Impact Assessment. Alternatives 1, 4 and 5 would entail some effort on the part of the Licensee to adopt the statistical guidelines provided for these alternatives. It is estimated that 4 person-weeks by one statistician will be needed. Thus, a one-time implementation cost of $5K is estimated for each of these alternatives. No incremental operating costs above the status quo would be incurred.

Alternatives 2 and 3 would involve facility-specific system analysis and adaptations of statistical techniques to incorporate sources of variability not currently included in the ID model, including process variability and human errors. One-time implementation costs are estimated to be 27 person-weeks or $30K. Once the alternative is implemented, only minor incremental operating costs will be incurred.

3.2.2.2 Deficiencies B3 and C4

These deficiencies are concerned with including all sources of error in the calculation of the variance of ID. They are resolved by adequate modeling of the conceptual ID with expectation zero. Including "nonmeasurement errors" will inflate the true standard deviation of ID, and the associated uncertainties of such errors will be difficult to estimate. It may be better to
leave these "statistical outcasts" as a competing alternative hypothesis when ID is significantly different from zero. Such nonmeasurement errors are currently the first thing sought in ID backup data when a "significant" ID is observed by a licensee. The alternatives to these deficiencies recommend adding terms to the ID model. The added terms could be used in defining a stepwise approach in which the results based on the (current) measurement error approach are evaluated; then, process variability is included in the model, etc. This would lead to the "spectrum of tests" recommended for Deficiency A1 (see Section 3.2.2.3).

Most of the past effort on modeling ID variability has been limited to characterizing the components of error in the measurements that affect ID. The common viewpoint at the present time, and that of the writers of the NRC staff paper (SECY-80-514) and of the panel, is that nonmeasurement errors are a fact of life and should be included in modeling the variance of ID.

The principal difficulty of including process variability in the ID model is that a totally satisfactory method for estimating the variance due to such variability is not yet available. The method currently proposed involves applying some fairly restrictive assumptions, one of which is that no diversions or appreciable losses of SNM have occurred in the series of material balance periods used for the process variance calculation (Tingey, Lumb and Jones 1982). However, it is the best method available, and it is worth pursuing.

If a hypothesis test is adopted as the basis of ID action thresholds instead of the administrative limits now applied (see Deficiency A1), another disadvantage of including process variability in the model will occur. Large variability in the process will inflate the estimate of the standard error of the estimated ID, thereby reducing the power to detect a given amount of loss.

Recommendation (B3). PNL recommends that NRC adopt Alternative 1 and take the following actions:

- Continue support of the efforts to model ID variability to include all sources of error.
- Develop guidelines for licensees in modeling ID variability and applying the model (see the assessments of Deficiencies C3 and B5 in Sections 3.2.2.1 and 3.2.3.2).
- Initiate research on methods for estimating the variance of nonmeasurement errors in the accounting process.
- Periodically review plant-specific modeling and estimation of ID variability (see the assessment of Deficiency B5).

Recommendation (C4). The alternatives for Deficiency C4 are complementary and do not require a decision or selection to be made. The panel indicated that regardless of the overall requirements for modeling ID, any prior-period adjustment should be done in such a way as to make the actual process traceable.
PNL recommends that the NRC provide guidance to the licensees in making adjustments to the nuclear material accounting records after the end of the reporting period. A recommended accounting procedure is presented below:

- Adjustments or corrections to entries in the accounts made after the end of the reporting period which affect both inventory and noninventory accounts (shipments, receipts, discards, ID, etc.) should be entered in a "prior-period adjustment account." Prior-period adjustments affecting only inventory accounts can be made by adjusting the accounts to their proper balances (Korstad 1980).

- No adjustment or correction for a prior period is to be recorded in a manner that will cause error in the ID for the current period.

- All adjustments or corrections should be traceable and auditable. They should show accurately what occurred from a statistical and MC&A standpoint and provide reference information supporting the adjustment.

- All adjustments or corrections for a prior period, i.e., adjustments that affect the ID of that period, should be recorded in the licensee's historical records of IDs and cumulative ID in such a manner that the adjustment is traceable and the historical ID data can be corrected before statistical analyses are performed.

- Adjustments applicable to prior-period IDs should be reported to all organizations and offices that maintain records of the historical ID and cumulative ID for the facility. These adjustments must be considered when studying trends within facilities or between facilities.

Impact Assessment. For Alternative 1 of Deficiency 83, inclusion of all sources of variability in addition to measurements in the model for ID variability will result in some incremental cost over costs incurred under current practices. The implementation-cost elements include development or adaptation of statistical methodology on a facility-specific basis, and modification of computer programs to carry out the computations. Also, there will be additional operating costs for calculations and data evaluation. The estimated costs for Alternative 1 are $30K (implementation costs) and $7K/yr (annual operating costs). The implementation of Alternative 2 is similar in many respects to Alternative 1, and the costs are estimated to be substantially the same.

For Deficiency C4, implementation of Alternative 1 will entail some costs to the licensee to modify its accounting report forms, procedures and ID calculations so that adjustments that affect ID are applied to the proper period. We have assumed that bias corrections, shipper/receiver differences, and corrections arising from scrap recovery operations are now being determined properly and that the additional action required will be to maintain traceability, credit and debit the right ID accounts and recalculate and update ID. In addition, the effects of adjustments on prior-period and
cumulative IDs need to be reported to NRC along with the standard errors of
the estimates. The information needed for implementation is available. The
incremental costs will arise from the need to revise accounting procedures, to
design the needed record forms, and to revise existing computer programs.
Operating costs will increase slightly to accommodate record-keeping require-
ments. The estimated implementation cost is 3.5 person-weeks or $4K. Esti-
imated annual operating costs are 2 person-weeks or $2K.

3.2.2.3 Deficiency A1

This deficiency is concerned with the use of nonstatistical considera-
tions in setting ID action limits. The panel and PNL made a distinction
between Deficiencies A1 and A2. Deficiency A1 was viewed as encompassing the
data analysis and evaluation that is the end product of statistical practice.
Deficiency A2 addresses the next step of determining ways in which the statis-
tical results will be used by the statistician's clients (the licensees and
NRC). Defining statistical decision rules is a statistical problem, but the
method of applying these rules is more of a policy/administrative problem.
Obviously, policy should give direction to formulation of decision rules, and
the range of possible decision rules can have an impact on policy formula-
tion. Close communication between client and statistician is required to
create the necessary understanding of both aspects.

Alternatives 1, 3, 4 and 5 are all special cases of Alternative 2.
Generally, "statistical decision theory" covers a range of increasingly
complicated statistical procedures characterized in a summary article by Carl
Bennett (1981). The commonly used (Neyman-Pearson) hypothesis testing
"decision theory" is the least complicated. The decision is based on stating
a null and an alternative hypothesis about a population parameter (e.g.,
[ID]), deciding on a test statistic, collecting data to estimate the value of
the test statistic and rejecting the null hypothesis if the test statistic
exceeds some prespecified value of the distribution of the test statistic
under the null hypothesis. This describes the traditional approach used in
safeguards to decide whether [ID] is excessive. The null hypothesis is that
the [ID] is zero, and the alternative is that it is not zero. The LEID
criterion provides the test statistic under the assumption that [ID] is
normally distributed with mean zero and population standard error sd(ID);
i.e., is the sample and population standard errors are assumed to be equal.
If the absolute value of the observed ID is greater than LEID=2sd(ID), then
investigative action is taken. This approach is intended to cause the
licensee no more than a 5% chance of investigating excessive IDs observed when
the population [ID] is actually zero. Given the population [SD[ID]], the
normal distribution assumption and a specified probability α (traditionally
α = 0.05) of deciding [ID] is greater than zero when, in fact, it is not, the
power of the test to detect a population [ID] of a specified value can be
calculated. In fact, by calculating the power for a number of alternative
values for [ID], the power function of the test can be graphed to resemble an
inverted normal curve over the possible values of [ID]. The ordinate is the
probability with a minimum of α at [ID] = 0 and maxima approaching unity

3.24
The "probability of alarming" is the power of the test, and the "false-alarm rate" is the probability that the licensee will have to conclude that [ID] is not equal to zero when, in fact, it is. Under this traditional procedure the "decision theory" problem reduces to a simple two-way table shown by Table 3.5.

**TABLE 3.5. Hypothesis-Testing Decision Table**

<table>
<thead>
<tr>
<th>Available Conclusions</th>
<th>[ID] = 0</th>
<th>[ID] ≠ 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ID] = 0</td>
<td>No Error, (1 - α) = Confidence Level</td>
<td>β = Probability of Missing a True Difference</td>
</tr>
<tr>
<td>[ID] ≠ 0</td>
<td>α = False-Alarm Rate</td>
<td>No Error, (1 - β) = Power, Probability of Alarming</td>
</tr>
</tbody>
</table>

The more complicated decision theory procedures would involve characterizing many possible states of nature by specific diversion scenarios, quantifying the social or monetary loss given that the strategy is successful, associating probabilities of frustrating each scenario by using specific safeguards strategies, and then maximizing the probability of minimizing the loss over all, or a subset of, scenarios and strategies. The two-by-two Table 3.5 above would be expanded to an n_1-by-n_2-by-n_3 table of probabilities for combinations of diversion scenarios, control strategies and losses. Depending on the kinds of information available or collected, these more complicated decision procedures have been called "game-theory", "optimization of linear or quadratic programming", or "risk analysis." Such procedures might be useful for setting ID action limits, but were considered beyond the scope of this study.

Alternatives 3, 4, and 5 depend on Table 3.5. Alternative 3 recognizes that setting ID action limits is necessary, but not sufficient, for statistically valid application of the accounting (measurement) data. Consequently, it suggests the use of hypothesis testing for other indices of statistical control as well. For example, process yields for the current period could be compared with those from previous periods, shipper-receiver differences could be investigated for their impact on ID, and components of ID could be subjected to hypothesis tests. In order to "control the Type 2 error" (having probability β in Table 3.5), as suggested in Alternative 4, a family of α-level tests needs to be theoretically defined, and that α-level test which gives the desired power (1-β) selected. This is because the α-level and the standard error of ID determine the allowable magnitude of ID, i.e., the critical value of ID at which there is 100(1-α/2) percent of the distribution of ID, centered on zero, left in each tail of the distribution. The Type 2 error rate is the probability β cut off by the critical value for the α-level test.
in the tails of the distribution centered on some alternative value. Call this value \([10(\beta)]\). Thus, \(\beta\) is a function of \(a\) and the common standard error of the distribution centered on zero and \(10(\beta)\). In order to reduce \(\beta\), i.e., increase the power \((1-\beta)\), the false-alarm rate \((\alpha)\) must be increased if \(10(\beta)\) and the standard error are fixed. This leads to the suggestion of Alternative 5, which is to develop a tradeoff strategy (implicitly between the NRC and the licensee) for setting \(a\) and \(\beta\) levels that will be acceptable.

Statisticians develop models so that the data collection can be efficiently directed and the results of the data analysis meaningfully evaluated. There are many ways for a statistician to analyze and interpret (evaluate) the data. Selection of specific tools is based on the goals of the data collections. Hypothesis testing is used, but the complexity of the alternatives to the null hypothesis that \(10 = 0\) is recognized. In line with the hypothesis testing versus statistical control concepts discussed in Section 3.2, the statistical goal of modeling is to determine what caused any significant estimated \(10\). The advantage of fuller modeling is that the search for causes can be better directed. The goal is to determine what the data tell us. For example, the data can tell us that the \(10\) is significant because process hold-up is low this period relative to other periods only if the modeling includes this capability.

The five alternatives to Deficiency A1 reflect a dichotomy between hypothesis testing and statistical control. Alternatives 1 (a and b), 4 and 5 emphasize hypothesis testing. Alternatives 2 and 3 allow for statistical control aspects. A basic assumption for the hypothesis testing approach is that the quantities tested are generated by measurement systems which are in a state of statistical control.

Alternative 1, proposed in SECY-80-514 (p. 2), has two parts. If Part 1a, to adopt the framework of hypothesis testing for evaluating \(10\), is accepted, then Part 1b logically (statistically) follows. No test of hypothesis should be applied, particularly those routinely applied, without consideration of the probability of detecting a specified difference (amount of material lost or diverted). This means that a power or operating characteristic (OC) curve should be developed for the test. Then "probability of alarming" can be determined for hypothesized differences from the null hypothesis that \(10 = 0\). In order to attain the required protection against a "specified amount" of difference, it may be necessary to change (increase) the false-alarm rate. In all of these considerations of false-alarm and detection rates regarding \(10\), the probabilities obtained depend on the models and estimators used in specific accounting systems. The results are useful for designing better safeguards systems, but the specific probabilities obtained may be far from reality since the statistical distribution of the estimator of \(10\) cannot be determined without many simplifying assumptions.

One problem with this approach in current practices is that it assumes that the simple alternative hypothesis to the null hypothesis that \(10 = 0\) is that a loss or diversion has occurred. As currently modeled, the alternative is not simple; the loss or diversion alternative has a number of competitors. The historical behavior of \(LEID\) as the statistic for testing the
significance of ID results in an alarm rate much greater than the nominal 5 percent rate predicted for the hypothesis test based on the action limit of LEID. Such a difference could easily be due to the number and diversity of the alternative hypotheses competing with loss of material as the assignable cause.

The causes of significant IDs (i.e., logical alternatives to the null hypothesis that ID is zero) can be categorized as follows:

1. Material Defects
   1.1 actual losses of material (theft or diversion, leaks, unobserved loss in filling or emptying)
   1.2 change in composition (evaporation, water absorption, oxidation, formation of nonreactive species)

2. Measurement Defects
   2.1 measurement or clerical mistakes (wrong scale factor, numerical transposition)
   2.2 measurement bias (including sampling bias)

3. Testing Defects
   3.1 an extreme observation from the distribution
   3.2 error distribution not adequately characterized

4. Modeling Defects
   4.1 failure to include important sources of error
   4.2 failure to model dependencies among sources of error.

Not all of these defects can be subjected to statistical tests. Some can even cover up a diversion or loss by causing a non-significant result. All have an impact on extensiveness of modeling. Another deficiency of the Alternative 1 approach is that the distribution of ID cannot be adequately characterized from the detailed model. This means that we do not know the functional form of the statistical distribution generated by the measurements and calculations which produce the estimated ID and its associated standard error of estimation. Years of experience have shown that the assumption that the estimated ID has a normal distribution with mean zero and standard deviation equal to the expected value of the estimated standard error of the estimated ID is not true, and may not even be an adequate approximation of the true situation. We have the testing defect situation 3.2 above: the error distribution is not adequately characterized. We cannot assign absolutely meaningful probabilities to false-alarm rates nor to detecting specified differences from the hypothesized ID of zero.
Some statisticians have found useful the intuitive assessment that the above normality assumption is not too far wrong. Besides, it is the only simple approach we have if constrained to developing a test criterion based on ID and LEID. Alternatives 4 and 5 are subject to the same problem of assuming probabilities that may have little basis in strict statistical analysis. The advantage of these approaches is that they provide comparatively simple decision rules. Administrators only have to look at one statistical test, and reviewers and inspectors only have to evaluate the procedure used to make that test.

Alternative 3 goes beyond the single test of ID. It allows for a spectrum of tests in the context of a fuller model for ID. The panel also envisioned the use of different degrees of action for different false alarm rates. A significant ID (at some nominal significance level) would initiate testing both the measurement-error part of the ID (effectively the current ID model), and the process-variability part. It would be a good statistical practice to consider these tests as indicators of loss of statistical control, and continue down the hierarchy of parameters and modeling levels until the level is reached where distributions can be meaningfully characterized. The investigation would not be totally statistical. For example, differentiating a leak from a diversion would require inspection of piping. But the statistics could indicate that the piping should be checked before the glove boxes.

This stepwise approach has the problem of complexity. There are many possible directions that can be taken. It would not be easy to prespecify a logical order for the investigation, but statistics could help give priorities to the process stages. The stepwise approach would suffer from lack of meaningful absolute probabilities, but relative probabilities should be comparable. That is, if the nominal relative probabilities of a defect in the piping and glove box are 0.93 and 0.85, respectively, the "true" probabilities should have the same order.

Statistical decision theory (Alternative 2) is a general approach which allows the use of costs of data collection and monetary or other quantifiable risks to design optimal (cost-wise) decision rules. This approach is complex and is not expected to result in better loss detection at the present state of ID variability knowledge.

Recommendation. PNL supports the view that a spectrum of tests and actions that monitor the state of statistical control of the nuclear material accounting system of the licensee should be utilized. (A "state of statistical control" means that the statistical parameters important to the system, e.g., the means and variances of the measurements and the processes affecting the variance of ID, are staying within their control limits.) These would serve the purposes of 1) demonstrating that the goals and requirements for nuclear material control and accounting, which are prescribed by the NRC (see Section 3.2.3.1 Deficiency A2), are being met; and 2) alerting the licensee when corrective action is necessary to locate and remove the cause of a loss of statistical control or an indication of an undesirable trend in the magnitude of ID and the variance of ID. The spectrum of tests and actions should be based on hypothesis tests, and both the power of the tests and the

3.28
probability of false alarms should be monitored. The licensee should establish action limits for specific remedial responses and maintain records of the test and action results as evidence of continuing statistical control over ID and its variance.

**Impact Assessment.** Alternative 1 would involve adapting statistical methodology to the specific needs of the facility and devising statistical procedures for the tests of hypotheses. This alternative would result in a one-time implementation cost to a licensee. Since the calculations for the hypothesis test are not more time-consuming than existing practices and no additional data are needed, the additional operating costs will be insignificant. The estimated one-time implementation costs are 20 person-weeks or $25K.

Alternative 2 would require use of considerably more complex methodology than Alternative 1, and the one-time implementation costs are estimated to be higher. The annual operating costs are, as with Alternative 1, not expected to be significantly different from the status quo. The estimated implementation cost is 27 person-weeks or $30K.

Alternative 3 involves the application of statistical methodology that is less complex than that used in Alternative 2 but will have similar one-time implementation costs due to the need to define more extensive testing protocols. In addition, we believe there would be an incremental operating cost to the licensee for additional computation and analysis. The estimated costs are 27 person-weeks or $30K implementation costs and 5 person-weeks or $6K annual operating costs. (a)

Alternatives 4 and 5 would have negligible incremental cost to the licensee over current practices.

3.2.3 Implementation Deficiencies (A2, B5)

3.2.3.1 Deficiency A2

This deficiency pointed out that "non-statistical considerations" are used to set alarm criteria. Alternative 1, proposed by SECY-80-514, is to use ID variability to set alarm criteria. Alternative 2, posed by the panel, is to use administrative limits developed "within an appropriate statistical framework." The panel did not think that the variability of ID should be the sole basis for alarm criteria. In fact, the variability of ID need not necessarily be used for setting alarm criteria, but it could be used to "alarm" as an indicator that investigative action should commence. Other data bases (e.g., the historical performance extractable from "White Book" data) could be used.

(a) This assumes six physical inventories/year. If the Reform Amendments for Category 1 licensees are adopted, the costs would be lower because of fewer inventories.
Recommendation. PNL supports Alternative 2; that is, administrative limits for ID should be developed and applied within an appropriate statistical framework. Administrative limits could be based on other than statistical considerations. Examples are controls on target amounts of loss that must be detected with high probability, minimum required actions for specified occurrences, and acceptable control limits for the variance of ID. The recommendation is that a statistical approach based on the variability of ID be used in conjunction with such controls. The preferred approach from a statistical standpoint is to utilize statistical hypothesis tests to determine whether ID limits are being achieved and whether the estimated ID and its standard error are in statistical control. If an alarm threshold is established on the basis of a specified power of detection of a hypothesized true loss, the variance of the ID should also be monitored to collect data for demonstrating that the variance being used in setting the alarm threshold is correct and, therefore, that the required power of detection is being achieved. Whether an action limit on the variance of ID is also imposed by regulation would be a policy decision by NRC. Three ways of estimating the variance of ID can be considered: 1) propagating the estimated variance of the measurements affecting a single ID value, 2) calculating the variance from historical observed IDs, and 3) pooling historical observed ID values.

Impact Assessment. Either Alternative 1 or 2 would involve estimated costs to the licensee nearly identical with Alternative 1 of Deficiency A1, which are $30K for implementation. There would be no significant additional operating costs.

3.2.3.2 Deficiency B5

This deficiency was concerned with the evaluation of "ID variability models" used by licensees. The two-part Alternative 1, given in SECY-80-514, stresses evaluation of model assumptions and the effects of modeling errors. It seems to be implied that NRC would provide guidance in how to do this evaluation and that the licensees would carry it out. The panel recommended Alternative 2, i.e., that a "third party" (similar to third-party measurement control audits) evaluate the facility variability models and their mode of application. The evaluations could be done by the NRC if adequate statistical expertise is available.

Recommendation. PNL recommends periodic independent evaluation of facility-specific ID variability models and their mode of application. This evaluation should be performed with the aid of a competent statistician having safeguards experience. If preferred by the NRC, it may be done by a third party, possibly as part of a requirement for an external audit of the licensee MC&A effectiveness. The advantage of the alternative is that it assures that competent statistical expertise will be enlisted to evaluate the licensee practices and recommend improvements in modeling and applications (parameter estimations) that are applicable in the specific facility. This approach will also accomplish the purposes of Alternatives 1a and 1b.
Impact Assessment. It is assumed for both Alternatives 1 and 2 that the licensees bear the full costs of analyzing and evaluating their ID variability models and performing the sensitivity tests. Also, it is expected that the exercise would be repeated whenever a change in the system occurs that could have important consequences that alter the model. For our analysis, we assumed this will occur every five years; therefore, it included an annual operating cost equal to one-fifth of one-time implementation costs.

The incremental cost to the licensee for Alternative 1 includes the cost of testing the assumptions used in the model and applying guidelines from the NRC for controlling modeling errors. The estimated incremental licensee cost associated with Alternative 1 is 18 person-weeks or $20K for one-time implementation costs and $4K annual operating costs.

Alternative 2 will require having a contractor or consultant review the licensee's MC&A system in detail, review the ID variability model, and determine whether the model is appropriate and correctly applied. The estimated costs per facility are $15K one-time implementation costs and $3K annual operating costs. The contractor cost estimates are for performing evaluations only and do not involve revising models developed by the licensee.
4.0 RECOMMENDED NOTATION

The notation recommended for use is based on the listing in ANSI N15.5- (1982 Draft), Table 3.4, and textual material in ANSI N15.16 (1974). Table 4.1 lists the statistical concepts and the notation to be used for each concept when the concept applies to a population, sample random variable, or to sample values. If the concept does not apply, nothing is entered in the column. Recommended notation is given for both typed final reports (R) and other (O) types of communication (drafts, computer printout, appendices). The limited set of symbols available to upper-case computer printers is used for these "other" types. Sample random variables are usually distinguished and discussed in theoretical development of estimators. Since such discussion would appear in more formal reports or detailed methodological appendices, the need for "other" notation would seldom occur, so none is suggested.

Several general rules exist for formal report notations:

1. Use upper-case English letters for random variables.
2. Use lower-case Greek letters for parameters.
3. Use lower-case English letters for observed values of random variables.
4. Use upper-case $E$ with brackets to designate the expected value of the bracketed random variable.
5. Subscript or append parenthesized notation to indicate the random variable or parameter intended when distinctions need to be made.

The following are modifications to these rules when only a limited set of printing characters is available:

6. Use brackets to designate population characteristics.
7. Use parentheses to designate sample values.
8. Substitute upper-case English letters for Greek letters and lower-case English letters.

Table 4.1 follows ANSI N15.5 Draft, Table 3.4, except that it does not carry the concept of a function of a random variable into the notation for mean and variance. Recognizing that a function of a random variable, or a set of random variables, is itself a random variable simplifies the notation. Since the use of $ID$ is fixed in the literature, lower-case "id" would be awkward, so rules for "other" notation are recommended at the end of Table 4.1.
### TABLE 4.1. Notation for Some Statistical Concepts

<table>
<thead>
<tr>
<th>Printer Type</th>
<th>Population</th>
<th>Sample Random Variables</th>
<th>Sample Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Univariate Random Variables</td>
<td>( X, Y, Z )</td>
<td>( X_1, X_2, \ldots, X_n )</td>
<td>( x_1, x_2, \ldots, x_n )</td>
</tr>
<tr>
<td>Parameter</td>
<td>( \theta, \mu, \sigma )</td>
<td>( T(X_2, X_2, \ldots, X_n), T_x )</td>
<td>( t(x_1, x_2, \ldots, x_n), T_x )</td>
</tr>
<tr>
<td>Degree of Freedom</td>
<td>( v )</td>
<td>( N-P )</td>
<td>( n-p(a) )</td>
</tr>
<tr>
<td>Distribution Function with Parameter</td>
<td>( F(x; \theta) )</td>
<td>( F(X_1, X_2, \ldots, X_n) )</td>
<td>( cdf(x_n) )</td>
</tr>
<tr>
<td>Corresponding Density Function</td>
<td>( f(x; \theta) )</td>
<td>( f(X_1, X_2, \ldots, X_n) )</td>
<td>( pdf(x_n) )</td>
</tr>
</tbody>
</table>

(a) \( p \) is the number of df lost in estimation; \( p = 1 \), for estimating the mean, in the calculation of the variance of a sample of size \( n \) from a population.
TABLE 4.1. (contd)

<table>
<thead>
<tr>
<th>Function of a Random Variable(b)</th>
<th>Population</th>
<th>Sample Random Variables</th>
<th>Sample Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>g(X)</td>
<td>g(Xn)</td>
<td>g(xn)</td>
</tr>
<tr>
<td>O</td>
<td>G(X)</td>
<td>G(X</td>
<td>M)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expected Value of a Function of a Random Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
</tr>
<tr>
<td>O</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean of a Random Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
</tr>
<tr>
<td>O</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variance of Random Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
</tr>
<tr>
<td>O</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Standard Deviation of a Random Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
</tr>
<tr>
<td>O</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bivariate Random Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
</tr>
<tr>
<td>O</td>
</tr>
</tbody>
</table>

(b) The identity function, g(X) = X is included.
<table>
<thead>
<tr>
<th>Printer Type</th>
<th>Population</th>
<th>Sample Random Variables</th>
<th>Sample Values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Covariance of Two Random Variables</strong></td>
<td>( \sigma_{XY}, \sigma_{X_1, X_2} )</td>
<td>( \text{Cov}(X_1, X_2) = \frac{\sum (X_{1i} - \bar{X}<em>1)(X</em>{2i} - \bar{X}_2)}{(n - 1)} )</td>
<td>( \text{Cov}(X_1, X_2) )</td>
</tr>
<tr>
<td><strong>Correlation Coefficient (product-moment) between two random variables</strong></td>
<td>( \rho, \rho(X_1, X_2) )</td>
<td>( \rho = \frac{(n-1)\text{Cov}(X_1, X_2)}{\sqrt{\sum (X_{1i} - \bar{X}<em>1)^2 \sum (X</em>{2i} - \bar{X}_2)^2}} )</td>
<td>( \rho, \rho(X_1, X_2) )</td>
</tr>
<tr>
<td><strong>Bias Associated with an estimator of ( \theta )</strong></td>
<td>( \theta, \theta(T; \theta) )</td>
<td>( \theta(T) = T_X - T_Z )</td>
<td>( \theta(t) - \theta(t) )</td>
</tr>
<tr>
<td><strong>Inventory Difference</strong></td>
<td>( \theta, \theta(T) )</td>
<td>( \theta(T) = T_X - T_Z )</td>
<td>( \theta(t) - \theta(t) )</td>
</tr>
<tr>
<td><strong>Standard Error of Estimated ID</strong></td>
<td>( \text{ID} )</td>
<td>( \text{SD}(	ext{ID}) )</td>
<td>( \text{sd}(	ext{ID}) )</td>
</tr>
</tbody>
</table>

(c) Here the random variable \( Z \) indicates data generated by measuring standards are to be used in the estimator \( T \).
5.0 REFERENCES


Bennett, C. A. 1981. "Introduction to Decision Analysis" ESARDA Lecture (To be published.)


APPENDIX

ALTERNATIVES TO CORRECT INVENTORY DIFFERENCE STATISTICAL TREATMENT DEFICIENCIES - A PANEL DISCUSSION
I. INTRODUCTION

On July 8 and 9, 1982, a meeting on the subject of the statistical treatment of inventory differences (ID) in nuclear material accounting was held at the Battelle Human Affairs Research Centers (HARC) facilities in Seattle, Washington. The meeting was attended by a group of people expert in the development and application of statistical methods to the problems of nuclear materials safeguards.

The specific intent of the meeting was to acquire opinions from the experts on alternative means of correcting a selected set of deficiencies in the statistical treatment of inventory differences as contained in the NRC regulations and regulatory guides. The deficiencies (commonly referred to as the "thirteen deficiencies") considered by the group had been previously identified through an NRC staff study mandated by the Commissioners. In addition to the identification of alternative remedies, the group of experts also provided suggestions as to how the alternatives could be evaluated.

This report is a summary of the meeting results. The report is comprised of five sections. The first section provides background information on the deficiencies and on a NRC-sponsored program at Pacific Northwest Laboratory (PNL), which is to be the immediate user of the meeting results. The second section describes the panel and its objectives as viewed prior to the meeting. The third section describes the alternative remedies to each deficiency as suggested by the panel. Each set of alternatives is accompanied by a brief narrative summarizing some of the panel's comments. The last two sections are devoted to the panel's views on the relative importance of the deficiencies and the attributes through which the alternatives can be evaluated and compared. The scope of this report is limited to material directly pertinent to PNL's program objectives. This report is not a proceedings in the sense of providing a summary of all discussions.

II. BACKGROUND

In March 1980, the NRC Commissioners requested that the Office of Management and Program Analysis (MPA) and the Office of Nuclear Material Safety and Safeguards (NMSS) jointly prepare a paper identifying possible changes to the current treatment of ID to make it more statistically valid. A paper—Report on the Statistical Treatment of Inventory Differences (SECY-BO-514)—was completed in November, 1980. It identified thirteen deficiencies in the areas of ID evaluation, determination of ID variability and ID definitions and modeling. The deficiencies, per SECY-BO-514, are tabulated below.
DEFICIENCY

A. ID Evaluation

A1 - The procedure for setting the current ID action limits reflects some non-statistical considerations and is not based on the framework of the statistical test of hypotheses. As a result, this procedure does not control the false alarm rate nor the probability of alarming and initiating an investigation when a specified amount of material is lost or diverted. In addition, the current procedure does not address evaluation of multiple-period inventory differences.

A2 - Current alarm criteria are based on non-statistical considerations and are set independently of the variability of ID.

A3 - The current hypothesis testing approach does not consider the trade-off between the false alarm rate and the probability of alarming and initiating an investigation when a specified loss or diversion has occurred.

B. ID Variability

B1 - The term "limit of error" has several inconsistent definitions.

B2 - The regulations and guides fail to distinguish among the true value of a measure of variability, the formula for computing it (the estimator) and the numerical value obtained by substituting observed data into the estimator (the estimate).

B3 - Not all sources of error are reflected in the calculation for LEID, nor otherwise substantially addressed in any guidelines.

B4 - Concepts and definitions of error which are in wide use in the industry may be overly restrictive and are not always consistent with definitions in other disciplines; this could be a source of confusion and could lead to inaccurate evaluation of ID variability.

B5 - Licensee ID variability models potentially contain weaknesses with respect to model structure and parameter estimation.

B6 - NRC does not routinely obtain data and other information necessary to independently verify all license LEID determinations.

C. ID Definition and Modeling

C1 - Inventory difference is referenced in official regulations as MUF. Lack of uniformity in terminology is a potential source of misunderstanding.
C2 - Regulations and guides fail to distinguish between true ID and observed ID. This distinction is essential for complete understanding and modeling, and for rigorous statistical treatment of ID.

C3 - The regulations and guides do not explicitly state the error terms in the ID model, and the model does not label random and fixed errors.

C4 - As defined, ID may reflect impacts of other than current-period operations. In particular, the regulations do not require the adjustment of ID to eliminate prior-period contributions caused by shipper-receiver differences, waste and scrap operations, and possible other operational factors.

Two of the deficiencies (A3 and B6) are being studied by the Office of Research and the Licensing Policy and Planning Branch, NMSS, respectively. The remaining deficiencies are the subject of an NMSS program with the Pacific Northwest Laboratory (PNL).

The PNL program is generally directed to the evaluation of the benefits that may accrue from removing the deficiencies and the costs associated with developing and implementing the required changes. The program elements as put forward in the Statement of Work are as follows:

Task 1 - Identification of the alternatives available to NRC to remedy the deficiencies identified in SECY-80-514.

Task 2 - Development of effectiveness criteria against which to compare the alternatives identified in Task 1. Development of evaluation techniques for estimating the cost or level of effort required to implement the alternatives is to be part of the effort on Task 2.

Task 3 - Performance of the value-impact analysis. This will be an assessment of the significance of each deficiency and the cost and benefits associated with correction. Included in the results from this analysis shall be an identification of the most important issues, those issues most easily resolved and those issues requiring long-range efforts to resolve.

The use of a recognized panel of experts was highly recommended to aid in identifying alternative remedies (Task 1) and for assisting in the value-impact analysis (Task 3).

The remainder of this report is devoted to the results of a meeting of experts conducted in conjunction with the above described program.
III. THE PANEL OF EXPERTS

On July 8 and 9, 1982, a group of experts met at the Battelle Seattle Research Center to discuss deficiencies in the statistical treatment of inventory differences in nuclear material accounting. The members of the panel of experts and their organizational affiliation are as follows:

<table>
<thead>
<tr>
<th>Name</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carl Bennett</td>
<td>Battelle Human Affairs Research Centers</td>
</tr>
<tr>
<td>Bob Easterling</td>
<td>Sandia National Laboratories</td>
</tr>
<tr>
<td>Yvonne Ferris</td>
<td>Rockwell International (Rocky Flats)</td>
</tr>
<tr>
<td>John Jaech</td>
<td>Exxon Nuclear Co.</td>
</tr>
<tr>
<td>Jim Johnston</td>
<td>Pacific Northwest Laboratory</td>
</tr>
<tr>
<td>Ralph Lumb</td>
<td>NUSAC, Inc.</td>
</tr>
<tr>
<td>Richard Mensing</td>
<td>Lawrence Livermore National Laboratory</td>
</tr>
<tr>
<td>Richard Post</td>
<td>San Jose State University</td>
</tr>
<tr>
<td>Jim Shipley</td>
<td>Los Alamos National Laboratory</td>
</tr>
<tr>
<td>Fred Tingey</td>
<td>University of Idaho</td>
</tr>
</tbody>
</table>

In addition, Rich Hooper (Battelle HARC) served as meeting moderator and Mark Mullen (PNL) served as rapporteur. The NRC/PNL program that provided the impetus for the meeting was represented by Mary Miller (NRC Program Monitor), Ken Byers (PNL Program Manager) and Bob Sorenson (Manager - Safeguards and Regulatory Analysis Unit/PNL). These individuals were not participants in the meeting.

The objectives of the meeting were described to the panel as follows:

1. Development of Alternatives. The panel will develop a set of alternatives appropriate for alleviating each of the eleven identified statistical deficiencies.

2. Identification of Areas of Applicability. As an aid in evaluating the significance of alleviating the deficiencies, the panel will be asked to identify the specific relevance of each deficiency to the licensees and to safeguards.

3. Identification of Effectiveness Attributes. The panel will be asked to suggest relevant attributes for evaluating the effectiveness of the alternatives in alleviating each deficiency.

4. Identification of Impacts. The panel will be solicited for ideas concerning techniques useful for estimating costs and level of effort required to implement the alternatives.

Generally, most of the meeting time (one and one-half days) was devoted to the development of alternatives. The remaining time was given to characterizing the alternatives in terms of value-impact attributes that may prove useful in choosing among them. Overall, it was an excellent meeting with extensive participation by all members of the panel.
IV. THE DEVELOPMENT OF ALTERNATIVES

The deficiencies and corresponding remedies suggested by the panel are presented here in the order that they were addressed by the group. The deficiencies identified under the heading "ID definition and modeling" were considered first followed by those under the heading "ID variability." The ID evaluation deficiencies were considered last. This is a natural arrangement where the deficiencies dealing with problem definition, analysis and evaluation were considered in that order. The importance of complete and well-defined models was a central issue throughout the meeting. This is viewed as an area of statistical methods can make a large contribution. Within each category, the deficiencies in terminology and definitions were considered ahead of those that concerned operational matters.

The order in which the deficiencies were addressed proved to be an important factor in the identification of some of the alternative solutions. Deficiencies are not independent; the manner of resolution of one deficiency may well affect the relative viability of alternative solutions to other deficiencies. In fact, the manner of resolution of one deficiency can affect the nature of problems that some of the other deficiencies were intended to address. This certainly complicates the value-impact analysis to follow in that some alternatives can only be evaluated conditioned on the implementation of specific remedies to one or more related deficiencies.

Included in the summary of alternatives put forward by the panel are the suggested solutions offered by the authors of SECY-80-514. This is done for completeness, and the SECY-80-514 suggestions are clearly identified as such. No meaning should be attached to the order in which the alternatives are listed.

DEFICIENCY C1

Inventory difference is referenced in official regulations as MUF. Lack of uniformity in terminology is a potential source of misunderstanding.

ALTERNATIVES

1. Replace the term MUF by ID in all regulations and official documents (SECY-80-514).

2. Replace ID with MUF to be consistent with world-wide usage.

3. Do nothing, but explain terms carefully.

Most of the discussion dealt with some of the history of the terms. Material accounting in the Manhattan Project emphasized accounting for valuable materials, not detection of diversion. Most of the people involved were accountants; hence, the term "material unaccounted for" (MUF) was natural. The
notions of an adversary and purposeful diversion came along later. There have been a succession of terms—MUF, UFM, BPID and ID. The name itself, be it MUF or ID, has no statistical implications; thus, the choice must be supported by non-statistical reasons. Public relations is seen as the principal concern. Among members of the panel, the term ID seemed to be generally accepted and commonly used. Once the discussion had moved on to other deficiencies, the term MUF was never used. This is not surprising in view of the training and experience represented by the panel members. This group's adaptability to changes in notation is certainly not representative of the safeguards community.

DEFICIENCY C2

Regulations and guides fail to distinguish between "true ID" and "observed ID." This distinction is essential for complete understanding and modeling, and for rigorous statistical treatment of ID.

ALTERNATIVES

1. Modify regulations and guides to be consistent with the ANSI Standard N15.16 definitions of "parameter," "estimator" and "estimate" for "true ID," "ID estimator" and "ID estimate," respectively (SECY-80-514).

2. Communicate the distinctions among "true value," "estimator" and "estimate" from fundamental statistical principles.

3. Reinforce the distinctions among "true value," "estimator" and "estimate" through suitable notation.

4. Develop the distinctions among "true value," "estimator" and "estimate" through education.

The discussion on this deficiency began with the acknowledgment that the communication of the distinctions among "true value," "estimator" and "estimate" to the statistical layman has haunted statisticians in all areas of application. Part of the difficulty is that it is not enough to talk about "true ID" and "observed ID." The distinction must specify models and components of ID. To this extent, deficiency C2 and deficiencies C3 and C4 are related. Several specific points were made:

- In practice, the problem is relatively minor because people are usually talking about observed ID.
- Choices for notation must accommodate the limitations of modern word-processing equipment.
- The language "estimated ID" is preferable to "observed ID."
• The distinction must be drawn from the concepts of safeguards and statistics and not referenced to some general description of statistical concepts alone.

• It is very important in communications to consistently reinforce the distinction.

The alternative remedies to deficiency C2 are not alternatives in the sense that one would be chosen. The panel strongly supported the implementation of all four as a remedy to C2.

**DEFICIENCY C3**

The regulations and guides do not explicitly state the error terms in the ID model, and the model does not label random and fixed errors.

**ALTERNATIVES**

1. Improve ID modeling by more explicit consideration of true ID and identification of random and fixed error components; modify regulations and guides accordingly (SECY-80-514).

2. Model ID such that $E(ID)=0$; describe errors in terms of the effects they have on the observations (i.e., errors described in the context of the analysis of variance).

3. Model ID such that $E(ID)=0$; the definition of errors (long-term, short-term systematic, random) should be relative to a particular data set.

4. Limit the terms in the ID models to those for which measurements are available; describe errors in terms of the effects they have on the observations (i.e., errors described in the context of the analysis of variance).

5. Limit the terms of the ID models to those for which measurements are available; the definition of errors (long-term, short-term systematic, random) should be relative to a particular data set.

This is probably the single most important deficiency, and the means of resolution chosen by the NRC will affect other deficiencies. As might be expected, this deficiency elicited considerable discussion. There were two fundamentally different notions with respect to modeling ID. The first was to require that the estimated ID be defined such that its expected value is zero (in the absence of any losses due to diversion). The second was to limit the terms in the ID model to those for which direct measurements are available. There was much discussion on the subject of modeling ID. A requirement that ID be modeled such that it has expectation zero would seem to require the treatment of all true effects, regardless of replication, as random. Thus, process deviations, measurement error—in fact, any deviation from the current
material balance model—are all chance variables. This leads to a simple hypothesis (null hypothesis), but it also leads to the most complex estimation problem. If the components in the ID model are limited to those for which direct measurements are available, then the estimation problem is straightforward but the null hypothesis is complicated. Many of the current evaluation problems derive from difficulties in developing and testing an appropriate hypothesis. Whatever route is chosen, it is mandatory that individual components in the ID model be described in terms of the sources of error that affect them and the propagation of the effects in computing the error in the estimated ID.

There was general agreement that the regulations and guides cannot realistically state error terms or even specific terms in the ID models. However, the need to clearly distinguish between random and fixed errors and between different components of true ID is well recognized. A suggestion was made that a clear checklist be developed for NRC reviewers to use in assessing a facility's models and calculations. The discussion of random and fixed errors led immediately to a restatement of some of the terminology difficulties addressed by deficiency B4. The notions of random and fixed effects per the analysis of variance do not, without additional language, allow a distinction between random and systematic errors as contained in J. L. Jaech's book, Statistical Methods in Nuclear Material Control (1973). There was considerable discussion of what is meant by "random" and "fixed" effects and a recognition that in safeguards application the distinction is often difficult to maintain in practice. Specific comments on the terminology issue are summarized in the narrative that accompanies deficiency B4. The AMASS methodology was briefly mentioned in connection with deficiency C3, but a detailed description was deferred (AMASS was described during the discussion of deficiency B3).

DEFICIENCY C4

As defined, ID may reflect impacts of other than current period operations. In particular, the regulations do not require the adjustment of ID to eliminate prior-period contributions caused by shipper-receiver differences, waste and scrap operations, and possibly other operational factors.

ALTERNATIVES

1. Modify regulations and guides to incorporate statistically acceptable procedures to correct for bias, shipper-receiver differences, and waste and scrap operations (SECY-80-514).

2. From an accounting viewpoint, the important thing is that the books balance; from a statistical viewpoint the important thing is that the accounts reflect what actually happened. The regulations should recognize the distinction. A requirement that ID be estimated such that $E(ID)=0$ resolves this deficiency.
The important point here was that regardless of the overall requirements on modeling ID, any prior-period adjustments should be done in a way that reflects what actually happened. This may, in fact, result in two sets of books: one that would be closed at the end of an accounting period where accounting variances are handled according to accounting principles and another which would remain available for making prior-period adjustments. The real issue seems to come down to the value and impact of requiring prior-period adjustments.

DEFICIENCY 81

The term "limit of error" has several inconsistent definitions.

ALTERNATIVES

1. Adopt the standard deviation as the measure of ID variability. Eliminate the use of the term "limit of error" (SECY-80-514).

2. Use standard error of the estimate as a measure of ID variability. Eliminate the term "limit of error." Find different words for the instances where "limit of error" is used in the context of evaluation.

3. Use estimated mean square error as a measure of ID variability. Eliminate the term "limit of error." Find different words for the instances where "limit of error" is used in the context of evaluation.

4. Use standard error of the estimate or estimated mean square error as a measure of ID variability. Eliminate the use of the term "limit of error" as indicating a measure of ID variability but retain the term where it is a defined interval for a specific evaluation problem.

The discussion began with a review of the history of the term "limit of error." The term is a holdover from Shewhart control charting that has come to be used to specify standards of quality of the overall measurement system. There was general agreement that either standard error of the estimate or estimated mean square error are appropriate measures of ID variability (being sure that the language reflects the distinction between true value and estimate referenced in deficiency 82). If the "limit of error" term is to be retained at all, its meaning should be confined to the designation of a specific interval and evaluation problem.

DEFICIENCY 82

The regulations and guides fail to distinguish between the true value of a measure of variability, the formula for computing it (the estimator) and the numerical value obtained by substituting observed data into the estimator (the estimate).
ALTERNATIVES

1. Modify the regulations and guides to define and distinguish among the true standard deviation, its estimator, and its estimate (SECY-80-514).

2. No alternative to that contained in SECY-80-514 was offered.

This deficiency and that described by C2 are conceptually equivalent. The need to emphasize communication remains.

DEFICIENCY B4

Concepts and definitions of error which are in wide use in the nuclear industry may be overly restrictive and are not always consistent with definitions in other disciplines; this could be a source of confusion and could lead to inaccurate evaluation of ID variability.

ALTERNATIVES

1. In the development of future guidance for licensees, eliminate concepts and definitions of error that are overly restrictive and confusing. In particular, investigate various approaches to the treatment of bias and its impact on ID variability, and clarify the notions of random and systematic error (SECY-80-514).

2. Use the terms "random effects," "residual effects" and "bias" in place of current terminology. Clearly specify sources of error.

3. Replace the terms "random" and "systematic" with "uncorrelated" and "correlated" errors and "bias."

4. Do not change the terminology but make sure that current terms are understood.

Discussion on various aspects of deficiency B4 took place at several points during the two days. The expected controversy regarding the terms "random" and "systematic" errors and, in particular, the terms "short-term" and "long-term" systematic errors did not develop. The discussion began with an attempt to define the problem with the current terminology and the extent of the problem.

The difficulty with the terminology seems to come down to the use of the word "systematic." Some people believe that, in many contexts outside of safeguards, the terms "systematic error" and "bias" are synonymous. Since bias (and, thus systematic error) is viewed as a constant, it does not have a distribution. If this assertion is accepted, then the term "systematic error variance" common in the safeguards literature is contradictory. It is important to note that among the members of the panel there was no confusion.
regarding the concepts of random and systematic errors and the fact that they propagate differently. The problem, where it exists concerns the labels attached to the concepts, not the concepts themselves. All agreed that the most important matter is a correct or proper analysis.

The discussion then turned to defining the extent of the problem. The assertion was made that the users in the safeguards community have grown up with and are comfortable with the current terminology. The difficulty seems to reside within a portion of the statistics community. This assertion was not challenged. The point was then made that while the word "effects" (per the analysis of variance) may be preferable, the effort required to change the practice of using the term "systematic" will be substantial and may be more trouble than it is worth.

Two alternatives for changing the terminology were described. The first draws from the language of the analysis of variance:

- "Random effect" would be used to designate any effect that is not known exactly and where the source of the effect can be clearly specified. The words "random effect," or "possible random effect" (over time), could be used in place of "systematic error" as long as the affected measurements are identified.

- "Residual effects" would be taken as the sum total of all remaining random effects where the source of the effects has not been identified.

- A "bias" would designate a known and constant effect -- known magnitude and which observations are affected.

The words "random effect" as used above need additional clarification. Many people view a random effect as an effect that is drawn from a distribution that can be characterized probabilistically in terms of the distribution. They would regard randomness or the lack thereof as an intrinsic property of the underlying physical process by which the effect was selected. They maintain that an effect either is or is not random, and knowledge about it does not change its basic nature. What happens in safeguards is that an effect that is not known exactly is often treated as if it were random, regardless of whether the underlying selection process is random or deterministic. The treatment of rounding errors is an example. The process of rounding a result to few significant figures is completely deterministic. For a given result, the rounding error is always the same. However, for purposes of error propagation it is convenient and empirically sound to treat rounding errors as if they were random, i.e., assign a "variance" due to rounding. The choice of a language that satisfies the philosophical needs and the empirically justifiable approaches to meeting practical needs is the fundamental problem.

The second alternative suggests language that, in the sense intended, is most familiar to people experienced with stochastic processes/time series. It does however, provide a clear alternative to the current terminology or that described above.
"Uncorrelated errors" would be used in place of "random errors."

"Correlated errors" would be used in place of "systematic errors."

"Bias" would be used in the same sense as above.

The terminology "random errors" and "related errors" was also suggested.

DEFICIENCY B3

Not all sources of error are reflected in the calculation for LEID, nor otherwise substantially addressed in any guidelines.

ALTERNATIVES

1. Improve guidelines to better model ID variability and to identify and characterize, to the extent possible, on a facility-specific basis, all contributors to the variability of inventory differences, including non-measurement errors (SECY-80-514).

2. The variance calculation that accompanies the modeling of ID such that it has expectation zero resolves B3. Specifically require the ID model to have a term for "process" variability where this term is carefully defined.

The discussion of this deficiency began with a brief description of the AMASS methodology and some of the results obtained through its application. The AMASS methodology is a stepwise linear model approach that attempts to partition the variability in ID observed at a facility over time into three components:

- variability due to measurement errors of all kinds
- variability due to the process
- a "catch-all" term that is intended to reflect losses from the process over time.

This analysis has been applied to several facilities with general conclusions that some facilities are dominated by process variability while others are dominated by measurement variability (stable process) and, to the extent that the loss term reflects true losses from the process, all facilities examined had non-negligible losses.

The general reaction to AMASS was that the concept was good, there was however, no attempt to acquire any sort of group endorsement regarding the general utility of AMASS per se or its value as a remedy to specific deficiencies. Discussion on specific methodological and application issues was limited to questions.
Other points made were that ID and LEID are based on different random variables—a distinction that is often overlooked—and the need to distinguish between changes in the ID variability models to better reflect sources of error and the proper treatment of the propagation of errors.

DEFICIENCY B5

Licensee ID variability models potentially contain weaknesses with respect to model structure and parameter estimation.

ALTERNATIVES

1a. Assure that the model assumptions are appropriately tested and that sensitivity analyses to assess the consequences of wrong assumptions in the modeling of the variability of random errors are performed (SECY-80-514).

1b. Provide guidelines to control the magnitude and effect of modeling errors, including those stemming from nonmeasurement contributors (SECY-80-514).

2. Utilize third-party opinion to evaluate facility-specific variability models and their mode of application.

The general position of the group was that remedies to deficiencies C3 and B3 would certainly improve facility-specific ID variability models but that there is really no substitute for making knowledgeable people a part of the evaluation process. The opinion was expressed that violations of normality assumptions are generally not of much consequence in this context.

DEFICIENCY A1

The procedure for setting the current ID action limits reflects some non-statistical considerations and is not based on the framework of the statistical test of hypotheses. As a result, this procedure does not control the false alarm rate nor the probability of alarming and initiating an investigation when a specified amount of material is lost or diverted. In addition, the current procedure does not address evaluation of multiple-period inventory differences.

ALTERNATIVES

1a. For ID evaluation, adopt the framework and terminology of the statistical test of hypotheses (SECY-80-514).

1b. Shift emphasis in the regulations and guides pertaining to statistical tests from false alarm rate to the probability of alarming and initiating an investigation when a specified amount of material is lost or diverted (SECY-80-514).
2. Use statistical decision theory.

3. Develop a spectrum of tests and accompanying actions that correspond to evaluations of measurement error, measurement error plus process variability, and ID.

4. Control the Type 2 error rather than the Type 1 error.

5. Control both types of error according to a defined tradeoff strategy.

Depending upon the remedies to deficiencies C3 and 33, a number of evaluation possibilities become available. Even within the formal framework of statistical tests of hypotheses, the appropriate procedures, action limits, error rates, and the like, are dependent on NRC objectives. There was general agreement that confining the evaluation solely to statistical tests of hypotheses involving ID where either the Type 1 or Type 2 error is controlled is too narrow an approach. A battery of tests with the flexibility to incorporate tradeoffs in selecting Type 1 and Type 2 errors and the flexibility to associate different action levels with different decisions (involving components of ID variability as well as ID) was promoted as being useful to the NRC. A decision-theoretic approach to evaluation can accommodate a more extensive specification of the formal decision problem.

Two other specific points were made. Multiple tests involving IDs from successive periods are complicated by the fact that the tests are not independent, which results in a significance level different from that selected for each test (analogous to the multiple comparisons problem from the analysis of variance). The second point was that alternative 1b from SECY-80-514 may have tremendous implications for plant operators.

**DEFICIENCY A2**

Current alarm criteria are based on non-statistical considerations and are set independently of the variability of ID.

**ALTERNATIVES**

1. Apply statistical principles to the construction of the test of hypotheses. In particular, establish alarm thresholds (or ID action limits) based on ID variability (SECY-80-514).

2. Develop and apply, within an appropriate statistical framework, administrative limits.

The discussion here was generally directed at the question: What is wrong with administrative limits? Is it really a deficiency that action limits reflect non-statistical considerations? These are policy questions that the NRC must resolve. However, there are issues of a practical significance in establishing action limits that go beyond statistical considerations.
argument was not that the setting of action limits should be devoid of statistical realities, but rather that action limits derived purely from statistical considerations may be too narrow.

V. RELEVANCE OF THE DEFICIENCIES

The panel was asked to evaluate the significance of alleviating each deficiency to the licensees and to safeguards. This exercise ran into immediate problems. The panel contended that the deficiencies are not independent and that there is a natural order in which the deficiencies would be addressed irrespective of individual impacts on licensees and safeguards. Further, the means chosen by the NRC to alleviate a deficiency is a crucial and, at this point, unspecified element in any evaluation of the significance of removing a deficiency. The alternative chosen to alleviate some of the deficiencies (e.g., C3) will alter the nature of other deficiencies. With the conclusion that it is not possible to assess the significance of individual deficiencies in isolation, the discussion turned to methods of grouping and ordering the deficiencies as a way of aiding the PNL value-impact analysis.

The deficiencies were grouped as follows:

(1) Deficiencies C1, C2, B1, B2 and B4 refer to terminology or definitional problems. To some extent, they can be dealt with independent of the other deficiencies.

(2) Deficiencies C3, B3 and A1 refer to analysis problems and, as such, have an impact on determining the correct action.

(3) Deficiencies A2 and B5 refer most directly to implementation problems, and it is here that the greatest impacts are likely to occur.

The correction of the definitional deficiencies does not affect the remaining deficiencies. However, the analysis and implementation deficiencies could, presumably, be more consistently dealt with if the definitional deficiencies are handled first. If this is the case then an "importance" ranking of remaining deficiencies is

\[ C3, B3, A1, A2, B5 \]

Deficiency C4 was assumed to be remedied by the alternative chosen to alleviate C3.
The resolutions to deficiencies C3 and B3 were seen as having the greatest potential to affect safeguards positively. The action levels selected as part of the remedy to deficiency A2 were judged to potentially have the greatest impact on the licensees.

VI. VALUE-IMPACT ATTRIBUTES

The panel made several suggestions regarding value and impact attributes of the alternatives that may prove useful to the PNL analysis. The value attributes were discussed in the context of effectiveness to safeguards:

1. Technical effectiveness as measured by such things as false alarm rate, power of the tests, flexibility of response, increased feedback, nature and extent of assumptions, etc.

2. Effectiveness in motivating individuals responsible for implementing material accounting systems. Improved understanding often results in a willingness to do things correctly. In addition, there are communication benefits in dealing with the public.

The impact or cost attributes were generally discussed in terms of who bears the cost:

- costs to the NRC
- costs to the licensees
- costs to other parties (DOE, international community, etc.).

Costs were viewed as being measurable in dollars or effort (man-years).

Some specific comments regarding value-impact attributes of the alternatives were made with respect to the deficiencies grouped as described in the previous section.

Definitional Deficiencies (C1, C2, B1, B2, B4)

- Most costs to the NRC are one-time costs involved in changing a vast amount of paperwork. A distinction between educational costs and implementation costs may prove useful.

- The bulk of the total costs will be borne by the NRC. The cost to licensees will be for review, comment and language changes in procedures guides.

- There will be costs associated with disparities that may develop with DOE and IAEA practice.

- The point was made that gradual changes made in concert with other changes to the regulations and regulatory guides would be much less expensive than attempting to make the modifications all at once.
• There will be communication and public relation benefits in correcting language inconsistencies.

• The motivational benefit that accompanies improved understanding among all parties is the largest benefit.

• The suggestion was made that the metric conversion problem might provide some useful analogies.

### Analysis Deficiencies (C3, C4, B3, A1)

• The alleviation of deficiency C3 was judged as most important in terms of the potential impact on the technical effectiveness of safeguards.

• The alternative remedy to C3 that would require that ID be modeled such that $E(ID)=0$ would be very expensive to implement. The benefit would be largely reckoned in terms of more effective safeguards and fewer investigations.

• There is a positive PR impact when the ID's are kept within limits a proper fraction of the time. A remedy to C3 (and B3) will result in fewer investigations.

• Proper modeling will inevitably have an impact on how the process is operated (feedback).

• Correct modeling, flexibility of response, power of tests, timeliness, false alarm rate, extent of feedback and the number of assumptions were all seen as measures of technical effectiveness.

### Implementation Deficiencies (A2, B5)

• The impact on the NRC to conduct the various reviews and the associated accommodations required of the licensees were seen as major cost elements.

• False alarm rates, action levels and the decision of which type of error to control in hypothesis testing are implementation attributes.

The panel recommended the following: in the value-impact evaluation, considerable care should be given to the distinction between those bearing the cost and those receiving the benefit.
### DISTRIBUTION

#### OFFSITE

<table>
<thead>
<tr>
<th>No. of Copies</th>
<th>U.S. Nuclear Regulatory Commission</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Division of Technical Information</td>
</tr>
<tr>
<td></td>
<td>and Document Control</td>
</tr>
<tr>
<td></td>
<td>7920 Norfolk Avenue</td>
</tr>
<tr>
<td>20</td>
<td>Bethesda, MD 20014</td>
</tr>
</tbody>
</table>

|                     | R. Miller                         |
|                     | U.S. Nuclear Regulatory Commission |
|                     | Office of Nuclear Regulatory      |
|                     | Research                          |
|                     | Washington, D.C. 20555            |

|                      | R. Easterling                     |
|                      | Sandia National Laboratories      |
|                      | Division 1223                     |
|                      | P.O. Box 5800                     |
|                      | Albuquerque, NM 87185             |

|                     | Y. Ferris                         |
|                     | Rockwell International            |
|                     | Rocky Flats Plant                 |
|                     | P.O. Box 464                      |
|                     | Golden, CO 80401                  |

|                      | J. Jaech                          |
|                      | Exxon Nuclear Company, Inc.       |
|                      | 600 - 108 Avenue N.E.             |
|                      | Bellevue, WA 98004                |

|                      | R. Lumb, President                |
|                      | NUSAC, Inc.                       |
|                      | 1650 Samuel Morris Drive          |
|                      | Reston, VA 22090                  |

|                      | R. Mensing                        |
|                      | Lawrence Livermore National       |
|                      | Laboratory                        |
|                      | P.O. Box 808                      |
|                      | Livermore, CA 94550               |

|                     | R. Post                           |
|                     | Math and Computer Science Department |
|                     | San Jose State University         |
|                     | San Jose, CA 95192                |

|                     | J. Shipley                        |
|                     | Los Alamos Scientific Laboratory  |
|                     | P.O. Box 1663                     |
|                     | Los Alamos, NM 87545              |

|                     | F. Tingey                         |
|                     | University of Idaho               |
|                     | P.O. Box 778                      |
|                     | Idaho Falls, ID 83401             |

#### ONSITE

<table>
<thead>
<tr>
<th>37 Pacific Northwest Laboratory</th>
</tr>
</thead>
</table>

| K. R. Byers                      |
| R. J. Brouns                     |
| J. W. Johnston                   |
| C. A. Bennett                    |
| F. P. Roberts                    |
| M. L. Mullen                     |
| L. C. Davenport                  |
| R. F. Eggers                     |
| B. W. Smith                      |
| R. M. Fleischman                 |
| N. L. Harms                      |
| R. J. Sorenson                   |
| Technical Information (5)        |
| Publishing Coordination (2)      |

Distr-1
This document presents an analysis of alternatives to correct deficiencies in the statistical treatment of inventory differences in the NRC guidance documents and licensee practice. Pacific Northwest Laboratory's objective for this study was to assess alternatives developed by the NRC and a panel of safeguards statistical experts. Criteria were developed for the evaluation and the assessment made concerning the criteria. The results of this assessment are PNL recommendations and are intended to provide NRC decision makers with a logical and statistically sound basis for correcting the deficiencies.