

CHAPTER 12 EVALUATION

Chapter Summary

This chapter discusses evaluation in emergency management, beginning with performance appraisals for individual members of the local emergency management agency (LEMA). Next, the chapter addresses the procedures for periodic evaluation of the local emergency management agency and local emergency planning committee (LEPC). The discussion then turns to procedures for evaluating drills, exercises, and incidents. The chapter concludes with a discussion of procedures for evaluating organizational training and community risk communication programs.

Personnel performance appraisals

Periodic performance appraisals make a critical contribution to the performance of any organization by providing a systematic review of an individual employee's performance on the job (Cascio, 1998; Schmitt & Klimoski, 1991). This evaluation is used to assess the effectiveness of his or her work. Performance appraisal serves four functions—development, reward, internal research, and legal protection. First, the developmental function is intended to improve the person's *ability* to do the job. In this context, performance appraisal can be used to guide decisions about training, reassignment, or termination. Training keeps the job constant, but changes the person. Reassignment keeps the person constant but changes the job, either laterally to another at the same level of authority in the organization, or vertically to one of greater (promotion) or lesser (demotion) authority and responsibility.

Second, the reward function is intended to improve the person's *motivation* to do the job. In this context, performance appraisal should have clear criteria that provide guidance to the employee about what is important to the organization. In addition, a good appraisal process can guide rewards and thereby improve productivity and satisfaction and decrease turnover.

Third, the internal research function provides essential data for the development of skills inventories and the validation of selection criteria. This function is typically performed by

industrial/organizational psychologists, so it is mostly a concern of large organizations with well staffed human resource departments. Fourth, legal protection is achieved when an organization conducts performance appraisals according to generally acceptable procedures and retains documentation that it has done so.

There are four principal questions that need to be addressed in the performance appraisal process. First, when should it happen? Second, who can do it? Third, what should be evaluated? Fourth, how should it be done? With respect to the first question, *when should it happen*, supervisors should recognize a formal appraisal is different from informal feedback about job performance (which should be frequent) and link performance appraisal to the task cycle. That is, if a person works on a lengthy project, an appraisal should be conducted soon after project completion. If there is no specific task cycle, or if the cycle exceeds one year, an appraisal should be conducted at least annually.

With respect to the second question, *who evaluates*, supervisors should consider who has the information on what is desired and what has been done. It also is important to consider who wants to control rewards in the organization and who can assist in follow up activities such as training. Typically, the immediate supervisor knows the goals of the unit and the individual's job description, and has a considerable amount of information about an individual's performance. In addition, supervisors generally want to maintain reward power and can assist in follow up activities such as training. Thus, supervisors are the most common evaluators, but there are also others who can make valuable contributions. For example, the employee's peers are a valuable source of information because they also know the goals of the unit and have much information about an individual's performance. Indeed, they sometimes have more accurate information than the supervisor because they have more frequent contact with the employee and observe a more

representative sample of behavior. However, they might downgrade their ratings of the employee if they think this will improve their own relative standing and, thus, their salary increases and promotion prospects. As a result, peer appraisals are more appropriate when there is a trusting, noncompetitive atmosphere or when special procedures are implemented to prevent competitive behavior from influencing peer ratings (Kane & Lawler, 1977).

Like peers, subordinates are also good sources of information because they also see aspects of an employee's behavior that are not seen by supervisors. Needless to say, an employee's subordinates will usually be concerned about divulging negative information because of the possibility the employee will retaliate if the source of negative information is divulged. Finally, employees themselves are good sources of information because they much more information about their performance than anyone else, especially when they work independently to produce a complete product. However, self evaluations often yield more positive evaluations than is warranted because employees tend to attribute their successes to their own efforts and their failures to external conditions in the workplace.

The diversity of information sources about what should be done and what has been done, as well as concerns about who controls rewards and who can implement change lead to a variety of different solutions to the question of who evaluates. In many organizations, the supervisor and employee both rate performance and seek to reconcile the (almost inevitable) differences in ratings through discussion of the information on which they based these ratings. Where possible, the sources of information are enlarged to include peers, subordinates, and even an organization's customers and suppliers.

With respect to the third question, *what should be evaluated*, supervisors should seek to rate performance on data that meet three conditions. First, the data must be available within the

time period in which the appraisal is being conducted, relevant to job performance, and comprehensive. Data availability can be a problem if a person works on projects that take years to show results. For example, a risk communication program might easily take more than a year to conduct and even longer to produce measurable changes in households' emergency preparedness. Consequently, performance must sometimes be evaluated on intermediate results rather than final outcomes. Second, supervisors should pay attention to the job relevance of the evaluation criteria to ensure behavior is being not considered that is extraneous to the job. It is obvious that an employee's choice of music and office decorations are irrelevant to job performance if they do not disrupt the workplace. However, it can sometimes be difficult to distinguish what is personally distasteful from what is actually disruptive, and to ensure only the latter influences performance evaluations. A more subtle aspect of job relevance is criterion contamination, which occurs when an evaluation is affected by factors other than personal performance. For example, two different employees might be given what seem to be equivalent risk communication assignments but one is assigned to an upper middle class neighborhood with long-term homeowners and an active community council, whereas the other is assigned to a working class neighborhood where there is substantial turnover among apartment renters who have no existing neighborhood organization. In this example, the second employee would implicitly be evaluated on factors that are beyond his/her control. Third, evaluation criteria should be comprehensive to ensure all parts of the job are being measured. In many cases, the short-term impacts of a person's behavior are readily recognized, but the equally important long-term consequences are not. For example, emergency managers might face short term pressure to meet with public safety personnel (fire, police, and emergency medical services) to ensure the EOP is updated, but meeting with land use or community development personnel to coordinate

the development of a preimpact disaster recovery plan can easily be overlooked.

With respect to the fourth question, *how should it be done*, supervisors should begin by ensuring that employee performance will be measured using an instrument that addressed the full range of job demands. In most cases, a jurisdiction's human resource department will have a set of performance appraisal criteria that have been devised for all civil service jobs. Typically, these instruments separately address *task* and *interpersonal* performance. In turn, task performance is frequently broken down into motivational and ability components. Sample performance appraisal criteria are displayed in Table 12-1.

Table 12-1. Sample Performance Appraisal Criteria

<ol style="list-style-type: none">1. <i>Time and project management:</i> Understands own job description and the function of the unit; well organized; sets and adjusts priorities in response to job impediments; delegates as appropriate; follows through on objectives; consistently produces work of a quality and quantity that is consistent with organizational needs.2. <i>Resource and knowledge management:</i> Understands budget processes relevant to the position; uses allocated resources wisely; understands and follows organizational procedures relevant to daily job operations; knows and uses sources of additional information and assistance as needed.3. <i>Decision making and problem solving:</i> Identifies problems, collects information, and weighs viable options; makes decisions and follows through4. <i>Innovation:</i> Generates new ideas5. <i>Personal management:</i> Initiates activity without awaiting direction from superiors; seeks additional responsibility; recognizes mistakes and adapts to them; perseveres until projects are completed.6. <i>Change management:</i> Accepts and supports new methods of job performance and organizational procedures.7. <i>Interpersonal skills:</i> Works well with supervisors, peers, subordinates, and customers; manages conflict effectively.8. <i>Communication:</i> Able to speak and write clearly but is diplomatic in dealing with others.9. <i>Quality of work life:</i> Demonstrates respect for individual differences, contributions, and family related responsibilities of others; supports and promotes organizational diversity initiatives.

Supervisors should recognize the difficulty of remembering all instances of effective and ineffective behavior each subordinate displays over the course of a year. Thus, it is advisable to take time at the beginning of each year to review the criteria that will be used to evaluate performance at the end of the year. Periodic review of the evaluation criteria will help the supervisor to notice examples of effective and ineffective behavior, and interpret these appropriately when they occur. In addition, supervisors should take time during the year to think

back over the course of each week to identify and record each employee’s typical level of performance, as well as any instances of notably effective and ineffective behavior. At the end of the year, the supervisor can draw on this information to rate each employee’s performance on each of the evaluation criteria.

Before the annual review meeting the supervisor should give employees copies of the performance appraisal form and ask them to review the past year (it is a good idea for employees to keep job diaries as well). The supervisor and the employee should then rate the employee on each of the performance criteria using a numerical scale such as the one listed in Table 12-2.

Table 12-1. Sample Performance Appraisal Rating Scale

	Performs far below job requirements	Performs below job requirements	Performs job requirements adequately	Performs above job requirements	Performs far above job requirements
	1	2	3	4	5
Supervisor					
Employee					

Both the supervisor and the employee should prepare to explain the reasons for their ratings in terms of specific examples of performance displayed during the rating period. However, supervisors often are *required* to provide a written explanation why they have given ratings of 1 (= Performs far below job requirements) or 5 (= Performs far above job requirements) because very low ratings are likely to lead to termination and very high ratings are likely to lead to promotion. Such high consequence actions necessitate a correspondingly high level of justification, although it certainly would be helpful to be specific about the reasons for intermediate ratings such as 2 (= Performs below job requirements, 3 (= Performs job requirements adequately), and 4 (= Performs above job requirements).

The supervisor should schedule a private meeting with each employee and open the

meeting with an acknowledgement of specific positive achievements and a frank discussion of specific performance shortcomings. The objective of the meeting should be to describe actual observable behaviors that indicate both good and bad (if any) instances of performance are being noticed. It is important for supervisors to focus on behavior, which can be changed, not personality characteristics, which are virtually impossible to change. The supervisor should emphasize good performance will be rewarded and poor performance will be corrected—by training if the problem is a correctable lack of ability, by withholding rewards if the problem is a minor lack of motivation, or by transfer or discharge if the problem is either an uncorrectable lack of ability or motivation. It is important for supervisors to avoid “sending the wrong message” in an otherwise positive appraisal by allowing the amount of time spent talking about negative aspects of a performance appraisal to overwhelm the amount of time spent on positive aspects. Dwelling on the negative is quite common and easy to understand because most supervisors want to save time by focusing on what needs to be fixed rather than “wasting time” talking about what is being done well. Nonetheless, praise is important because it lets the subordinate know that good work is being noticed and will have positive consequences. Explicit recognition of good performance during the appraisal especially important when, as is the case in most public sector organizations, there is little difference between the smallest and the largest salary increases within the organization. Of course, it is even better if praise during the annual performance appraisal is given in addition to recognition of good performance throughout the year.

The supervisor should also offer employees an opportunity to explain the basis for their self-ratings particularly if they can provide an explicit rationale for those ratings. Such openness provides supervisors with opportunities to share information and solve problems rather than “tell

and sell” their own assessments. Listening carefully to what employees have to say, allowing time for a full discussion, and focusing on documented instances of behavior rather than speculations about personality characteristics will go a long way toward decreasing the stress in a situation that is uncomfortable for both the employee and the supervisor.

It also is advisable for the supervisor to collaborate with the employee in setting objectives for the coming year. These objectives should be specific and measurable so both of them can determine at the end of the year if the objective was accomplished. Some objectives should be performance-oriented (e.g., tasks performed or projects completed) whereas others should be developmental (e.g., training courses taken). In addition, objectives should be set only if they are feasible for the individual to accomplish within the period of performance. Thus, the objective “Get an emergency operations plan *approved* by the end of the year” should be revised to “Get an emergency operations plan *submitted* by the end of the year” because the employee can’t control the approval process which, in any event, might take longer than the end of the year to complete. Setting objectives is an important way of showing high-performing employees how they can obtain promotions. Just as important, it is a way to keep poor-performing employees from giving up altogether because a good development plan, based on clear objectives, shows them how they can achieve better ratings.

Evaluating the LEMA and LEPC

Evaluations of the LEMA and LEPC are logical extensions of the procedures for conducting personnel performance appraisals. As noted in Chapter 3 on *Building an Effective Emergency Management Organization*, local emergency managers should work with the other members of the LEMA and LEPC to set specific, measurable objectives they can accomplish within the period of performance. These objectives should be developed collaboratively because

such goals elicit greater commitment than goals that a supervisor sets unilaterally. The goals for the LEMA and LEPC should differ from each other both because they are different organizations with different responsibilities and also because the emergency manager's control over the allocation of resources in the LEPC are much more limited than her or his control over the LEMA.

Specifically, as noted in Chapter 3, the local emergency manager should work with the LEMA staff to assess the current status of the jurisdiction's hazard/vulnerability analysis, hazard mitigation program, emergency preparedness program, recovery preparedness program, and community hazard education program. Next, the LEMA staff should review the capability shortfall identified in previous years and also the multi-year development plan that was designed to reduce the capability shortfall. If the goals and schedule set in that document are inappropriate, LEMA staff should work collaboratively to set revised goals in each of the major programmatic areas, based on an assessment of the LEMA's current capability. As is the case with the multi-year plan, the local emergency manager should work with the LEMA staff to set specific milestones (objective indicators of task performance) to determine if they are making progress at a satisfactory rate throughout the year. Clear assignment of authority and responsibility for task performance will not only simplify the process of individual performance appraisal at the end of the year, but it will also enhance the likelihood of successful task performance.

Evaluating performance of the LEPC is somewhat more complex, but follows basically the same procedures as are used for the LEMA. Each LEPC subcommittee should identify the specific tasks that must be accomplished in order to make progress in its functional area (e.g., hazard/vulnerability analysis; planning, training, and exercising; recovery and mitigation; public

education and outreach; LEPC management). In some cases, this will lead subcommittee members to set an objective of task performance, whereas in other cases the objective might be to acquire the resources needed to perform a task. For example, the Hazard/Vulnerability Analysis Committee might set an objective of acquiring a computer program and database and then getting a member trained to use them to conduct analyses.

As is the case with the LEMA, the LEPC should use a collaborative process to set specific, measurable objectives that it can achieve within the performance period. The subcommittees should coordinate their objectives with each other, either through the Executive Committee or in general meetings of the LEPC. Once all of the subcommittees have set objectives, they should review their performance informally throughout the year and more formally at the end of the year. Once the LEPC as a whole has reviewed its annual performance, there should be a discussion with senior elected and appointed officials to ensure they are aware of the LEPC's achievements during the previous year.

Evaluating Drills, Exercises, and Incidents

Evaluating drills, exercises, and incidents has some elements in common with employee, LEMA and LEPC performance appraisals, but also some significant differences. The primary difference is that performance in drills, exercises, and incidents is measured over a relatively short period of time. In drills, performance often is measured over a period of minutes, and in exercises and incidents performance is measured over a period of hours to days. This shortened time period makes evaluation easier in some ways because there is less performance to evaluate. On the other hand, task performance usually is measured much more intensively during drills and exercises and there are many people's performance to observe. Finally, incidents—especially those in which a loss of life or extensive destruction of property—has the potential for

generating lawsuits. In turn, these can stifle the free exchange of information needed to learn from experience and improve the state of community emergency management.

Drills. When planning to conduct drills, the first task is to specify clearly what are the objectives to be tested (National Response Team, 1990). Typically, drills are used to test people, facilities, and equipment on tasks that are *difficult*, *critical*, and are *performed infrequently*. The first two conditions are important because they make failures in task performance likely and they escalate the consequences when failure does occur. The third condition is important because lengthy time intervals between opportunities for task performance cause skill decay in people and deterioration (e.g., aging of batteries, corrosion of connections, or loss of calibration) in equipment. Exercises are more comprehensive than drills because they are used to test people's ability to perform both *taskwork* and *teamwork*. Where the former is obviously the ability to competently perform each separate aspect of the emergency response, the latter is the ability to allocate resources and schedule tasks to achieve a coordinated performance that is efficient, effective, and timely (McIntyre & Salas, 1995).

The evaluator must have a level of proficiency that meets or exceeds that of the person being evaluated and must identify any facilities, equipment (e.g., calculators, computers, or communication devices), and job performance aids (e.g., written procedures, tables, or figures) that are specified for use in performing the task to be tested. In most cases, the evaluator randomly selects one or more persons from a list of personnel listed as the principal or alternate performer in the position to be tested. The player is provided with information about a hypothetical situation, known as a *scenario*, that creates a need to perform the task. The player is asked to "walk through" task performance by performing each step in that task. In cases where there is a significant mental as well as physical component, the player might be asked to "talk

through” the scenario by identifying the information needed, the way in which the information is processed, and the final judgment or decision made. Drills usually involve only a single person, or at most a few people, who are located close together. Consequently, a single individual can provide the information from the scenario, observe the player’s performance, and note any deviations from the EOP or its procedures.

Functional exercises. A functional exercise differs from a drill in the larger number of personnel, usually from a single department, who are responsible for performing a single function within the EOP (e.g., protective action selection, hazmat spill control). A functional exercise involves more tasks than a drill, and thus more personnel and equipment, so the exercise objectives are more numerous and the scenario is usually more complex. Consequently, there is a division of labor in the management of the exercise. One person serves as a *controller*, a person who provides information from the scenario, and another as an *evaluator*, a person who observes the player’s performance and notes any deviations from the EOP or its procedures. Indeed, some functional exercises require multiple controllers (who maintain contact by radio or cell phone) to provide information to teams of players operating in different locations. Moreover, these complex functional exercises will also require multiple evaluators to evaluate teams in multiple locations. However, some remote teams who have modest performance demands might be assigned a single individual who serves a both controller and evaluator.

Table-top exercises. A table-top exercise differs from a drill or functional exercise because it involves a group of senior personnel, usually branch chiefs or department heads, who serve as the directors (either the principals or alternates) of their functions. The scenarios for table-top exercises vary in their complexity; some are as simple as open-ended questions designed to generate a free-ranging discussion about a particular problem the local emergency

manager (or the LEPC's planning, training, and exercising committee) has noticed. For example, a table-top exercise might address the criteria (e.g., minimum strike probability and storm category) for initiating an evacuation of local hospitals in advance of a hurricane, the resources available for providing transportation support to the hospital evacuation, and the methods of facilitating the return of ambulances and other vehicles against the flow of evacuation traffic. The very nature of the table-top exercise makes it amenable to staffing by a single controller who also serves as the evaluator.

Full-scale exercises. A full-scale exercise simulates a community-wide disaster by simultaneously testing multiple functions and, especially, the coordination among these functions. The complexity of full-scale exercises requires thorough planning of the scenario, as well as coordination among the many controllers and evaluators. There also is a need for training the controllers, who might need to make *ad hoc* adjustments if the players take unexpected actions during the exercise that deviate substantially from conditions listed in the scenario. In addition, the many evaluators are also likely to need training if the demand for players and controllers exhausts the locally available supply of highly qualified evaluators.

The magnitude of full-scale exercises varies considerably. Small ones might provide only a limited test of some functions (e.g., a single school might be selected to test the evacuation plan). However, large exercises conducted for nuclear power plants can involve thousands of players, and scores of controllers and evaluators. Unlike drills, table-top exercises, and functional exercises—whose schedule is usually announced in advance—some full-scale exercises are unannounced. Thus, the exercise scenario and also the time at which it will be initiated are unknown to the participants. The rationale for unannounced exercises is that this prevents agencies from deliberately rescheduling training, vacations and other conflicts of their best

trained (occasionally their *only* trained) personnel so they can participate in the exercise. Thus, unannounced exercises provide a more accurate assessment of community preparedness, especially when exercises begin on the evening and night “back shifts”. This is certainly a valid reason for scheduling unannounced exercises, but it is nonetheless a good idea for the LEPC to work up to unannounced exercises by first verifying satisfactory performance in announced exercises. Performing well in an unannounced exercise is a challenging goal and, as noted earlier, achievable goals should be set first to build confidence and motivation to improve. Poor performance can not only prove to be a public embarrassment to the participating agencies and a demoralizing experience for participants; a very large number of errors might make it difficult to identify clear cut “lessons learned” and to develop a consensus on how to improve.

Incidents. The evaluation of performance in an incident can be extremely informative because it provides an unscheduled test of the emergency response organization. This has the advantage of providing a realistic test of many incident management functions, such as organizational activation and notification, in a way that would not be done in an announced exercise. Of course, the disadvantage of actual incidents as evaluations of the emergency response organization is that they are uncontrolled. That is, both the magnitude of the event and the response functions that are tested are matters of chance. Of course, incidents also have no controllers or evaluators, so respondents must rely on their memories and any documentation they have produced to establish who did what, where they did it, when they did it, and why they did it.

Critiques. All three forms of exercises (tabletop, functional, and full-scale) and incident responses benefit from oral critiques by all involved players, controllers, and evaluators (National Response Team, 1990). However, some full-scale exercises have so many participants

that a smaller number of representatives must be selected from each responding unit. In an exercise critique, discussions should address whether the exercise objectives were met. In an incident critique, the question is whether the response was consistent with the EOP and procedures. If there were deviations from the EOP and procedures, the participants should discuss why this occurred. In some cases, the deviation is adaptive (i.e., responders used a more effective method of protecting public health, safety, property, and environment) and the conclusion will be that the EOP or procedures need to be revised. In other cases, the deviation will be judged to be maladaptive, so the solution will be to reassign personnel, improve training, or upgrade facilities and equipment. Whatever problems are identified, it is important to focus on changing the system, not blaming individuals. The results of the critique should be documented in a written report that contains specific recommendations for action, assignment of responsibility for implementation, and a schedule for completion.

Evaluating Training and Risk Communication Programs

The procedures for evaluating training and risk communication programs are distinctly different from the previous types of evaluations because both of these types of evaluations involve evaluations of the effects of some treatment on a sample of individuals (Cook, Campbell, & Peracchio, 1990; Goldstein, 1993; Schmitt, & Klimoski, 1991). However, procedures for evaluating training and risk communication programs are similar to the previous types of evaluations because it is essential to define in advance what are the criteria for defining the success of the program.

Criteria for defining program success

Criteria for judging the success of training programs are often classified into four groups—reaction, learning, behavior, and results (Goldstein, 1993). Reaction criteria consist of

trainees' judgments about the training program. These usually include trainees' evaluations of the trainers, the facilities and equipment, the amount of material they learned, their enjoyment of the class, and their willingness to take another class from the instructor. Learning criteria are defined by trainees' performance on written tests of knowledge or performance tests of skills acquired during the training program. Behavior criteria refer to trainees' ability to apply knowledge and skills they learned during training when they return to their jobs. In the context of emergency management, this means performance during drills, exercises, and incidents that take place after training is completed. Finally, results criteria refer to the consequences of trainees' performance on the job. That is, did the training make a difference in the overall performance of the organization? Someone who can flawlessly demonstrate a skill that is never used on the job (e.g., knowing how to conduct radiological monitoring in a community that has no exposure to radiation hazard) will probably never have an impact on the safety of the community.

The same four criteria can be used in the assessment of a community risk communication program. Reaction criteria are measured by participants' reactions to the speaker, setting, communication medium, and message content; learning criteria are measured by the participants' judgments about the hazard and hazard adjustments. Behavior criteria are measured by households' and businesses' implementation of hazard adjustments; results criteria are measured by reductions in casualties, damage, and disruption from disasters. Clearly, reaction criteria are the easiest to collect, whereas learning and behavior criteria are more difficult to obtain. The infrequency of disasters makes the collection of results criteria extremely difficult.

The logic of causal inference

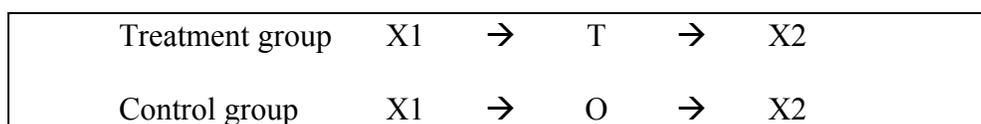
A treatment (e.g., a training program or a risk communication program) can be said to have had an effect on a dependent variable (e.g., a measure of reaction, learning, behavior, or

results criteria) if there is evidence that three conditions exist. Condition 1 is that the treatment (the potential cause) took place before the dependent variable (the potential consequence). Condition 2 can be satisfied in one of two ways. Either the scores of those who received the treatment have *changed* from what they were before they received the treatment or the scores of those who received the treatment are *different* from the scores of another group that has not received the treatment. It is crucial to the evaluation of Condition 2 that the changes or differences on the dependent variable are unlikely to have occurred by chance alone (the way this is determined will be explained below). Condition 3 is that the observed differences on the dependent variable cannot plausibly be explained by other systematic causes.

Simples study designs for assessing treatment effects

A study design is defined by the choices a program evaluator makes about the selection and assignment of participants to treatments and whether (and when) to measure variables. This section will discuss five study designs that can be used to test treatment effects. The two strongest designs are the *pretest-posttest control group design* and the *after-only control group design*, both of which are true experimental designs. Three other designs are quasi-experimental, meaning they provide some evidence for treatment effects, but it will not be as conclusive as that provided by true experimental designs. These are the *one group posttest-only design*, the *one group pretest-posttest design*, and the *posttest only design with nonequivalent groups*.

Pretest-posttest control group design. The diagram of this design shows the participants in the study are measured on a pretest at Time 1, after which they are assigned to treatment and control groups at Time 2, and then measured on the posttest at Time 3.

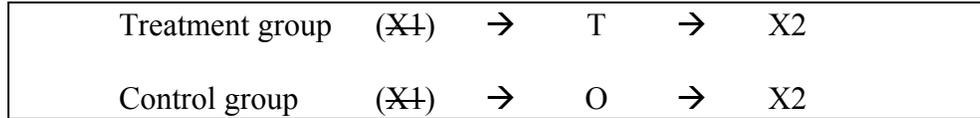


To meet Condition 1, participants are *randomly assigned* to groups before scores are collected on the posttest (X2). The treatment group receives the treatment (T = training or risk communication intervention) and the control group receives either no treatment or, preferably, a “filler” treatment that is expected to have no effect on the evaluation measures (O). A filler treatment is used if the members of the control group are likely to react negatively to believing the other group received a benefit they did not receive. Ideally, the control group should be treated in a manner that is identical in all respects other than the defining characteristic of the treatment (i.e., the particular aspect of training or risk communication that is of interest to the program evaluator). This is the same idea as giving a placebo (“sugar pill”) to the control group in a study of a drug’s effectiveness. Problems could arise in the interpretation of the study results unless the control group *appears* to the participants to be treated in the same way as the treatment group,

To meet Condition 2, participants’ scores on the pretest measure can be subtracted from their scores on the posttest measure to determine the difference between the two groups in their changes over time. Using the difference in the change scores as the dependent variable allows the researcher to verify the members of the treatment and control groups are equivalent on the measures used in the pretest and posttest and random assignment to conditions makes it unlikely the two groups differ systematically on any other characteristics that were not measured. This provides strong support for assuming Condition 3 has been met. If there are statistically significant differences between the treatment and control groups (the way statistical significance is determined will be explained later), this provides strong evidence in support of Condition 2.

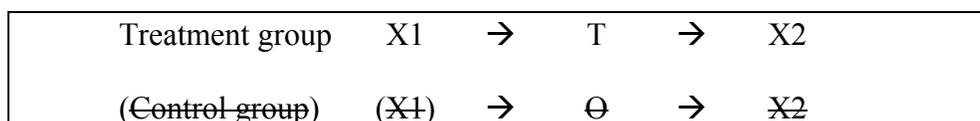
After-only control group design. The diagram of this design shows the participants in the study are *not* administered a pretest at time 1, designated by the strike through X1, but are

assigned to treatment and control groups, after which they are measured on the posttest.



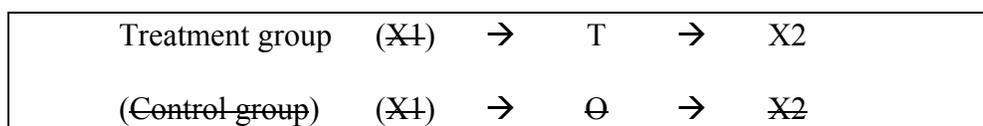
As with the pretest posttest control group design, the treatment group receives the treatment and the control group receives either a filler treatment or no treatment. Similarly, participants are *randomly assigned* to groups before scores are collected on the post-test (X2). As with the previous design, all participants are administered the posttest after the implementation of the treatment (to satisfy Condition 1), but here the between-group difference in *posttest* scores (not the between-group difference in *change* scores used in the previous design) is used to satisfy Condition 2. The absence of a pretest prevents the researcher from verifying the members of the treatment and control groups are equivalent on the measures of interest, but random assignment to conditions makes it unlikely the two groups differ systematically on this or any characteristics that were not measured. Consequently, we can have reasonable confidence that Condition 3 was met, but we cannot be as confident with this design as with the pretest-posttest control group design.

One group pretest-posttest design. The diagram of this design shows the study participants are given a treatment before the measurement of the scores on the post-test (satisfying Condition 1) and treatment group members are measured on the pretest at Time 1 and the posttest at Time 2. However, none of the participants in the study is assigned to a control group (designated in the study design diagram by the strike through all symbols associated with the control group)



This design does allow a researcher to address Condition 2 by determining if scores after the treatment have changed from those that existed before the treatment but not if they are also different from the scores of a group that has not received the treatment. The weakness of this design is its inability to rule out the possibility that any statistically significant change in mean scores from pretest to posttest is due to some other factor (Condition 3). For example, the use of this design to evaluate a risk communication program would be severely compromised if a tornado were to occur in another area of the same state or even if there were a particularly devastating tornado elsewhere in the country. Any change from pretest to posttest might be due solely to the treatment, solely to the reports of the tornado strike, or more likely some unknown combination of the two.

One group posttest-only design. The diagram of this design shows the participants in the study are *not* measured on the pretest at Time 1 (designated by the strike through X1), nor are they assigned (randomly or otherwise) to treatment and control groups (designated by the strike through the control group). Participants only receive the treatment followed by the posttest measure.



This design does meet Condition 1 because the treatment precedes the measure of effect. However, its obvious weaknesses are that the absence of a control group prevents the researcher from determining if there is a *difference* on the posttest measure and the absence of a pretest precludes determining if there has been a *change* on the posttest measure. These deficiencies usually prevent the program evaluator from testing Conditions 2 or 3, but these problems are sometimes addressed by asking the respondents to recall what they knew or had done before the

treatment occurred. When the “treatment” is an event that cannot be controlled (e.g., a hurricane or other disaster), there rarely is a feasible alternative to such “retrospective pretests”, but it is still possible to use control groups. In such cases, nearby communities with similar demographic characteristics that were unaffected by the disaster would be suitable choices for a control group. In any event, there is little excuse for neglecting a pretest before most training or risk communication interventions.

Nonequivalent control group designs. These designs occur when program evaluators are unable to randomly assign study participants to treatment and control groups, but nonequivalent control group designs look like experimental designs (e.g., the pretest-posttest control group design and the after-only control group design) in other respects. It is especially common for one pre-existing group to receive a treatment and another pre-existing group to be selected as a control group. As an example, a local emergency manager might want to use an *after-only control group design* to test a risk communication program administering posttests to two neighborhood associations after one received risk information brochures. This design would meet the requirements of Condition 1 because the treatment preceded the posttest. Even if the treatment group were to show significantly greater scores on the posttest in their risk perception and hazard adjustment scores (thus meeting the requirements of Condition 2), it is possible the differences between the two groups were due to pre-existing differences rather than to the effects of the treatment (thus, failing to meet Condition 3). To rule out this possibility of spurious causation, the program evaluator would need to rely on other information to establish the equivalence of the groups. To continue the example of the risk communication program, one might examine census data to determine whether the two neighborhoods were similar with respect to variables that have been shown to be correlated with hazard adjustment—such as

household income and the proportion of homeowners (see Lindell & Prater, 2000). Ruling out some plausible rival hypotheses for which measures are available cannot rule out other plausible rival hypotheses for which measures are *not* available, but ruling out at least some of the most plausible alternative explanations can give the program evaluator some degree of confidence the effect was, indeed, due to the treatment.

Statistical analysis of treatment effects

In order to draw conclusions about the effects of a treatment, it is necessary to demonstrate the differences between groups on the dependent variable, usually a posttest, are unlikely to have occurred by chance. This is accomplished by making one of three comparisons—*changes* in scores from the pretest to posttest, *differences* in scores between the treatment and control groups, or differences between the treatment and control groups in the *changes* in their scores from pretest to posttest. The rest of this section will describe the results of a hypothetical posttest-only design comparing the data from a treatment group and a control group in an evaluation of the effectiveness of a risk communication program in which the local emergency manager distributed information about seismic hazard to the treatment group but not the control group.

The data matrix. It is common for data to be collected on multiple items, which might be multiple questions from individuals responding to a questionnaire or multiple demographic characteristics about census tracts from the census data, or some other multiple attributes about multiple entities. Figure 12-1 displays the process by which data from a questionnaire are transferred to a data matrix in which each row represents a different entity (e.g., questionnaire respondent) and each column represents a different attribute (e.g., questionnaire item).

Figure 12-1. Data matrix



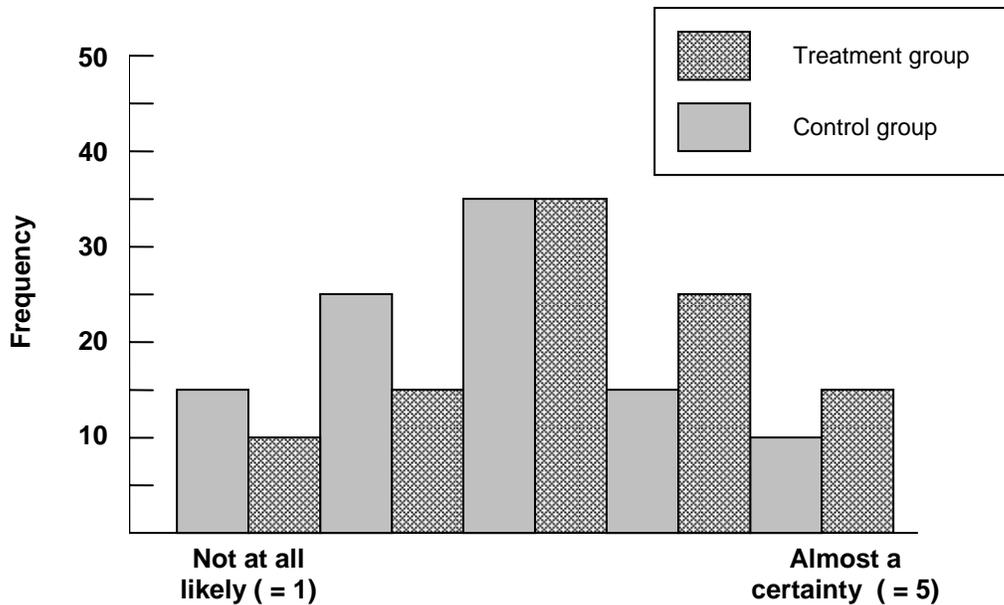
	1	2	...	j	...	N
1	X_{11}	X_{12}		X_{1j}		X_{1N}
2	X_{21}	X_{22}		X_{2j}		X_{2N}
...						
i	X_{i1}	X_{i2}		X_{ij}		X_{iN}
...						
M	X_{M1}	X_{M2}		X_{Mj}		X_{MN}

As Figure 12-1 indicates, each respondent from 1-M has a score on each item 1-N (although some of the scores might be missing because the respondent did not answer one or more of the questions. The data from a program evaluation might be further organized so the data from the control group are at the top of the matrix and the data from the treatment group are at the bottom of the matrix. This data matrix provides a method of neatly organizing the data, but it does not provide any clear insights about the differences between the treatment and control groups. To determine if there are systematic differences between the two groups, the data in the matrix are analyzed by displaying the distributions of scores on items and computing statistics on the item scores.

Distributions. When there multiple participants in a study, the responses to any item (i.e., a column in the data matrix) can be characterized by a *distribution* of scores. That is, some scores might be quite frequent, whereas other scores are rarely or never found. For example, Figure 12-2 shows two hypothetical distributions of scores on a question asking “To what extent do you think you will experience an earthquake in the next 10 years that will injure you or a member of your household?” The histogram shows the frequency with which respondents checked each of the five fixed-responses ranging from *Not at all likely* (= 1) to *Almost a certainty* (= 5). The frequency of each response can be determined by comparing the height of each bar to the scale on the left. The bars shaded gray show there were 15 members of the control group who checked Category 1, 25 who checked Category 2, 35 who checked Category

3, 15 who checked Category 4, and 10 who checked Category 5. By contrast, the crosshatched bars show there were 10 members of the treatment group who checked Category 1, 15 who checked Category 2, 35 who checked Category 3, 25 who checked Category 4, and 15 who checked Category 5.

Figure 12-2. Hypothetical distribution of scores on a questionnaire item.



This distribution of scores is informative because it is immediately obvious that the control group has more scores below the midpoint of the scale (40 of the 100 responses are in categories 1 or 2) than above the midpoint (25 responses are in categories 4 or 5). The reverse is true for the treatment group, which has fewer scores below the midpoint (25 responses are in categories 1 or 2) and more of them above the midpoint (40 responses are in categories 4 or 5). The difference in the pattern of responses suggests the risk communication program had a positive effect in increasing people's perceptions of risk. If there were only a few items on the questionnaire, such histograms would provide information about program effectiveness. However, if there are many items (questionnaires frequently have over 100 items), the large number of histograms would provide an overwhelming amount of information, so distributions

need to be summarized by *statistics*, which are numbers that summarize specific aspects of a distribution. The next sections describe statistics measuring three different aspects of distributions. These are the central tendency and dispersion of a distribution of scores on a single variable and the degree of association between the distributions of two different variables.

Measures of central tendency. Three different statistics can be used to measure a distribution's central tendency, which is the tendency of scores to bunch toward the middle of a distribution. Note that the middle of the distribution is not necessarily the same as the midpoint of the response scale because the middle of a distribution might be near one end of the response scale (e.g., the middle of the distribution might be at 3 on a scale ranging 1-10). The *mode* (M_o) is the most frequently response; in Figure 12-1, the mode for the control group is 2, whereas the mode for the treatment group is 3. The *mean* (M), which is more familiarly known as the average, is defined as the sum of the scores divided by the number of scores. That is, $M = (\sum X) / N$, where the expression " $\sum X$ " indicates the sum of the scores and the expression " $/ N$ " indicates the sum should be divided by the number (N) of scores. For the data in Figure 12-1, the mean for the control group is $M = 2.8$ and the mean for the treatment group is $M = 3.2$. The *median* (M_d) is the score for which half of the scores in the distribution fall below it and the other half fall above it (if there is an even number of scores, the median is the mean of the two middle scores). For the data in Figure 12-1, the median for the control group is $M_d = 3$ and the median for the treatment group is also $M_d = 3$. The medians for the two groups are the same because there are many responses ($N = 100$ in each group), but there are only five categories. Indeed, the median is quite useless in this type of situation, but it is much more useful for variables having many categories and especially for variables that can be measured on a continuous scale. In some cases, such as income, there are many fewer cases in the distribution with low values (e.g.,

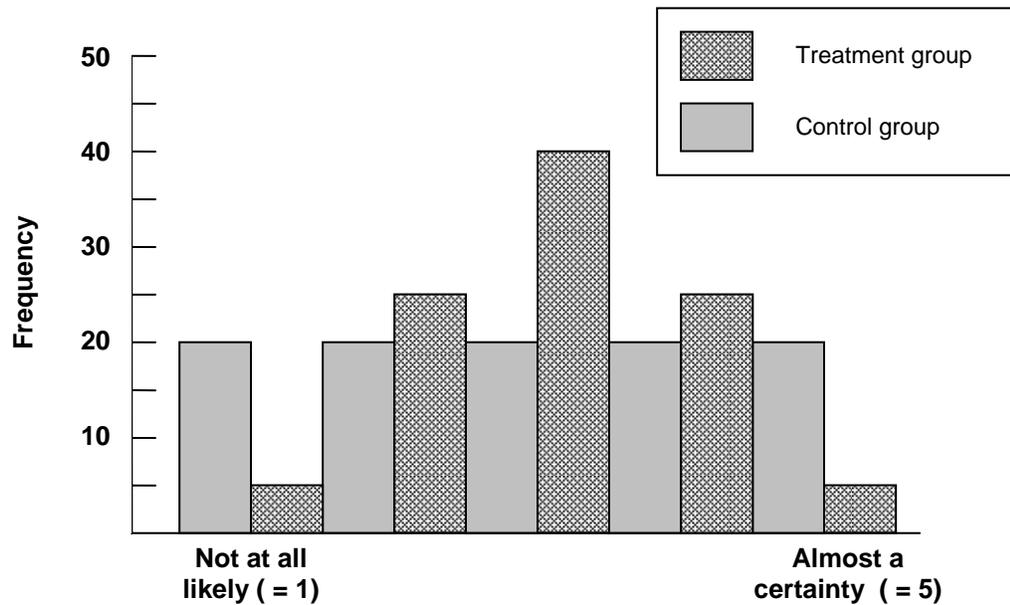
incomes less than \$50,000/year) than with high values (e.g., incomes greater than \$500,000/year). In such instances, the median can provide a much better estimate of a distribution's central tendency than can the mean. For example, the mean of the following five incomes—\$15,000; \$20,000; \$25,000; 30,000; and \$150,000—is $M = \$48,000$. This is a very misleading representation of the distribution's central tendency. However, $M_d = \$25,000$, which provides a more representative estimate of the central tendency in this distribution than does the mean.

Measures of dispersion. The dispersion, or variability, of the scores around the center is also an important aspect of a distribution because two different distributions could have the same mean, but that statistic would be much less representative of the observed values in one distribution than in the other. For example, Figure 12-2 shows a pair of distributions that have the same means (in both cases, $M = 3.0$) but different dispersions.

One way of measuring dispersion is the *range* of scores, which is computed by subtracting the lowest score from the highest score, but this index is not very useful for variables having few categories. In Figure 12-2, both distributions have the same range ($5 - 1 = 4$) even though the treatment group clearly has more scores near the center of the distribution. A more useful measure, the *variance*, is defined as the mean squared deviation from the mean of the distribution. Algebraically, this is expressed as $s^2 = (\sum X - Mn)^2 / N$, where the variance of the control group scores is $s^2 = 2.0$ and the variance of the treatment group scores is $s^2 = 0.9$. These results are consistent with an intuitive meaning of dispersion because the variance is smaller in the treatment group ($s^2 = 0.9$) than in the control group ($s^2 = 2.00$) just as the treatment group's scores are more closely clustered around the center of its distribution than the control group's scores are clustered around the center of its distribution. In practice, the square root of the

variance, known as the *standard deviation*, is more commonly used than the variance as a measure of dispersion because the standard deviation is measured in the same units as the original scores. The standard deviations for the control group and treatment group are 1.41 and 0.95, respectively.

Figure 12-2. Hypothetical distributions differing in dispersion.



Measures of association. We frequently want know whether there is some association between one variable and another. The simplest case occurs when two variables are measured as dichotomies; that is, they are measured as one of two states—either present or absent. To continue the example from the previous sections, suppose the evaluation of the risk communication program included one question assessing the respondents’ seismic risk perceptions as either high or low and another question asking if they had purchased earthquake insurance. We can judge the degree of association between these two variables in Table 12-3, which shows the percentage of respondents purchasing hazard insurance rises from 35% to 53%

when risk perception changes from low to high. Conversely, the percentage of respondents who fail to purchase hazard insurance decreases from 65% to 47% when risk perception changes from low to high. These results indicate a positive association between risk perception and hazard insurance purchase because an increased level of risk perception is associated with an increased level of hazard insurance purchase.

Table 12-3. Hypothetical association between seismic risk perception and hazard insurance purchase.

		Insurance purchase	
		No	Yes
Risk perception	High	47%	53%
	Low	65%	35%

However, suppose there are five items added to measure respondents' seismic risk perception and 16 emergency preparedness and hazard mitigation items added to measure their adoption of seismic hazard adjustments (Lindell & Prater, 2000). If a program evaluator wanted to know whether respondents' scores on the risk perception variable are associated with the hazard adjustment index, s/he could use a measure of association known as *Pearson's product moment correlation coefficient* (more simply known as Pearson's *r*). Pearson's *r* is equal to +1.0 when two variables are perfectly positively related, 0.0 when they are completely unrelated, and -1.0 when they are perfectly negatively related.

Statistical testing. Testing the statistical significance of the change in the mean scores between the pretest and posttest or of the difference between the treatment and control group is a topic that is an entire course in itself (e.g., Utts, 1999; Newton & Harvill, 1997). Nonetheless, the fundamental principles involved, and their relevance to program evaluation, can be explained in a few paragraphs. The basic aim of statistical significance testing is to address Condition 2—can the differences be explained by chance alone? In the case of an after-only control group design,

we have collected data from N_T members of the treatment group and N_C members of the control group. It is easy to calculate the mean of the treatment group (M_T) and the control group (M_C), and it is almost certain the means of the two groups will not be *exactly* the same. To determine if the observed difference could have occurred by chance alone, it is necessary to make some assumptions. The first assumption is members of the treatment and control groups are *randomly sampled* from a *population* of individuals and that the results are to be generalized to this population. The population to which the results are to be generalized is simply the set of all people to whom the conclusions are expected to apply. For example, a program evaluator might define the population as the set of all households within a particular jurisdiction. The assumption of *random sampling* means each member of the population has an approximately equal probability of being selected. It is important to recognize random sampling from a population is different from random assignment to either the treatment or control group. Ideally, random sampling takes place first and then those who have been randomly selected are randomly assigned to either the treatment or control group.

In theory, we could draw conclusions about the probability of chance findings by using a computer to generate samples of data from a single population. We could then draw repeated samples of individuals from the population, randomly assign them to either the treatment or control group, and measure each group's mean score on the posttest (M_T and M_C). We could then compute the difference (D_{T-C}) between the mean for the treatment group and the mean for the control group and, if we did this a large number of times, generate a distribution of these differences between means. We could then compare the difference (D_{T-C}) we obtained from the study we actually conducted with the distribution of differences that was obtained from the

computer simulation experiment. If the obtained difference (D_{T-C}) was larger than 95% of the values in the simulated distribution, we could conclude that any difference in means that is this large or larger is unlikely to have occurred by chance.

In practice, this procedure is unnecessary because statisticians have identified conditions under which it is possible to use a *theoretical* sampling distribution of the differences in means rather than generating an *actual* sampling distribution. In fact, it is possible to determine the statistical significance of the difference between two means by using the formula

$$t = \frac{M_T - M_C}{SE_d}, \quad (1)$$

where M_T is the mean of the treatment group and M_C is the mean of the control group. In addition, there is SE_d , which is called the standard error of the difference. This is simply the standard deviation of the sampling distribution of the difference in the means (D_{T-C}). After subtracting M_C from M_T and dividing by SE_d , the obtained value of t (called the t -statistic) is compared to a value in a statistical table to determine if the obtained value exceeds the tabled value. If this is the case, the difference in means is said to be statistically significant, meaning it is unlikely (but not impossible) the obtained difference between the two means could have occurred by chance.

Although the computation of the two means is quite simple, the formula for calculating the standard error of the difference, SE_d , is substantially more complex. Indeed,

$$SE_d = \sqrt{\frac{(N_T - 1)s_T^2 + (N_C - 1)s_C^2}{N_T + N_C - 2} \left(\frac{1}{N_T} + \frac{1}{N_C} \right)}, \quad (2)$$

where s_T^2 is the variance of the treatment group scores and s_C^2 is the variance of the control group scores. The complexity of this formula is not a problem in practice because spreadsheets such as

Microsoft EXCEL and statistical packages such as the Statistical Package for the Social Sciences (SPSS) perform all the calculations once the data have been entered into a data matrix (see Appendix A for an illustration of the use of SPSS to perform a statistical test).

The formulas for t (Equation 1) and SE_d (Equation 2) reveal three points that are important in any program evaluation. First, large values of t are desirable because they indicate the difference in means is unlikely to have occurred by chance. Thus, Equation 1 implies SE_d should be as small as possible. Second, Equation 2 implies small values of SE_d will be achieved when the variances are small within the treatment and control groups. This means treatment effects are more likely to be classified as statistically significant if the members of the treatment and control groups are *each* relatively homogeneous in their scores on the dependent variable. Third, Equation 2 implies small values of SE_d will be achieved when the sample sizes are large within the treatment and control groups. This means program evaluators should seek large samples to test the effects of their programs.

In practice, samples of the size used as examples in this chapter are usually satisfactory but there might be cases in which larger samples are needed, or other cases in which smaller samples will suffice. Any local emergency manager seeking to evaluate training or risk communication programs should seek the assistance of an experienced program evaluator such as a statistician or applied social scientist.

References

- Cascio, W.F. (1998). *Applied psychology in human resource management*, 5th ed. Upper Saddle River, NJ : Prentice Hall.
- Cook, T.D., Campbell, D.T. & Peracchio, L. (1990). Quasi experimentation. Pp. 491-576 in M.D. Dunnette and L.M. Hough (eds.) *Handbook of industrial & organizational psychology*, 2nd ed., Vol 3. Palo Alto CA: Consulting Psychologists Press.
- Goldstein, I.L. (1993). *Training in organizations: Needs assessment, development and evaluation*, 3rd ed. Pacific Grove CA: Brooks/Cole.
- Newton, H.J. & Harvill, J.L. (1997). *StatConcepts: A visual tour of statistical ideas*. Pacific Grove CA: Duxbury.
- Lindell, M.K. & Prater, C.S. (2000). Household adoption of seismic hazard adjustments: A comparison of residents in two states. *International Journal of Mass Emergencies and Disasters*, 18, 317-338.
- McIntyre, R.M. & Salas, E. (1995). Measuring and managing for team performance: Lessons from complex environments. Pp. 9-45 in R. A. Guzzo, E. Salas and Associates (eds.). *Team effectiveness and decision making in organizations*. San Francisco: Josey-Bass.
- National Response Team. (1990). *Developing a hazardous materials exercise program: a handbook for state and local officials, NRT-2*. Washington DC: Author.
- Schmitt, N.W. & Klimoski, R.J. (1991). *Research methods in human resources management*. Cincinnati OH: South-Western.
- Schutt, R.K. (2001). *Investigating the social world: The process and practice of research* (3rd ed.). Thousand Oaks CA: Pine Forge Press.
- Utts, J.M. (1999). *Seeing through statistics*, 2nd ed. Pacific Grove CA: Duxbury.

Appendix A: Evaluation of a risk communication program

A local emergency manager has implemented a risk communication program in two neighborhoods within her county. One neighborhood received a brochure describing the community's exposure to earthquakes and actions homeowners could take to protect themselves and their property from earthquakes. She has examined the census data for the community and confirmed both neighborhoods have very similar demographic characteristics. Both are stable middle income neighborhoods with a majority of married homeowners having families. A random sample of 200 households was selected in each neighborhood and 50% of each sample returned completed questionnaires. The distributions of control and treatment group responses to a question asking "To what extent do you think you will experience an earthquake in the next 10 years that will injure you or a member of your household?" was displayed in Figure 12-1. As reported earlier, the mean for the control group was $M_C = 2.80$ and the mean for the treatment group was $M_D = 3.20$. The means are in the direction that would be expected if the treatment increased people's perception of earthquake hazard, but a statistical test is needed to determine if the difference is likely to have occurred by chance. A *t*-test conducted using SPSS shows the following results.

Table A-1. Group Statistics

Variable	Group	N	Mean	Std. Deviation
Riskper	Control	100	2.80	1.172
	Treatment	100	3.20	1.172

Table A-2. Independent Samples Test

t	df	Significance (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
					Lower	Upper
2.413	198	.017	.40	.166	.073	.727

Table A-1 lists the variable analyzed (*Riskper*, which is a variable label that provides an abbreviated description of the item content, risk perception), the names of the two groups (on separate lines), the number of respondents from each group, and the mean and standard deviation for each group. Table A-2 lists the value of the *t*-statistic ($t = 2.413$), the number of degrees of freedom ($df = N_D + N_C - 2 = 198$), the significance level (the probability that the observed value of the *t* statistic could have arisen by chance alone, $p = .017$), the difference between the means of the two groups ($MD = M_D - M_C = .40$), the standard error of the difference ($SE_D = .166$, which was calculated using Equation 2), and a 95% Confidence Interval of the Difference, which says—with 95% confidence—that the two groups are not drawn from the same population because the confidence interval does not include a difference of zero.