

An Analysis of Chemical Stockpile Emergency Preparedness Program Exercise Results

Volume 2: Preliminary Evaluation and Analysis of CSEPP Exercise Database

**Decision and Information
Sciences Division
Argonne National Laboratory**

Operated by The University of Chicago,
under Contract W-31-109-Eng-38, for the
United States Department of Energy

An Analysis of Chemical Stockpile Emergency Preparedness Program Exercise Results

Volume 2: Preliminary Evaluation and Analysis of CSEPP Exercise Database

by D. Wernette and K. Lerner

Decision and Information Sciences Division,
Argonne National Laboratory, 9700 South Cass Avenue, Argonne, Illinois 60439

June 1998

Work sponsored by U.S. Department of the Army, Chemical and Biological Defense Command

Prepared by:

Decision and Information Sciences Division Argonne National Laboratory Argonne, Illinois 60439-4832 Telephone (630) 252-5464 <http://www.dis.anl.gov>

Prepared for:

U.S. Army
Chemical and Biological Defense Command Aberdeen Proving Ground, Maryland 21010-5423

Disclaimer

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, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

CONTENTS

ABSTRACT	1
1 INTRODUCTION	1
1.1 Description of the Database	2
1.2 Scope and Organization of this Report	2
2 METHODOLOGY	4
2.1 Intercoder Reliability Methodology	4
2.2 Multivariate Analysis Methodology	6
2.3 Limitations on Analysis	7
3 FINDINGS AND ANALYSIS	8
3.1 Intercoder Reliability Findings and Analysis	8
3.2 Multivariate Exploratory Analysis and Findings	9
3.2.1 Characteristics Statistically Associated with Below-Average Outcomes	11
3.2.1.1 Association with the Tab Variable	11
3.2.1.2 Association with Objective Variable	13
3.2.1.3 Association with Site Variable	15
3.2.1.4 Summary of Characteristic Association Findings	17
3.2.2 Characteristics of Exercises with Concentrations of Below-Average Outcomes	20
3.2.2.1 Selection of Exercises -- Site Distribution	21
3.2.2.2 Tab Analysis	22
3.2.2.3 Objective Analysis	22
3.2.3 Temporal Analysis of Below-Average Outcomes	24
3.2.3.1 General Trends	24
3.2.3.2 Selected Pairs of Consecutive Exercises	27
4 CONCLUSIONS AND ADDITIONAL POSSIBLE ANALYSES	31
4.1 Intercoder Reliability Findings and Potential Follow-up Studies	31
4.2 Multivariate Analysis Findings and Potential Follow-up Studies	32

TABLES

1 Mapping of Outcome Measures from Old and New Classifications.....	7
2 Multiplicity of Codings.....	9
3 Accuracy of Coding, by Coder.....	9
4 Accuracy of Coding, by Objective.....	10
5 Cross-Tabulation of Report Tab and Outcome for Time 1 Period.....	12
6 Cross-Tabulation of Report Tab and Outcome for Time 2 Period.....	12
7 Cross-Tabulation of Objective and Outcome for Time 1 Period.....	14
8 Cross-Tabulation of Objective and Outcome for Time 2 Period.....	16
9 Cross-Tabulation of Site and Outcome for Time 1 Period.....	18
10 Cross-Tabulation of Site and Outcome for Time 2 Period.....	19
11 Cross-Tabulation of Report Tab by Outcome for 1992, BCA Exercise.....	21
12 Percentages of Total Below-Average Outcomes, by Year and Site.....	22
13 Distribution of Below-Average Outcomes by Tab for Selected Exercises.....	23
14 Distribution of Below-Average Outcomes by Objective for Selected Exercises in Time 1 Period.....	24
15 Distribution of Below-Average Outcomes by Objective for Selected Exercises in Time 2 Period.....	24
16 Numbers of Below-Average Outcomes, by Site and Year, 1992-1996.....	26
17 Percentages of Below-Average Outcomes, by Site and Year, 1992-1996.....	26
18 Below-Average Outcomes by Tab for 1992-1993 BCA Exercises.....	28
19 Below-Average Outcomes by Objective for 1992-1993 BCA Exercises.....	28
20 Numbers, Percentages, and Changes in Below-Average Outcomes by Tab and Objective for Selected Pairs of Exercises.....	30

**AN ANALYSIS OF CHEMICAL STOCKPILE EMERGENCY PREPAREDNESS
PROGRAM EXERCISE RESULTS**

**VOLUME 2: PRELIMINARY EVALUATION AND ANALYSIS OF
CSEPP EXERCISE DATABASE**

by

D. Wernette and K. Lerner

ABSTRACT

This study investigated the quality and usefulness of the information in the Chemical Stockpile Emergency Preparedness Program (CSEPP) exercise database. It incorporates the results of two separate analytical efforts. The first effort investigated the process of assigning standardized codes to issues identified in CSEPP exercise reports. A small group of issues was coded independently by each of several individuals, and the results of the individual codings were compared. Considerable differences were found among the individuals' codings. The second effort consisted of a statistical multivariate analysis, to investigate whether exercise issues are evenly distributed among exercise tabs, sites, and objectives. It was found that certain tabs, sites, and objectives were disproportionately associated with problem areas in exercises. In some cases, these problem areas have persisted over time, but in other cases they have undergone significant shifts over the time span of the investigation. The study concludes that the database can be a useful resource for analyzing problem areas and setting priorities for CSEPP program resources. However, some further analyses should be performed in order to more fully explore the data and increase confidence in the results.

1 INTRODUCTION

The Chemical Stockpile Emergency Preparedness Program (CSEPP) has been evaluating emergency response exercises at the eight Chemical Stockpile storage sites since 1992. The results of these exercises have been entered into a user-friendly database, described in *An Analysis of Chemical Stockpile Emergency Preparedness Program Exercise Results, Vol. 1: The CSEPP Exercise Results Database ("Analysis, Vol. 1 ")*. The purpose of the database is to provide a method for CSEPP managers to track and analyze the exercise results.

To provide a reliable basis for analysis, the database must meet two conditions. First, the data entered into the database must be of high quality in terms of accuracy and reliability. Second, it must be possible to analyze the data and extract meaningful results that will be useful to CSEPP managers in making policy and budgetary decisions. This report presents the results of a preliminary investigation of these two conditions.

1.1 DESCRIPTION OF THE DATABASE

The CSEPP exercise database contains information about all the issues that were identified in the exercise reports reviewed. For each issue, the database contains such basic information as the site and year of the exercise, the type of issue (strength or weakness, with weaknesses subdivided into various categories), and which tab of the exercise report it appeared in (i.e., whether the issue pertained to on-post, off-post, or joint activities).

In addition, to classify the issues, CSEPP and contractor staff used a standardized list of substantive topics that was based on the CSEPP exercise evaluation form. Each issue has been "pigeonholed" according to this form, at three successively finer levels. The first level is the applicable exercise objective element. Each objective is then divided into more specific "criteria," and each criterion into yet more specific "functional components." A table of the objectives, criteria, and functional components is provided in Appendix A of *Analysis, Vol. 1*. There are over 300 different functional components.

The process of classifying issues according to the categories described above is referred to as "coding." A given issue may be assigned one or more codes, depending on how many separate points it contains.

1.2 SCOPE AND ORGANIZATION OF THIS REPORT

The first topic investigated in this report is the reliability of the coding process. Since the process of assigning codes to an issue involves the application of judgment on the part of the coders, it is important to determine how well the judgments of individual coders agree for a given issue. Does the coding process represent a reliable system for classifying exercise issues, or does it reflect the individual biases of the persons doing the coding? This issue is addressed below under the heading of "intercoder reliability" (Section 2.1). Considerable differences were found in the way that particular individuals classified particular issues.

- 1 The CSEPP standard exercise objective elements are found in Appendix C to the CSEPP Exercise Program Guidance. As an example, objective element 1.3 is "Facility Activation."

The second topic investigated in this report involves the relationships among the basic pieces of information recorded about each issue: the site, type of issue, tab, etc. The purpose of this investigation is to determine whether there are meaningful patterns in the data that can be used to identify particular sites, areas, or topics that are consistently weak or strong in the exercise reports. Significant relationships were found that may be of interest for policymakers in the CSEPP. However, these results are preliminary in two respects. First, not all aspects of data quality are examined and evaluated. Second, the analytical techniques employed are limited to those of most direct relevance. To cite one example, exercise issues were taken "as-is"; no attempt was made to investigate or compare whether the exercise evaluation process itself had been consistent from site to site or over time. This and other limitations of the study, as well as potential remedies for them, are addressed in the final section of the report.

The report is organized in four sections. In Section 2, we describe the methods employed in evaluating intercoder reliability for some elements of the data, and in initially analyzing the relationships between variables in the data. Section 3 presents the results of the intercoder reliability evaluation and of the analysis of relationships between the variables. Section 4 addresses possible additional steps to improve both the quality and usefulness of the data for the CSEPP.

2 METHODOLOGY

2.1 INTERCODER RELIABILITY METHODOLOGY

Intercoder reliability was examined by conducting a controlled test of the coding process, in which several persons coded the same issue independently, and their selections were recorded for later comparison. The controlled test was conducted as part of a coding session at the CSEPP Office in Edgewood, Maryland, during the week of October 6-10, 1997. During the coding session, a group of CSEPP and contractor personnel met to encode the issues identified in several 1996 and 1997 CSEPP exercises. Each issue was ultimately encoded according to a consensus of the group. All told, several hundred issues were encoded during the session. Of these, 24 issues were included in the controlled test.

For each issue included in the test, each person in the group (referred to as "coders") developed and recorded his or her own coding independently. After they had done so, the group discussed these "candidate" codings and arrived at a consensus coding for each issue, which was recorded and included in the database. Efforts were made to ensure that the issues selected for the test represented a reasonable cross section of issues generally: they were drawn from various exercise reports from various sites, from all three exercise tabs, and included both strengths and weaknesses.

The group of coders also represented an informal cross section of the community of possible end users of the database, including both Army CSEPP and contractor personnel. It included personnel with a variety of experience levels, ranging from those who were familiar with the database coding process, to those familiar with CSEPP but not with the database, to those completely unfamiliar with CSEPP.

Each coder's independent codings were collected and entered into a spreadsheet program² for comparison with the final (consensus) codings. The results of this analysis are presented in Section 3.1. For each coder, two numbers were derived: multiplicity of codings and accuracy of codings. These terms are explained below:

- *Multiplicity of Codings.* The coding process allows for the fact that a given issue, as stated in an exercise report, may contain more than one element or point. For example, an issue related to communications problems experienced by medical personnel may contain a communications equipment component and a medical services component. Thus, the issue may properly be assigned two different codes. Of the issues included in the coding test, the final codings ranged from a single code number up to four different ones. The term

2 Microsoft® Excel™

"multiplicity" of coding refers to the average number of codings arrived at per issue.

- *Accuracy of Codings.* The term "accuracy" is used here to refer to the level of agreement between a particular coder's coding choices and the final coding(s) for a particular issue. For example, if Coder A assigned two codes to an issue, and there were two final codes for that issue of which only one matched one of Coder A's choices, then that would be characterized as a 50% accuracy rating for Coder A for that issue. The accuracy ratings presented for each coder thus represent the average level of agreement between their codings and the final codings for those issues. Separate accuracy ratings were developed for each level of the coding process: objective, criterion, and functional component.³ In reviewing these ratings, it should be noted that the element/criterion/functional component levels are "nested" in the following sense: a mismatch at one level implies that all downstream levels also do not match. For example, if a coder's "element" choice does not match the final coding, then neither the criterion nor the functional component choice will match either. If the element matches but the criterion does not match, then the functional component will not match. Conversely, a match at the functional component level implies that all upstream levels (criterion and element) also match. Thus, reading left to right across the accuracy tables (Tables 3 and 4), the match percentages must always either stay the same or drop; they can never increase.

In addition to the numbers computed for the coders, we also compiled accuracy ratings for each objective. These numbers represent the overall accuracy of codings for issues related to each CSEPP objective. For example, of the issues included in the test, three were related to Objective 1 (Initial Alert and Activation). For those three issues, the codings of all of the individuals were compiled to yield an overall accuracy rating for codings related to Objective 1. This was done in order to determine whether some objectives were inherently harder to code than others.

- 3 The process of computing the accuracy ratings was somewhat complicated by the fact that the number of codes assigned to a given issue was not always the same between an individual coder and the final coding. For example, Coder A may have assigned only one code number to a given issue, whereas the final coding for that issue may have included two numbers. Where an individual coder assigned fewer codes than the final, they were considered to have "missed" the additional codings; in the example, Coder A's accuracy would be either 50% (1 of 2) or 0% (if the one code number did not match either of the finals). Where individual coders assigned a greater number of codes to an issue than the final, their codes were sorted according to how closely they matched the final code(s). The best matches were counted, and then the "extra" codings were counted as "misses." For example, if Coder A assigned three codes to an issue where the final coding was singular, the coder's accuracy would be 33% at best.

2.2 MULTIVARIATE ANALYSIS METHODOLOGY

Data from the CSEPP Exercise Database were transferred to a statistical package (SPSS) for purposes of analysis. Analyses were then performed on the relationships between the following issue characteristics (or "variables"):

- Site (the relevant chemical installation and year of the exercise).
- Tab (whether the issue concerns on-post, off-post, or joint activities).
- Objective (the applicable CSEPP exercise objectives).
- Outcome (strengths vs. weaknesses).

The tables in Section 3.2 summarize the findings of this analysis. For example, Table 5 shows the relationship between tab and outcome for exercises in 1992 and 1993; it shows which tabs had more strengths than weaknesses and vice versa.

Unlike the intercoder reliability analysis, which dealt with a small subset of the issues in the database, all of the data were used in the multivariate analyses.

For some of the multivariate analyses, the database was essentially divided into two parts, and each part was analyzed separately. This was done because in 1994, changes were introduced in the coding of two of the key variables: the outcome measure and the exercise objectives. Issues from exercises held in 1992 and 1993 were classified according to the "old" outcomes and objectives, and issues from exercises from 1994 on were classified according to the "new" schemes.⁴ While there is considerable overlap between the old and new schemes, it is not a simple one-to-one mapping; for example, in the old scheme there was one medical objective, whereas in the new scheme there are separate objectives for medical transport and hospital treatment. Therefore, one cannot directly compare, say, the distribution of strength findings among objectives for exercises held in 1993 and 1995.

The main differences between the old and new classification schemes are as follows:

1. Under the old scheme, the outcome measure could take one of six values, listed from relative worst to best: Area Needing Remedial Action (ANRA); Area Needing Correction (ANC); Area Needing Improvement (ANI); Issue (ISS); Observation (OBS); and Strength (STR). The newer system is simpler,

4 Actually, two of the exercises conducted in 1994 were evaluated according to the old system: the exercises at Deseret Chemical Depot and the Blue Grass Chemical Activity.

employing only three categories: finding (FND), observation, and strength. For purposes of our analysis, we lumped all outcomes into three categories, labeled "below average," "average," and "above average." Table 1 shows how old and new issue classifications were mapped into these categories.

2. As mentioned, the objectives were changed. The number of objectives was reduced from 18 to 15, with a somewhat different mix of activities among the objectives from the old to the new.

The "Site" and "Tab" variables remained the same across the entire time span represented in the database (i.e., they did not change in 1994.)

2.3 LIMITATIONS ON ANALYSIS

As noted in the report title and introductory section, the findings and analyses presented below should be viewed as preliminary in nature. Section 4 contains a discussion of additional analyses that could be performed in order to confirm these findings and produce further results.

**TABLE 1 Mapping of Outcome Measures
from Old and New Classifications**

Old	New	This Study
ANRA, ANC	Finding	Below Average
ANI, ISS, OBS	Observation	Average
STR	Strength	Above Average

3 FINDINGS AND ANALYSIS

3.1 INTERCODER RELIABILITY FINDINGS AND ANALYSIS

The findings of the intercoder reliability study are presented in Tables 2-4 below. Table 2 presents the average multiplicity of codings for each individual coder and the final codings. Table 3 presents the accuracy numbers for each coder, for each level of coding (element, criterion, and functional component). These numbers represent the degree of agreement between each coder's individual codings and the final codings. Table 4 presents the average level of accuracy for all coders, for issues related to each CSEPP objective. Analysis of these results is included after each table.

Table 2 Analysis. The number of final codings assigned to a particular issue ranged from one to four. On average, 1.66 final codes were assigned per issue. There were rather large differences among the coders in terms of their average number of codes per issue; some only rarely assigned more than one code to an issue, whereas others were relatively close to the final numbers of codings. The final codings were more multiple than any individual's codings. This may be explained as a reflection on the process of developing the final consensus codings: in considering the various candidates proposed by the individual coders, the group tended to be inclusive.

As shown in the table, the number of issues coded varied for the individual coders. During the coding session, particular coders occasionally had to be absent for brief periods to attend to other tasks. Thus, they missed coding some of the test issues.

Table 3 Analysis. The percentages presented in this table represent the chance that a given code number, assigned by an individual coder after reading an issue in an exercise report, would agree with the final code number assigned after the group had reached consensus. For the complete coding (out to the functional component level), that chance was somewhat less than 1 out of 2, on average. Most of the coders had about a one-third to one-half chance of hitting the final coding, with one outlier achieving a two-thirds rate. Somewhat higher success rates were experienced at the criterion and element levels. For example, an individual coder had a 61% chance, on average, of having his or her assignment of an element code to an issue agree with the final code.

These results indicate that the process of assigning codes to an exercise report issue is subject to differences in individual approach and perspective.

TABLE 2 Multiplicity of Codings

Coder	No. of Issues Coded	Total No. of Codes Assigned	Average No. of Codes per Issue
Coder A	19	23	1.35
Coder B	23	36	1.56
Coder C	24	38	1.58
Coder D	16	17	1.06
Coder E	23	28	1.22
Final	24	40	1.67

TABLE 3 Accuracy of Coding, by Coder (%)

Coder	Accuracy of Element Codings	Accuracy of Criterion Codings	Accuracy of Functional Component Codings
Coder A	54	49	49
Coder B	64	55	43
Coder C	79	71	69
Coder D	42	35	35
Coder E	56	44	36
Totals	61	52	47

Table 4 Analysis. Of the 15 CSEPP exercise objectives, 12 were represented by at least one issue in the coding test. The full-code accuracy rates compiled by objective range from a low of 14% to a high of 78%. This indicates a significant variation among objectives; the data appear to suggest that issues related to the first several objectives were "harder to code" than those related to the last several objectives. There may be features of the wording or scope of those objectives that make them more prone to disparate interpretations. However, it should be noted that the study sample size was relatively small for this aspect of the analysis (only one to three issues per objective).

3.2 MULTIVARIATE EXPLORATORY ANALYSIS AND FINDINGS

The multivariate analysis that follows is organized into three topical subsections. Section 3.2.1 addresses the question of whether some tabs, sites, or objectives are inherently more

TABLE 4 Accuracy of Coding, by Objective

Objective	No. of Issues	No. of Codes	Accuracy of Element Codings (%)	Accuracy of Criterion Codings (%)	Accuracy of Functional Component Codings (%)
1. Alert/activate (initial)	3	18	61	44	44
2. Hazard assessment	1	7	29	14	14
3. Protective action decisions	0	--	--	--	--
4. Command/control	2	11	45	36	27
5. Public notification	2	18	72	39	39
6. Communications	3	28	46	46	39
7. Special populations	1	20	35	35	35
8. Traffic/access control	0	--	--	--	--
9. Public affairs	3	30	77	67	60
10. Medical - first response	2	13	61	61	61
11. Medical - transport	2	15	73	67	47
12. Medical - facilities	2	12	83	75	67
13. Field response	2	9	78	78	78
14. Evacuee care	1	3	67	67	67
15. 24-hr operations	0	--	--	--	--

"error-prone" or difficult for exercise participants to accomplish successfully. In other words, are they statistically associated with below-average outcomes? The results of this analysis could potentially be used by CSEPP managers to focus such resources as training, staff effort, etc. where they are most needed. The analyses and tables in Section 3.2.1 represent compilations across all of the data in the database.⁵

In Section 3.2.2, we examine those exercises that had particularly high concentrations of below-average outcomes, and we analyze them separately in an effort to identify the particular problem areas or "hot spots" associated with a disproportionate amount of poor outcomes. The specific sites, tabs, and/or objectives that proved problematic are identified, as measured by high concentrations of below-average outcomes. This approach is applied to the exercises in each of the five years for which data are available. The results are then summarized, to the extent that the

The data used in the multivariate analyses were edited as follows: only those issues for which Record Identification numbers (RIDs) are coded with a "1" in the "toggle" category were used. The toggle numbers were assigned during the coding process to preserve the original count of issues extracted from exercise reports (see *Analysis, Vol. 1*, Section 2.2.1.2 for further explanation). In this study, only records with a toggle value of 1 were used, to avoid using duplicate records in the data analysis. This may, however, have introduced other biases in the findings of which we are not aware.

identified exercises have characteristics in common. These results could help CSEPP policymakers target resources to the exercise traits most closely associated with below-average outcomes in the most problem-plagued exercises.

The last section, Section 3.2.3, examines the exercises either prior to or subsequent to those identified in Section 3.2.2, in order to determine whether the problems identified in the first exercises tend to persist from one exercise to the next. The results in Section 3.2.3 are presented for pairs of exercises that occurred at the same location and within the same time period (either 1992-1993 or 1994-1996). To the extent that problems do persist, they might warrant increased resources and attention.

3.2.1 Characteristics Statistically Associated with Below-Average Outcomes

This section contains tables and discussions bearing on the relationships between below-average outcomes and three explanatory variables: site, tab, and objective. Separate findings are presented for the two time periods (1992-1993 and 1994-1996), for reasons explained above. It appears that there are significant relationships between each of these variables and issue outcomes.

3.2.1.1 Association with the Tab Variable

Time 1 Period (1992-1993): Table 5 below shows the association between the report tab and outcome variables for the Time 1 period, 1992-1993. Overall, 30.3% of all outcomes were classified as below average during this time period. In contrast, 38.6% of the joint activities were below average, and 34.1% of the off-post activities were below average. On-post activities had a noticeably lower percentage of below-average outcomes (20.9%).⁶

Thus, it appears that the report tab variable clearly influences the likelihood of a below-average outcome. Specifically, joint location activities are most likely to produce below-average outcomes, followed by off-post activities. On-post activities, in contrast, are least likely to produce such below-average outcomes.

Time 2 Period (1994-1996): Table 6 presents the pattern of statistical association between report tab and outcome for the Time 2 period, 1994-1996. The bottom row of the table indicates that only 15.8% of all the exercise activities were below average during this time period (compared

⁶ The Pearson chi-square value for the relationship between these two variables is 18.1, which is statistically significant beyond the 0.001 level. This largely reflects the large number of cases (813) included in the table. Unless otherwise noted, the reader may assume that all similar relationships discussed below will be significant at or beyond the 0.01 level in this section of the report.

**TABLE 5 Cross-Tabulation of Report Tab
and Outcome for Time 1 Period**

Report Tab	Measure	Outcome		Total
		Below Average	Average & Above	
On-post	Count	58	219	277
	% of report tab	20.9	79.1	100
Joint	Count	44	70	114
	% of report tab	38.6	61.4	100.0
Off-post	Count	144	278	422
	% of report tab	34.1	65.9	100.0
Total	Count	246	567	813
	% of report tab	30.3	69.7	100.0

**TABLE 6 Cross-Tabulation of Report Tab
and Outcome for Time 2 Period**

Report Tab	Measure	Outcome		Total
		Below Average	Average & Above	
On-post	Count	96	516	612
	% of report tab	15.7	84.3	100.0
Joint	Count	48	182	230
	% of report tab	20.9	79.1	100.0
Off-post	Count	102	617	719
	% of report tab	14.2	85.8	100.0
Total	Count	246	1315	1561
	% of report tab	15.8	84.2	100.0

with 30.3% for the Time 1 period); this may reflect the aforementioned change in the classification system. As in the Time 1 period, the greatest percentage (20.9%) of below-average outcomes was found in the joint activities. However, the below-average percentages for the on-post (15.7%) and off-post (14.2%) activities were only slightly lower in this time period. These relatively small differences in below-average percentages among the report tab categories suggest that for the second time period, this explanatory variable has relatively less connection to the outcome, in terms of its predictive power. This is also reflected in the relatively weak (0.05) level of statistical significance for the relationship in this table.

3.2.1.2 Association with Objective Variable

Time 1 Period (1992-1993): Table 7 shows the number of issues associated with each exercise objective for exercises in the first time period, along with the percentages that had below-average versus average or above-average outcomes. Clearly, different objectives had very different percentages of below-average outcomes. The highest percentages of below-average outcomes were found for the "secure accident locations" (71.4%), "assess & classify" (60%), "public alert & PAR dissemination" (50%), "protective action decision" (42.4%), and "implement protective actions" (41%) objectives. In the first two cases, the numbers of records in the rows are very low (7 and 5, respectively), so great emphasis should not be placed on these objectives' high percentages of below-average outcomes. The numbers of exercise records for the other three objectives, however, are much larger; 46 for "public alert & PAR dissemination," 33 for "protective action decisions," and 39 for "implement protective actions." The numbers of records suggest that these three objectives may warrant additional attention and/or resources.

Time 2 Period (1994-1996): Table 8 presents the cross-tabulation between objective and outcome for the Time 2 period. As was the case for the Time 1 period, there is considerable variation in the percentage of below-average outcomes for the various objectives. Protective action implementation stands out with the highest such percentage: 43.1%. Four objectives ("protective action decisions," "alert & notification," "emergency worker exposure control," and "congregate care") are clustered fairly closely to one another, in the 25-30% range. The highest percentages for specific objectives in this time period are well below those in the earlier period. This is not surprising, given that the overall percentage of below-average outcomes in this period is about half that of the earlier period.

TABLE 7 Cross-Tabulation of Objective and Outcome for Time 1 Period

Objective	Measure	Outcome		Total
		Below Average	Average & Above	
Command/control	Count	14	67	81
	% of objective	17.3	82.7	100.0
Assess & classify	Count	3	2	5
	% of objective	60.0	40.0	100.0
Alert/mobilize/activate	Count	25	46	71
	% of objective	35.2	64.8	100.0
24-hr staffing	Count		2	2
	% of objective		100.0	100.0
Communications	Count	55	146	201
	% of objective	27.4	72.6	100.0
Assess hazard & PAR	Count	24	67	91
	% of objective	26.4	73.6	100.0
Protective action decisions	Count	14	19	33
	% of objective	42.4	57.6	100.0
Public alert & PAR dissemination	Count	23	23	46
	% of objective	50.0	50.0	100.0
Public information	Count	37	70	107
	% of objective	34.6	65.4	100.0
Implement protective actions	Count	16	23	39
	% of objective	41.0	59.0	100.0
Congregate care	Count	3	21	24
	% of objective	12.5	87.5	100.0
Medical response activities	Count	7	22	29
	% of objective	24.1	75.9	100.0
Emerg. worker contamination control	Count	9	23	32
	% of objective	28.1	71.9	100.0

TABLE 7 (Cont.)

Objective	Measure	Outcome		
		Below Average	Average & Above	Total
Mitigate hazards	Count	4	10	14
	% of objective	28.6	71.4	100.0
Operate/maintain response equipment	Count	2	3	5
	% of objective	40.0	60.0	100.0
Secure accident locations	Count	5	2	7
	% of objective	71.4	28.6	100.0
Army legal advice/response	Count	5	21	26
	% of objective	19.2	80.8	100.0
Total	Count	246	567	813
	% of objective	30.3	69.7	100.0

3.2.1.3 Association with Site Variable

Time 1 Period (1992-1993): Table 9 shows the outcomes by site for the first time period. During this period, some sites generated many more issues (both below and above average) than others. In fact, half of all 813 exercise issues for this time period occurred at two sites: ACA (202) and BCA (207).⁷ The remaining records are spread relatively evenly over the other six sites. The two sites in Table 9 with the highest frequencies of exercise records are also characterized by unusually high percentages of below-average outcomes: 38.6% below average for ACA and 39.1% below average for BCA. These are not the sites with the highest percentages of below-average outcomes, however. For the UCD site, 88.6% of the activities were classified as below average, followed by NCA with 45.2% below-average activity records. These four sites together (ACA, BCA, NCA, and UCD) accounted for virtually all of the below-average outcomes for this time period.

⁷ Three-letter abbreviations for the site names are used throughout: Anniston Chemical Activity (ACA); Blue Grass Chemical Activity (BCA); Deseret Chemical Depot (DCD); Edgewood Chemical Activity (ECA); Newport Chemical Activity (NCA); Pine Bluff Arsenal (PBA); Pueblo Chemical Depot (PCD); and Umatilla Chemical Depot (UCD).

TABLE 8 Cross-Tabulation of Objective and Outcome for Time 2 Period

Objective	Measure	Outcome		Total
		Below Average	Average & Above	
Alert/activate	Count	10	78	88
	% of objective	11.4	88.6	100.0
Hazard assessment	Count	22	103	125
	% of objective	17.6	82.4	100.0
Protective action decisions	Count	14	41	55
	% of objective	25.5	74.5	100.0
Command/control	Count	9	188	197
	% of objective	4.6	95.4	100.0
Alert/notify	Count	34	83	117
	% of objective	29.1	70.9	100.0
Communications	Count	25	182	207
	% of objective	12.1	87.9	100.0
PA implementation	Count	22	29	51
	% of objective	43.1	56.9	100.0
Traffic control	Count	33	140	173
	% of objective	19.1	80.9	100.0
Public affairs	Count	26	183	209
	% of objective	12.4	87.6	100.0
Medical - first response	Count	7	42	49
	% of objective	14.3	85.7	100.0
Medical - transport	Count	6	41	47
	% of objective	12.8	87.2	100.0
Medical - facilities	Count	12	90	102
	% of objective	11.8	88.2	100.0
Emergency worker exp.	Count	44	104	148
	% of objective	29.7	70.3	100.0

TABLE 8 (Cont.)

Objective	Measure	Outcome		Total
		Below Average	Average & Above	
Congregate care	Count	15	43	48
	% of objective	31.3	68.8	100.0
24-hr operations	Count	1	11	12
	% of objective	8.3	91.7	100.0
Total	Count	290	1437	1727
	% of objective	16.8	83.2	100.0

Time 2 Period (1994-1996): Table 10 presents the relationship between site and outcome for the 1994-1996 (Time 2) period. The pattern identified in the first period does not carry over. Only one site has a below-average outcome percentage well above the over-all average of 15.8%: ACA (28.8%). The below-average percentages for the other sites are close to or below the overall average. Although a statistically significant relationship between site and outcome clearly exists in this time period, it is not a major factor, especially in comparison to the first time period.

3.2.1.4 Summary of Characteristic Association Findings

The major findings of the associational analyses can be summarized in the following points:

Time 1 Period:

- 41.3% of joint activities had below-average outcomes, followed by off-post activities (37.7%), and on-post activities (23.1%).
- Five objectives had unusually high below-average percentages in this time period: "secure accident locations" (75%), "assess & classify" (60%), "public alert & PAR dissemination" (59%), "implement protective action" (40%), and "public information" (37%). The last three of these objectives account for slightly over one-third of all below-average outcomes.

**TABLE 9 Cross-Tabulation of Site and Outcome
for Time 1 Period**

Site Name	Measure	Outcome		Total
		Below Average	Average & Above	
ACA	Count	78	124	202
	% of site name	38.6	161.4	100.0
BCA	Count	81	126	207
	% of site name	39.1	60.9	100.0
DCD	Count	5	75	80
	% of site name	6.3	93.8	100.0
ECA	Count	18	39	57
	% of site name	31.6	68.4	100.0
NCA	Count	33	40	73
	% of site name	45.2	54.8	100.0
PBA	Count		91	91
	% of site name		100.0	100.0
PCD	Count		68	68
	% of site name		100.0	100.0
UCD	Count	31	4	35
	% of site name	88.6	11.4	100.0
Total	Count	246	567	813
	% of site name	30.3	69.7	100.0

**TABLE 10 Cross-Tabulation of Site and Outcome
for Time 2 Period**

Site Name	Measure	Outcome		Total
		Below Average	Average & Above	
ACA	Count	46	114	160
	% of site name	28.8	71.3	100.0
BCA	Count	20	168	188
	% of site name	10.6	89.4	100.0
DCD	Count	27	147	174
	% of site name	15.5	84.5	100.0
ECA	Count	35	156	191
	% of site name	18.3	81.7	100.0
NCA	Count	19	190	209
	% of site name	9.1	90.9	100.0
PBA	Count	49	246	295
	% of site name	16.6	83.4	100.0
PCD	Count	23	117	140
	% of site name	16.4	83.6	100.0
UCD	Count	27	177	204
	% of site name	13.2	86.8	100.0
Total	Count	246	1315	1561
	% of site name	15.8	84.2	100.0

- Four sites had unusually high below-average percentages in this time period: UCD (92.9%), NCA (51%), ACA (44%), and BCA (38.7%). The last two of these account for half of all below-average outcomes in this time period.
- In short, all three explanatory variables have statistically significant relationships with the dependent variable, outcome, for this time period.

Time 2 Period:

- The percentages of outcomes are significantly different (better) in Time 2 than in the earlier time period. Below-average outcomes account for 33.4% of the Time 1 outcomes, but only 15.8% in Time 2. Average and above outcomes are correspondingly more frequent in Time 2 (84.2%) than in the earlier time period (62.3%). These differences may well reflect the changes in the coding format for this variable between these two periods, as noted earlier.
- The patterns of association between the three independent variables and outcome are similar in the Time 2 period to those in Time 1, but with some minor differences. Below-average outcomes remained concentrated in the joint report tab, although the concentration was less in Time 2 than in Time 1. At most sites, the percentage of below-average outcomes declined from the Time 1 to the Time 2 period; however, UCD ran counter to this trend. Among objectives, command and control had high rates of below-average outcomes in both periods, while the percentages of below-average outcomes increased from Time 1 to Time 2 for the following objectives: communications, congregate care, and 24-hour operations.

3.2.2 Characteristics of Exercises with Concentrations of Below-Average Outcomes

This section presents another approach to identifying problem areas in the CSEPP exercise program. In this approach, we seek to identify problem areas or "hot spots" that have occurred in previous exercises and to investigate whether they are connected by common factors or threads. Like the previous section, this analysis relates issue outcome to tab, site, and objective, but in a different way. This analysis focuses on finding those tabs, sites, and objectives that account for the greatest proportion of the total below-average outcomes for the relevant year or exercise. In contrast, the previous section focused on the variables that had the greatest percentage of below-average outcomes as opposed to average or above. To see the difference, consider Table 11 below.

Table 11 shows the distribution of issues with below-average and average and above outcomes, among the three tabs, for the 1992 BCA exercise. The joint tab had 11 out of 15 issues

TABLE 11 Cross-Tabulation of Report Tab by Outcome for 1992, BCA Exercise

Report Tab	Measure	Outcome		Total
		Below Average	Average & Above	
On-post	Count	19	39	58
	% of outcome	31.7	52.7	43.3
Joint	Count	11	4	15
	% of outcome	18.3	5.4	11.2
Off-post	Count	30	31	61
	% of outcome	50.0	41.9	45.5
Total	Count	60	74	134
	% of outcome	100.0	100.0	100.0

with below-average outcomes, or 73.3% below average. This agrees with the findings in Section 3.2.1, that in general the joint tab had the greatest percentage of below-average outcomes. However, looking at the "below-average" column from the top down, it is apparent that the greatest number of below-average findings for this exercise occurred in the off-post tab rather than the joint tab. In fact, the off-post tab accounted for 50% of the below-average findings overall. In order to identify the areas where the greatest number of problems occur, it should be useful to consider this type of analysis which focuses on finding the greatest concentrations of below-average outcomes.

3.2.2.1 Selection of Exercises -- Site Distribution

The first step in this analysis is to select, for each year, the exercises that had the largest concentrations of below-average outcomes. Table 12 shows, for each year from 1992 to 1996, the percentage that each site contributed to the total tally of below-average outcomes for that year. For example, there are four exercises in the database from 1992, in which a total of 79 below-average outcomes were recorded. The BCA exercise that year had 60 of those 79, or 75.9% of the total. The DCD exercise had 3 of those 79, or 3.8% of the total. For each year, the top two exercises were chosen for further analysis, as indicated in Table 12 by the bolded figures. Looking across the table from left to right, one can see that the selected exercises are distributed quite evenly across the sites; each site has at least one and no site has more than two.

TABLE 12 Percentages of Total Below-Average Outcomes, by Year and Site

Year	ACA	BCA	DCD	ECA	NCA	PBA	PCD	UCD
1992	-	75.9^a	3.8	0	20.3	-	0	-
1993	46.7	12.6	1.2	10.8	10.2	0	0	18.6
1994	13.0	13.0	7.3	14.6	7.3	26.8	7.3	10.6
1995	38.5	-	23.1	15.4	6.4	10.3	-	6.4
1996	-	8.9	-	11.1	11.1	17.8	31.1	20.0

^a For each year, bolded values are the two highest percentage values.

3.2.2.2 Tab Analysis

Table 13 shows, for each of the exercises selected above, the distribution of below-average outcomes among the three exercise tabs.

For each exercise, the tab with the largest percentage of below-average outcomes is bolded. The bolded figures show a pronounced pattern: high percentages of below-average outcomes occurred only in the off-post locations for the exercises in the years 1992 through 1995. In the two 1996 exercises, in contrast, the below-average outcomes occurred on-post. This raises two obvious questions:

1. What accounts for this change between the earlier years and 1996? Additional analyses will be required to answer this question.
2. Is this change the beginning of a lasting trend, or is it only a one-year phenomenon? Data for additional years will be necessary to answer this question.

3.2.2.3 Objective Analysis

Table 14 shows the percentages of below-average outcomes by objective for the selected exercises in the first two years of the program: 1992 and 1993. The bolded figures represent all of the cases where a given objective accounts for 14% or more of the below-average outcomes for the exercise in question. (An even distribution among all 15-18 objectives would be between 5% and 7% per objective.) In other words, the bolded numbers show the greatest concentrations of problems in the most problem-plagued exercises. The "Tally" column on the right shows the number of bolded

TABLE 13 Distribution of Below-Average Outcomes by Tab for Selected Exercises

Report Tab	1992		1993		1994		1995		1996	
	BCA	NCA	ACA	UCD	ECA	PBA	ACA	DCD	PCD	UCD
On-post	31.7	12.5	24.4	16.1	33.3	18.2	30.0	33.3	85.7	88.9
Joint	18.3	6.3	24.4	3.2	22.2	12.1	26.7	22.2	7.1	11.1
Off-post	50.0	81.3	51.3	80.6	44.4	69.7	43.3	44.4	7.1	0

figures in each row (i.e., the number of times that each objective accounted for 14% or more of the below-average outcomes).

The table shows some pattern of concentrations of below-average outcomes associated with specific objectives. Specifically, communications is over-represented as an objective, creating problems in all four of the exercises shown in this table. The alert/mobilization and public alert objectives are over-represented as problems in two of the four exercises, while PI (public information) is over-represented as a problem in only one exercise.

Table 15 presents the comparable percentages of below-average outcomes associated with the objectives for the six exercises in the last three years for which data are available: 1994-1996. Communications is again a problem, as indicated by the three bolded percentages in its row. Emergency worker contamination exposure is even more problematic in this time period, however, with four bolded percentages. Traffic control is likewise a problem, with three percentages indicating over-representation. Finally, alert/activation and hazard assessment produce concentrations of problems in one exercise each.

Comparing the patterns for the two time periods shows general continuity in problem objectives: communications, public affairs/ public information, hazard assessment, and alert/ activate/mobilize are identified as problem areas in at least one exercise in both time periods. The differences between the time periods in problematic objectives are much less evident. Of the five objectives with disproportionately high below-average percentages in the first time period, only one (public alerting) does not also appear in the second time period. Table 15 shows that in the second time period there are eight objectives with one or more bolded percentages, indicating they were over-represented in below-average outcomes in the exercises. Of these, four (traffic control, public affairs, emergency worker exposure control, and congregate care) had no bolded percentages in the first time period. In one of these cases (traffic control), however, the objective was not listed among the first time period objectives. In short, both continuity and differences in problematic objectives exist between the two time periods; of the two, continuity seems somewhat more pronounced.

TABLE 14 Distribution of Below-Average Outcomes by Objective for Selected Exercises in Time 1 Period

Objective	1992		1993		Tally
	BCA	NCA	ACA	UCD	
Command/control	1.7	6.3	10.3	3.2	0
Assess/classify	3.3	-	-	-	0
Alert/mobilize/activate	16.7	-	14.1	6.5	2
24-hr staffing	-	-	-	-	0
Communications	21.7	25.0	15.4	22.6	4
Hazard assessment & PAR	6.7	37.5	7.7	9.7	1
Prot. action decisions	6.7	-	6.4	6.5	0
Public alert/PAR dissem.	5.0	18.8	9.0	22.6	2
Public information	11.7	-	9.0	25.8	1
Implement prot. actions	3.3	12.5	9.0	3.2	0
Monitor pub. cont.	-	-	-	-	0
Congregate care	-	-	3.8	-	0
Medical response	6.7	-	1.3	-	0
Emergency worker contam. control	8.3	-	2.6	-	0
Mitigate hazards	5.0	-	1.3	-	1
Equipment	-	-	2.6	-	0
Secure locations	-	-	3.8	-	0
Army legal advice/response	3.3	-	3.8	-	0

3.2.3 Temporal Analysis of Below-Average Outcomes

This section examines trends in the database over time. Specifically, we first examine the general trends in numbers and percentages of below-average outcomes in the total set of exercises. We then turn to selected pairs of consecutive exercises at the same site, to identify patterns of changes at individual sites.

3.2.3.1 General Trends

Table 16 presents the numbers of below-average outcomes, by site, for each of the five years under study. The bottom row of this table presents the total numbers of below-average outcomes for all exercises in each of the years. The largest number of such outcomes occurred in 1993, followed by 1994. Numbers of below-average outcomes in the other years were well below

TABLE 15 Distribution of Below-Average Outcomes by Objective for Selected Exercises in Time 2 Period

Objective	1994		1995		1996		Tally
	ECA	PBA	ACA	DCD	PCD	UCD	
Alert/activate	-	-	3.3	11.1	21.4	-	1
Hazard assessment	-	18.2	10.0	-	7.1	11.1	1
Prot. action decisions	-	6.1	3.3	-	7.1	11.1	0
Command/control	11.1	9.1	-	-	-	11.1	0
Alert/notify	-	6.1	10.0	5.6	-	-	0
Communications	11.1	3.0	23.3	16.7	14.3	-	3
Prot. action implementation	5.6	21.2	-	-	-	-	1
Traffic control	27.8	18.2	3.3	22.2	-	11.1	3
Public affairs	-	3.0	16.7	-	7.1	-	1
Medical - 1st resp.	-	-	-	-	7.1	11.1	0
Medical - transport	11.1	-	3.3	-	7.1	-	0
Medical - facilities	11.1	9.1	-	11.1	-	-	0
Emergency worker exposure	5.6	3.0	20.0	16.7	28.6	44.4	4
Congregate care	11.1	3.0	3.3	16.7	-	-	1
24-hr operations	5.6	-	-	-	-	-	0

the two peak years, with the smallest number occurring in 1996. In part, the total numbers of below-average outcomes reflect the number of exercises conducted in each year. This appears to be especially the case for 1994, in which all sites conducted exercises. This is not the whole story, however, for six exercises were conducted in each of the years 1993, 1995, and 1996, yet the last two exercises had fewer than half as many below-average outcomes as the first. It is also important to remember that the coding scheme for outcomes was changed between 1993 and 1994, so the numbers for the 1992 and 1993 years are not directly comparable with those for 1994 and later.

Table 17 presents the percentages of all outcomes at each exercise, and for all exercises, that were below average in each of the years under study at each of the CSEPP sites. Examination of the "Total" row of this table suggests a somewhat different picture than that painted in Table 16: the percentage of outcomes that was below-average peaked in 1992, and declined thereafter. Examination of the trends by site reveals three general patterns. Some sites, such as BCA, ECA, NCA, and UCD, have downward trends in their percentages, indicating that their over-all performances were improving with time. Other sites, such as ACA, DCD, and PBA, have essentially flat trends. Finally, one site (PCD) has an upward trend in the percentages, shifting from no below-average outcomes in its first two exercises to 25% in the last exercise. It is important to remember

TABLE 16 Numbers of Below-Average Outcomes, by Site and Year, 1992-1996

Site Name	Count, by Year					Count, Total
	1992	1993	1994	1995	1996	
ACA		78	16	30		124
BCA	60	21	16		4	101
DCD	3	2	9	18		32
ECA		18	18	12	5	53
NCA	16	17	9	5	5	52
PBA			33	8	8	49
PCD			9		14	23
UCD		31	13	5	9	58
Total	79	167	123	78	45	492

TABLE 17 Percentages of Below-Average Outcomes, by Site and Year, 1992-1996

Site Name	Percentage, by Year					Percentage, Total
	1992	1993	1994	1995	1996	
ACA		39	23	33		34
BCA	45	29	15		5	26
DCD	15	3	12	19		13
ECA		32	27	13	15	21
NCA	55	39	7	15	9	18
PBA		0	25	7	15	13
PCD	0	0	14		19	11
UCD		31	17	14	11	24
Total	36	28	17	17	12	21

in examining these trend patterns that the outcome coding scheme changed between 1993 and 1994. We have not as yet examined what possible effects this may have had on these results.

3.2.3.2 Selected Pairs of Consecutive Exercises

This analysis follows up on exercises that were identified as problem areas in Section 3.2.2, by looking, in each case, at the immediately subsequent exercise at the same site. The purpose of this analysis is to examine whether trouble spots identified in an exercise tend to recur in subsequent exercises at the same site. Four pairs of exercises were analyzed to illustrate the value of this technique and provide some preliminary information on the question of recurrence.

To compare sequential exercises at a site involves comparing the numbers and percentages of below-average outcomes in the two exercises according to tab and objective. Such comparisons should only be made within the two time periods, 1992-1993 and 1994-1996. We examined two pairs of exercises for each of the time periods: the BCA and NCA exercises in 1992-1993, and the PBA and ECA exercises in 1994-1995. These exercises were selected to maximize the numbers of below-average outcomes in the first half of each pair.

Separate tables for one of the pairs (the 1992-1993 BCA exercises) are presented and discussed below to illustrate the comparison concept (Tables 18 and 19). Results from all four pairs are summarized in Table 20 and discussed below.

Analysis of Table 18: Table 18 presents the data for the comparison of report tab and outcome relationships for the 1992-1993 BCA exercises. The number of below-average outcomes at this site decreased from 60 in 1992 to 21 in 1993. To what extent did this decline reflect improved performance at the off-post locations, which contributed 50% of such outcomes in the 1992 exercise? The number of off-post below-average outcomes declined from 30 in 1992 to 14 in 1993. That decline of 16 is almost half the total decline of 39 from 60 below-average outcomes in 1992 to 21 in 1993. Clearly, the 1992-1993 improvement in exercise performance at the BCA site is due in part to this decline in off-post below-average outcomes. However, comparison of the equivalent percentages in these two cells shows that the proportion of below-average outcomes occurring off-post actually increased, from 50% in 1992 to 66.7% in 1993. So while performance was improving off-post between these two exercises, it was improving even more at the on-post and joint locations. The result is an increase in the relative concentration of below-average outcomes off-post.

Analysis of Table 19: Table 19 presents the relevant cells of the cross-tabulation of objectives and outcome for the two BCA exercises under study. Alert/mobilize/activate and communications were identified earlier as the two objectives with disproportionately high concentrations of below-average outcomes in the 1992 exercise. To what extent did improvements

TABLE 18 Below-Average Outcomes by Tab for 1992-1993 BCA Exercises

Report Tab	Count/Percentage	
	1992	1993
Tab A: On-post	19 / 31.7%	2 / 9.5%
Tab B: Joint	11 / 18.3%	5 / 23.8%
Tab C: Off-post	30 / 50.0%	14 / 66.7%
Totals	60 / 100%	21 / 100%

TABLE 19 Below-Average Outcomes by Objective for 1992-1993 BCA Exercises

Objective	Count/Percentage	
	1992	1993
Alert/mobilize/activate	10 / 16.7%	0 / 0%
Communications	13 / 21.7%	11 / 52.4%
Totals	60 / 100%	21 / 100%

in these areas contribute to the overall improvement at this site between 1992 and 1993? Comparison of the numbers of below-average outcomes in these two years shows a mixed picture: the number of below-average outcomes for alert/mobilize/activate went from 10 in 1992 to zero in 1993. This indicates a dramatic improvement in performance concerning this objective, assuming it was tested in both exercises. The results for the communications objective are less striking: a decline from 13 below-average outcomes in 1992 to 11 in 1993. Because the total number of below-average outcomes is so much less in 1993 (21) than in 1992 (60), the 11 below-average communications outcomes in 1993 are a much higher percentage (52.4%) of the total. This shows the importance of using the actual count numbers, as well as percentages, in such comparisons.

The number of below-average outcomes at the BCA site decreased from 60 in 1992 to 21 in 1993, a decline of almost two-thirds. To what extent did declines in the numbers of below-average outcomes for the most problematic objectives contribute to this over-all decline in below-average outcomes? Ten out of sixty (or 16.7%) of the below-average outcomes in the 1992 BCA exercise were associated with the alert/mobilize/activate objective. This objective had no below-average outcomes in 1993, a dramatic improvement. Such is not the case for the second problematic objective, communications, the below-average outcomes of which drop from 13 in 1992 to 11 in

1993. With the total number of below-average outcomes in 1993 so much smaller than in 1992, the communications objective in 1993 contributed over half (52.4%) of all below-average outcomes in this year, compared to the much lower 21.7% of below-average exercises in 1992. (The results for the other objectives in these two years have been omitted from Table 19 to make it easier to read.)

Discussion of Table 20: Table 20 presents data and comparisons for all four pairs of exercises studied. For each pair, the table shows the tab and one or more objectives where concentrations of below-average outcomes occurred in the first exercise of the pair. The far-left column identifies the site and years of comparison. The second column identifies the variable (tab or objective) for which the outcomes are being compared. The third and fourth columns present the numbers (count) of below-average outcomes for this variable at this site in the first and second years, respectively. The "Delta # Year 1-2" column presents the change in the number of below-average outcomes for this variable characteristic at this site from the first to the second year. The next three columns present the identical information, in terms of percentages of all below-average outcomes for that exercise. The two Delta columns are highlighted in bold, since they are the major points of interest in this table.

The "Delta # Year 1-2" column shows the improvements (positive numbers of below-average outcomes) or declines in performance between consecutive years for the variable characteristics, sites, and years identified in the two far-left columns. With one exception, all of the values in this column are positive, which shows that performance between these two years was improving, in general. Half of the entries in this column are small, ranging from -1 to 3; the other half are larger, ranging from 6 to 18. It should be pointed out that in each of these latter cases the number of below-average outcomes in the second year was at or near zero, indicating a near-complete correction of the earlier problem with this variable characteristic's performance.

The far-right column in Table 20 shows the change in percentages (Delta %). These numbers indicate whether or not the percentage of all below-average outcomes associated with the tab or objective in the first year increased (reflected in negative values) or decreased (reflected in positive values) in the second year. For example, in the first row, 50% of the below-average outcomes at the BCA site in 1992 occurred off-post, but in 1993 this percentage increased to 66.7%, indicating a relative increase in the concentration of problems off-post. In general, however, we find positive numbers in the Delta % column, indicating that improvements in "trouble-spot" areas between the consecutive exercises occurred not only absolutely, in terms of fewer below-average outcomes, but also relative to other tabs or objectives. As noted, however, these results are not conclusive, given the exploratory nature of the analysis and the limited number of cases involved.

TABLE 20 Numbers, Percentages, and Changes in Below-Average Outcomes by Tab and Objective for Selected Pairs of Exercises

Site/Years	Tab or Objective	Year 1 Below- Avg. #	Year 2 Below- Avg. #	Delta # Year 1-2	Year 1 Below - Avg. %	Year 2 Below - Avg. %	Delta % Year 1-2
BCA/1992-93	Off-post	30	14	16	50	66.7	-16.7
	Alert/mobilize/activate	10	0	10	16.7	0	16.7
	Communications	13	11	2	21.7	52.4	-30.7
NCA/1992-93	Off-post	13	14	-1	81.3	82.4	-1.1
	Communications	4	3	1	25	17.6	7.2
	Assess hazard & PAR	6	4	2	37.5	23.5	14.0
PBA/1994-95	Off-post	23	5	18	69.7	62.5	7.2
	Alert/mobilize/activate	6	0	6	18.2	0	18.2
	Public alert & PAR dissem.	7	1	6	21.2	12.5	8.7
	Public information	6	0	6	18.2	0	18.2
ECA/1994-95	Off-post	8	7	1	44.4	58.3	-13.9
	Public information	5	2	3	27.8	16.7	11.1

4 CONCLUSIONS AND ADDITIONAL POSSIBLE ANALYSES

As noted in the introductory section, this report is preliminary in a number of respects. The points outlined below address the major limitations of the study and identify additional steps that could be taken to improve the potential value of the study and increase confidence in the data and their interpretation.

4.1 INTERCODER RELIABILITY FINDINGS AND POTENTIAL FOLLOW-UP STUDIES

In terms of intercoder reliability, the investigation revealed substantial differences among the individual coders as to how particular issues should be classified. This finding is significant, because it means that an individual using the database to find issues related to a particular (fine-level) topic may not have the same "mental map" of what the different topics mean, and thus may have difficulty finding all of the issues related (in his or her mind) to that particular topic. A number of possibilities might be investigated to improve on the coding process, or to find alternatives to the process, as follows:

- Devise a different coding scheme that would be more intuitive and lead to greater intercoder reliability.
- Truncate the coding process at the element level. At present, this level has the greatest reliability rate. Truncating the classification process at this level might increase its reliability even more by eliminating cases where difficulties at the fine level of classification (functional component) drive disagreements over the correct element.
- Determine the potential uses for the database; this should "drive" or determine the level of accuracy required in the coding. The 61% average accuracy presented in Table 3 may be adequate for some purposes, but inadequate for others. Identifying the various possible uses for the data should be a first step in determining the desired level of accuracy in the data coding.
- Emphasize the keyword search function of the database over code-based searching. Keyword searching is an alternative way for database users to find issues related to a specific topic. Keyword searching could be enhanced by such measures as developing more detailed instructions, devising lists of suggested key words, or attaching entries from a standard keyword list to the issues in a separate record or a related database.

- Compare the results of keyword searches to the coding results. If the two are highly correlated, that might suggest that the coding function could, at least in some instances, be automated. If, in contrast, the keyword search results are not highly correlated with the codes entered for those reports, this may call into question either the value of the keywords or the value of the codes.

Improving the codes used for the exercise database does not necessarily require the recoding of all entries already in the database. One could recode only those records containing the old codes that are being improved, which in many instances would be a small subset of the entire record set. Consequently, one does not necessarily need to view code improvement as an expensive or overwhelming task. Code improvement should be undertaken only when the promise for greater accuracy justifies the resources required.

4.2 MULTIVARIATE ANALYSIS FINDINGS AND POTENTIAL FOLLOW-UP STUDIES

The multivariate analysis found that site, report tab, and objective are each independently associated with the outcome variable for both time periods under study. There is also considerable continuity from one time period to the next. Specifically, the ACA site, the joint report tab, and the communications, public information, hazard assessment, and alert/activate/mobilize objectives have high percentages of below-average outcomes in exercises in both time periods. In terms of trends over time, four sites (BCA, ECA, NCA, and UCD) have downward trends, three sites (ACA, DCD, and PBA) have essentially flat trends, and one site (PCD) has an upward trend in percentage of below-average outcomes.

Comparison of sequential exercises at the same site produces mixed results: in some cases, improvements in performance appear to be due to some change in the problematic objective or the problematic report tab. In other instances, the relative performance of the earlier problem areas actually becomes worse. Clearly, both actual performance (measured by number of below-average outcomes) and relative performance (measured by percentage of below-average outcomes for the objective or exercise tab) are important and should be examined in future exercises.

Further refinements of the multivariate analysis are recommended, to address limitations in this study. In particular:

- This study is limited in that only those activities that were recorded in the exercise evaluations have been included in the analyses presented here. Examination of the "thresholds" at which activities are or are not included in evaluation reports, to the extent this has not already been addressed, could increase confidence in the data.

- The multivariate analysis presented above is preliminary, in that it does not address either (1) the relative importance of the objectives, in terms of protecting the public health and safety, or (2) how information from these analyses can be used to improve performance in the exercises. Additional work in both areas could be explored. One possible application would be to interview staff involved in the "more difficult" objectives (i.e., those with high percentages of below-average outcomes) at the sites with low percentages of below-average outcomes and high percentages of above-average outcomes, to learn their "tricks-of-the-trade." These ideas/methods could then possibly be shared, through training, with staff at other sites that have had less positive outcomes.
- Another uncertainty pertaining to the multivariate analysis is the degree to which it is affected by the results of the intercoder reliability study. The results dealing with the site and tab variables should not be affected. However, the results concerning the explanatory powers of the objective variable may be affected by the subjective nature of the coding process, as revealed in the intercoder reliability study. The mere fact that different coders do not always classify issues the same way does not necessarily cast doubt on the analytical findings with respect to objectives. It may, however, introduce biases that could be eliminated (or better characterized) upon further study.
- The database could be used to study associations among issues, as a way to research causal connections. For example, communications is among the objectives that comes up most frequently among exercise issues. In exercises where communications issues are found, does one also find disproportionate numbers of issues related to staff activation, command and control, or public alerting? If such an association were found, the text of the issues could be examined to determine whether these problems in fact resulted from communications problems.