CDC

CENTERS FOR DISEASE CONTROL AND PREVENTION

Guide to Evaluating the Effectiveness of Strategies for Preventing Work Injuries

How to show whether a safety intervention really works

National Institute for Occupational Safety and Health
NIOSH

SCIENTIFIC COMMITTEE ON ACCIDENT PREVENTION

INSTITUTE FOR WORK & HEALTH
INSTITUT DE RECHERCHE SUR LE TRAVAIL ET LA SANTÉ

Guide to Evaluating the Effectiveness of Strategies for Preventing Work Injuries:

How to Show Whether a Safety Intervention Really Works

Lynda S. Robson, Harry S. Shannon, Linda M. Goldenhar, Andrew R. Hale

DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Centers for Disease Control and Prevention
National Institute for Occupational Safety and Health

April 2001

DISCLAIMER

Mention of any company name or product does not constitute endorsement by the National Institute for Occupational Safety and Health.

Copies of this and other NIOSH documents are available from NIOSH.

For information about occupational safety and health topics contact NIOSH at:

1-800-35-NIOSH (1-800-356-4674)
Fax: 513-533-8573
E-mail: pubstaft@cdc.gov
www.cdc.gov/niosh

National Institute for Occupational Safety and Health
Publications Dissemination
4676 Columbia Parkway
Cincinnati, OH 45226-1998

For information about the Institute For Work & Health and its research:

416-927-2027
Fax: 416-927-2167
E-mail: info@iwh.on.ca
www.iwh.on.ca

DHHS (NIOSH) Publication No. 2001-119

Table of Contents

Acknowledgements

The idea for this guide arose in a meeting of the Scientific Committee on Accident Prevention (SCOAP) at the International Commission on Occupational Health (ICOH) meeting in Tampere, Finland, July 1997.

The co-authors thank the many individuals and organizations which gave valuable feedback to draft versions of the Guide. These include: other members of the NORA Intervention Effectiveness Team (in particular, Larry Chapman, Catherine Heaney, Ted Katz, Paul Landsbergis, Ted Scharf, Marjorie Wallace); other SCOAP Committee members (Tore Larsson, Jorma Saari); additional academic colleagues (Donald Cole, Xavier Cuny, Michel Guillemin, Richie Gun, Gudela Grote, Per Langaa Jensen, Urban Kjellén, Richard Wells); individuals within Ministries or Departments of Labour (Brian Zaidman, Barry Warrack), the Workplace Safety Insurance Board of Ontario (Richard Allingham, Marian Levitsky, Kathryn Woodcock), Workers' Compensation Board of BC (Jayne Player), and Workplace Health, Safety and Compensation Commission of New Brunswick (Susan Linton); representatives of Ontario Safe Workplace Associations (Dave Snaith (Construction), Linda Saak and R. Stahlbaum (Electrical & Utilities), James Hansen (IAPA), Louisa Yue-Chan (Services), Mark Diacur (Transportation)); Irene Harris and members of the Occupational Health and Safety Committee of the Ontario Federation of Labour; Mary Cook, Occupational Health Clinics for Ontario Workers. Institute members who shared reference materials are also thanked - John Lavis and Dorcas Beaton - as is Linda Harlowe (IWH), who constructed all of the detailed graphical figures.

The final word of acknowledgement goes to the existing body of evaluation expertise that we drew upon and adapted to a safety application. We are especially indebted to the classic works of Cook and Campbell [1979] and Patton [1987, 1990].

Preface

Our aim in this book is to provide students, researchers and practitioners with the tools and concepts required to conduct systematic evaluations of injury prevention initiatives and safety programs. Successful evaluations will advance the discipline of occupational safety by building a body of knowledge, based on scientific evidence, that can be applied confidently in the workplace. This knowledge will provide a solid foundation for good safety practice, as well as inform the development of standards and regulations. Building such a knowledge base will help practitioners avoid the temptation of adopting safety procedures simply because they appear "intuitively obvious" when no scientific evidence actually exists for those practices.

Users of the guide are encouraged to demonstrate the strongest level of evidence available for an intervention by measuring the effect on safety outcomes in an experimental design. Even when this level of evidence is not obtained, much useful information can still be gained by following the recommendations in the book. In doing so, the safety field will become current with other disciplines, such as clinical medicine, where evaluation information is increasingly available and allows for evidence-based decision-making.

We hope that this guide will assist safety specialists to meet the challenge of effectiveness evaluations. Please let us know if you found it useful by completing the evaluation form provided at the end to the document.

Information About Authors

Linda M. Goldenhar

Linda Goldenhar is a Research Psychologist at the National Institute for Occupational Safety and Heath. She received her Ph.D. in Health Behavior at the University of Michigan. Her interests include intervention research, quantitative and qualitative data collection methods, work-organization, job stress, and women's health. She is currently the team leader for the NIOSH NORA Intervention Effectiveness Research team. The mission of the team is to educate occupational researchers about intervention research issues and to encourage the development, implementation, and evaluation of occupational safety and health interventions. Linda has published numerous peer-reviewed articles and delivered conference presentations covering a wide array of occupational health-related and behavioral topic areas. She is on the editorial board of Health Education and Behavior and Journal of Safety Research.

Andrew R. Hale

Andrew Hale is professor of Safety Science at the Delft University of Technology in the Netherlands and editor of the journal Safety Science. His background is as an occupational psychologist. He worked for 18 years in the UK at the National Institute of Industrial Psychology and later at the University of Aston in Birmingham. His experience in research in safety and health began with studies of accidents and human error in industry and moved later into studies of risk perception, safety training and more recently safety management and regulation. Since moving to Delft from Aston he has expanded his area of research applications from industry to include transport (road, water and air), hospitals and public safety.

As editor of Safety Science and as a reviewer for both this journal and a number of others he is responsible for assessing the articles presented for publication. In that capacity he has seen at close quarters the need for improvement in the quality and application of the research and analysis methods used in safety science.

Lynda S. Robson

Lynda Robson is a Research Associate at the Institute for Work & Health, Toronto, Canada. She obtained her Ph.D. in Biochemistry from the University of Toronto, working afterwards in research labs. More recent experience includes graduate-level training in epidemiology and evaluation methods, as well as collaboration in economic evaluation projects. Her current areas of interest are the measurement of "healthy workplace" performance and safety intervention effectiveness evaluation methodology. She is the non-management co-chair of the Institute's Joint Health & Safety Committee.

Harry S. Shannon

Harry Shannon is currently a Professor in the Department of Clinical Epidemiology Biostatistics and Director of the Program in Occupational Health and Environmental Medicine at McMaster University, Hamilton, Canada. He is also seconded to the Institute for Work & Health, Toronto, Canada as a Senior Scientist.

Harry holds a B.A. in Mathematics from Oxford University, an M.Sc. from Birmingham University and a Ph.D. in Applied Statistics at the University of London, U.K. He has authored over 75 peer-reviewed scientific papers, mainly in occupational health.

Harry's current research focus is on work-related injuries and musculoskeletal disorders, and psychosocial conditions at work. Recent research includes studies of back pain in auto workers and upper extremity disorders ("RSIs") in newspaper workers. He has also studied the relationship between workplace organizational factors and injury rates. Harry is a member of the Scientific Committee on Accident Prevention (SCOAP) of the International Commission on Occupational Health (ICOH). He is on the editorial board of Safety Science; and is a member of the NIOSH NORA group on the Social and Economic Consequences of Occupational Illness and Injury.

Chapter 1

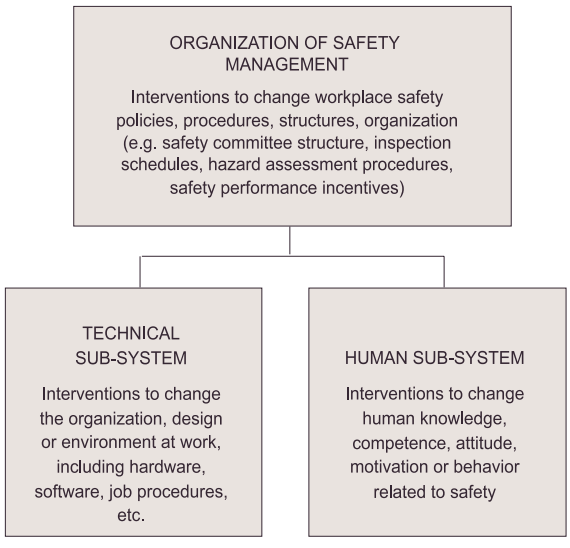
Introduction: Safety Intervention Effectiveness Evaluation

1.1 What is a safety intervention?

A *safety intervention* is defined very simply as an attempt to change how things are done in order to improve safety. Within the workplace it could be any new program, practice, or initiative intended to improve safety (e.g., engineering intervention, training program, administrative procedure).

Safety interventions occur at different levels of a workplace safety system (Figure 1.1), including the level of safety management and various human and technical sub-system levels in the organization that management can influence. An additional level of the system in Figure 1.1, above the organizational, pertains to the laws, regulations, standards and programs put in place by governments, industries, professional bodies, and others. Examples of interventions at this level include the Safe Communities Incentive Program (Safe Communities Foundation, Ontario, Canada), the Safety Achievers Bonus Scheme (South Australian Workcover Corporation) and small business insurance pooling (CRAM, France). This guide does not deal with interventions at the community level, although some of the issues discussed are applicable.

Figure 1.1 Levels of intervention in the workplace safety system



1.2 Effectiveness evaluation

We are focusing here on *effectiveness evaluation* (also known as outcome evaluation or summative evaluation), which determines whether a safety initiative has had the intended effect. For example, such an evaluation might answer the question, does the new incident investigation process instituted three years ago (for the purpose of decreasing injuries) actually prevent injuries in subsequent years? This type of evaluation is the "CHECK" portion of the PLAN-DO-CHECK-ACT (PDCA) continuous quality improvement cycle.[4]

Although injuries are often measured in an effectiveness evaluation to determine whether the initiative has had an effect or not, there are two situations where this might not be the case. One of them arises from injury outcome data that is unreliable or invalid - (e.g., while evaluating an initiative in a small workplace). In this case a surrogate measure of safety could be used (e.g., a checklist of safety conditions), if shown to be valid for the circumstances in which it will be used. The other situation is where the program's explicit objective is not to decrease injury incidence, but rather, some other objective such as improve worker or management competence or attitudes.

However, if the purpose of the program is to ultimately affect injury incidence by targeting competence or attitudes, it would be beneficial to include a measure of injuries or a valid surrogate.

1.3 Overview of the evaluation process and the guide

Figure 1.2 provides an overview of the effectiveness evaluation process. Much of the activity in evaluation precedes the point where the intervention or initiative is introduced. Although evaluations can be done retrospectively, appropriate data are typically unavailable. The guidelines contained in this guide assume that a safety need in the workplace and the nature of the intervention have been identified, but you are at a point prior to introducing a new intervention in the workplace. Even if this is not the case, and you have already implemented your intervention, you should find the guide relevant to the evaluation of your intervention.

Chapter 2 identifies the decisions required before people can work out the details of designing an evaluation. They include the available resources, time lines and, of course, the purpose of the evaluation. A few of the broader methodological issues, such as the use of qualitative and quantitative methods and choice of outcomes, are also introduced.

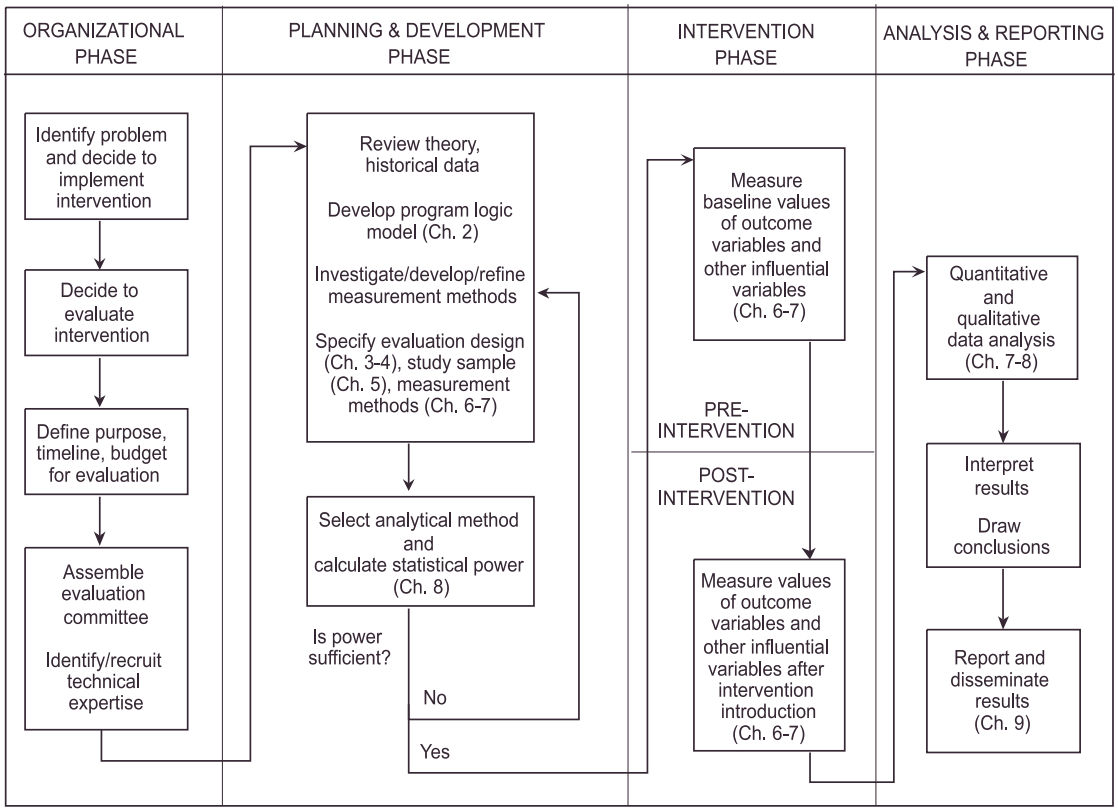
Chapters 3 and 4 introduce several evaluation designs (i.e., methods of conducting evaluations in conjunction with delivering an intervention). These designs specify the groups of people or workplace to be evaluated, as well as what will be measured and when it will occur.

Chapter 5 explains in more detail who to include in the evaluation - a choice which affects the generalizability of the evaluation results and the types of statistical analyses to be used.

The next two chapters (6 and 7) consider quantitative (Chapter 6) and qualitative (Chapter 7) data collection methods. Quantitative methods yield the numeric information necessary to determine the size of an intervention's effect and its trustworthiness - determined through statistical testing (Chapter 8). Qualitative methods yield conceptual information which can inform the evaluation design at the beginning, and the interpretation of the results at the end. The guide ends with a summary of the practice recommended in previous chapters (Chapter 9).

[4] PDCA cycle mentioned in many references on quality improvement concepts. For example: Dennis P [1997] Quality, safety, and environment. Milwaukee, Wisconsin: ASQC Quality Press.

Figure 1.2 Overview of the effectiveness evaluation process



1.4 Other types of evaluations

Other types of evaluation, besides effectiveness evaluation, are useful in the process of improving safety in the workplace. They will only be described briefly here. A *needs assessment* can be carried out to determine exactly what type of intervention is required in a workplace. Analyses of injury statistics, incident reports or employee surveys, as well as interviews with key workplace personnel (e.g., safety manager, disability manager, union representative, etc.) can identify particular safety issues. This determines what type of intervention(s) should be chosen or designed to address an identified need.

After choosing and introducing a new safety initiative to a workplace, a *process evaluation* (also known as a formative evaluation) can be used to determine whether the new initiative is being implemented as planned. It assesses to what extent new processes have been put in place and the reactions of people affected by the processes. Furthermore, the refinement of a new initiative and its implementation before its effectiveness can be measured. If the process evaluation determines that the initiative was not implemented as planned, the time and trouble of conducting an effectiveness evaluation might be spared, or at least delayed until it becomes more meaningful.

Finally, economic analyses can be used to evaluate workplace interventions, including *cost-outcome, cost-effectiveness* and *cost-benefit analyses*. They also depend on effectiveness information. The first two analyses estimate the net cost of an intervention (i.e., the cost of the intervention

minus the monetary saving derived from the intervention) relative to the amount of safety improvement achieved. (Monetary savings include reductions in workers' compensation premiums, medical costs, absenteeism, and turnover, etc.) This yields a ratio such as net cost per injury prevented.

In a cost-benefit analysis, monetary values are assigned to all costs and outcomes resulting from an intervention, including health outcomes. Furthermore, a net (monetized) benefit or cost of the intervention is calculated.

Drummond et al. [1994], Haddix et al. [1996] and Gold et al. [1996] are useful introductions to economic evaluations.

Table 1.1 Types of intervention evaluations

Types of Evaluations	Purpose
Needs assessment	Determines what type of intervention is needed
Process evaluation	Assesses the quality of the intervention delivery and identifies areas for improvement
Effectiveness evaluation	Determines whether an intervention has had the effect intended on outcomes, and estimates the size of the effect
Cost-outcome analysis	Determines the net cost of an intervention relative to its health effect
Cost-effectiveness analysis	Compares different intervention alternatives using cost-effect ratios
Cost-benefit analysis	Compares different intervention alternatives using net benefits

Chapter 2

Planning Right from the Start

2.1 Introduction

One golden rule of intervention evaluation is that the intervention and its evaluation should be planned simultaneously. Decide on your evaluation design and methods **before** the intervention is introduced. If you do not, you risk missing the only opportunity for collecting important data before the intervention.

This chapter gives some information about organizing and carrying out an intervention evaluation, including who to involve to oversee the evaluation and how to use models to assist with planning. It also highlights key issues to consider during the planning stage.

2.2 Defining the scope of the evaluation

Some basic decisions are required before a detailed evaluation strategy can be specified. They should be done through a collaborative process among those who will use the evaluation results, and others either implementing the intervention and evaluation project or funding the evaluation. In a workplace setting, these decisions should involve senior management. The types of things to be determined at the outset should include: 1) overall purpose of the evaluation; 2) the main questions that the evaluation should answer; 3) available resources (financial, personnel, in-kind assistance); and 4) the deadline for the evaluation results. These have to be in place early on, since they will influence the methodological deliberations. In particular, the rationale for the evaluation will influence the strength of the evidence sought. For example, you might want a more rigorous evaluation design if the result of the evaluation is meant to inform a decision with larger resource or policy implications.

2.3 Who should be involved with the evaluation?

2.3.1 Evaluation committee

In all but the simplest evaluation, it is advisable to assemble a committee to oversee the evaluation. Ideally, this same group would also interact with the evaluators in the selecting or designing of the safety intervention. The larger the potential resource and personnel impacts of any conclusion drawn from the evaluation, the greater is the need to have an evaluation committee. The following factors can be considered in deciding whether or not to set up a committee.

People with different backgrounds and skill-sets will be included in an evaluation committee and any individual might play more than one role. Key are those who will use, or could influence, the use of the evaluation results. Their inclusion will ensure the following: the evaluation will address decision-makers' concerns; the evaluation and its results will be legitimized; and communication of the evaluation results will be initiated early in the evaluation process. You need at least one person who can understand and critique the methodological issues of both the intervention and the evaluation.

Also important is the expertise of those directly affected by the intervention, who have special insight into the reality of the workplace conditions. Senior management and labor representatives enhance the decision-making capacity of the committee, as well as facilitate required changes in workplace practices and mobilize the cooperation of all involved. Intervention skeptics are put to good use, as their input will likely result in a more rigorous evaluation. On the other hand, individuals involved in either choosing or developing the intervention should also be included, otherwise they might be reluctant to accept any evidence that their intervention is ineffective.

Evaluation committee representation ideally includes:
• Stakeholders key to utilization and dissemination of results • Key management representatives (e.g., relevant decision-maker) • Key worker representatives (e.g., union representatives, opinion leaders, intervention participants) • Evaluation expertise • Diversity of disciplinary perspectives (e.g., engineering, safety, human resources, etc.) • Diversity of workplace divisions / departments • An intervention critic, as well as intervention proponents

There is a wide range of capacities in which evaluation committees can function. At one extreme, committee members can serve in a relatively passive fashion, giving or denying approval to what the core evaluation team develops. At the other end of the spectrum would be an *action research* model: all workplace representatives and methodological experts collaborate to formulate evaluation questions, design the evaluation and interpret results. An example of this model is the article by Hugentobler et al. [1992]. The choice of where to place the committee between the two extremes involves weighing the complexity of the intervention and how widespread the buy-in has to be, as well as the time and resources available.

2.3.2 Internal vs. external evaluators

From a scientific point of view, it is preferable that an intervention is evaluated by an independent party with no vested interest in showing that the intervention is effective (or not effective). In spite of all efforts to be open-minded, it is simply human nature to put more effort into finding and drawing conclusions from information which confirms our expectations than contradicts them. However, practitioners often find they have to choose between carrying out the evaluation themselves or having no evaluation at all. Although the bias inherent in an "in-house" evaluation is never entirely removed, it can be diminished. This is achieved by inviting others, especially those with opposing views, to comment on analyses and reports of the evaluation, and by being very explicit in advance about what will be measured and regarded as evidence of success or failure.

2.3.3 Technical or methodological expertise

It is quite possible that you might have to look outside your own organization to tap into some of the specialized skills required for certain evaluations. Common areas requiring assistance are questionnaire development and statistical analysis. Consider contacting the local academic institutions, for expertise in one of several departments: biostatistics, occupational health & safety, management, (social or occupational) psychology, public health, education. Some universities even have consulting services geared to assisting community residents. Other means of contacting experts would be through safety research organizations, safety consultants or safety professional organizations. In any case, the rule is as before: involve these people early in the evaluation.

2.4 Models to assist planning

Certain types of models can assist in planning all but the simplest intervention and evaluation. They diagrammatically depict the important relationships among the workplace conditions, interventions and outcomes, as well as identify what should be measured in the evaluation. The process of constructing models will often reveal critical gaps in thinking and identify critical issues. They also provide a means of generating common understanding and communicating among those involved in the intervention and evaluation, as well as serving as an efficient aid to communicating with others.

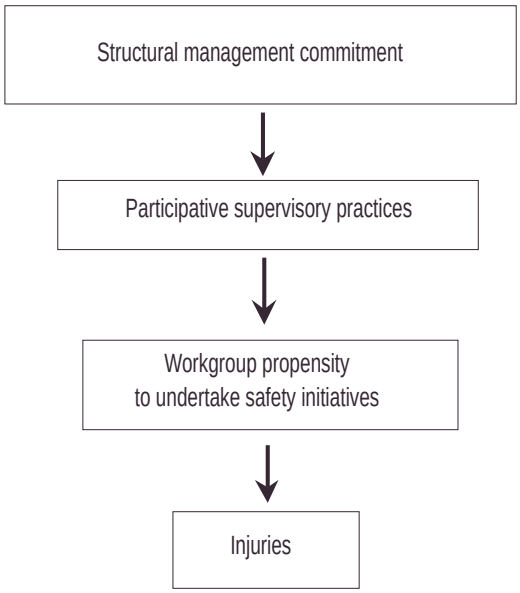
One type of model is a conceptual model; another is a program logic model. Both are somewhat different and complementary: the conceptual model tends to be more comprehensive in its inclusion of factors which affect the safety outcomes of interest and the program logic model often includes more information on the intervention itself. Conceptual models are used by researchers in many disciplines, while program logic models are used by program evaluators - principally those evaluating social or public service programs.

2.4.1 Conceptual models

A conceptual model typically uses arrows and boxes to represent the causal relationships (arrows) among important concepts (boxes) relevant to an intervention. An example follows (Figure 2.1).

The arrows indicate that structural management commitment affects participative supervisory practices, which in turn affect the workgroup's propensity for undertaking safety initiatives; and these initiatives affect the likelihood of injuries. The relationship represented by an arrow in a model can be either positive or negative. We expect a greater structural management commitment will lead to more participative supervisory practices - a positive relationship; while greater workgroup propensity to undertake safety initiatives would lead to fewer injuries - a negative relationship. More complex models show multiple relationships, involving several arrows and boxes. Examples of more complex models are included in an appendix to this guide.

Figure 2.1: An illustration of a simple conceptual model[5]



We recommend that a *conceptual model* relevant to the intervention be developed[6] while the intervention is being chosen or designed. Ideally, any model would be based as much as possible on existing research and theory. You might adapt an existing model to your own situation. The conceptual model helps clarify what the intervention hopes to change and the mechanism by which that should happen. This clarification could even reveal certain causal factors not yet addressed by the intervention.

[5] This model is based on some of the relationships reported in Simard M, Marchand A [1995]. A multilevel analysis of organizational factors related to the taking of safety initiatives by work groups. Safety Sci 21:113-129.

[6] For guidance in constructing conceptual models, see Earp and Ennett [1991].

The conceptual model tells us what should be measured in an evaluation. What we measure using quantitative methods are known as *variables;* i.e., attributes, properties or events which can take on different values and correspond to concepts in the model. Using the conceptual model above, the numerical score from supervisors completing questionnaires on safety practices could be the variable corresponding to the concept "participatory supervisory practices".

Independent and dependent variables

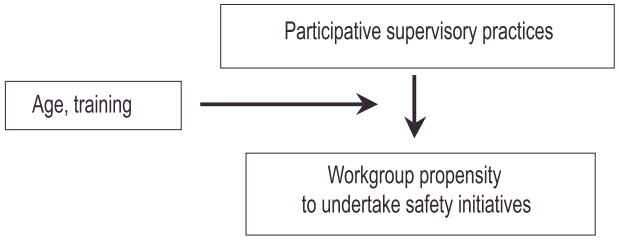
As demonstrated, the conceptual model does not describe the safety intervention itself, but rather, depicts the variables expected to change during an intervention. For example, the above model might be applicable in an intervention where structural management commitment was going to be changed by establishing a joint management-labor health-and-safety committee. The variable being manipulated as part of the intervention, i.e., presence or absence of a joint health-and-safety committee in this case, is known as the *independent variable.*

In contrast, the variables affected by the intervention, i.e., the variables corresponding to the other three concepts in the model, are known as *dependent variables*. These include the variable(s) corresponding to the *final outcome,* "injuries" in this case. They also include the variables corresponding to the concepts which mediate an intervention's effect, e.g "participatory supervisory practices" and "workgroup propensity to take safety initiatives". The latter are known as *mediating, intervening* or *intermediate variables.*

Effect-modifying and confounding variables

Effect-modifying variables[7] are sometimes important to include in a conceptual model. While these variables are not the focus of the intervention, they often need to be considered when data is collected and interpreted. An effect-modifying variable is one which modifies the size and direction of the causal relationship between two variables. For example, "participative supervisory practices" might have a greater effect on "workgroup propensity to take safety initiatives" if the workgroup has more training in safety or is younger. An effect modifying variable is depicted by an arrow extending from it to the relationship which it modifies, as Figure 2.2 shows.

Figure 2.2: Depiction of an effect-modifying variable in a conceptual model

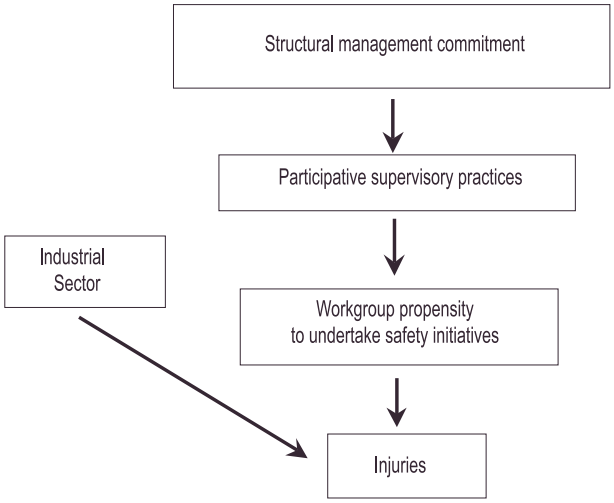


Another type of variable to include in a conceptual model is a *confounding variable.* This is a variable which is related to the independent variable (i.e. presence of intervention or not), as well as the dependent variable of interest, but is not a mediating variable. To build on the earlier illustration, "industrial sector" could be a confounding variable if one is looking at the effect of changing "structural management commitment" through the legislated introduction of joint health-and-safety committees (see Figure 2.3).

Suppose that you decided to evaluate the effect of the legislation by comparing the injury rate of organizations with established committees to those without them. Suppose too, that the industrial sectors with the higher injury rates are more likely to have committees because of greater inspector activity. If you were to then compare injury rates of companies with

[7] In some disciplines, such variables would be referred to as moderator variables.

Figure 2.3 Depiction of a confounding variable in a conceptual model



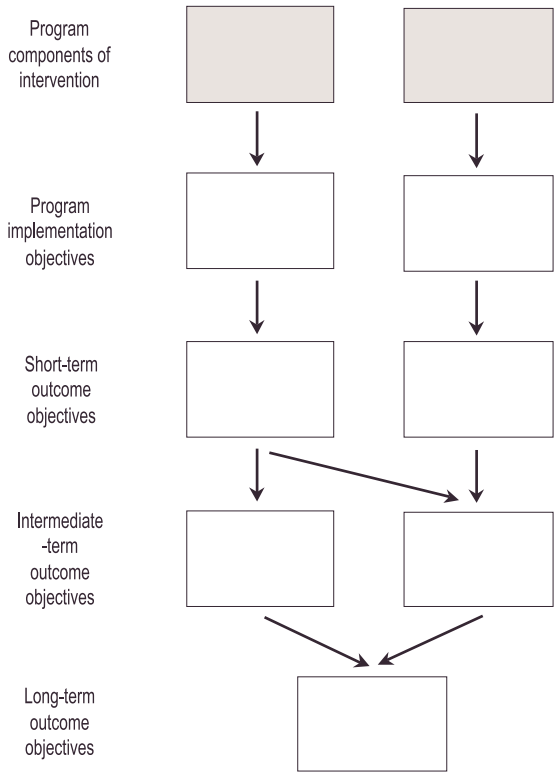
committees (intervention group) versus those without committees (control group), you would find that the companies with the committees had the higher injury rates. However, the result is most likely due to the specific industrial sector, rather than the intervention. Thus, a conclusion that committees result in higher injury rates would be inaccurate. You would form a different conclusion if you took the sector into account in the study design or analysis. To limit the influence of confounding factors, take them into account in your study design (preferable) or your analysis.

2.4.2 Program logic models

Program logic models are highly regarded by evaluators as a means of assisting evaluation planning[8]. They distinguish short-, intermediate- and long-term outcome objectives. Figure 2.4 depicts a generic program logic model. Implementation objectives are concerned with the desired outputs of the intervention (e.g., provide training sessions). They are distinct from outcome objectives, which are concerned with the effects of the program. Continuing the training example, a short-term objective could be improved knowledge; an intermediate objective, changed behavior; and a long-term objective, decreased injuries.

A program logic model, developed for a particular safety intervention, could have more or less boxes than shown in Figure 2.4, depending on the number of intervention components and objectives. Also, they could be linked in a different pattern of arrows.

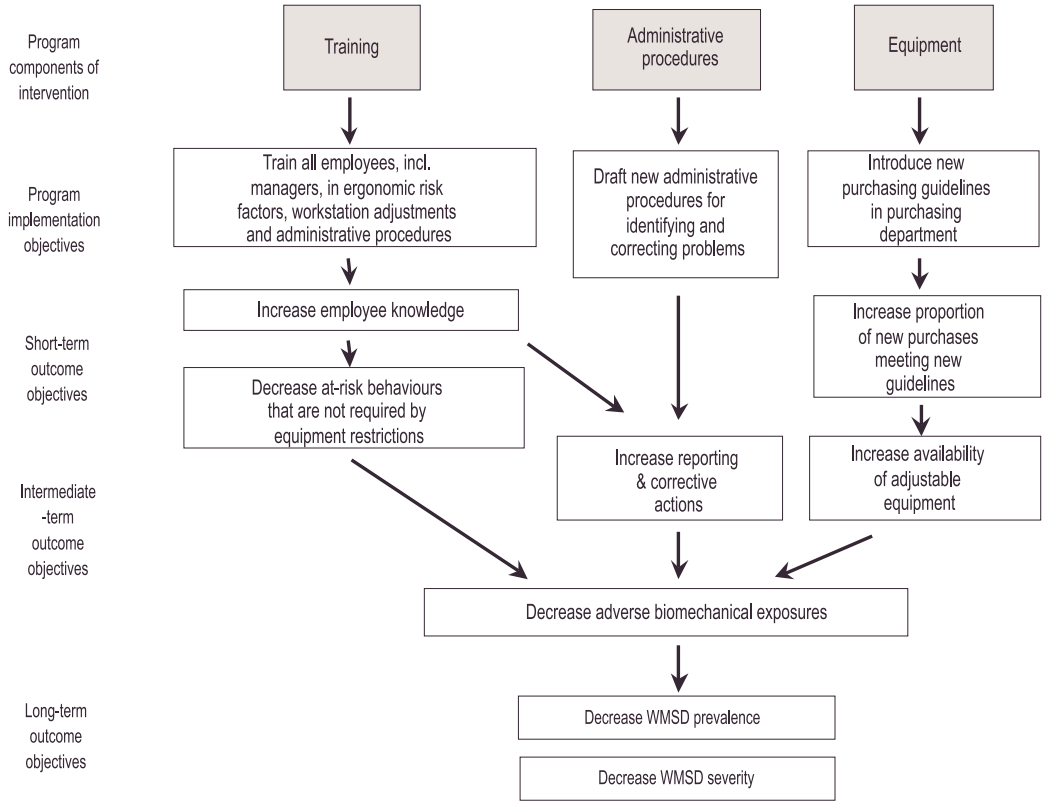
Figure 2.4 Generic program logic model



Like a conceptual model, the program logic model gives insight into what should be measured in an evaluation. Ideally, one selects a means of measuring the achievement of each of the objectives identified in the model. Figure 2.5 depicts an example of a program logic model for a hypothetical ergonomic intervention.

[8] Program logic models are explained by Rush and Ogborne [1991]

Figure 2.5 Example of a program logic model for an ergonomic program



A limitation of program logic models is that they often do not depict additional variables, such as confounding and effect-modifying variables. The advantage is that they more explicitly indicate the components of the intervention, and link them with the intervention objectives. Both types of models can be used to identify potential *unintended outcomes* of the intervention. These outcomes are not intended to result from the intervention; but they nevertheless might. As such, these outcomes can be difficult to anticipate, but consideration of the models can help.

Using the models, one looks at the changes that are supposed to happen following the intervention. You then try to think about what other effects could possibly result from the changes. For example, an intervention to reduce needle injuries by eliminating recapping prior to disposal into containers might not only have the intended effect of decreasing recapping injuries, but also an unintended effect of increasing disposal-related injuries if disposed into poorly designed containers.

2.5 Quantitative vs. qualitative methods for collecting evaluation data

Tradeoffs are made in choosing which outcomes to measure. These include consideration of the quality of evidence required, available resources and quality of data potentially available. In a very complete effectiveness evaluation, you collect pertinent data on all concepts represented in the intervention models, some using qualitative methods and others quantitative. Quantitative methods involve measuring *indicators* or *variables* related to the concepts in

the models, such that they yield numerical data. Qualitative methods, on the other hand, do not yield numerical data and rely instead on interviews, observation, and document analysis.

To clearly demonstrate intervention effectiveness, it is almost mandatory to use quantitative techniques and measure an outcome variable (e.g., injury rate). A demonstration of statistically significant change or difference in this variable (e.g., significant decrease in injury rate) and its unambiguous attribution to the presence of the intervention provides very good evidence of intervention effectiveness. However, when both qualitative and quantitative methods are used in an evaluation, an especially rich source of information is generated because the two can provide a check and a complement for each other. Whereas quantitative methods are used to answer - "How big an effect did the intervention have on the outcome(a) of interest?" - and - "Was the effect statistically significant?" - qualitative methods help answer - "How did the intervention have that effect?" - and - "What was the reaction of participants to the intervention?". Qualitative methods can also be used at the earlier stage, when developing the quantitative methods for the study, by answering - "What would be important to measure when determining the size of the effect of the intervention?"

2.6 Choosing the evaluation design

Using the conceptual and / or program logic models as a guide for measuring, you will then choose an *evaluation design*. This is the general protocol for taking measurements; i.e., from how many group(s) of workers / workers will measurements be taken and when will they be taken? Choosing the design involves a compromise among a number of considerations. Some are summarized below and will be discussed throughout the guide.

Considerations in choosing an evaluation design

1) What is the strength of evidence required to address the purpose of the evaluation?
   - What evaluation design will convince the decision-makers that the intervention has succeeded or failed?
2) Are there any ethical and legal considerations?
   - Cannot assign people to a situation if one it is likely to be more harmful
   - Some forms of data collection might require consent of individuals
3) What data collection and analysis is possible with the resources (financial, personnel, expertise) available?
   - Some forms of data collection are more expensive than others.
4) Has the intervention been introduced already? If so, has some data collection already taken place?
   - Circumstances may limit choices regarding the evaluation design
5) What is the time frame required and / or available for evaluation?
   - Does demand for results preclude long-term measurement?
   - What does preliminary research predict regarding the timing of the maximum program effect, as well as any erosion of that effect
6) What does the conceptual model for the intervention suggest should be measured and when?
7) How does the organization of the workplace influence the design?
   - Is randomization of workers / workplaces or the use of comparison groups possible?
   - Can an "intervention" group receive an intervention without other groups being aware?
   - How many workers / workplaces are available for the study?
8) Does the design offer sufficient *statistical power*?

2.6.1 Strength of evidence provided by different evaluation designs

The goal of occupational safety intervention research is to be able to say that a specific intervention either enhanced, or did not enhance, worker safety. The degree of confidence with which you can legitimately draw this type of conclusion depends on the strength of evidence provided by the study. Usually, an *experimental design* provides the strongest evidence of a causal link between the intervention implementation and observed effects. This design strength arises from: 1) the use of a *control group*, which does not participate in the intervention and is compared with the group which does participate; and 2) the assignment of people or workplaces to either intervention and control groups through an unbiased process (i.e., by *randomization*).

However, the logistical requirements of experimental designs often cause them to be unfeasible, especially for single, smaller worksites. In such cases, *quasi-experimental designs*, should be considered. They represent a means of compromising between the practical restrictions of workplaces and the rigour required for demonstrating intervention effectiveness. They often include a control group, albeit one created through a non-random process; and in all cases, yield more information than *a non-experimental design*. This last group of designs, including the common *before-and-after* design, are often necessary due to time and circumstances. Although it provides weaker evidence, using a before-and-after design is better than no evaluation at all.

Table 2.1: Characteristics of different types of evaluation designs

Type of design	Characteristics of design			
	Inherent in design		As commonly used in workplace evaluations	
	Strength of evidence of effectiveness	Randomization of workers/ workplaces	Control or comparison group	Pre-intervention measurements
Non-experimental	Weak	No	Sometimes	Sometimes
Quasi-experimental	Moderate	No	Sometimes	Yes
Experimental	Strong	Yes	Yes	Yes

2.6.2 Ethical considerations

Potential for harm vs. likelihood of benefit

Ethical considerations might be another reason for not choosing to use an experimental design. If there is preliminary evidence of a certain type of intervention being beneficial or if it could be presumed to be beneficial, you might not want to plan an evaluation where some workers are purposely put into a control group in which there is a chance of great harm. For example, the addition of machine guarding to a cutting machine could promise to be very beneficial, especially if there has been past severe injuries that probably could have been prevented by guarding. You might want to use only a before-and-after design in this situation.

Individual consent

Different cultures have various views on what types of research are ethically acceptable. Modern Western societies tend to emphasize individual autonomy and demand "fully informed consent" from study subjects. For example, in medical research, such consent is required to enroll a patient in a study.

The situation in North American workplaces is somewhat different than in the medical field. Management is considered to have certain prerogatives, as well as legislated responsibilities regarding safety, which allows it to make decisions about what happens in the workplace. Thus, employers can impose a new safety intervention upon their work force, and people might be assigned to an intervention group or control group, without first seeking their consent. This is less likely to occur in some European countries, particularly Scandinavia, where more worker participation exists in work-related decisions. In addition, workplace research carried out under the aegis of an academic or federal research institution usually requires approval by ethics committees, which will certainly insist upon individual consent to obtain data of a more personal nature (e.g., health records, employee opinions).

2.7 Practical tips

2.7.1 Time management

The evaluator has to be careful about the time set aside to properly conduct the evaluation. Often, the need to implement the intervention and resolve any unexpected problems can end up monopolizing what hours are available for the evaluation. The less obvious, but nevertheless important, demands of consulting with the evaluation committee and communicating with workplace parties regarding the evaluation can too easily fall by the wayside unless time is allocated for them.

2.7.2 Dealing with reaction to interim results

Be prepared for the reaction by workers, employers, etc. to interim results. If the results are encouraging, the group designated as a control may apply strong pressure to be allowed to receive the intervention immediately and not continue to incur possible risk until scientific proof has been accumulated. Your evaluation committee may come up with all sorts of interim suggestions about ways to modify the intervention, which could destroy the scientific purity of the study. In the face of proposed alterations to the intervention or evaluation design, always keep in mind what effect the proposed changes will have on your ability to analyse, interpret and draw conclusions from results. There is no easy single answer to these dilemmas and the decision made will depend partly on the importance of the evaluation compared to the value of the change the intervention is designed to produce.

2.7.3 Intervention diary

We strongly recommend keeping a diary or log-book of the intervention and the evaluation, in order to supplement formal data collection

methods. It can be exceptionally valuable to look back after the evaluation has revealed unexpected results. The diary provides a convenient place to record information about any potential influences on the intervention outcomes that were not taken into account by the experimental design, including what is taking place in the workplace - such as a change in senior personnel or work processes. Evaluators can also track decisions made over the course of the evaluation which might not be documented elsewhere.

2.7.4 Getting cooperation of workplace parties

A critical step in a successful intervention and evaluation is obtaining the cooperation of all involved. This includes the people in the workplace participating in the intervention and evaluation, their supervisors or managers, those using the results of the evaluation and anybody else interacting with the individuals or activities just mentioned. Evaluation can be threatening. It implies someone is checking, observing, and questioning. It is vital to explain to those involved what the evaluation is for, what will happen to the information collected (particularly if this is personal or confidential) and what will be done with the results.

Often, it is necessary to stress that the evaluation aims to learn about and improve the intervention, and not engage in criticizing or fault-finding. In particular, individuals with a stake in the success of the intervention will be sensitive, as will those who chose or designed it, or invested time in implementing it. Ongoing, repeated, clear, and, ideally, interactive communications with all involved is recommended. Communication can help allay fears, reduce resistance to change, increase acceptance of the evaluation results and encourage adoption of the intervention, if shown to be successful.

2.8 Summary

This chapter has introduced some of the initial decisions and considerations that must be taken into account in planning an evaluation. These include the evaluators' terms of reference and the selection of an evaluation committee. It was also shown how conceptual models and program logic models can help identify what concepts should be assessed for the purpose of an evaluation. Such assessment can take place using quantitative or qualitative methods. In addition, there are a number of ethical considerations in an evaluation.

Once you have undertaken the broader decisions outlined in this chapter, you will be ready to choose the evaluation design. Options for designs will be discussed further in Chapters 3 and 4.

Key points of Chapter 2

- Decide objectives and scope of the evaluation.
- Involve all parties relevant to the intervention (expertise, influence, etc.) in planning and at subsequent stages of the evaluation.
- Make a conceptual model relevant to the intervention, identifying concepts which must be measured or accounted for in the design.
- Make a program logic model.
- Start designing the evaluation, considering several aspects.
- Consider using a quasi-experimental or experimental design, if permitted by feasibility and ethical considerations.
- Keep an intervention diary.

Chapter 3

Before-and-after design: A simple evaluation design

3.1 Introduction

Chapters 3 and 4 will discuss three basic types of evaluation designs: *non-experimental; quasi-experimental;* and *experimental*.[9] These designs differ in the strength of evidence they provide for intervention effectiveness. We can be more confident in a result from an evaluation based on an experimental or quasi-experimental design than with a non-experimental design; but often we have no choice. We either have to use a non-experimental design or not attempt any evaluation. In this case, we advise that a non-experimental design is better than none at all.

Chapter 3 therefore focuses on the *before-and-after design*, a type of non-experimental design commonly used in safety studies. We outline some of the reasons the before-and-after design must be used with caution. These are called *threats to internal validity;* i.e., circumstances which threaten our ability to correctly infer whether or not the intervention had the desired effect. Chapter 4 will cover quasi-experimental and experimental designs.

3.2 Design terminology

In design terminology, "before" refers to a measurement being made before an intervention is introduced to a group and "after" refers to a measurement being made after its introduction. Equivalent terms for "before" and "after" are "pre" and "post".

3.3 Non-experimental designs

The before-and-after design is the most important *non-experimental design* for our purposes, since it is a reasonable option for an evaluation. Although it suffers from many threats to internal validity, it can, in many cases, provide preliminary evidence for intervention effectiveness, especially when supplemented with complementary information. We will consider it in detail in Sections 3.4 and 3.5. There are also two other types of non-experimental designs: after-only and after-only-with-a-non-randomized-control-group. As implied, measurement is made only after the intervention's introduction in both of these designs and, hence, they are less desirable.

An example of an after-only design is where you measure safety knowledge in a test only after you train a work group. The weakness of this approach is that you cannot be sure whether the test score would have been any different without training. Even if the average score after training was 90%, the training might actually have been ineffective, since the group might also have scored 90% on the test if it had been given before the training. Thus, the only acceptable use for this type of design is to ascertain if a certain standard has been met. It is not useful for effectiveness evaluation.

An example of the after-only-with-a-non-randomized-control-group design is a case where you measure knowledge in one group following training, and also measure it in another group which did not have the training. If the score is higher in the first group than in the second, for example, 90% compared with 80%, you might be tempted to think that the training had been effective. However, once again, you cannot tell if the training had any impact, because you do not know what the "before" values would have been. If they had been 90% and 80% respectively, your conclusion regarding program effectiveness would have been wrong. Thus, the after-only-with-non-randomized-control-group design is also not useful for effectiveness evaluation.

[9] We follow the classification described by Cook and Campbell [1979] and Cook et al. [1990].

3.4 Before-and-after design

O X O

The before-and-after design offers better evidence about intervention effectiveness than the other non-experimental designs just discussed. Consider our example of training. Suppose a group had been given a test of knowledge in the morning and scored 50%, and following a day of training, the group repeated the same test and scored 80%. [We illustrate this in the above figure by each "O" representing a measurement (in this case, the test), and the "X" representing the introduction of the intervention (training).] In this situation, the evidence would be strong that the training caused the increase in test score. Another way of saying this is that the evidence of *causality* would be strong. Besides the training, little else over the course of the day could have caused the observed increase in knowledge (provided of course that we do not use an identical test on the two occasions and give the group the answers in the meantime).

The before-and-after design is most useful in demonstrating the immediate impacts of short-term programs. It is less useful for evaluating longer term interventions. This is because over the course of a longer period of time, more circumstances can arise that may obscure the effects of an intervention. These circumstances are collectively called *threats to internal validity*.

3.5 Threats to internal validity of before-and-after designs

Threats to internal validity are possible alternative explanations for observed evaluation results. The more threats to internal validity, the less confident we are that the results are actually due to the intervention. A good evaluation identifies the possible threats to its internal validity and considers them one-by-one as to the degree of threat they pose.

One way to minimize the threats to validity is to consider other data or theory. In doing so, you might be able to render a particular threat to validity highly unlikely. Or, you might be able to show that it poses a minimal threat and thus does not change the conclusions of the evaluation. Some of the later examples illustrate this approach.

The ideal way of dealing with internal validity threats is by using a quasi-experimental or experimental design. Chapter 4 will show that you can eliminate most of the threats we are about to discuss by including a good "non-randomized control group" in your evaluation design. However, in the following we will assume that it has only been possible to undertake a before- and-after design. Our suggestions for dealing with threats to internal validity are made with this limitation in mind.

Table 3.1: Threats to internal validity

Threat to internal validity	Description of threat
History	Some other influential event(s), which could affect the outcome, occurs during the intervention
Instrumentation/Reporting	Validity of measurement method changes over course of the intervention
Regression-to-the-mean	Change in outcome measure might be explained by a group with a one-time extreme value naturally changing towards a normal value
Testing	Taking measurement (e.g. test) could have an effect on the outcome
Placebo	Intervention could have a non-specific effect on the outcome, independent of the key intervention component
Hawthorne	Involvement of outsiders could have an effect on the outcome, independent of the key intervention component
Maturation	Intervention group develops in ways independent of the intervention (e.g. aging, increase experience, etc.), possibly affecting the outcome
Dropout	The overall characteristics of the intervention group change due to some participants dropping out, possibly affecting the outcome

3.5.1 History threat

A "history threat" occurs when one or more events, which are not part of the intervention but could affect the outcome, takes place between the "before" and "after" measurements. Common history threats include changes in the following: management personnel; work processes, structure or pace; legislation; and management-labor relations. Clearly, the longer the time between the "before" and "after" measurements, the more opportunity is there for an extraneous, interfering event to happen.

There are two history threats in the accompanying example (opposite), one from outside the company and the other from inside. Either the community campaign or the human resource initiatives - or both - are alternative reasons for the observed decrease in injury rate.

Example of a history threat

You are trying to evaluate a new ergonomic intervention for nurses in a hospital. An educational program about back health and lifting techniques was provided, a program of voluntary stretches introduced and lifting equipment purchased. It was found that the injury rate for the two years before the intervention was 4.4 lost-time back injuries per 100,000 paid hours and for the two years following it was 3.0. Thus, you conclude that the ergonomic intervention has been effective.

But what if one month after the education program, a government ministry launched a year-long public awareness campaign aimed at reducing back injury? And, what if the president of the hospital was replaced two months after the in-house ergonomic program and her replacement introduced human resource initiatives to improve communication among staff? This would make you less confident about concluding that it was the intervention alone that made the difference in back injuries.

How to deal with history threats

The opportunities for history threats to arise in safety intervention evaluations are considerable because of the complex nature of the workplace and its environment. Careful consideration should be given to the events which could affect the safety outcome of interest. Earlier, an intervention diary was recommended as a means of keeping track of such events throughout the intervention and evaluation. Interviews with key personnel, a form of *qualitative* investigation, can also identify significant events that have taken place. [Qualitative methods are discussed in Chapter 7]. Even if none have occurred in the broader environment or the workplace itself, it is important to be able to confidently state this in the report of findings from a before-and-after design.

If you do identify a history threat, try to estimate how large an effect the threat would likely have had. The following example illustrates how you can use other data to estimate these possible effects.

Example of dealing with a history threat

Consider the preceding example of the hospital ergonomic program which appeared to have decreased back injury rates. To reduce the threat of the community campaign, you could try to obtain statistics on changes in back injury rates for other hospitals in the same community. If the injury rate in the other hospitals had remained constant or increased, you could conclude that changes in your hospital (i.e., either the ergonomic intervention or human resource (HR) initiatives) had an effect on injury rates beyond any effect of the community education program.

As far as considering the effect that the HR initiatives might have had on injury rates, you could look at other HR-related outcomes, e.g., non-back related absenteeism, non-back related health care claims or turnover. You reason that if the HR initiatives were powerful enough to have an effect on injury rates, then they should also affect other employee-related health indicators. If these other outcomes show little change, you could be more confident that the observed decrease in injury rates was due to the ergonomic program alone and not the president's initiatives.

3.5.2 Instrumentation/reporting threat

An instrumentation threat to validity occurs whenever the method of measuring safety outcome changes between the "before" and "after" measurements. There are many ways this could happen, depending on the type of outcome measured, as the following examples illustrate (Exhibit 3.1).

Exhibit 3.1 Examples of instrumentation threats

1) You are evaluating a new, redesigned workstation by taking electromyographic measurements of muscle activity in workers using their old workstations and then again in the same workers with the new workstations. You did not realize that at the time the measurements were taken at the old workstations, the equipment was malfunctioning, leading to an underestimate of the true values.

2) You give a multiple choice pre-test of knowledge with four possible choices: unlikely, somewhat unlikely, somewhat likely and likely. Afterwards, someone comments to you that they had a hard time choosing among the four choices and would have liked a "don't know" option. In the post-test you decide to give five possible choices, the four above, plus "don't know".

3) You ask workers to fill out a safety climate questionnaire before and after the introduction of a joint health and safety committee in the plant to evaluate the committee's effectiveness in changing the safety climate. The first time, employees filled out the questionnaire on their own time; the second time, they were given time off during the day to complete it.

An instrumentation threat of special concern in safety evaluations using injury statistics is any change in injury reporting over the course of the evaluation. Sometimes, this arises simply through changes in administrative policies or procedures: e.g., the definition for an injury's recordability changes or the recognition of an injury as work-related changes. A particularly tricky situation arises when the intervention itself affects the reporting of incidents, as the following examples portray (Exhibit 3.2).

Exhibit 3.2 Examples of reporting threats arising from intervention

1) You are evaluating the effect of undergoing a voluntary safety audit on the work-site by looking at injury statistics before and after introduction of the audit. You realize, however, that the audit includes a section on data collection procedures which could affect the reporting of incidents on the work-site. The increased reporting could cancel out any reduction in actual injuries. Thus, any change in injury statistics following the audit, might be due to the new reporting process alone, and not to true change in the injury rate.

2) A mandatory experience rating program was introduced by the government. This gives companies financial incentives and penalties based on past injury statistics. Following the introduction of the program, it was found that lost-time injury frequency for the jurisdiction had declined. Critics of the program cite anecdotal evidence of workers with serious injuries being pressured to accept modified work. No one can be certain that the decline in injury frequency results from enhanced injury prevention or from simply a suppression of reporting.

Dealing with instrumentation and reporting threats

The best way to deal with most instrumentation threats is to avoid them in the first place. Keep "before" and "after" measurement methods constant. Make sure all measuring equipment is functioning properly. Give questionnaires in the same format and under the same conditions for all measurements. Keep processes for generating and collecting injury statistics constant.

In the cases where the intervention itself might have affected the reporting of injuries, especially minor injuries, take a close look at the ratio of major-to-minor injuries. The greater the suppression of reporting, the higher this ratio will be, because minor injuries are easier to not report than major injuries. If the ratio is constant, then the likelihood of an instrumentation threat is reduced. You can also interview key informants to learn if the intervention affected reporting.

3.5.3 Regression-to-the-mean threat

Regression-to-the-mean is an issue when the basis for choosing the intervention group is a greater apparent need for the intervention (e.g., higher injury rates). This situation is not unusual as the following examples illustrate (Exhibit 3.3).

An alternative explanation for the apparent success of the safety initiatives is "regression- to-the-mean". This concept can be understood as follows. From year-to-year a group's or company's injury rates will fluctuate - sometimes they will be higher and sometimes lower. Any group with a lower than usual injury rate in any given year is therefore more likely to have their rates increase than decrease in the following year, assuming workplace conditions do not change. Similarly, the odds are that any group with a higher than usual injury rate in any given year is more likely to have their rates decrease than increase in the subsequent year. Thus, if the intervention includes only groups with high injury rates, a part of any decrease observed may have nothing to do with the intervention itself. Rather, the rate is simply moving closer to the average rate.

Exhibit 3.3 Examples of regression-to-the-mean threats

1) Division A of the manufacturing company had an unusually high rate of slip and fall injuries last year. The president is concerned and an enhanced inspection schedule is therefore implemented. Rates are lower the following year and so it appears that the intervention was successful.

2) The government labor ministry has decided to implement a new educational initiative with companies whose injury rate in the previous year was twice as high as the average for their industrial sector. Personal contact was made with officials of those companies, at which time penalty-free on-site inspection and advice was offered. The group of companies had, on average, a lower injury rate the following year. Thus, it appears that the ministry program was successful.

Dealing with regression-to-the-mean threats

There is nothing that can be done to deal with a regression-to-the mean threat if you have a single measure of frequency or rate before the intervention and one after, for a single workplace. However, with historical data, you can see whether the before measurement is typical of recent data or if it is an outlier. If it is the latter, then regression-to-the mean does threaten the validity of the conclusions and you might want instead to use more of the historical data to calculate the pre-intervention frequency or rate. Hauer [1980, 1986, 1992] has developed statistical approaches that correct for this phenomenon in data from multiple work-sites. However, an alternative approach would be a quasi-experimental or experimental design, where some high injury groups receive the intervention and others are kept under control conditions for comparative purposes.

3.5.4 Testing threat

A *testing threat* to internal validity is a concern when the act of taking the before measurement might itself affect the safety outcome used to evaluate the intervention. This threat is only an issue for such outcomes as worker safety knowledge, attitudes or practices. Any of these could be affected by the act of taking initial measurements by methods involving questionnaires, interviews or observations. This contrasts with injury outcomes which can usually be measured without interacting with workers.

Dealing with the testing threat

If you always plan to give a pre-test before giving the intervention, you do not really need to know whether any observed effect was due to the pre-test, the intervention, or the combination of both. However, you then must continue to include the pre-test as part of the intervention package. Removing the pre-test risks decreasing the overall effect of the intervention. If you want to do away with the pre-test, you should at first continue to include a post-test. You can then check to see whether the post-test results are affected by the pre-test's removal. If not, and if the groups with which you are intervening are similar over time (and so similar pre-test scores can be assumed), then you can conclude that a testing effect was unlikely. However, a truly definitive investigation of a testing effect requires a quasi-experimental or experimental design.

Example of a testing threat

You want to evaluate a training intervention designed to increase worker participation in plant safety. You use a questionnaire to assess pre-intervention worker attitudes, beliefs and practices concerning participation. You administer a second, post-intervention questionnaire after a three month program of worker and supervisor training. Comparison of the questionnaire results show a significant change in scores, indicating that participation has increased.

Upon reflection you are not really sure what accounts for the improvement in the score. You reason that it could be any of the following: a) an effect of the training program alone; b) an effect of having awareness raised by completing the first questionnaire; or c) a combined effect of completing the questionnaire and then experiencing training. Either of the latter two possibilities involves a testing threat to the internal validity of the evaluation.

3.5.5 Placebo and Hawthorne threats

The "placebo effect" is a concept from clinical studies of medical treatments. It has been observed that a percentage of the study subjects treated with a placebo (i.e., an inactive substance), instead of a medical treatment, will show an improvement of symptoms beyond that expected of the normal course of their medical condition. It seems that the placebo operates through a psychological mechanism which results in an alleviation of symptoms. The patients believe that the treatment will be successful and this has an effect in itself on the outcome.

Example of a placebo threat

Due to an increasing prevalence of "repetitive strain injury" in a telecommunications firm, the management agreed to purchase new keyboards for one division. A survey of employee upper extremity symptoms was conducted the week before the keyboards were introduced and then three weeks afterwards. Everyone was pleased to find a significant decrease in reported symptoms between the "before" and "after" measurements. Management was on the verge of purchasing the same keyboards for a second division, but there was concern about a "placebo effect" of the new keyboard.

The "Hawthorne effect"[10] usually refers to an effect of the involvement of researchers or other outsiders upon the measured outcome. This term arose from a famous study of factory workers at Western Electric's Hawthorne Works in the 1920s. A subset of workers was moved to a different section of the factory, their working conditions manipulated and the effect of this on their productivity observed. It turned out that workers were more productive under any of the work conditions tried - even uncomfortable one like low lighting conditions. It is believed that the effect of the new psychosocial working conditions (i.e., increased involvement of workers) in the experimental situation actually overshadowed any effect of the changes in the physical environment.

Example of a Hawthorne threat and one way to deal with it.

A work-site decides to implement and evaluate a new training program focused on changing safety practices by providing feedback to employees. A consultant examines injury records and, with the help of workers and supervisors, develops a checklist of safety practices. The list will be used by the consultant to observe the work force and provide feedback to the employees about their practices. The consultant realizes that his presence (and the taking of observations) could make workers change their normal behavior. To avoid this potential Hawthorne effect, he makes baseline observations on a daily basis until his presence seems to no longer create a reaction and the observations become constant.

Dealing with Hawthorne or placebo effects requires somehow "separating" them from the effect of changing an injury risk factor as part of the intervention. In the example above, the effect of the consultant (and that of taking observations) was separated from the effect of providing feedback by having the consultant take baseline measurements prior to starting the feedback.

[10] Some discourage the continued use of this term. See Wickström G, Bendix T [2000]. The "Hawthorne effect" - what did the original Hawthorne studies actually show? Scan J Work Environ Health 26:363-367

3.5.6 Maturation threat

A *maturation threat* to internal validity occurs when the apparent change in safety outcome could be due more to the intervention group changing naturally (i.e., employees growing older, or becoming more knowledgeable, and more experienced) rather than to the intervention itself.

Example of a maturation threat

A shipping company instituted annual medical screening for their dock-workers in order to identify and treat musculoskeletal problems early. Injury statistics after four years of the program indicated that the incidence of injuries remained about the same, but the length of time off work per injury had been increased. It appeared that the program had been detrimental. But an occupational health nurse pointed out that there is a tendency for older workers to return to work more slowly than younger workers, following an injury. Because there had been few new hires at the company, this maturation threat was a real possibility.

Dealing with maturation threats

In the example above, we need to consider aging of the work force as a maturation threat. A statistician might eliminate this maturation threat by using appropriate statistical techniques in the analysis. With such a correction in the above example, we might find that the program actually had either no effect or even a positive effect, instead of the apparent detrimental effect.

3.5.7 Dropout threat

The dropout threat to internal validity arises when enough people drop out of the study to alter the characteristics of the intervention group. Furthermore, these characteristics are statistically related to the safety outcome of interest. This would not matter if you could still measure the safety outcome on all people who started the program. For instance, some individuals might drop out of a company back belt program. Yet, if they continue to work for the company, you could still evaluate the intervention if the company injury rate data for the entire intervention group is available. Not surprisingly, safety outcome data on drop-outs is not always accessible, as the following example illustrates.

Example of a dropout threat

An evaluation of an intervention to reduce farmer injuries takes “before” and “after” measurements using a questionnaire. The questionnaire includes questions about injuries requiring hospital treatment experienced over the past year. Unfortunately, over the course of the intervention, a large number of farmers withdraw and do not fill out the questionnaire again. Thus, “after” measurements are not possible for a large percentage of participants.

You find that the average self-reported injury rate for the group decreased and so it appears that the intervention had an effect. But you can not be sure whether this was actually due to the intervention or that those with higher injury rates dropped out of the study earlier.

Dealing with a dropout effect

If you have access to final outcome data for the individuals or groups who dropped out of the intervention, be sure to include their data with that of the others assigned to the intervention at the start of the study (as long as it does not contravene the conditions described in their consent form). This estimate of intervention effect will give a minimal, conservative estimate of the program's potential since not everyone was exposed to the program for the full length of time.

If you do not have access to final outcome measures for intervention dropouts, an important thing to do is compare the "before" measurement of dropouts, as well as their other characteristic (e.g., age), with those who continued with the intervention. If those who continued and those who dropped out are similar in these measurements, then the threat of dropout on the outcome is reduced. You can assume that if the dropouts and others were the same before the intervention, they would also be similar afterwards, with the exception of having completed the intervention. If the "before" measurements or other characteristics of the dropouts are different, then the threat of dropout persists. You could confine the estimate of the program's effectiveness to those individuals who participated for the entire length of the program. However, the results would not be generalizable to the entire *target population* and likely overestimate the program's effectiveness for that population.

3.6 Summary

This chapter focused on the before-and-after design, which is considered to be a type of non-experimental design. A before-and-after evaluation design was shown to suffer from several threats to internal validity: history, instrumentation, regression-to-the-mean, testing, placebo, Hawthorne, maturation and dropout. Fortunately, as the illustrations showed, these threats to internal validity can be handled to some extent by additional data collection or analysis.

We also showed the inherent vulnerability of a before-and-after design to internal validity threats, especially for long-term evaluation periods. The longer your intervention and evaluation, the more you will want to consider a quasi-experimental design as an alternative. We discuss both quasi-experimental and experimental evaluation designs in the next chapter.

Key points of Chapter 3

- If you have no choice but to use a before-and-after design, try to eliminate the threats to internal validity.
- Identify other changes in the workplace or community that could have an effect on the outcome (history threats) and estimate their possible effect.
- Ensure that before and after measurements are carried out using the same methodology (to avoid instrumentation or reporting threats).
- Avoid using high-injury rate groups, or other such extreme groups, as the intervention group in a before-and-after study (to avoid regression-to-the-mean threats).
- Allow for the fact that taking a test can have an effect of its own (testing threat).
- Identify possible placebo or Hawthorne threats and try to minimize them.
- Identify any natural changes in the population over time which could obscure the effect of the intervention (maturation threat), and possibly correct for their effect during the statistical analysis.
- Identify the effects of intervention participants dropping out and allow for this in the analysis.

Chapter 4

Quasi-experimental and experimental designs: more powerful evaluation designs

4.1 Introduction

In Chapter 3 we described the simplest type of evaluation design for intervention effectiveness evaluation, the before-and-after or pre-post design. We showed how its strength is inherently limited by several threats to internal validity.

In this chapter, we discuss several types of quasi-experimental and experimental designs. All offer some advantages over the simple before-and-after design, because some of the threats to internal validity are eliminated. In the first section we show how a quasi-experimental design evolves from the addition of one or more design elements to a before-and-after design. After this, we describe experimental designs. Although the latter offer the greatest strength of evidence, quasi-experimental designs are often more feasible in workplace situations. We close this chapter with the discussion of various threats to internal validity that arise with a control or comparison group.

4.2 Quasi-experimental designs

Design strategies which change a before-and-after design into a quasi-experimental design
Strategy 1: add a control group
Strategy 2: take more measurements before and after the intervention implementation
Strategy 3: stagger the introduction of the intervention among groups
Strategy 4: add a reversal of the intervention
Strategy 5: use additional outcome measures

There are five basic strategies to improving upon a before-and-after design. This section describes common approaches to adopting one or more of these strategies.

4.2.1 Strategy #1: Add a control group (e.g., pre-post with non-randomized control)

O X O
O O

The *pre-post with non-randomized control design* mimics a simple experimental design. Like the experimental design, there is at least one group which receives the intervention (intervention group) and one group which does not (control group)[11]. The difference lies in the way participants are assigned to groups for the purpose of intervention implementation and evaluation. In an experiment participants are randomly assigned;[12] in quasi-experimental designs, they are not. Often assignment of participants to a group is predetermined by the work organization. For example, you might deliver an intervention to one company division. Another division, which is similar, acts as a *non-randomized control group* by not receiving the intervention. In the example below, the assignment of reindeer herders to intervention and control groups was determined by geographical location.

[11] The terminology varies regarding the use of the term “control group”. Some use it only in the context of experimental designs, in which the intervention and control groups are formed through randomization. Others, including ourselves, also use the term control group in the context of quasi-experimental designs, in which groups are formed through a non-random process. In this case, the quasi-experimental control group is referred to as a “non-randomized control group”. “Comparison group” is sometimes a synonym for “control group”, but in other cases is reserved to describe the non-intervention group in a quasi-experimental design.

[12] Random assignment of participants to groups is discussed in Section 5.4.

Advantages of the "pre-post with non-randomized control group" design

By adding a non-randomized control group to the simple before-and-after design, you automatically reduce some of the threats to internal validity discussed in Chapter 3. In particular, interference by external circumstances (i.e., history effects) is reduced, because they will often apply to both the control group and the intervention group. It therefore allows a separation of the effect of the intervention from that of other circumstances. The following example illustrates this.

Example of a pre-post with randomized control group design

Due to the high rate of injuries among reindeer herders, preventive measures were developed. In intervention group A, letters describing possible preventive measures were sent to district leaders and contacts, who were asked to pass on the information to herders in their district. In intervention group B, occupational health personnel trained in prevention measures passed on information during medical examinations. There was also a control group C, which received no intervention. Pre-post statistics for the three groups are shown below.

Statistical analysis confirmed that the groups did not differ in terms of a decrease in accident rate. The authors had to conclude that the intervention efforts were ineffective.

Number of accidents/working days for reindeer herder groups[13]

Year of accident data	Intervention groups		Non-randomized control group
	A	B	C
1985 (pre)	18.7	21	19.2
1987 (post)	15.1	14.9	14.6

The above example demonstrates that it is possible to conclude that an intervention is ineffective, even though fewer accidents are seen after the intervention. The control group showed the evaluators how much change to expect in the absence of the intervention. These changes were likely due to history, and possibly, testing and Hawthorne effects, according to the original report by Pekkarinen et al.[13] Thus, we see how the presence of the control group allowed one to examine the intervention effect, free from the influence of internal validity threats.

On the other hand, a new threat to validity - selection effects - arises from using a non-randomized control group. This threat occurs when the intervention and control groups differ with respect to the characteristics of group participants and these differences influence the measures used to determine an intervention effect. Selection effects will be discussed further at the end of the chapter.

[13] Data from Pekkarinen et al. [1994] with permission of the Arctic Institute of North America.

4.2.2 Strategy #2: take more measurements (time series designs)

O O O X O O O

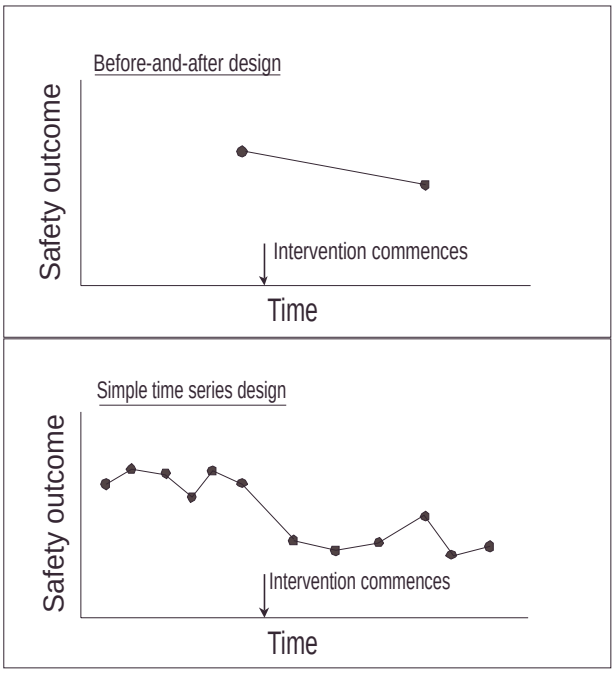
A *simple time series design* differs from the simple before-and-after design by taking additional measurements before and after the intervention. A baseline time trend is first established by taking several outcome measurements before implementing the intervention. Similarly, in order to establish a second time trend, several of the same measurements are made after introducing the intervention. If the intervention is effective, we expect to find a difference in outcome measures between the two time trends.

Advantages of simple time series design

Figure 4.1 illustrates how much easier it is to interpret the results of a time series evaluation design than a simple before-and-after design. In the first panel we see that there has been a drop in our safety measure from the period before the intervention to the one afterwards. As discussed in Chapter 3, several possible alternative explanations for this come to mind, e.g., history, maturation, instrumentation or Hawthorne effects. By adding measurements, as shown in the second panel, we can reduce the likelihood of some of these alternative explanations.

The maturation threat is eliminated because we observe that the change between the baseline time trend and the second time trend is abrupt. In contrast, changes due to maturation, such as increasing age or experience, are more gradual. Regression-to-the-mean or testing effects have also been eliminated as possible threats because we can see that safety outcomes are repeatedly high before and repeatedly low afterwards. Placebo and Hawthorne effects are less likely explanations because they tend not to be sustained once people have adapted to a change in their conditions. The threat of a history effect is somewhat lessened because the window of opportunity for a coincidental event is narrowed by the more frequent measures taken. Dropout and instrumentation both remain as threats, without consideration of additional information.

Figure 4.1 Comparison of before-and-after and time series designs



How many measurements are needed for a time series design?

The number of measurements you need for a time series design depends on the amount of random fluctuation (noise) that may occur in the outcome being measured and how much of an impact the intervention is expected to have. Somewhere between 6 to 15 measurements to establish a baseline and the same number again to establish the trend afterwards are typically required.[14]

Because of the necessity for many measurements,

[14] Several workplace examples can be found in Komaki and Jensen [1986].

the time series design is suitable for only some situations. For example, a time series design, using injury rate as the outcome measure would likely not be suitable for a small workplace. It simply takes too long - a year or more - for a small workplace to establish a reliable injury rate.

On the other hand, the design could be quite suitable in a group of small workplaces, a bigger workplace, or if observed work-site conditions were measured instead of injury rate. These situations permit more frequent and reliable measurement.

Even when it is not possible to take as many measurements as are needed for a time series analysis, taking additional measurements over time is still a good idea. It gives you a better sense of the pattern of variability over time and whether the last "before" measurement is typical of the ones preceding and the first "after" measurement is typical of the ones following. You are better informed of potential threats to internal validity and the sustainability of the intervention's effect. It may allow you to better estimate the effect of the intervention more accurately by pooling data.

Multiple time series designs

O O O X O O O
O O O O O O O

Even better than using basic strategy #1 or #2 alone, you can strengthen the before-and- after design even more, by combining both approaches (adding a control group and taking more measurements).

4.2.3 Strategy #3: Stagger the introduction of the intervention (e.g., multiple baseline design across groups)

O O O X O O O O O O
O O O O O O O X O O O

A special type of multiple time series design is known as "multiple baseline design across groups". With this design, all groups eventually receive the intervention, but at different times. As a result, all groups also serve as a comparison group to each other.

Advantages of the multiple baseline across groups design

The advantage of the multiple baseline across groups design is that it markedly reduces the threat of history effects. When an intervention is given to only one group, you can never really be sure that something else did not coincidentally occur at the same time to cause the measured effect. Even when you are using a control group, something could still happen to only the intervention group (besides the intervention itself) that affects the outcome.

When the intervention's introduction is staggered, with the apparent effects correspondingly staggered, history effects are an unlikely explanation for the result. This is because one coincidence of the intervention and an extraneous event happening close together in time is plausible, but two or more such coincidences are much less likely.

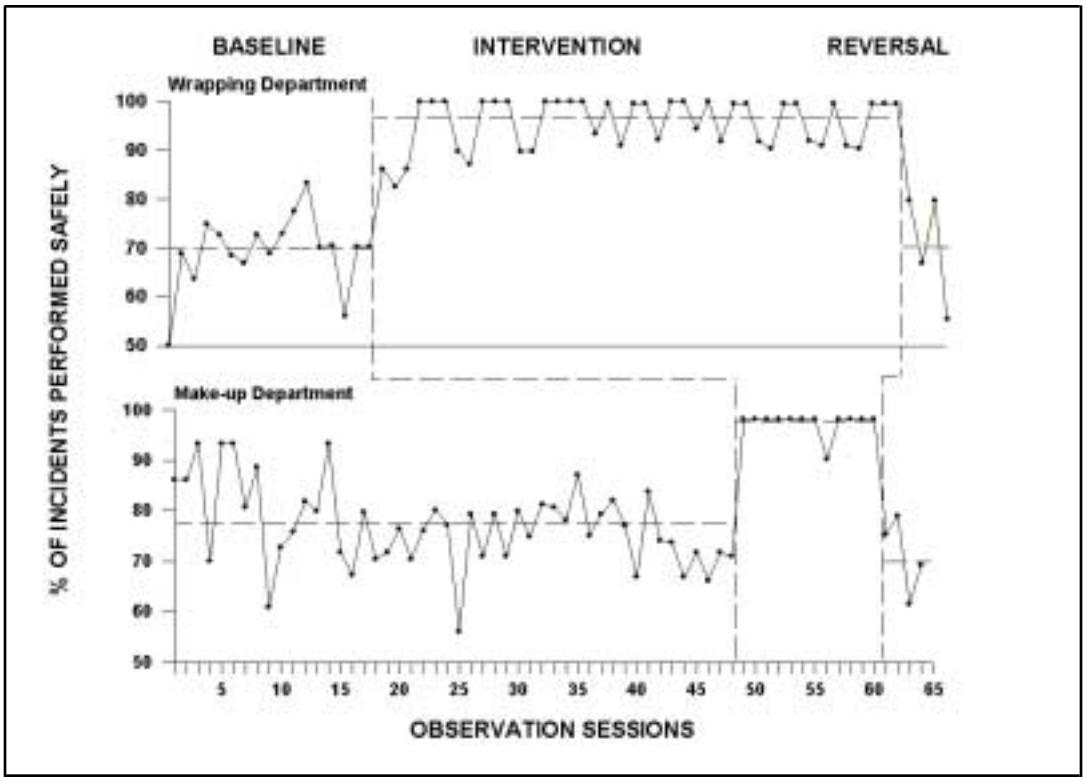
Whenever a workplace or jurisdiction has more than one division or group, a staggered introduction of the intervention should be considered as an alternative to introducing it to all divisions or groups at the same time. This staggered arrangement can also allow an interim assessment and, if appropriate, modification of the intervention or its implementation, before it is introduced into other divisions (though such modifications should be considered in the analysis and interpretation of results).

Example of a multiple baseline across groups design[15]

A safety behavior training intervention was undertaken at a food manufacturing plant. The intervention was first introduced in the wrapping department and then in the make-up department. The intervention started with an educational session on safety behaviors, after which a list of safety behaviors was posted. From then on the group was given feedback by posting the results of behavioral observations.

Safety behaviors were measured by a trained observer (three to four times a week). The observer used a checklist which gave an estimate of the percentage of incidents performed safely. Baseline measurements of safety behaviors were taken prior to introduction of the intervention.

You can see how, in each department, the change in safety behaviors followed implementation of the intervention. Having this sequence of events happen not only once, but twice, bolsters the causal link between intervention and behavior change. Further, because implementation occurred at different times, we really end up with two separate estimates of the amount of change caused by the intervention. [The reversal part of the intervention will be discussed in Section 4.2.4]



[15] Example from Komaki J, Barwick KD, Scott LR [1978] A behavioral approach to occupational safety: pinpointing and reinforcing safety performance in a food manufacturing plant. Journal of Applied Psychology 63:434- 445. Copyright © 1978 by the American Psychological Association. Adapted with permission.

4.2.4 Strategy #4: Reverse the intervention

O O O X O O O -X O O O

One way of strengthening a before-and-after or even a time series design is to follow the introduction of an intervention with another phase of the project in which the intervention is removed. In the simplest case, you end up with three phases: a baseline phase; an intervention phase; and a reversal or withdrawal phase. The rationale here is that if you remove the intervention conditions, you should correspondingly see a change in the outcome back towards the baseline condition.

Of course, this design is clearly not suitable for all situations, because it is hoped that the effect of an intervention will last and therefore is not easily reversed. However, as the Figure in section 4.2.3 shows, it has been found useful when behavior is the safety outcome being measured. In this case, the intervention was "reversed" by no longer giving the posted feedback.

Advantages and disadvantages of designs with a reversal phase

If you can demonstrate the effect of a reversal phase, you will have markedly reduced several of the internal validity threats discussed in Chapter 4 - in particular history, maturation, testing, dropout and Hawthorne (assuming researchers / outsiders are still present during reversal phase). Instrumentation and placebo effects may still remain as issues and should be considered. After demonstrating the effect of intervention reversal, you are then free to reinstate the intervention.

The downside to the reversal design feature is that repeated changes in safety programming could create confusion, stress and resentment among those affected. As well, if an intervention has looked promising following its introduction, subsequent removal could be considered unethical. Thus, use this design feature with caution.

4.2.5 Strategy #5: Measure multiple outcomes

O_1/O_2 X O_1/O_2

The final strategy for increasing the strength of an evaluation design is to use more than one type of outcome measure. We describe two approaches to doing this.

4.2.5.1 Add intervening outcome measures

We pointed out, using models in Chapter 2, that there can be a number of outcomes intervening between an intervention and the final outcome. We should ideally try to measure as many of these different intervention outcomes as is feasible, in order to bolster the strength of evidence provided by the evaluation design. This includes measurement of the intervention's implementation, as well as short- and intermediate-term effects of the intervention.

Measures of intervention implementation, such as the documentation of equipment purchases and work task modification in the following example, are especially important. In instances where a program has failed, you want to be able to distinguish between an inherently ineffective program and a flawed implementation. If an intervention has not been implemented as intended, measuring effectiveness by measuring changes in outcome will likely underestimate the intervention's potential impact. Thus, if inadequate implementation is found by the evaluation, you might try first to improve this part of the intervention, instead of discarding the intervention altogether.

Example of adding intervening outcome measures

A company plans to implement a participatory ergonomics program. Plans involve forming a labor-management committee, assessing employee needs, purchasing new equipment, modifying work tasks and providing worker education. The health and safety coordinator plans to measure the ultimate impact of the program by comparing self-reported symptoms and injuries before and after the intervention is implemented.

However there are concerns that a change in symptom and injury rates could have a number of alternative explanations, such as staffing changes, the business cycle, management changeover and Hawthorne effects, etc. To deal with this concern, the health and safety coordinator plans some additional measurements: records of equipment purchases; and self- reports of work tasks, practices and stressors. These all measure outcomes intervening between the intervention and the final outcome of changes in symptoms and injuries.

Illustration of the value of measuring intervention implementation

Mason [1982] tried to evaluate the effectiveness of a train-the-trainer kinetic handling training course, by looking at the change in the rate of back and joint injuries in the companies of instructors who had taken course. When practically no change was found after a year, it was valuable to know that this was probably because few of the instructors had organized and carried out in-company courses based on their own training during that year. Furthermore, those who did run courses had failed to retain most of their training and therefore could not pass on the handling techniques. The lack of any measurable effect of the intervention on injuries was therefore no proof that the kinetic handling technique itself was not effective, but rather that an improvement in the training methods for trainers were needed.

4.2.5.2 Add a related but untargeted outcome measure

The second approach to adding outcome measures involves measuring an outcome which is similar to the main outcome measure, but not targeted by the intervention. The additional outcome measure should be similar enough to the main outcome measure so that it is susceptible to the most important threats to internal validity. However, it also needs to be different enough that it should be unaffected by the intervention. The following examples show how this approach works.

Examples of adding related but untargeted outcomes

1)[16] The effect of new equipment on oil-drilling platforms was primarily evaluated by changes in the rate of tong-related injuries, a type of injury which should have been reduced by using the new equipment. The rate of non-tong-related injuries, a related but untargeted outcome measure, was also tracked. Although this second type of injury should have been unaffected by the intervention, it would likely be similarly susceptible to any history or reporting effects threatening the internal validity of the evaluation. Thus, including this untargeted injury measure in the evaluation reduced these threats, since any history or reporting effects on tong-related injuries would also be detected by changes in the non-tong-related injuries.

2)[17] An ergonomic intervention among grocery check stand workers was primarily evaluated by measuring self-reported changes in musculoskeletal discomfort. The intervention appeared successful because of significant change in reported symptoms in the neck / upper back / shoulders and lower back / buttocks / legs, the two areas predicted to benefit from the ergonomic changes. This conclusion was bolstered by a finding of no significant changes in symptoms in the arm / forearm / wrist, which were not targeted by the intervention. This made history, maturation, instrumentation, placebo, Hawthorne and instrumentation effects a less likely explanation for the improvement in the upper extremity and lower back areas.

4.3 Experimental designs

Two key features of an experimental design are 1) the use of a *control group* and 2) the assignment of evaluation participants to either intervention or control groups through *randomization*, a process in which participants are assigned to groups in an unbiased manner.[18] Thus, an experimental design uses an approach similar to strategy #1 in quasi-experimental designs (Section 4.2.1).

The use of randomization gives the experimental design greater strength. We can be more certain that any differences between the intervention group and the control group, with respect to the apparent effect of the intervention, can be attributed to the intervention, and not to group differences. Although it is often not feasible to use an experimental design, it has been used in several occupational safety situations.

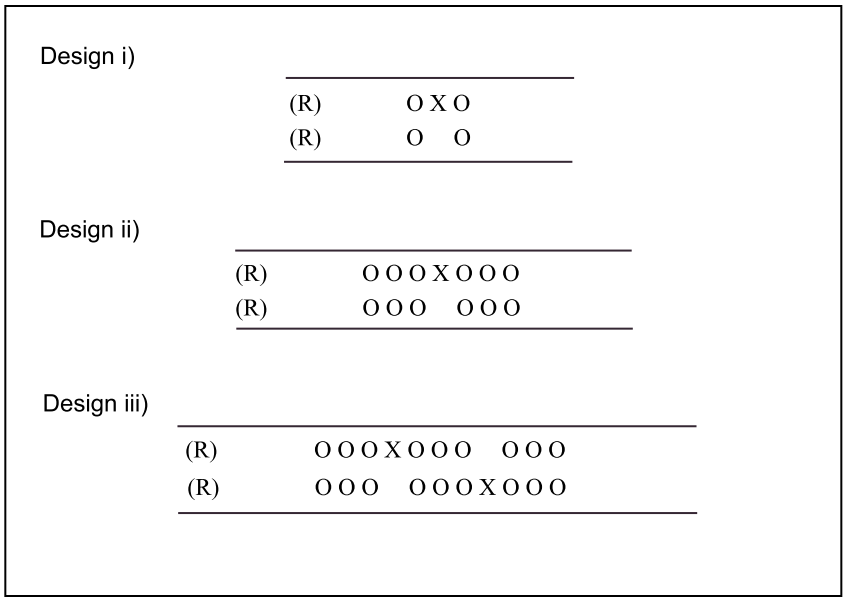
4.3.1 Experimental designs with "before" and "after" measurements

Earlier, three types of quasi-experimental designs were discussed that use non- randomized control groups: pre-post with non-randomized control group (Section 4.2.1), multiple time series (4.2.2) and multiple baseline across groups (4.2.3). These same design approaches can turned into experimental designs by using randomization to create the groups.

The first design shown in Figure 4.3, "pre-post-with-randomized-control" has been used in the subsequent examples. The first example involves randomizing work-sites into groups, and the second, randomizing individuals into groups.

[16] Based on Mohr and Clemmer [1989]
[17] Based on Orgel et al. [1992]
[18] Randomization is discussed in Section 5.4.

Figure 4.3: Experimental designs with "before" and "after" measurements



Example of an experimental design (1)[19]

An intervention for principle farm operators and their farms consisted of an on-site farm safety check with feedback and a one-day educational seminar. Potential participants in the intervention were identified from a list of all farms in the Farmers Association, using a random selection process. Of these, 60% of farm operators agreed to participate in the study. They were then assigned to either an intervention or control group, using a *randomization* procedure. To evaluate the intervention, these groups were compared on measures taken before and after the intervention: self-reported injuries and near-injuries (final outcome) and safety perceptions, practices and attitudes (intermediate outcomes).

Table 4.3 Example of an evaluation of a farm safety intervention using an experimental design

Group	Pre-intervention measures	Intervention	Post-intervention measures
	Injury self-report + perceptions, practices, attitudes questionnaire	Safety check + education	Injury self-report + perceptions, practices, attitudes questionnaire
A (intervention)	O	X	O
B (control)	O		O

[19] Adaptation of intervention described in Glassock et al. [1997]

Example of an experimental design (2) [20]

Two interventions for the prevention of back injury were evaluated with an experimental design involving warehouse workers for a grocery distribution center. Ninety workers with the same job classification were randomly selected from among the 800 employees at a warehouse. The ninety workers were then randomly assigned to one of three groups. One group was given one hour of training on back injury prevention and body mechanics on the job. A second group was also given the training, as well as back belts to wear. The third group served as a control group, receiving neither training, nor back belts. Both "before" and "after" measurements were taken: knowledge (short-term outcome); injuries and days absent as reported in health records (final outcomes). Abdominal strength was also measured in case it decreased as a result of wearing the belt (unintended outcome).

Table 4.4 Example of an evaluation of back belt and training interventions using an experimental design

Group	Pre-intervention measures	Intervention components		Post-intervention measures
	(Abdominal strength + questionnaire + injuries + absenteeism)	Training in body mechanics	Black belt use	(Abdominal strength + questionnaire + injuries + absenteeism)
A (intervention 1)	O	X		O
B (intervention 2)	O	X	X	O
C (control)	O			O

4.3.2 Experimental designs with "after"-only measurements

(R)	X O
(R)	O

One advantage of randomization is that in some situations it may allow for not having "before" measurements. This can be especially advantageous if you are worried about the measurement influencing the outcome of interest ("testing effect", section 3.5.4). It is also advantageous if taking a before measurement is costly (e.g., the administration of a questionnaire).

The disadvantage of not obtaining "before" measurements is that it will not be possible to see if the groups differed initially with respect to the outcome measure. You would therefore not be able to make any allowance in the analysis for these group differences.

[20] Based on Walsh and Schwartz [1990]

4.4 Threats to internal validity in designs with control groups

We discussed how designs that use control groups can markedly reduce the threats to internal validity discussed in Chapter 3. However, using control groups introduces some new threats to internal validity which we consider below. In spite of these, control groups are still strongly recommended. On balance, they strengthen the evaluation design far more than they weaken it.

4.4.1 Selection threats

A *selection threat* occurs when the apparent effect of the intervention could be due to differences in the participants' characteristics in the groups being compared, rather than the intervention itself. For this reason, control and intervention groups should be similar, especially with respect to any variables that can affect the measured outcome(s).[21]

Whenever you compare groups created through a non-random process, as in the quasi-experimental designs, you must consider how selection could affect your results. In what way do the people in the groups differ? Do they differ in their initial value of safety outcome measure or other characteristics (e.g., age, level of experience, level of education, etc.) which could influence the way groups respond to the intervention? If so, you need to collect information on these differences and make allowances for these differences in your statistical analysis.

Even by using a randomization procedure to create groups, as in a true experiment, you can have a selection threat.

4.4.2 Selection interaction threats

We just described how it is important for groups to be similar in their characteristics at the outset of an evaluation. It is also important that they remain similar and are treated similarly over the course of the evaluation. Otherwise, *selection interaction-effects* threaten the legitimacy of your evaluation conclusions. Recall that there are a variety of threats to internal validity in before-and-after designs, e.g., history, instrumentation, dropout, etc. In many cases having a control group - especially a randomized control group - can reduce or eliminate these threats to internal validity. The exception to this situation is when something happens to one group (e.g., history, instrumentation, maturation, etc.) and not to the other, resulting in selection interaction threats; i.e., selection-history, selection-instrumentation, selection-maturation, etc.

For example, a *selection-history effect* could occur if you are comparing two different divisions in a "pre-post with non-randomized control group" design. What if the supervisor of only one of these divisions changed during the course of the evaluation? You could not be sure whether between-group differences in the "before" to "after" changes was due to the effect of the intervention on the intervention group - or due to a change in the leader in one group. Selection-history interaction threats to internal validity are often beyond the evaluator's control, as in the example above. If they should arise, they are dealt with as was described for history threats (Section 3.5.1).

A *regression-to-the-mean-interaction threat* to internal validity arises if you deliver an intervention to units with high injury rates and compare their results to units with lower injury rates. Even if there was no intervention effect, the high injury group would tend to have a decrease in rates, and the others might have even shown an increase. The proper control group

[21] Depending on the type of evaluation design and the context, these characteristics or variables are sometimes called *confounders;* other times they are called *effect modifiers* or moderating variables.

would be a second group with similarly high injury rates.

A *dropout interaction threat* arises if one group has a greater rate of dropout than the other, especially if it results in the two groups having different characteristics. Characteristics of particular concern are those which could affect how the study participants respond to the intervention (e.g., age, level of experience, level of education), as well as differences in the initial value of the safety indicator used to measure outcome. While these differences are sometimes taken into account in the statistical analysis, it is preferable to avoid selection-dropout threats to internal validity altogether by taking steps to ensure that people continue participating in the intervention project and its evaluation.

Most other selection interactions, i.e., selection-instrumentation, -testing, -placebo, - Hawthorne, or -maturation effects can be minimized by treating the control group as similarly as possible to the intervention group with the exception of the intervention itself. Ideally, the evaluators should have just as much contact with individuals in the control group as those in the intervention group. In practice, such an arrangement may not be feasible.

4.4.3 Diffusion or contamination threat

A *diffusion threat* to internal validity (also known as a contamination threat) occurs when the intervention delivered to one group "diffuses" to the other. This can easily happen when the intervention is educational in nature, since workers naturally share information with one another. It is even possible for new equipment given to the intervention group to be shared with the control group. Diffusion is most likely to occur when the intervention is perceived as beneficial. It is undesirable for an evaluation because it reduces the differences observed between groups in their "before" to "after" changes. Thus, you might conclude that an intervention was ineffective when it really was not. The best way to reduce the threat of diffusion is by keeping the intervention and control groups as separate as possible.

4.4.4 Rivalry or resentment threat

Finally, threats to validity can arise when people in the control group react to not receiving the intervention. Suppose a safety incentive program has been introduced to encourage safe behaviors. The control group could react by not reporting injuries so its safety performance ends up looking good compared to the intervention group. Or the opposite might be done. Injuries could be "over-reported" to demonstrate that the group needs an incentive program as well. In both cases we could say that the control group has changed its behavior due to rivalry. Resentment effects are also possible. The control group, for example, could resent not being given the opportunity to participate in an incentive program. This souring of labor-management relations in the division could cause an increase in injury rates.

Rivalry or resentment threats can affect the evaluation's conclusions in either direction. Depending on the situation, they can either increase or decrease the differences between groups in "before" to "after" changes. The effects just described can sometimes be avoided by communicating well with groups or promising that if the intervention is shown to be effective, the control group will receive the intervention afterwards. If interventions are conceived and introduced through a participatory process, unexpected reactions are less likely. However, it is impossible to anticipate every reaction to a program. This is one area where qualitative investigation can be very helpful. Interviews with a few knowledgeable people in the control group should give insight into whether rivalry or resentment dynamics are an issue. As with the diffusion threat, the rivalry or resentment threats might be avoided if groups in different locations are compared and communication between the groups does not occur.

4.5 Summary

A quasi-experimental or experimental design is more likely to give a truer estimate of the effect of an intervention than a non-experimental design. You can change a (non-experimental) before-and-after design into a quasi-experimental one through one or more of the following design strategies: adding a control group; taking more measurements; staggering the introduction of the intervention; reversing the intervention; or using additional outcome measures. By adding these design elements you can increase the strengthen the design and reduce or eliminate the threats to internal validity discussed in Chapter 3.

Experimental designs differ from quasi-experimental designs by always involving a control group and by assigning subjects to intervention and control groups under a randomization scheme. Otherwise, many of the elements of quasi-experimental and experimental designs are the same. Although some new threats to internal validity need to be considered when using designs with control groups - selection, selection interactions, diffusion, rivalry, resentment - the use of control groups is almost always recommended whenever feasible.

Key points of Chapter 4

- Improve upon a simple before-and-after design, and use a quasi-experimental design, through one or more of five strategies:
    - adding a control group
    - taking more measurements
    - staggering introduction of the intervention among groups
    - adding a reversal of the intervention
    - using additional outcome measures.
- Improve upon a quasi-experimental design, and use an experimental design, by assigning participants to intervention and control groups through randomization.
- Check that intervention and control groups receive similar treatment throughout the evaluation period, apart from the intervention itself.
- Avoid (but check for) diffusion, rivalry or resentment effects.

Chapter 5

Study sample: Who should be in your intervention and evaluation?

5.1 Introduction

Having decided on the evaluation design, you should choose which individuals, work groups or workplaces will be included in the evaluation project. They comprise the *study sample.* This chapter will discuss how to select the study sample from a larger group of possible participants, and how to form different comparison groups within the study sample, through randomization and other means.

5.2 Some definitions

Let us start by distinguishing three terms:

1) The *target population*[22] consists of the people, groups or workplaces which might benefit from the safety intervention. For example, if you identify a safety need for construction workers and conduct an intervention among a participating group of workers, you want it to apply to all construction workers. The target population is "all construction workers" in this case.

2) The *sampling frame* is a segment of the target population - e.g., construction workers in a given company, union, or city.

3) A further subset, the *study sample* includes those people, work groups or workplaces chosen from the sampling frame.

In summary, the study sample is a sub-group of the sampling frame, which in turn is a sub-group of the target population.

5.3 Choosing people, groups or workplaces for the study sample

The people or workplaces included in the evaluation may be determined by circumstances. For instance, if your concern is a single workplace, and a decision has been made to introduce the intervention to all its 50 employees, then your study sample has been pre-determined. However, in dealing with a large workplace, it might not be necessary to include everyone in the evaluation, if lower numbers (i.e., a smaller *sample size*) provide you sufficient *statistical power* (see Chapter 8) to detect an intervention effect. This is especially worth considering if a measurement on larger numbers of people increases the cost of data collection or otherwise makes it unfeasible. Thus, situations will arise where you need to select a study sample for evaluation purposes from among all participants in the intervention.

The study sample should be representative of the sampling frame and / or target population. This is because more is required than just knowing whether the intervention or program worked only for the particular group selected for the study and their particular circumstances (e.g., time, place, etc.). A bigger concern is whether the intervention is generally effective. In other words, you want the results to have generalizability - also known as *external validity*. Safety managers will want to know that the evaluation results, obtained with a sample selected from among their employees, applies to their workplace as a whole. Someone with a multiple workplace perspective - i.e., a researcher, corporate safety director or policy-maker - will want the results to apply to the whole target population.

How can the sample be made representative of the sampling frame? Suppose a training intervention is being implemented for safe work practices, and you will evaluate it by observing workers doing their job before and after the intervention. You have determined it is not feasible to observe all of them going through the program; so you limit yourself to a smaller sample.

You could ask for volunteers, but they will

[22] Note that the term target population is not always defined this way. Some people define it as what we are calling the sampling frame.

probably be unrepresentative of all the workers, since they are more likely to have particularly good work practices. Another alternative is to choose everyone who works in one area (making it easier to do the observations). But again, this group is unlikely to represent the whole workplace, given the wide variety of jobs.

The best method is to choose a *random sample*, which increases the chance of a representative sample from the target population. This avoids choosing subjects on any other basis (e.g., volunteering).

Using a control group, as well as an intervention group, in your evaluation design, will increase the required sample size.

Considerations when choosing the study sample

- To what target population will the results be generalized?
- How many participating individuals, workgroups or workplaces are potentially available for the intervention and evaluation?
- What sample size will give sufficient statistical power?
- What is the marginal cost per participant of data collection?
- Will a control group be used in the evaluation design?
- Will the sample be stratified?

5.3.1 How to choose a (simple) random sample

Random selection involves choosing your sample where each person in the sampling frame has a known probability of being selected. With *simple random* sampling, the most common type of random sampling, the probability of being selected is equal for everybody. The process is similar in many ways to tossing a coin for each person in the sampling frame and choosing those people for whom the coin comes up heads. But with coin-tossing the probability of getting heads is 0.5 and of getting tails is also 0.5. We may only want to choose 50 out of 200 people, so the probability of selection in this case is one in four.

There are several different ways to choose a random sample. One of the simplest is to use *random number tables*, which typically show many thousands of randomly selected digits. This means that when the tables are generated, there is exactly a one-in-ten chance (probability) that each of the digits 0,1,...,9 would be selected for any position on the table. The digits are usually shown in groups of 5 - this has no particular meaning - it simply makes the table easy to use. Random number tables are often included in the appendices of statistics text books. Alternatively, many statistical or spreadsheet software packages have a function which generates random numbers. We used one to generate Table 5.1.

Table 5.1 Random number table

5 8 5 0 1	4 2 5 4 9	8 1 2 0 0	7 9 0 5 6	1 1 0 7 0	7 6 0 8 1
9 9 3 7 1	6 1 6 9 0	8 6 9 5 5	6 4 2 0 4	5 7 2 0 9	8 5 8 5 1
4 6 0 0 6	2 9 0 9 7	9 6 1 2 8	6 3 3 5 6	5 8 2 1 9	6 3 7 4 3
9 8 0 6 3	0 0 9 1 3	1 3 5 1 7	4 4 4 5 2	4 4 1 9 6	9 2 4 0 5
1 9 0 0 0	6 5 3 9 1	7 8 9 6 2	1 7 4 0 0	1 7 3 7 3	5 2 6 9 0
5 8 0 5 8	7 3 5 2 6	4 4 7 9 9	0 2 3 5 5	6 8 1 3 4	5 5 1 2 1
2 4 3 4 8	5 0 1 3 4	2 6 7 3 4	4 0 0 5 0	4 6 3 6 0	5 3 4 5 7
2 4 4 3 3	2 2 1 2 3	0 6 2 9 6	3 0 4 4 2	1 2 3 5 3	5 0 5 0 4
3 7 2 6 7	6 8 2 8 5	0 9 7 1 7	1 9 8 6 8	4 8 3 6 6	0 3 1 6 7
0 9 3 0 4	1 7 6 9 9	4 6 3 6 5	7 1 5 9 0	8 5 5 7 7	0 7 1 9 3
2 1 4 7 7	3 9 8 0 5	4 9 8 1 2	5 2 9 0 0	5 4 7 6 9	5 3 4 1 1
8 6 2 4 4	6 6 9 5 1	4 6 3 3 1	7 6 1 2 4	2 6 8 2 5	4 5 5 1 8
0 7 5 1 0	0 1 8 3 9	9 5 7 2 5	2 2 1 3 4	4 5 7 5 2	8 8 2 0 3
9 5 8 4 0	5 3 0 1 7	8 2 1 3 1	7 4 4 8 7	4 2 2 8 3	6 8 6 3 7
0 5 4 8 4	6 4 9 6 8	4 0 2 9 8	7 1 9 1 8	3 4 5 5 3	3 2 4 8 5
8 6 0 7 0	8 3 1 2 7	0 1 1 2 3	0 2 1 3 3	0 8 4 6 9	6 8 2 9 0
8 7 5 7 5	3 0 3 7 4	3 3 7 3 0	0 9 4 4 1	9 2 5 1 9	4 1 6 6 5
6 8 5 4 4	7 6 7 4 6	3 4 0 6 3	5 5 2 1 9	4 5 7 6 5	4 7 2 3 0
3 3 6 7 9	7 8 4 7 6	4 7 8 6 7	1 9 4 4 8	9 4 2 1 8	2 9 5 3 6
8 5 1 4 3	9 6 1 2 2	0 5 7 4 5	7 7 2 6 0	7 4 0 9 2	4 8 7 5 3
3 1 8 9 4	6 7 5 2 2	8 2 2 8 6	7 7 4 1 4	1 5 3 7 2	3 5 7 7 9
7 8 6 8 3	4 9 3 2 9	4 5 4 8 2	5 7 8 2 6	5 5 1 4 2	8 6 0 7 6
7 0 1 3 5	6 1 5 6 3	7 1 8 8 5	3 8 8 1 5	5 1 2 7 5	7 1 4 1 0
7 7 3 1 0	9 9 0 9 6	9 4 5 5 6	2 7 8 7 5	0 6 9 3 9	6 7 1 2 5
3 6 9 6 8	5 1 9 9 1	9 7 2 7 4	2 9 2 7 0	8 0 4 8 6	2 4 4 5 6
3 8 2 8 7	8 0 7 5 4	5 1 4 2 2	4 1 3 9 0	4 9 8 4 3	5 1 9 3 9
9 6 3 6 7	3 1 2 7 7	5 5 9 5 8	5 1 1 7 5	4 7 0 9 6	1 3 5 7 4
6 0 4 8 8	3 5 6 1 9	9 5 3 7 4	2 6 4 4 4	0 3 7 9 3	6 4 2 8 4
5 7 9 7 5	8 3 6 9 8	8 0 5 2 1	7 0 3 5 3	2 6 2 5 1	5 7 2 6 2
0 7 3 6 4	3 7 1 9 4	9 9 1 5 6	3 5 1 7 0	9 0 9 4 1	2 9 5 5 8

This table was generated using a spreadsheet software with a random number generation function. It can be used for selecting a random sample from a sampling frame and for randomizing sample subjects to intervention and control groups. The groups of three digits indicated above are used in the illustration of random sampling in Section 5.3.1.

How to select a random sample using a random number table

How do you randomly select, say 50 people (the study sample) from a group of 839 workers (the sampling frame)? To do this, you can use a random number table, such as the small one provided here (Table 5.1). Typically, you start using the table at an arbitrary point, determined by, for example, rolling a die. You can roll to see which of the large groups of rows to select, and then roll three more times to similarly select the exact row, the group of columns and the exact column. If you rolled 3, 4, 5, 4, you would start at the third group of rows, 4th row, 5th group of columns and 4th column ending up at the number 8.

Since the sampling frame contains 839 workers, number them from 1 (or rather 001) to 839. The number 839 is less than 1000, so you can go through the table, using three digits at a time. Reading from the table then, you would read off the following sequence of numbers to start: 836, 863, 705, 484, 649, 684, 029, 871, 918, 345, 533, 248, 586. Ignore digit triplets lying between 840 and 999, as well as 000. This means you would select for your sample, the workers numbers 836, 705, 484, 649, 684, 029, 345, 533, 248, 586. You could continue until you have the 50 people required. If the random number table should yield a repeated triplet of digits, ignore this and use the next valid triplet.

5.3.2 How to choose a stratified random sample

Random sampling is not guaranteed to produce a sample representative of the sampling frame, although it does work "on average." That is, if we repeat the procedure many times, the average proportion of women in the samples is the same as their proportion in the sampling frame as a whole. When a fairly large sample is chosen, representativeness is more likely. However, with small samples, this may not be the case. To avoid the problem of lack of representativeness, select a *stratified random sample.* This allows you to deliberately select a sample with the same proportion of men and women as in the total group.

Do this by *stratifying* the group into the two sexes. Then choose a sample of men, and a sample of women, applying the same process as in simple random sampling with each stratum. The first sample has the number of men you want; the second, the number of women. Opinion polls typically use some form of stratified sampling, though one that is rather more complex than has been described.

(Caution: with stratified sampling, the statistical approach you use to analyze the data must be modified.)

Another reason for stratifying would be if there are important differences in the reaction of sub-groups to an intervention, and one of the sub-groups in the sampling frame is quite small. For instance, suppose you want to look at the effect of flexible work hours on both men and women in a manufacturing environment by means of a survey, yet women comprise only 5% of the working population. You would end up with about 10 women, if 200 workers were selected at random from the total work force, making your estimate of the effect of the intervention on females in your sample imprecise.

However, the precision could be greatly improved if you first stratify by sex and then choose 50 women and 150 men, using random sampling from each stratum.

5.4 Randomization - forming groups in experimental designs

You have seen how to randomly select a single sample from a bigger group. Suppose, you do an experiment, with one intervention group and one control group. In this case, you *randomize* study subjects (i.e., workers, work groups, or workplaces) to either group and make sure they all have the same chance (probability) of being assigned to each one. Typically, these probabilities will be one half, or equal for each group. It is rather like tossing a "fair" coin - heads you go in the intervention group, tails you become a control.

5.4.1 Why randomize?

Why randomize subjects into intervention and control groups? The primary purpose is to avoid selecting only particular types of people. For example, we do not want only volunteers for the intervention group, leaving other people for the control group. Volunteers differ in many ways from non-volunteers. Similarly, we do not want all the men in one group and women in the other. We even want to avoid the tendency for men to be in one group rather than another.

You might argue that you can certainly tell if subjects are men or women, and thus check for any imbalance of the two sexes in the treatment and control groups. But what about factors you do not measure, or do not even know about or cannot measure? The answer is that with randomization it does not matter! This is because *on average* these factors balance out if you randomize. When we say *on average*, we mean: if we repeat the randomization many times, and each time calculate the resulting proportion of men in the treatment and control groups, the average of all these proportions for the intervention group would be the same as that for the control group. Similarly, the average proportion of women in the intervention and control groups would be equal, following many randomizations. This is true of variables we have not measured.

How to randomize

Suppose you want to randomize people into two groups, with an equal probability of selection into either. As with random selection, there are several ways we can proceed. Using the random number table (Table 5.1), you could start at the point where you left off in choosing your random sample and read off the sequence: 070 83127 01123 02133 08... Taking single digits, if the digit is even (including 0) you allocate the person to the intervention group. If the digit is odd you allocate to the control group. Alternatively, our rule could be that if the digit is between 0 and 4, the subject goes in the intervention group; if between 5 and 9, the subject becomes part of the control group.

We will illustrate this using the odd / even rule and allocate 20 people into two groups, 10 per group. First, number the 20 people 01 to 20. The first digit in the random number sequence is 0, so subject 01 is assigned to the control group; the second digit is 7 so subject 02 is in the intervention group. Continuing, you see that subjects 02, 05, 06, 08, 10, 11, 13, 14, 16, 17 are put into the intervention group. Since you now have ten people in this group, you can stop the randomization and put the three remaining subjects in the control group. Sometimes, you will decide in advance to randomly allocate people to groups without guaranteeing equal numbers in each group. If you did this here, you would keep selecting single digits so that subject 18 would also go into the intervention group and subjects 19 and 20 would go in the control group. This means that out of 20 people, eleven are in the intervention group and nine are in the control group. There is always the risk of an imbalance like this, particularly with small samples.

Sometimes, you might see a study where, even though proper randomization techniques have not been used, it seems that there is no biased pattern of selection into the treatment or control group. Why is this still not as good a design? The problem is that there may in fact still be some selection bias. For example, someone may have deliberately (or even sub-consciously) put into the intervention group people considered more interested in safety. This will mean the groups are not truly comparable.

5.4.2 Randomized block design and matching

You may want to ensure that a characteristic, such as sex, is balanced between the groups, in order to avoid a selection threat to internal validity. Thus, in this case you want equal numbers of men in both intervention and control groups; and, similarly, equal numbers of women in each group. How can you guarantee this?

The answer is to stratify subjects and randomize within the strata (or "block" in the jargon of experimental design). What you do is list all the men to be randomized and assign them in equal numbers to intervention and control groups. Then do the same for women, and you will have a similar distribution of the sexes in each of the groups.

Another possibility is to *match*. First, pair up (match) subjects according to characteristics like sex, age, duration of employment and so on. You can then (randomly) allocate one member of the pair to the intervention group, and the other to the control group. (This process is really like the randomizing within blocks, with each block reduced to just two people). In practice, it can be difficult to get exact matches. So instead of taking people with the same year of birth, you may have to match pairs to within two or three years in age.

5.5 Forming groups in quasi-experimental designs

It might be difficult or even impossible to match or randomize subjects to one group or another, given the requirements of a particular organization of people at a workplace. Or a group (e.g., department, work-site, etc.) might have already been chosen to participate in the intervention, thereby preventing any randomizing of participants to intervention and control groups. In such cases, you can still choose another group, which will serve as the control group, as in a number of the quasi-experimental designs discussed in Chapter 4.

The overriding principle in choosing non-randomized groups for a quasi-experimental design is to make them truly comparable. They should be similar in all respects *apart from the intervention*. In comparing two departments, you want them to be similar in their work activities. You would not compare an accounts department with a maintenance department. Of course within workplaces there may be no directly comparable groups - so aim to select ones that closely resemble each other. You might even try similar departments in other workplaces, preferably from the same type of industry.

The actual choice you make depends on your local situation. We cannot say specifically what group would be best, but several characteristics can be considered.

Characteristics to consider when choosing a non-randomized control group

- worker characteristics (e.g., age, sex, experience)
- nature of job tasks
- work environment (i.e., exposure to hazards, safety controls)
- workplace organization (e.g., structures for decision-making and safety, work flow)
- contextual factors (e.g., health & safety culture; management support for safety)
- past safety record

5.6 Summary

We have described how to randomly select participants from a sampling frame. This is done so that the study sample is representative of the sampling frame and the intervention results will be applicable to the larger group. We also described how the process of randomization can be used to create intervention and control groups in experimental designs. For situations in which groups are formed non-randomly, some considerations were given.

Key points of Chapter 5

- Choose your sampling frame, so that it is typical of the target population to which you want to generalize your evaluation results.
- Select a study sample from your sampling frame using random sampling.
- Whenever possible, use an experimental design with randomization to assign participants to intervention and control groups.
- In quasi-experimental designs, select intervention and control groups so that they are similar.

Chapter 6

Measuring outcomes

6.1 Introduction

Chapters 3 and 4 described the various study designs used in evaluation. In those chapters, we referred to taking measurements. This chapter will discuss those measurements. We will first introduce the concepts of reliability and validity - two key characteristics to consider when choosing a measurement technique. We will then review several common ways of measuring safety outcomes, examining reliability and validity. Finally, we will list a wider range of considerations in choosing your measurement method(s).

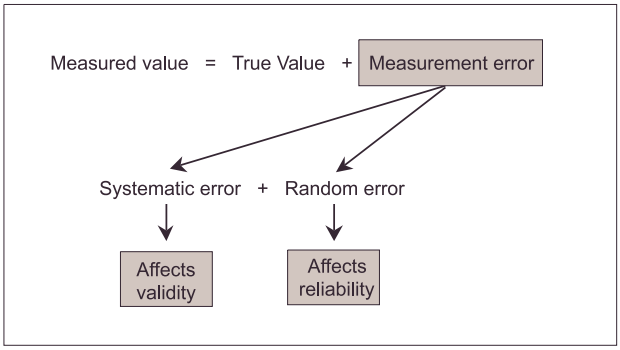
Here, we will be discussing only quantitative methods; i.e., those which yield numerical information. The next chapter deals with methods which yield qualitative information. A comprehensive evaluation should include both types of data.

6.2 Reliability and validity of measurements

A measured value consists of two parts: the true value plus a certain amount of *measurement error* (i.e., the error we make in measuring). It is this measurement error which makes a particular measured value either higher or lower than the true value. Thus, measurements are more *accurate* when measurement error is minimized. Imagine using a ruler to measure the legs of a table. Since we do not measure them perfectly, each estimate of leg length will consist of the true value plus or minus a small amount of error.

Measurement error, in fact, consists of two parts. One part is called *systematic error*, also known as *bias*. This type of error exists when we consistently make an error in the same direction. This would happen, for example, if we always looked at the ruler from an angle, causing us to consistently underestimate the table leg length. The other part of measurement error is *random error*. As the name implies, it fluctuates randomly, sometimes leading to overestimation, and sometimes to underestimation. These two types of measurement error affect the reliability and validity of a measurement method. While evaluating the effectiveness of your intervention, apply measurement methods which minimize both types of measurement error. In other words, these methods should be valid and reliable.

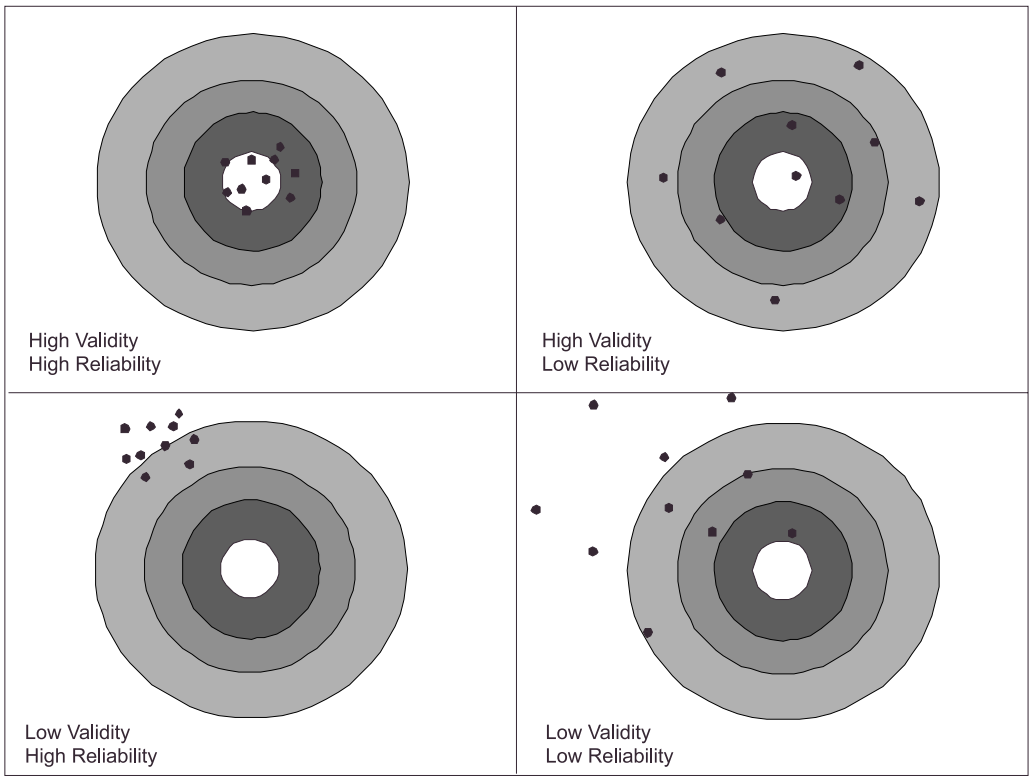
Figure 6.1: Types of error in measurement



Measurements that are *valid* have a low degree of systematic error and measurements that are *reliable* have a low degree of random error. In other words, a valid method means we are measuring what we had hoped to measure. A reliable method gives us consistent answers (while measuring the same thing) on numerous occasions. If a measurement method is both valid and reliable, it is considered to be *accurate*.

Figure 6.2 illustrates these concepts of reliability and validity with the analogy of a shooting target. Consider the center of the target as the true value and the bullet holes as the actual measured values. Reliable measurement has a low degree of scatter to the values (as in the left-hand panels in Figure 6.2). Valid measurement is centered on the true value (as in the top panels).

Figure 6.2: Illustration of the effects of reliability and validity on measurement



Why reliability and validity are important

Poor reliability is a problem for evaluation because it makes it harder to detect any effect of a safety intervention, even if it is truly effective. This is because it is hard to tell whether any changes in critical safety outcome measures are due to the intervention or simply to random fluctuation. Unfortunately, lost-time injury rate data, except in larger workplaces, often has low reliability. For this reason, alternative or additional outcome measures must often be used to measure the effect of an injury reduction intervention.

When methods are reliable, but have poor validity, your conclusions drawn from evaluation results might be wrong, especially if you are measuring a concept different from the one you thought you were measuring.

Specific types of reliability and validity

There are several specific types of reliability which may be of concern to evaluators: analytical equipment reliability or precision, inter-rater reliability, test-retest reliability and internal consistency of questionnaires. We will elaborate on these in our discussion about particular types of outcomes measures.

Similarly, several types of validity can be at issue. Major types are criterion, content and construct.

- *Criterion validity* is the extent to which the measurement predicts or agrees with some criterion of the "true" value or "gold standard" of the measure. For example, you would establish the criterion validity of work-site observation measurements by showing a correlation of these measurements with injury rates in a given workplace.

- *Content validity* is applicable only to instruments like questionnaires or checklists. It is concerned with whether or not they measure all aspects of the concept(s) of interest. This type of validity can be established by having experts review the instrument for its completeness of content.

- *Construct validity* is the hardest to establish, but the most important. It pertains to how well the measure accurately indicates the concept (or *construct*) of interest. This is not an issue when you are dealing with a concrete concept like injuries. You already know that lost-time injury rates are accepted indicators - assuming the data are free from biases. On the other hand, construct validity becomes more of an issue whenever you are measuring abstract concepts. For example, if the objective of the intervention is to change the safety climate of the workplace, you might want to measure the outcome (safety climate) with a safety climate questionnaire. This is a good idea, but you need to determine how the construct validity of the questionnaire was established. In other words, how was it established that the questionnaire truly measures "safety climate" and not something more limited like safety policy and procedures.

6.3 Different types of safety outcome measures

Common safety outcome measurement methods
1. Administrative data collection - injury statistics
2. Administrative data collection - other statistics
3. Behavioral and work-site observations
4. Employee surveys
5. Analytical equipment measurement
6. Workplace audits

In the following, we discuss some of the more common outcome measurement methods, with a focus on the reliability and validity.

6.3.1 Administrative data collection - injury statistics

Several types of injury statistics have become standard reporting measures for companies. They are often determined by legislative requirements. The most common measure, *injury frequency rate*, is equal to the number of injuries per unit of exposure. There are different categories of injuries where rates can be calculated: e.g., *lost-time or disabling injuries; recordable injuries* (i.e., those required by law to be recorded); *medical treatment injuries*; and *first-aid only injuries*. Although a less commonly accepted standardized measure, *near-injury* rates can also be calculated.

Various units of exposure are used to calculate frequency rates. Worker-hour units of exposure - typically 100,000 worker-hours or 1,000,000 worker-hours - yield relatively precise frequency rate estimates. However, the number of workers, a cruder measure of exposure, is also used. The choice between the two depends on the state of record-keeping, with the former requiring good records of worker-hours, including lay-offs, overtime and lost time. The *number of injuries* can also be used to compare two time periods of equal length, but the equivalence of the worker-hours of exposure during the time periods must be confirmed.

Severity rate is another widely used injury statistic. It is calculated by taking a ratio of lost-time hours over the corresponding units of exposure - the higher ratios corresponding to greater severity. Severity rate is a useful complement to frequency rate, since some interventions can have an impact on severity, but not on frequency. This could result, in some cases, from the interventions affecting severity more than frequency. It could also result from it being easier to (statistically) detect an effect on severity than frequency, for a given effect size.

Claims data collected by workers' compensation systems are useful for evaluating interventions delivered to a multiple workplaces in a jurisdictional area.

Validity and reliability concerns with injury statistics

The major concern in using injury and claims data involves the reporting *biases* that may exist and stem from systematic errors which cause injury records to be consistently different from the true injury occurrence. Such biases can enter at any point during the transfer of information - from the time a worker experiences an incident to when that incident becomes part of national statistics. On the one hand, certain compensation or work environments may encourage over-reporting of injuries by workers. On the other hand, incentives to individuals or organizations to minimize injuries may encourage underreporting of injuries or a reclassification to a lower level of severity. In particular, incentives for early return-to-work might result in the recording of a medical-aid only incident, which in the past or in a different location would have been considered a lost-time injury.

The degree of underreporting can be a great source of bias. One study of hospital workers' data from self-report questionnaires showed that 39% of those who had experienced one or more injuries did not report them.[23] Although the main reason for not reporting was that they considered the injuries too minor, in fact, 64% of them involved medical treatment and 44% lost work time. In another study in the U.S.,[24] OSHA 200 forms (U.S. workplaces are required to record injuries and illnesses meeting certain severity criteria for the Occupational Safety & Health Administration) from several companies were compared with company clinic records. This showed that the OSHA logs had captured only 60% of the reportable injuries.

A filter model of the injury reporting process[25] has been developed that can help identify the places at which biases can influence injury reporting and the reasons for these biases (Table 6.1). A filter in this model is anything which prevents some of the reportable injury data at one reporting level from passing to the next level of reporting. For example, level one is considered to represent the true injury rate, with the first filter being the worker's decision-making process about whether to report the injury or not. The second, third and fourth filters operate at the workplace; the fifth at the transmission of company-level data into aggregate data at the jurisdictional level. The filters operate differently for injuries of differing severity. The less severe the injury the more effective and less consistent are the filters. Accordingly, near-injuries are especially prone to biased and inconsistent reporting.

Often the presence of some reporting biases can be tolerated in an evaluation - if they continuously affect the data in the same way. A problem arises if they operate differently, either over time or between groups, for any of the measurements being compared. Just how differently they operate has to be estimated and taken into account when interpreting results.

One special evaluation challenge in using injury statistics are those cases where the intervention itself has an impact on the way the filters operate. Management audits, for example, tend to improve reporting, while safety incentive programs that reward workers or managers for low injury rates discourage reporting. In such situations it is important to include methods for verifying the injury data, as well as to incorporate supplementary outcome measurement methods, not subject to the same type of biases.

[23] Weddle MG [1996]. Reporting occupational injuries: the first step. J Safety Res 27(4):217-223.
[24] McCurdy SA, Schenker MB, Samuels SJ [1991]. Reporting of occupational injury and illness in the semiconductor manufacturing industry. Am Public Health 81(1):85-89.
[25] Webb et al. [1989]

Table 6.1: The filter model for work injury reporting: six levels and five filters[26]

	Worksite			Company Administration		Govern-ment
	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
	Total injury rate in work-place →	Worker-defined injury rate →	Supervisor-defined injury rate →	Medical centre-defined injury rate →	Company-defined injury rate →	State or national work injury statistics
	First filter	**Second filter**	**Third filter**	**Fourth filter**	**Fifth filter**	
Character-istics of the filters	Illness behaviour	Referral to medical centre	Completion of accident report form	Completion of company accident statistics report	Compilation of aggregate accident statistics	
Key Individuals	Worker	Supervisor	Medical center staff	Safety officer	Government agencies	
Factors in operating on key individuals	Severity and type of symptoms; stress; learned patterns of illness	Training; attitudes; workload; conse-quences of referral, e.g., increased paperwork	Training; attitudes; workload	Training	Stringency of surveillance	
Other factors operating	Attitudes of fellow workers	Physical proximity of supervisor to worker; proximity of medical center; company policy; production demands; characteristics of worker	Severity and type of injury; company policy; characteristics of worker	Company policy; insurance company requirements; trade union effectiveness; legislative requirements	Efficiency of data collection methods	

[26] Reprinted from Accident Analysis & Prevention 21, Webb et al. Filtering effects in reporting work injuries, 115-23, Copyright 1989, with permission from Elsevier Science.

Misclassification errors can cause a problem in injury statistics. Misclassification can be attributed to errors or disagreement in the judgement of people involved in abstracting information from incident or clinical records, and depends in part on the individual coders, as well as on the classification scheme used. A common method of classifying injuries and illnesses, the International Classification of Diseases, 9th revision (ICD-9) has been shown to perform poorly for soft-tissue disorders, since several different ICD-9 codes can be assigned for the same condition.[27] As well, the ability to code incident descriptions consistently has been shown to depend on which aspect of the incident is being coded. Table 6.2 shows the percent agreement in coding for two trained people coding the same 100 injury reports using a standardized coding scheme[28].

Table 6.2: Reliability of coding injury descriptions

Item being coded	*Reliability*
Sex	0.98
Year of birth	0.89
Industry classification	0.64
Injury location	0.64
Type of injury	0.92
Part of body injured	0.92
Injury legally notifiable or not	0.84
Agent of injury	0.79
Event precipitating injury	0.44
Contributory factors	0.61

Coding is considered unreliable when scores are less than 0.7. As you can see from Table 6.2, many of the important items being coded fell below this cut-off point. To improve the inter-rater reliability, the coders may need more training or the coding scheme may need revising. In addition, you may want to maintain the same coders through the data collection phase of the evaluation, because a change in personnel could have a large impact on the results. You should also check for any changes in classification practices over time or differences in the coding practices between intervention and control groups.

Checking the validity of injury statistics before and during data collection

If you use injury statistics as an outcome measure to evaluate the intervention, consider the potential biases at the outset. Investigations of data validity can be made beforehand, and, if necessary, corrective steps taken to improve data collection. However, if that is done, wait until their impact on the resulting statistics has stabilized.

Also, check the validity of the injury statistics by comparing them with data obtained at a lower level of reporting - e.g., a comparison of the frequency of incidents in summary statistics with the medical or safety records on which the statistics are based. Sometimes, several sources are used to identify the complete universe of incidents, after which you can determine how well any single source captures them. You might use supervisor, clinic and claims reports to identify the universe of incidents in a workplace and then see what percentage is captured by each type of report.

Checking the validity of injury statistics after data collection

Even after ensuring that the collected statistics are consistent with respect to any biases, reality might differ from expectation. Thus, it is a good idea to check the data after it has been collected as well. A good indicator of underreporting is

[27] Buchbinder R, Goel V, Bombardier C, Hogg-Johnson S [1996]. Classification systems of soft tissue disorders of the neck and upper limb: do they satisfy methodological guidelines? J Clin Epidemiol 49:141-149.

[28] Adapted from Glendon AI and Hale AR [1984]. A study of 1700 accidents on the youth opportunities programme. Sheffield: Manpower Services Commission.

the *ratio of minor to major injuries*. Since minor injuries are more likely to be "filtered out" than major ones, a constant ratio indicates the stability of any underreporting biases (or lack of them). This method does depend on major injuries' occurring frequently enough in the period measured that they perform as reliable measures.

Small numbers issue and injury statistics

Lost-time injuries and other severe injuries might not be useful in effectiveness evaluations in smaller workplaces. The small number of injuries found in a smaller workplace leads to statistical tests with low *power*. This means it will be more difficult to detect an intervention effect, even if the intervention is truly effective.

For this reason, people sometimes use the frequency of less severe injuries, e.g., first-aid-only injuries, to evaluate safety intervention effectiveness. However, less severe injuries are more susceptible to underreporting biases than more severe injuries. Thus, the opposing considerations of statistical power versus validity must be weighed when choosing measurement methods. Examine both severe and minor injuries if the data are available.

Another alternative or supplementary approach to the small numbers issue, is to measure intermediate outcomes, e.g., work practice observations, work-site observations or safety climate. Power will likely be greater in the statistical tests on these data than on injury data. Another advantage of this approach lies with the location of the intermediate outcomes, which are situated closer to the intervention than injuries in the causal chain. The intervention effect is therefore less "diluted" or attenuated by other factors. The challenge in using upstream proxy measures for injury outcomes is that a good correlation has to be established between this proxy measure and a more established measure of safety such as injuries, in order to demonstrate the validity of the proxy measure. We will discuss some other measures below.

6.3.2 Administrative data collection - other statistics

As discussed, there are some problems with relying upon injury statistics for evaluating workplace interventions. For this reason you might choose other types of data found in administrative records. In doing so you can adopt the strategy, referred to earlier, of using an intermediate outcome as a proxy for the final outcome in the evaluation. Even when the injury statistics serve as a final outcome measures, additional administrative data can provide insight on the way the intervention brought about change. Finally, these data are often useful for demonstrating that the intervention is being implemented as planned.

As described in the above section, consider whether any biases are entering the process of collecting these data; and conduct validity checks if needed.

Example of the use of other administrative data[29]

Consider an example in which an incident investigation intervention is being evaluated. A new consultative group has been formed to assist supervisors investigating incidents, with an aim to improving the quality of actions resulting from the investigation. The following quantitative measures derived from administrative records are included in the intervention evaluation: 1) time between incident occurrence and formal reporting of it; 2) the number of near-incidents reported; 3) percentage of incidents for which corrective measures were suggested.

Change is seen in all these intermediate outcomes in the direction that suggests that the intervention has been effective. It is also found that the number and severity of incidents (final outcomes) showed a decrease, although only the latter is statistically significant. Thus, even though the injury frequency did not show statistical significance, changes in the intermediate outcome measures (as well as some qualitative evidence not discussed here) *together* suggested that the intervention had been effective.

[29] Example from Menckel and Carter [1985]

6.3.3 Behavioral and work-site observations

Observations of worker safety-related behavior are increasingly being used to measure the effect of safety interventions, especially behavior-based interventions. To use this method, you first develop a checklist of behaviors that the intervention is trying to influence. Targeted behaviors are identified from incident investigation reports, as well as from the opinions of supervisors and workers. Be prepared to first develop a trial checklist and make adjustments according to feedback. This will ensure that the list is comprehensive (improving its validity) and the items are unambiguous (improving its reliability).

The checklist is then used by an observer (trained supervisor, evaluator or worker) who visits the work area at a randomly selected time of day and makes observations for approximately half an hour. For each behavioral item on the list, the observer marks either "performed safely", "performed unsafely" or "not observed". Following the observation period, the proportion of safe behaviors is calculated. It consists of the ratio of the number of items performed safely over the number of items observed.[30]

Another checklist approach uses observations on the work-site conditions (e.g., housekeeping, emergency equipment). Details of such a method for the manufacturing industry can be found at a site of the Finnish Institute of Occupational Health:

www.occuphealth.fi/e/dept/t/wp2000/Elmeri1E.html.

Advantages and disadvantages of behavioral and work-site observational measures

Observational measures offer several advantages. First, they are "leading indicators" instead of a "trailing indicators", meaning you do not have to wait until incidents happen to get a measure of safety. Rather, measurement is "upstream" of incidents in the causal pathway. Second, you can take observations frequently, as often as several times a week. This yields data sensitive to changes caused by interventions and can be analyzed for time trends. Third, there is some evidence that behavior serves as a valid proxy for injuries as a final outcome measure. Reviews of injury records show that the majority of injuries are associated with unsafe acts. [This is not to say that responsibility for such injuries lies exclusively with the worker carrying out these unsafe acts. Conditions leading to these unsafe acts are typically the responsibility of management.]

Further, evaluations of behavioral interventions - at least the ones that have been published - tend to find that an improvement in behaviors is correlated with a decrease in injury rate. Validation of work-site checklists using injury rates as a criterion has also been achieved [Laitinen et al. 1999a,b].

A drawback with behavioral observations is that they are sometimes regarded unfavorably by those being observed, and in some environments could be considered unethical or otherwise unacceptable. This is less of a problem with work-site observations, because there is less emphasis on the behavior of individuals. As well, observations of the work-site can be carried out in a less intrusive manner, thereby interfering less in the measurement of the intervention's effect.

[30] See Krause [1995] for more details on this methodology.

Additional validity and reliability issues when using behavioral and work-site checklists

There are additional methodological issues to consider when using observational checklist techniques. One is the *inter-rater reliability* of the list. This is determined by having more than one person observe the same area simultaneously and independently record their observations. The amount of agreement between the two raters can then be compared. Typical ways of reporting inter-rater reliability are the percent agreement on items (the percentage of list items categorized the same) or, better, by using a special statistic known as a *Kappa coefficient*.[31] This statistic takes into account the percent agreement that would occur simply by chance before the percent agreement is calculated. Kappa values between 0.40-0.59 indicate agreement which is moderate, 0.60-0.79 substantial and 0.80-1.00 very good. A high reliability ensures that there will be little variation in the data as a result of having different observers carry out the observations. If this reliability is low during the piloting of the checklist, you should be able to improve it by removing any confusing language or better specifying criteria.

In order for the checklist to be able to measure change, avoid what are called *ceiling effects*. This refers to the situation where you are unable to detect an improvement of the index score, because the initial measured values were already high and have limited room for improvement. Thus, you ideally want your pre-intervention index scores to be around 50%. If they are higher than 80% during the testing phase, see if you can modify or substitute items so that they will be more difficult to achieve and still be true to the intervention's aims.

Before accepting observations as the definitive final outcome measure in evaluating an intervention, you would want to determine if a statistically significant *correlation* exists between the behavioral index scores and injury rates. If this is not possible beforehand, then collect some kind of injury data during the evaluation, along with the observations, in order to calculate such a correlation. To get data yielding sufficient statistical power during analysis, you might require a measure of minor injuries, provided that any underreporting biases remained constant throughout the study.

6.3.4 Employee surveys

Employee surveys often measure what cannot otherwise be observed. They examine the knowledge, attitudes, beliefs, or perceptions held by individuals. Occasionally, they assess group phenomenons such as climate or culture. They can also be used to measure (self-reported) safety practices and injuries, both of which can also be quantified by the methods discussed above.

Surveys of knowledge are an appropriate final outcome measure if the intervention is only designed to change knowledge. A similar statement could be made about surveys of attitudes, beliefs, perceptions, practices, culture or climate. However, if the intervention is designed to ultimately affect worker safety and injury rates, then one must be cautious about using surveys which measure only knowledge, attitudes, beliefs or perceptions as a proxy for injury rates as a final outcome measure. Such questionnaires have not usually been sufficiently validated through correlation with injury rates to justify their use in this way.

Tips for better questionnaire data

In choosing from a number of possible questionnaires, Exhibit 6.1 has some questions that will assist you in that selection. They are based on the assessment of the validity and reliability of the proposed questionnaire. The more questions to which you can answer "yes", the more likely the questionnaire is suitable for

[31] Percent agreement and Kappa statistics are used with categorical data; correlation coefficients are used to measure the inter-rater reliability of continuous data.

Exhibit 6.1 Questions for assessing questionnaire suitability

1) Do the questions seem applicable to your environment?
2) Was the questionnaire developed on a similar population in a similar environment or, alternatively, has its validity in a diverse range of populations / environments been demonstrated?
3) Was a process followed for ensuring that the content of the questionnaire was complete? (i.e., were all important aspects of what is being measured included in the questionnaire's content?)
4) Do the questions measure what your intervention is designed to change?
5) If you are measuring an intermediate outcome with a questionnaire as a proxy for a final outcome, has a statistically significant correlation between the proxy and the final outcome measure been established?
6) Has the sensitivity of the questionnaire been shown by its detection of changes over time, ideally in response to an intervention similar to yours?
7) Has good "test-retest reliability" of the questionnaire been demonstrated?[32]
8) Is there a high "internal consistency" among groups of questions forming scales?[33]

the effectiveness evaluation.

If you want to develop a questionnaire, consult some specialized books on the subject, such as the one by Aday [1996]. Better yet, consult with a specialist in the area (e.g., an organizational psychologist) to assist you. Developing a good questionnaire requires a significant investment of resources; so whenever possible we suggest you use existing questionnaires, scales or items.

Administering an employee survey

Devise a method of distributing and collecting questionnaires so that you will know your response rate; i.e., how many have been returned out of the number given to potential respondents. It is important to achieve high response rates in surveys so that the results can be considered representative of the entire group to which the questionnaires were given. Another check on representativeness - especially important in the case of a poor response rate - involves looking for any differences (e.g., age, department, etc.) between those who responded to the questionnaire and others who did not. The greater the difference between the two groups, the more cautious you must be in drawing conclusions from the survey. Sending potential respondents one or two follow up reminders about completing the questionnaire is a good idea. Participation is also more likely if you can assure people of the confidentiality of their responses, publicize the survey through the workplace and obtain support from influential management and worker representatives.

[32] Test-retest reliability, usually measured by a reliability coefficient, which ranges between 0 and 1, is the consistency of questionnaire scores on the same group of people given the same questionnaire on more than one occasion. The period between occasions should be long enough that people forget the answers they gave on the first test and short enough that what is being measured has not changed. This depends on what is being measured, but is typically 2 weeks to 3 months. A reliability coefficient of 0.8 is considered good; most consider a value of 0.7 to be minimally acceptable.

[33] Internal consistency, usually measured by "Cronbach's alpha", is a measure of the reliability of scales. Scores on scales are made up by combining the scores of answers to a group of related questions. Higher values of alpha show that the questions in the scale are in fact related and can be considered a scale for measuring a construct. Alpha values between 0.70 and 0.90 are optimal.

6.3.5 Analytical equipment measures

Data collected with analytical equipment will not serve as final outcome measures for evaluating a safety intervention, but they might serve well as good intermediate outcome measures. For example, to evaluate an intervention involving a workstation or task redesign to decrease musculoskeletal injury, you could indirectly measure changes in biomechanical exposures indirectly through electromyography of muscle activity or videotaped task performance.

Validity and reliability of analytical equipment measures

The major issues related to the use of analytical equipment are largely those related to experimental control of the study conditions. You need to ask: is the instrument being used under the conditions for which it was intended (e.g., temperature)? Are proper calibration and maintenance procedures being followed? Is there anything present in the environment that could interfere with the equipment so that it gives false readings? Is the equipment operator properly trained and using standard procedures? Any of these sources of error could potentially affect results in either a systematic or random way.

The reliability and validity of measurements taken with analytical equipment can be improved by minimizing variation in the operation of the equipment - both in the environment and in those operating the equipment. Reliability can also be improved by taking multiple measurements and using the average as the data point. However, the additional cost of taking multiple measurements has to be balanced with the gains in reliability realized, especially if the equipment is not a major source of unreliability in the study.

6.3.6 Workplace audits

Workplace safety audits are another way to assess safety interventions. They focus on safety elements upstream of injuries, such as safety policy, management practices, safety programs, and, sometimes, workplace conditions. Audits have been developed by both commercial and non-profit organizations; and large companies have even developed in-house company-specific audits. Sector-specific audits also exist. Often, they are designed to give qualitative information, but some yield a quantitative measure or score. These summary scores can then be used as an outcome measure in evaluating certain interventions at the organizational level.

Validity and reliability considerations when using workplace audits

Before using an audit, consider the same questions already raised regarding employee questionnaires (section 6.3.4). In particular, make sure that the content of the audit is appropriate, given the intervention's goals. If the findings from the audit are being used as a proxy for a final outcome measure, such as injuries, you will need data that validates its use in that manner. Data could be from similar workplaces and show a statistically significant correlation between audit scores and injuries.

6.4 Choosing how to measure the outcomes

Your choice of outcome measures will depend on many things, including the nature of the intervention and its objectives, the setting, and your resources. While injury rates might be a suitable choice in one case; it might not be in another.

6.4.1 Evaluation design and outcome measures

Final outcome measures

Consideration of the safety intervention's objectives should help in deciding what is the ideal type of outcome to assess the intervention's effect. If the intervention is ultimately meant to reduce injuries in the workplace, then the ideal outcome measurement is an (unbiased) measure

Considerations when choosing the outcome measures

Evaluation design and outcome measures

1. Which measures should be included to address the objectives of the safety intervention (final outcome)?
2. Which, if any, measures should be included to provide an understanding of how the intervention works or bolster the strength of the design (intermediate and implementation outcomes)?

Measuring unintended outcomes

3. Which measures should be included to detect possible unintended outcomes of the intervention?

Characteristics of measurement

4. Do the methods really measure the outcomes they are intended to measure, from a conceptual point of view (construct validity)?
5. Is the outcome measurement method free of systematic biases (validity)?
6. Is the measurement method reliable?
7. Have the measurement methods been used on a group similar to the one under study before?

Statistical power and measurement method

8. Will there be sufficient statistical power during analysis with the method chosen and the number of evaluation participants?

Practical considerations

9. Is the measurement method feasible (i.e., cost, administrative requirements)?

Ethical aspects

10. Can the measurements be carried out in an ethical manner (i.e., fully informed consent)?

of injuries. Unfortunately, the frequency of lost-time events in a given month or year in many workplaces is too low to show clearly an intervention's effect. Sometimes, inconsistent biases in the injury statistics are of concern. If you cannot collect useful injury data, then you need a good substitute - with evidence that it is indeed a good substitute. For example, if you want to evaluate your intervention using observations as a substitute for injury rates, you need to show that a strong correlation exists between the observational measure and (unbiased) injury statistics.

Choosing intervening outcome measures

The choice of intervening outcome measures will depend on an understanding of how the intervention works, as shown in the conceptual model or program logic model relevant to your intervention. We already discussed how you can strengthen your evaluation by including measurements of the steps between the intervention and the final outcome (Section 4.2.5.1). This provides insight into how the intervention worked (or did not work), which is useful for planning future interventions. It can also bolster any evidence of an effect on the final outcome. For example, you might find that a decrease in injuries followed a training intervention. There is a temptation to think that the intervention had been successful. However,

if you also measured the practices targeted by the intervention and found that they did not change, the question arises as to whether the change in injury rate was due to something else besides the intervention. On the other hand, if both injuries and unsafe practices showed a decrease following the intervention, you could be more confident that the intervention indeed caused the cause of the decrease in injuries.

A thorough effectiveness evaluation determines the extent to which the intervention was implemented as planned. This information will be especially valuable if the intervention appears to have no effect. You want to distinguish between the two possibilities: 1) the intervention is inherently ineffective, even when implemented as planned; or 2) the intervention is potentially effective, but was poorly implemented. Such information is valuable to those who might want to repeat the intervention or who design interventions. Program implementation can be assessed using both qualitative and quantitative measures.

6.4.2 Measuring unintended outcomes

Unintended consequences of the intervention, by their very nature, are difficult to anticipate, and hence, difficult to measure. It is possible that an intervention successfully decreases one type of injury but increases another type. This increase could occur in the same group of personnel or it could involve another group within the same workplace. Safer conditions, for example, for equipment operators might mean more danger for maintenance workers. The basic principle in measuring unintended outcomes is to include measurements apart from the ones most directly related to the intervention.

Other unintended outcomes may arise through compensatory behavior in response to the intervention. A particular engineering intervention to reduce exposure to a hazard, for instance, could result in a decrease in the use of personal protective equipment because people feel safer. By including a measure of personal protective equipment use in your evaluation, one could see if this was happening. If so, it could explain why an intervention which looked promising failed.

An example of unintended outcomes measurement

Interventions to decrease needlestick injuries in hospitals have typically involved recommendations to avoid recapping. The primary indication of success of this intervention is a decrease in the frequency of recapping injuries. However, it has also been important to track other types of needlestick injuries. In some cases an increase in injuries during disposal has been detected, which has led to replacing disposal receptacles with ones of a safer design. You also need to confirm that the decrease in needlestick injuries in health personnel has not been achieved at the expense of an increase in injuries to those who handle disposal receptacles and garbage.

6.4.3 Characteristics of measurement method

We already discussed the very important considerations of reliability and validity at the beginning of this chapter. You also need to consider these characteristics of the measurement method within the context of the group and setting where it is applied. For example, a questionnaire developed to be reliable and valid with a white-collar, native-speaking working population could perform poorly with a blue-collar, immigrant population. Thus, measurement methods developed with a different work population might need modification before they can work well in another situation.

Consider also the conditions under which the

measurement method is used. If they are not the same as those for which the method was designed, then determine whether they perform well under the new conditions. A delicate instrument performing well under laboratory conditions might not do so well in a workplace where there is vibration, dust, etc. The results from a self-administered questionnaire could be quite misleading if, because of workplace politics, you are not allowed to ensure that everyone has actually received the questionnaire. Such issues of appropriateness and the implications for reliability and validity of data might lead you to choose a different measurement method than what might be used under other conditions.

6.4.4 Statistical power and measurement method

In choosing an outcome measure, consider whether there will be sufficient *statistical power* during analysis (see section 8.4). Thus, a power calculation should be carried out **before** the intervention is implemented. Calculations might show that the outcome measure initially chosen will not yield sufficient power, given the planned number of evaluation participants. You might then choose a different measurement method or measure a different outcome in order to increase power. For instance, you might decide to use a self-reported symptom survey or reported minor injuries instead of reported lost-time injuries.

6.4.5 Practical considerations

It is also important to be practical when choosing measurement methods. What is the cost of taking these measurements in terms of time and material resources? How much disruption of work processes is involved in taking these measurements? Is the necessary expertise available to carry out these measurements properly? Are data already being collected for other reasons available for the purposes of the evaluation?

6.4.6 Ethical aspects

There might be ethical issues about the use of some measurement methods. For instance, behavioral observations may be inappropriate in some environments - at least without the consent of those being observed. It is customary, actually required, that researchers in an academic environment obtain written consent from participants before using personal data (e.g., health records, employee opinions).

6.5 Summary

This chapter highlighted two important measurement concepts: reliability and validity. Several methods of measuring safety outcomes, including administrative data collection, behavioral and work-site observations, employee surveys, analytical equipment measures and workplace audits were reviewed with a focus on reliability and validity issues. Additional issues which influence the choice of evaluation measurement methods besides measurement properties were also discussed: outcomes indicated by the evaluation design; detecting unintended outcomes; statistical power; practicality; and ethics.

Key points of Chapter 6

- Choose implementation, intermediate and final outcome measures based on the intervention's conceptual model or program logic model, as well as the objectives of the intervention.
- Consider injury statistics, other administrative records, behavioral/work-site observations, employee surveys, analytical equipment and workplace audits as means of measuring outcomes.
- Use measurement methods which are valid, reliable and practical.
- Try to eliminate bias in injury statistics or keep such bias constant. If there is change, report on its estimated effect on the results.
- Choose measures which will yield sufficient statistical power during analysis.
- Consider the ethical aspects of measurement methods.
- Anticipate and try to measure unintended outcomes of the intervention.

Chapter 7

Qualitative methods for effectiveness evaluation: When numbers are not enough

7.1 Introduction

Qualitative methods play an important role in safety intervention evaluation. Although in most situations, numbers are necessary to prove effectiveness, qualitative methods can yield information with a breadth and depth not possible with quantitative approaches.

We first describe four methods of gathering qualitative information: 1) interviews and focus groups; 2) questionnaires with open-ended questions; 3) observations; and 4) document analysis. We identify and illustrate several different ways in which these types of data can be used in an effectiveness evaluation. We follow with some details of how to select study subjects, analyze the collected data, and ensure good quality data.

Methods of collecting qualitative data

1) Interviews and focus groups
2) Questionnaires with open-ended questions
3) Observations
4) Document analysis

7.2 Methods of collecting qualitative data

7.2.1 Interviews and focus groups

A major means of gathering qualitative information is through in-depth interviewing. This involves *open-ended questions*, where interviewees can answer questions on their own terms and in as much detail as they like. This is in contrast to the typical questions found on employee surveys, that prompt for yes/no, multiple choice or very short answers. For example, a truly open-ended question asks "what do you think about the new safety program?". In contrast, only a limited range of answers is allowed if you ask, "how useful was the new safety program?" or "was the new program useful?"

The types of questions used in interviews will depend on the purpose of the data- gathering. They could be about any of the following:

- knowledge (e.g., What did you learn about in the training?)
- experience (In what ways, if any, have things changed in the way safety is done around here, since the program began?)
- practices (In what way, if any, has the training program influenced your safety practices on the job?)
- opinions (What do you think of the program?)
- beliefs (What do you think the company's goals are in providing you this program?)
- feelings (How do you feel about participating in the program?).

A good interviewer is sensitive to the mood and feelings of the interviewee(s), listens well, and encourages them to elaborate on the topic discussed. Better interviews will result from someone who has been trained to conduct interviews and has practiced with the interview questions. There are a number of approaches for collecting interview data.

Structured interviews

Structured interviews contain a standardized means of getting information. The same set of carefully worded and ordered set of questions are used with each respondent. This technique reduces the level of skill required for the interviewer to do a good job and curtails the influence of any particular interviewer on the results. Structured interviews are useful where several people are conducting the interviews or if the interviewers are inexperienced. On the other hand, there is less opportunity to learn about individual subject differences and circumstances while using the structured approach.

Semi-structured interview

A semi-structured approach to interviewing represents a compromise between standardization and flexibility. Here, an interview guide is used, which is basically a checklist of the issues explored during the interviews. There is no set order to the topics, and specific questions are not necessarily worked out ahead of time. However, before ending the interview, the interviewer makes sure all the items have been covered, through the natural course of the conversation. Any topics not yet covered can then be addressed. As with the structured interview, this method ensures that the same type of interview data are gathered from a number of people.

Unstructured interview

The unstructured interview is more like an informal conversation, where questions are generated during the natural flow of conversation. Although certain topics are covered, there are no predetermined questions. The data varies with each person interviewed. This makes the analysis more challenging. As well, more opportunity exists for an interviewer's bias to influence the results. The strength of this approach though is that the interviewer can tailor the approach and line of questioning to each individual.

Focus group interview

A focus group is like an interview with a small group of people rather than only one person. A semi-structured approach is most useful. About six to ten people can be interviewed together and the interviews usually last from one-and-one-half to two hours. This allows time for participants to discuss about eight to ten questions.

The focus group technique is a highly efficient way to collect data. You receive the opinions of several people at the same time. The social setting provides a measure of validation for the information, since extreme or false views tend to be challenged by others in the group. A skilled facilitator can guide the group's dynamics so that the participants stay on topic and people who are either shy or have less popular opinions are encouraged to speak.

Exert some caution in selecting individuals for a focus group. First, this format is not advisable if sensitive information of either a personal or organizational nature is sought. People might be reluctant to speak up and could be vulnerable to repercussions if they do. For similar reasons, and, depending on the subject of the interview, you should probably group people together with similar positions within the organizational hierarchy. In particular, consider separating labor and management; and supervisors and those they supervise. In some cases, you might want to group men and women separately.

Guidelines for obtaining good interview data

1. Let the subject(s) know at the outset how long the interview will last, its purpose and general outline. Explain how confidentiality will be observed.
2. Obtain consent (preferably by signing a consent form) for participating before starting the interview
3. Start off the interview with non-controversial questions that require minimal recall. More sensitive topics, including questions on knowledge, should be covered once a rapport has been established.
4. Create an atmosphere of having a conversation. You do not want people to feel as if they are being examined.
5. Ask clear, truly open-ended, questions.
6. Be nonjudgmental.
7. Be attentive. Indicate interest through your actions and offer verbal feedback.
8. Tape record the interview in order to have a detailed record for analysis. Record important points in your notes.

7.2.2 Questionnaires with open-ended questions

Researchers do not consider structured questionnaires - even with truly open-ended questions - to be the most effective way to gather qualitative information. It is assumed that many people do not want to take the time to write out a response. As well, it cannot be sensitive to interviewee differences, since everyone gets the same question. The depth of responses is limited because there is no opportunity to follow up on an interviewee's statement with other questions.

On the other hand, if you are using a questionnaire to measure the quantitative objectives in the project, you can then quite economically expand the breadth of the results by including a few open-ended questions. These can be useful for gauging participant reactions, identifying program barriers, bringing out unintended consequences of the intervention, and verifying the picture obtained from quantitative measures. Furthermore, the results of this initial screen can help you decide on the nature and extent of any follow-up qualitative studies.

7.2.3 Observations

Another way of collecting qualitative data is to actually go on-site and observe what is going on. Depending on your needs for the evaluation, everything can be captured, including the physical environment, social organization, program activities, as well as behaviors and interactions of people. Or you can take a more narrow focus. The type of observational data used in qualitative analyses can be different than that used in quantitative analyses. In the latter, specific observations are always being sought: e.g., whether a particular procedure is being done correctly or if a particular work-site condition is observed. In contrast, for the purpose of qualitative analysis, specific types of observations might not be defined beforehand.

Observational data is especially helpful in evaluating safety programs as an external evaluator. An understanding of the physical and social environment will be increased. You will catch issues that might go unreported during the interviews because the insiders are too close to their situations. As well, people might not speak freely during interviews in fear of reprisal from co-workers or management. Finally, an on-site visit can be the best way to verify that intervention activities are occurring as described.

If you are an internal evaluator planning to use observations, be aware that one's view of things is influenced by one's background and position within the organization. Thus, if observations are going to play a large role in an evaluation, consider bringing in an external, more neutral observer. Similarly, you might have to choose between being an observer or a participant, or something in between. The more you participate, the more first-hand your knowledge will be. The disadvantage is that it becomes more difficult to maintain "objectivity" and your presence could influence those around you.

Tailor the length and frequency of observations to your requirements. This can range from a single two-hour site visit to verify program implementation to a full-time, year-long presence to fully understand, for example, a change in safety climate. Field notes are the primary means of recording observational information. This can be supplemented with photographs or videos, although such methods are often obtrusive. Good field notes require a selectivity that can focus on the important details, yet not severely bias the notes.

7.2.4 Document analysis

Documents of interest in workplace safety intervention evaluations can include material containing policies or procedures related to the intervention, safety records, committee minutes, correspondence, memoranda, or reports. They can suggest topics to include in interviews or

questionnaires and offer evidence of intervention implementation, barriers to implementation, or other events in the workplace that could threaten the evaluation's internal validity.

Be aware that documents are never more than a partial reflection of reality. Some are normative; such as procedures documents. They tell what should be done, not whether it actually is done. Some documents are descriptive - e.g., minutes of meetings. However, they can reflect one person's view (e.g., the minute-taker or chair of the meeting), more than the collective view.

7.3 Ways to use qualitative methods in effectiveness evaluation

Interviews, questionnaires, observations and documents are used alone or in combination towards several purposes in safety intervention evaluations. Here, we elaborate on five ways in which they can contribute to an effectiveness evaluation.

Ways to use qualitative methods in effectiveness evaluation

1. Identifying implementation and intermediate outcomes
2. Verifying and complementing quantitative outcome measures
3. Eliminating threats to internal validity
4. Identifying unintended outcomes
5. Developing quantitative outcome measures

7.3.1 Identifying implementation and intermediate outcomes

Qualitative data can help elucidate the steps between the intervention and the final outcome, including implementation and intermediate outcomes. They can identify results not captured in the quantitative measures. This can be an important addition to an evaluation, since it is not usually possible to quantitatively measure every pertinent intermediate effect of the intervention. It can be difficult to anticipate them all and measure them quantitatively. You especially want to find out the extent to which the intervention was implemented as planned. Document analysis, observations and interviews can be used to check on program activities.

Example of how qualitative methods can be used to identify intermediate outcomes

Let us return to an earlier example[34] where an intervention consisted of a workplace- based incident investigation team assisting supervisors in their investigation of incidents. Quantitative data included final outcome measures (frequency and severity of injuries) and some intermediate outcome measures (length of time between incident and its report and percentage of incidents generating corrective action proposals). Interviews helped fill in the picture further of how the intervention could have led to the observed decrease in injuries and their severity. The interviews revealed that supervisors and safety representatives found the incident investigation teams helpful and felt that better corrective actions were conceived. Thus, better quality corrective actions - an intermediate outcome - has been identified as a plausible means by which the frequency and severity of injuries were decreased.

7.3.2 Verifying and complementing quantitative outcome measures

Qualitative measures are used to verify quantitative measures. Through an approach of "triangulation", two or more different methodological approaches can measure the same thing in order to establish consistency. You

[34] Menckel and Carter [1985]

might undertake a broad-based safety initiative to change the "safety climate" in the workplace. Certainly, you could use a safety climate questionnaire, which typically consists of close-ended questionnaire items, to assess a change in safety climate. Also valuable are open-ended questionnaire items or interviews completed by key personnel regarding observed changes in the workplace atmosphere concerning safety. If the methods are consistent in their portrayal of change in safety climate, then a "cross-validation" of the methods has been achieved and you can present your conclusions with more confidence.

Sometimes the methods are complementary in that they might measure different aspects of the same concept. Open-ended questions or interviews might detect aspects of change missed by a questionnaire containing only close-ended items.

7.3.3 Eliminating threats to internal validity

Interviews with key officials can provide information crucial for addressing potential threats to internal validity.

Example of how qualitative information helps reduce threats to internal validity

In the evaluation example just discussed on the previous page, interviews and analysis of safety committee minutes revealed the following information which helped eliminate threats to internal validity. The workplace physical plan, products, production techniques and activities, as well as the safety-related policies, purchases and activities (apart from the creation of the incident investigation committee) had remained constant over the six-year evaluation period - suggesting no evidence of history threats. There was also no evidence for an instrumentation or reporting threat, since there were no changes in the incident reporting criteria, nor in safety-related policies, purchases and activities (apart from the creation of the committee).

7.3.4 Identifying unintended outcomes

Interviews and, possibly, observations are useful ways to identify unintended outcomes. Although some unintended outcomes can be assessed quantitatively, such as an increase in an untargeted type of injury, others would be better detected through qualitative inquiry.

Interviews are an especially good at gauging the reactions of intervention participants and others involved in the intervention, including supervisors, union leaders and managers. Their reactions and those of others involved with the intervention, are important, since a poor response by an influential individual or group of individuals at a work-site could have a big effect on the program. It might explain the lack of success of a promising intervention. Unintended outcomes can also be more positive. In one evaluation, for example, interviews with workers and foremen showed that several people believed that the recent decrease in the number of labor grievances could be attributed to the improved industrial relations resulting from the participatory ergonomic program.

7.3.5 Developing quantitative measures

Data collected using qualitative methods in the planning stage of the evaluation can provide the basis for the subsequent development of relevant quantitative measurement instruments. Here are three examples.

Examples of how qualitative studies can help develop quantitative instruments

1) Interviews, observations and document analysis can lead to the development and inclusion of certain items on questionnaires. For example, say that opinions expressed in interviews had a repeating theme that safety is for sissies. If your intervention is in part designed to change this attitude, then it would be a good idea to develop a questionnaire that includes questions which measure such safety attitudes.

2) People have used the details of incident records, a qualitative information source, to develop workplace-specific checklists of work practices or work-site conditions used in quantitative assessment. They review records to find which unsafe practices and conditions are associated with incidents. Interventions are then developed which encourage the practice of safer alternatives. As well, checklists of these safe practices and work-site conditions are developed and used in evaluation. Quantitative measurement consists of making (random) observations and recording whether the safe or unsafe version of the practice or work-site condition was observed.

3) Menckel and Carter[35] described a new safety initiative in which a group assisted workplace supervisors in their investigation of incidents within their division. Preliminary interviews and document analysis showed that there was often a long delay between incident occurrence and its formal reporting. As a result, corrective safety measures were correspondingly delayed in their implementation. Thus, one of the ways evaluators chose to measure the effect of a new workplace incident investigation group was by how long it took for incidents to be reported.

7.4 Selecting a sample for qualitative purpose

Once you have decided to use qualitative data collection methods as part of the program evaluation, you need to decide from whom, or about what, data should be collected. This might include collecting data from specific employee work groups, manager groups, female or male workers, or different classifications of workers. Additionally, you might want to collect data about a particular event, occurrence, or incident.

Rather than surveying the entire work force, use what is called *purposeful sampling*. Here, one selects information-rich cases to study in-depth. They are purposefully selected so that the investigator can learn, in detail, about issues of central importance to the program. For example, you might want to ask different employee workgroups about their experience in a particular occupational safety program. Then compare quotes across groups to see if there are differences in experiences which might influence the intended goals of the program. Furthermore, you might separately ask male and female workers about any problems in participating in the program. Again, comparisons can be made to see if both females and males similarly received the program as intended.

We describe eight different purposeful sampling strategies that may be used.

Extreme or deviant case sampling

Identify unusual or special cases. It is possible that much can be learned from extreme conditions (good or bad) rather than the many possibilities which fall in the middle. For example, survey data collected after a safety program is over might show one or two people

[35] Menckel and Carter [1985]

who have made big changes. A follow-up with an interview could validate the responses as well as discover what in the program motivated them to make such big changes. By limiting the focus to extreme cases, this approach to sampling is economical in time and resources.

Heterogeneity sampling/maximum variation sampling

Identify cases with differing characteristics (e.g., age, gender, education, job classification) to provide diverse points of view. Any common patterns emerging from the variant cases can capture the core experiences and shared aspects of a program or event.

Homogenous sampling

Identify a small group of information-rich cases - similar in terms of background, employment level, experiences, etc. and explore the issue of interest in depth. It might be of interest to separate groups of management and then labor and compare their opinions about a safety program.

Typical case sampling

Identify "typical" individuals to describe the benefits of the program. Cases are selected with the co-operation of key informants such as program staff. This information can be used to help "sell" the program to others reluctant to participate.

Critical case sampling

Find individuals who could dramatically make a point about the program. They may be identified by asking a number of people involved with the program. A good bet are the leaders in the group who could provide suggestions about how to improve the program.

Criterion sampling

Identify and study cases that meet some predetermined important criterion. Even if all employees at the work-site receive the training, you might interview only those most exposed to the particular hazard targeted by the training. They may reveal major system weaknesses that could be targeted for improvement.

Politically important case sampling

Identify, and select (or not) politically sensitive individuals. You might want to interview a union steward who supports the program, and thereby can enrich the information obtained.

Convenience sampling

The most common method in selecting participants for qualitative data collection lies with picking cases that are easiest to obtain and those most likely to participate. This is also the least desirable method. The problem is that in the end, it is difficult to know exactly who was interviewed and if their opinions are consistent with others possibly affected by the program.

7.5 Qualitative data management and analysis

A variety of methods are used to analyze qualitative data. The process is described here in very general terms and appears as a sequence of steps, which in actual practice can occur simultaneously or may even be repeated. First, all raw information, if not already in a written form, is converted to text. Thus, taped interviews are transcribed and visual material is summarized using words, etc. This body of textual material is reviewed to identify important features and, possibly, summarize them. A coding system of keywords, or some other data reduction technique, is developed to facilitate this process. The data, either in summarized form or not, is then reviewed to identify patterns. These

patterns are concerned with the following: similarities or differences among groups or events; repeated themes; and relationships among people, things or events.

Identification of patterns leads to some generalizations or tentative conclusions regarding the data. Depending on the scope of the investigation, you might examine the trustworthiness of these generalizations by testing them with the results of further data collection or comparing them with existing theory.

Success at the data analysis stage requires that good data management practices are observed from the beginning of data collection. Use systematic methods for collecting, storing, retrieving and analyzing data. People have developed various techniques to help highlight, organize or summarize data. A useful reference in this regard is the book by [Miles and Huberman 1994]. This reference also reviews the various software developed to assist in both the data reduction and pattern recognition stages of analysis.

7.6 Ensuring good quality data

Concerns about reliability and validity apply to qualitative data, just as they do to quantitative data. Thus, anyone reading a report of a qualitative investigation wants to know that the stated methods have been used consistently throughout the study (reliability concerns). They also want to know that there are no hidden biases in the data collection, the data analysis nor the conclusions drawn (validity concerns).

The following contains considerations and suggestions for ensuring that good quality data is collected.

Minimizing evaluator bias

The product of a study no doubt bears the personal mark of the people conducting it. However, researchers generally try to reduce their effect on their research by using concepts and methods agreed upon by other researchers. Ways to guard against bias include the following: outlining explicit methods for data collection and data analyses; adhering to these methods; having more than one researcher collect data; having a second, non-biased person summarize and/or draw conclusions from the data; and letting the data speak for themselves and not forcing them into a framework designed by the researcher.

Appropriate sampling

Someone reading your evaluation wants to be sure that the right sample has been selected for the stated purpose. For example, you could not claim to be truly representing workplace perceptions of the effectiveness of an intervention, if either management or employee representatives are not represented. Thus, the rationale and method of sampling must be explicit and justified with respect to the study's aims.

Validation by subjects

One of the best ways to determine whether or not you "got it right" in your study, is to check with the subjects you are studying. This involves confirming the accuracy of the data collected, the reasonableness of the method used to summarize it, and the soundness of the conclusions. Of course the potential biases of the subjects consulted must be kept in mind when weighing their opinions.

Thorough methods of drawing conclusions

Avoid drawing conclusions too soon. This can be caused by researcher bias or pressure to come up with answers quickly. In contrast, well-grounded conclusions require time for at least some of the following activities: 1) reviewing collected data to identify anything which has been overlooked; 2) searching for evidence which

contradicts preliminary conclusions, either by reviewing data already collected or by gathering new data; 3) confirming important data or conclusions through "triangulation", i.e., finding agreement when using a different data source, methodology or researcher; and 4) exploring alternative explanations for patterns observed in the data.

Conduct a pilot study

Conducting a pilot study or trial run with your proposed research methods is often of great value. Feedback from those involved in the pilot study can be used to refine a sampling strategy, interview guide, other data collection procedures, and even procedures for data management.

7.7 Summary

We have reviewed four major methods for gathering qualitative information: interviews; questionnaires with open-ended questions; observations; and document analysis. Qualitative data can be used in several ways to complement quantitative methods: identifying implementation and intermediate outcomes; verifying and complementing quantitative outcomes; eliminating threats to internal validity; identifying unintended outcomes; and developing quantitative measures. In contrast to quantitative methodology, qualitative methods usually employ one of several purposeful sampling strategies. We briefly discussed methods of analysis and methods to ensure good quality data.

Key points from Chapter 7

- Use interviews and focus groups, questionnaires with open-ended questions, observations, and document analysis to enrich your evaluation.
- Use qualitative methods for one or more purposes:
    - identify implementation and intermediate outcomes
    - verify and complement quantitative measures
    - eliminate threats to internal validity
    - identify unintended outcomes
    - develop better quantitative measures.
- Use one of several purposeful sampling strategies.
- Collect and analyze data in ways which enhance their reliability and validity.

Chapter 8

Statistical Issues: Are the results significant?

8.1 Introduction

This chapter will not answer all your questions on statistics. It cannot cover all the possible approaches that intervention evaluations might require. Statisticians will maintain they should be consulted right from the design stage of a project. This is not just self-interest. Certainly, they want your business - including the consultation fee! But even the statistician among us (HSS), who typically provides free consultations for faculty colleagues, makes the same point: there are often aspects of the way the evaluation is designed, or the type of data collected among other factors, that mean the statistical analysis will not be entirely straightforward. You can avoid ending up with "messy data" by discussing the study in advance. The data could still turn out to be complicated and it may be best to find a statistician to do the analysis. After all, you have probably spent a lot of resources on ensuring the highest quality of intervention and data collection; so you do not want a second-rate evaluation.

This chapter provides an overview of the statistical concepts applicable to intervention evaluation. We start by explaining the need for statistical methods, followed by a discussion of the meaning of *p-values*. If you have ever read a scientific paper with quantitative information, then you have seen p-values. They show an equation like "$p<0.05$" along with a comment stating if the result is "statistically significant". Another way of presenting statistical results are *confidence intervals*. Also being introduced is the notion of *statistical power* and how that relates to the *sample size*.

Later in the chapter, we discuss what issues to consider in choosing a statistical technique. Two have already been mentioned - the type of data you have and the study design being used. No calculations are shown in this chapter, but some simple examples are included in Appendix B. They correspond to the evaluation designs outlined in Chapters 3 and 4.

8.2 Why statistical analysis is necessary

Surely, if the change in injury rate in an intervention group is greater than the change in injury rate in the comparison group, doesn't that prove that the intervention has worked? Why are statistics needed? The answer is that real-life data are subject to random variability. Suppose you have a perfectly balanced ("fair") coin. (Statisticians love using examples about coin-tossing to explain the principles they use.) Toss it ten times, and you can expect five heads and five tails. But it is also reasonably likely that you could get six tails and four heads, or four tails and six heads simply as a result of chance (random variability). You might also get a 7-3 split. As will be seen in the next section, the question from a statistical viewpoint becomes: how far from a 5-5 split do I have to get, before I become suspicious about the coin and question if it really is fair?

The analogy with the safety situation can be seen if we think about a study investigating whether back belts reduce back pain. You randomize half the people to receive back belts (the intervention group), while the other half (the control group) is left alone. After the intervention group has had back belts for a while, everyone is asked about levels of back pain. [This is an example of an experimental design with "after"-only measurements, Section 4.3.2] For each person, there is a pain score.

The average score in the group given back belts may be somewhat better than the average in the control group. Does this mean that back belts work - or is it that simply by chance, the randomization led to more people who have a lot of back pain ending up in the control group? And what if it is the opposite? Suppose, the back belt group does a little bit worse than the control group? Does that mean that these belts are actually harmful - or is it because there happened to be more people who get pain in the intervention group? Statistical analyses can

indicate how likely (probable) these possibilities are.

Statistical techniques can also address a selection threat to internal validity, i.e., when differences in participants' characteristics between the intervention and control group could be responsible for the measured intervention effect (Section 4.4.1). You notice, for instance, that people given back belts are on average somewhat older than those who do not get them. Furthermore, because older workers are less likely to use the belts, and the intervention group itself is older, you might not see much change in comparison with the control group. To reduce this type of threat, statisticians have developed techniques that *account* for or *control* the difference in ages (or other variables) between the two groups.

Caution: Having emphasized the vital importance of statistical analysis, we warn you about indiscriminately using statistical packages. Inexperienced researchers are sometimes tempted to feed their data into a standard software package and ask for everything to be compared with everything else. This is not good practice. Statistical testing should follow a clear definition of hypotheses; and the testing should not determine the theory. Instead, the hypotheses should come from the chosen intervention models.

8.3 P-values and statistical significance

A leading philosopher of science in the 20th Century, Sir Karl Popper, argued that science proceeds by the method of *refutation*. At any time, scientists have only the best theory (or hypothesis) at the moment to describe how the real world works. This hypothesis can always change based on new observations. What scientists must do, argued Popper, is to devise experiments aimed at disproving (refuting) their latest theory. As long as the experiments fail to do so we continue to regard the theory as at least a reasonable approximation to the real world. If an experiment disproves the theory, then a new one must be adopted. Classical statistical reasoning works in a similar fashion, basing the rejection of an initial hypothesis on probabilistic grounds.

With our example of back belts: start from a position of skepticism about their value, i.e., hypothesize that the intervention has no effect (*null hypothesis*). If the program is useless, then expect no difference in the back pain between the intervention and control groups. However, in

What do we mean by a "true"effect?

Several times in this chapter we refer to a "true" or "real effect" of a program or intervention. Surely, you may think, the true effect is what we find. How could the effect be considered not true?

Part of the answer is that the estimate of the effect is subject to variability. The groups studied may be comparable, but cannot be identical even if they are created through randomization. So if you repeat your evaluation study elsewhere the size of effect might be larger or smaller.

If you repeat the study many times, you could take the average (mean) of the effect sizes. This would balance out the times when the effect just happens to be larger with those when it happens to be smaller. In practice, studies are rarely repeated even once, let alone many times. But you can do this as "thought experiment". In fact, statistically we think about a (hypothetical) infinite number of replications. If we could actually do them, the average effect size over the replications would be what we have called the *true effect*.

practice, the intervention group sometimes does better and sometimes worse than the control group, i.e., the average scores show less pain or more pain. The question then becomes: how big a difference must exist between the groups before you start to doubt the null hypothesis and accept that the change comes from a real effect of the program?

Typically, the statistical approach takes into account the difference in the mean (average) levels of pain between the two groups, as well as how much variability exists between the scores of the people in the study and the *sample size* (how many people are included in the study). The analysis produces a *p-value*, which can be interpreted as:

> ***The probability that a difference at least as large as the one seen could have occurred simply by chance, if there really is no effect of the intervention.***

When this probability is small enough, you reject the hypothesis of no difference and start to believe (at least for practical purposes) that the back belts have changed the level of pain. When the p-value is larger, you cannot reject the hypothesis, so you would conclude the belts do not "work", at least in the way they were used. The cut point for the p-value, which represents the probability you are willing to allow of concluding the intervention works - or does harm - when it is really ineffective,[36] is known as α (the Greek letter *alpha*). When the p-value is less than this, the result is declared *statistically significant*.

How small does the probability have to be for you to reject the hypothesis and claim that the intervention works? Strictly speaking, there is no right or wrong answer here. It depends on your willingness to draw the incorrect conclusion that the belts work when they really do not. This in turn depends on the resource and policy implications of the evaluation results. The more expensive the intervention, the less likely you want to make this type of mistake, so you want the probability to be very low; and vice versa - if the intervention is cheap and aimed at very severe injuries, you may be willing to apply it even if the evidence of its value is less strong. In practice, though, alpha is usually taken as 0.05 (or 5%).

> **Important note:** In the interpretation of the p-value, the phrase "if there really is no effect of the program" is crucial. Ignoring it has led to many a misinterpretation of p-values. We now discuss the situation where there *is* an effect.

8.4 Statistical power and sample size

Now there is another side to all this. We have discussed the p-value based on the condition that the program is useless. But what if it works? If the intervention is truly effective, you want to be reasonably sure to reject the initial null hypothesis. Just as you can get six heads and four tails even with a fair coin, you could also get five heads and five tails even with a coin biased toward heads. Similarly, even if the program truly has a moderate effect, you might be unlucky in your study and only observe a very small difference, which may not be statistically significant. If you fail to reject (i.e., you accept) your initial hypothesis of no difference, and there really is one, the mistake is known as a *Type II error*. The probability of such a mistake, i.e., the probability that you fail to reject the hypothesis when it is false, is known as β (the Greek letter beta).

This means that the probability that you correctly reject a false hypothesis, i.e. you detect the program's effectiveness, is $1-\beta$, and this value is known as the *power* of the study. The importance of this is that you obviously do not want to do a

[36] This type of mistake is known as a *Type I error*

study that has little chance of demonstrating a real effect of an intervention. You want to be reasonably sure that you will conclude there is a difference if one really exists. *Thus, it is important to consider power before undertaking the intervention and evaluation.*

You can do this in two ways. You could set the power you want and then calculate the sample size needed; or you may have a certain number of workers who could be involved in a study, and you can estimate the power the study would have with that number.

The first approach is actually a preferable way to conduct an evaluation - indeed, clinical trials of new drugs do it this way round. Typically, researchers design evaluations so that the power is 80% (sometimes 90%); that is, if the intervention is truly effective, there is an 80% (90%) chance that the data you gather and the statistical test you use allows you to conclude that the intervention is effective.

In practice, workplace interventions usually involve a fixed number of employees, for example, all workers in a plant or a department. So you can't set power in advance - rather, you should check what power you will have. Several components go into the calculation of power: the *effect size* - how much effect you think the intervention will have (or should have in order to be worth replicating elsewhere); *sample size* (the number of evaluation participants or, more formally, experimental units); how much variability there is between the outcome measurements within the sample; and the values you set for α and β. The formula you use to calculate power, like your choice of statistical test, depends on the experimental design and type of data collected.

All other things being equal, the larger the sample size, the larger is the power. Similarly, the less variation in the outcome measure, the larger the power. If you should find that the intended plan would likely yield power much lower than 80-90%, you might want to change your evaluation design, choice of outcome measures, or number of people included in the evaluation. Cohen (1988) shows how to do power calculations and you can also use the statistical packages mentioned in Appendix B.

8.5 Confidence intervals

Another way of showing the degree of uncertainty in an estimate of the effect of an intervention is through *confidence intervals*. Suppose, that you do a study where people taking a program improve their knowledge score by an average of five points. It could be that the program is actually useless but, just by chance, this apparent benefit is found. Alternatively, the program may really be worth more than a five-point improvement; but by chance you happen to underestimate the real benefit. In other words, the real benefit of the program might be higher or lower than the observed value of five points. An obvious question is: how much higher or lower? You can often construct a confidence interval (as illustrated in Appendix B), within which you are reasonably sure the "true" level of benefit lies. (The interval is calculated, taking into account how much variability exists in individual knowledge scores.) Typically, you see 95% confidence intervals, which means that you are 95% sure (i.e., there is a probability of 95%) that the size of the true effect is within the interval.[37]

In many ways, the confidence interval is more useful than a p-value, which simply indicates whether a difference between groups is or is not statistically significant. With a confidence interval, you get a sense of just how high or low the benefit might reasonably be.

The narrower the interval the better, since the range of plausible values is smaller. Suppose, the confidence interval shows that the observed

[37] 95% confidence intervals are based on $\alpha = 0.05$; 99% confidence intervals are based on $\alpha = 0.01$; etc.

value of the average benefit of the program (five points) is from a one point benefit to a nine point benefit. (Confidence intervals can be, but are not always, symmetrical about the estimate of effect.) Although your best estimate of the program effect is still five points, it is also quite plausible that the benefit could be as low as one point - which we might consider trivial - or as high as nine points, a very useful improvement. Thus, we would be quite uncertain about the value of the program. A smaller interval, of between four and six points, would be preferable. All other things being equal, a narrower interval can be obtained with a larger sample size.

As a general rule, if a 95% confidence interval excludes the value of zero, you will know that if you tested the null hypothesis (i.e., no (zero) difference between the values being compared), you would be able to reject the hypothesis, using alpha = 0.05.

8.6 Choosing the type of statistical analysis

Up to this point we have simply referred to the interpretation of the p-value and confidence interval that result from a statistical analysis. Before you get there, you need to think about what type of statistical analysis to use. There are a number of issues to consider. Some are discussed in detail in this section; some later in the chapter.

Things to consider when choosing the type of statistical analysis

- What type of data is it - categorical or continuous?
- What type of evaluation design is used?
- What is the unit of analysis?
- What is the study sample size?
- Is correction for different group characteristics needed to avoid selection effects?

8.6.1 Type of data

The type of data being analyzed is important. If you determine whether or not someone had any injuries, you have what is called categorical or discrete data; and if you use the actual number of injuries occurring to each person or the counts for an entire workplace you still have categorical data. Continuous data can take essentially any value in a range (at least in principle). Age is a continuous variable, although in adults we usually simply round off the number of years and say that someone is 42, rather than 42.752 years old. In practice, the boundary between categorical and continuous variables can be a little fuzzy. Thus a measure of behaviors, which might take an integer value from 0 to 100, is typically considered to be continuous. Some statistical methods are *robust* which means that taking such variables as continuous is acceptable. Any variable that takes at least ten values can reasonably be considered continuous. Another situation in which the boundary between categorical and continuous is fuzzy is when you analyze injury rates. Although the rates are continuous data, statistical tests intended for categorical data are sometimes used. This is because the analysis in such cases uses the actual numbers of injuries, which are categorical variables.

8.6.2 Evaluation design

The choice of statistical analysis must also take account of the evaluation design. A simple comparison would be between the post-intervention scores of those who have experienced an intervention compared with those who have not. Similarly, scores before and after an intervention can be compared. The simplicity of these two types of studies makes them useful for distinguishing two situations which require different analyses. In the first case, the scores in the two groups all come from different individuals. In the second case, in contrast, each individual received two scores - before and after the intervention. This is an example of a *repeated measures design*. The two scores tend to be related, since those who scoring higher before the intervention will likely score high (or at least relatively high) after the intervention.

In analyzing repeated measures we can take advantage of this relationship. The techniques usually reduce the "noise" in the data, because they remove much of the variability between people. This allows the "signal", the differences over time within each person's scores, to become clearer.

Table 8.1 is a guide to statistical tests appropriate for the designs that we have discussed. As indicated, illustration of some of these tests can be found in Appendix B.

Table 8.1: Choice of statistical test based on evaluation design and type of data

Type of design	Type of outcome data	Statistical test	Section number
Before-and-after	Rate	Chi-squared test for comparing rates	B.1.1
	Continuous	Paired t-test	B.1.2
Pre-post with randomized or non-randomized control group	Rate	z-test for comparing rate ratios or rate differences	B.2.1
	Continuous	Two sample t-test (groups similar) or multiple regression (groups different)	B.2.2
Experimental designs with "after"-only measurements	Rate	Chi-squared tests for comparing rates	B.3.1-B.3.2
	Continuous	Two sample t-test (two similar groups), ANOVA (two or more similar groups), or ANCOVA (two or more different groups)	B.3.3-B.3.4
Simple or multiple time series; multiple baseline design across groups	Categorical, rate or continuous	Time series analysis techniques (e.g., ARIMA)	B.4

8.6.3 Unit of analysis

You also need to think about the *unit of analysis.* If an intervention is targeted at changing a workplace, rather than individuals, then each workplace might count as just one unit. Another possibility is that you conduct a randomized study aimed at changing individual behavior, but you randomize work groups, rather than individuals. For example, if you have a company with a number of small but geographically separated units, you might randomize work groups to receive or not receive the intervention. Then in any given work group either everyone gets the intervention or no one does. This sampling of "clusters" rather than individuals must be taken into account in the analysis. In essence, the issue is that individuals may display similar behavior within a given unit (i.e., a group effect). The greater this tendency, the lower the effective sample size. The concern about "cluster sampling" is real and relatively common - but often it is not accounted for in the analysis.

8.7 Avoiding pitfalls in data analysis

Data exploration

It is good practice before jumping into a (relatively) sophisticated analysis to look at the data in a fairly simple fashion. You can look at the means of groups, the proportion of subjects having injuries, frequency distributions or the range of values observed. Graphs or diagrams can often be helpful. These approaches give you a "feel" for the data, as well as help find errors arising from the data collection or processing. Failure to do these things may lead to the application of a technique inappropriate for the data, or even worse, analyzing incorrect data.

Changes to the designs

It is often tempting, for practical purposes, to "tweak" the intervention or evaluation designs, to make what at face value might seem to be minor changes. The result will generally have important implications for the type of analyses done. If you had originally planned the study with a particular type of statistical approach, you may not be able to use it. This is not to suggest that you should always be rigid in following a pre-planned design, but rather that you should make changes with caution.

Other pitfalls

We have already mentioned a few pitfalls and how to deal with them: choosing the right unit of analysis - especially when we engage in "cluster" sampling; ensuring studies are large enough to have adequate power; and ensuring we do not simply press buttons on our computer to produce an answer based on an incorrect analysis.

8.8 Summary

In this chapter, you have seen some basic concepts in statistical inference, including p-values, statistical power and confidence intervals. We pointed out some of the things to consider when undertaking an analysis: type of data; evaluation design; unit of analysis; sample size; and correction for group characteristics. Some examples of analyses, corresponding to the designs discussed in Chapters 3 and 4, can be found in Appendix B.

Key points from Chapter 8

- Discuss the analysis with a friendly statistician while designing the evaluation - do not wait until after you have collected the data.
- Check statistical power while designing the evaluation.
- Do an initial data exploration to get a "feel" for the data.
- Choose the type of statistical test according to the type of evaluation design and the type of data.
- If in doubt, discuss with a friendly statistician.

Chapter 9

Summary of recommended practices

9.1 Introduction

We have discussed the various methods of effectiveness evaluation in the context of evaluating safety interventions. The following section gives an overview of some of the key messages from the previous chapters. You likely will not be able to follow all of the recommended practices. As a whole, they represent an ideal. Even if you are not able to follow all of the practices outlined in this guide, it does not mean you should not proceed with your chosen intervention and some level of its evaluation.

You will no doubt need to summarize and report on the results of your evaluation. Some guidance on this aspect has been included in Appendix C.

9.2 Summary of recommended practice

Planning and development

- Identify resources available and information needs of the end-users of the evaluation results
- Involve all parties relevant to the intervention and evaluation in the planning stage, as well as at subsequent stages of the evaluation
- Seek external expertise on evaluation design, methodology, and analysis, if necessary
- Review relevant theory, research literature, methodology, historical data
- Develop a conceptual model and / or program logic model
- Keep an intervention diary

Method development

- Determine reliability and validity of measurement methods if not already known; pilot test when necessary
- Use qualitative methods to inform the use and design of quantitative methods
- Pilot test any qualitative or quantitative methods that are new
- Estimate statistical power based on planned methods - if insufficient choose new study sample size, evaluation design or measurement method

Study sample

- Choose a study sample representative of the target population
- Use random sampling methods in selecting a sample from a sampling frame
- Choose a sample size large enough to give sufficient statistical power
- Consider using randomized block or matching designs to avoid selection effects
- In quasi-experimental designs (non-randomized), choose intervention and control groups so that they are very similar
- In experimental designs, use randomization to assign participants to intervention and control groups
- In qualitative studies, choose a purposeful sampling strategy suitable for the evaluation purpose and intervention circumstances

Evaluation design

- If you have no choice but to use a before-and-after design, for reasons of feasibility or ethics, try to eliminate the threats to internal validity
  - identify other changes in the workplace or community which could effect the outcome (history threat) and measure their effect
  - ensure that before and after measurements are carried out using the same methodology, to avoid instrumentation or reporting threats
  - avoid using high-injury rate groups as the intervention group in a before-and-after study, to avoid regression-to-the-mean threats
  - allow for the fact that taking a test can have an effect of its own (testing threat)
  - try to minimize Hawthorne threats by acclimatizing workplace parties to researchers before measuring the intervention's effect
  - identify any natural changes in the population over time which could obscure the effect of the intervention (maturation threat), and try to allow for them in the statistical analysis
  - investigate whether the intervention participants' dropping out could have and effect (dropout threat)
- Use a quasi-experimental design whenever possible instead of a before-and-after design by using one or more of the following strategies:
  - include a control group
  - take additional measurements both before and after the intervention
  - stagger the introduction of the intervention to different groups
  - add a reversal of the intervention
  - use multiple outcome measures
- Use an experimental design whenever possible instead of a quasi-experimental design by assigning participants to intervention and control groups through randomization
- When using control groups, check that intervention and control groups receive similar treatment throughout the evaluation period, apart from the intervention itself; avoid, but check for, diffusion, rivalry or resentment effects
- Plan a measurement timetable to capture maximum intervention effect and characterize longer term effects
- Collect additional data to address threats to internal validity not addressed in the primary experimental design
- Try to triangulate and complement data collection methods by using multiple methodologies, especially qualitative and quantitative

Measuring intervention implementation

- Use both qualitative and quantitative methods to assess
  - degree of intervention implementation
  - problems with intervention implementation

Measuring intervention outcomes

- Measure intermediate and final outcomes
- Use both quantitative and qualitative methods
- Select reliable and valid methods, appropriate for the study sample and intervention
- Consider injury statistics, other administrative records, behavioral / work-site observations, employee surveys, analytical equipment and workplace audits as means of measuring outcomes.
- Select quantitative methods which give sufficient statistical power during analysis
- Consider the ethical aspects of measurement methods

Measuring unintended outcomes

- Use both qualitative and quantitative methods to assess any unintended outcomes

Statistical analysis

- Decide on the statistical methodology *prior* to undertaking evaluation
- Calculate power before data gathering begins – modify the design or measurement methods if power is inadequate
- Use appropriate techniques for analysis based on type of data and experimental design

Interpretation

- Try to use the results of qualitative enquiry to enhance the understanding of the quantitative results
- Identify all likely alternative explanations for the observed results apart from the true effect of the intervention (i.e., threats to internal validity)
- Examine the feasibility of alternative explanations, using a quantitative approach whenever possible and collecting additional data if necessary

Conclusions

- Address evaluation questions in your conclusions

Glossary[1]

Alpha (α): in statistical analysis, the probability you are willing to allow of concluding that the intervention works, or does harm, when it is really ineffective

Before-and-after design: (syn. pre-post design) a research design where measurements are taken both before and after the introduction of an intervention to measure its effect; permits less confident causal inferences than a quasi-experimental or experimental design

Conceptual model: diagram which represents the causal relationships among important concepts relevant to an intervention

Confidence interval: interval surrounding a point estimate, where the true value of the estimated parameter is found with a probability of (1-α)

Confounding variable: variable which affects both the independent variable (presence of intervention or not) and the dependent variable of interest; it is not a mediating variable

Control group: group for which there is no intervention; group which is compared to the group undergoing the intervention and the difference in group outcomes attributed to the effect of the intervention; created through randomization in experimental designs; created using non-random means in quasi-experimental designs

Effect modifying variable: variable which modifies the size and direction of the causal relationship between two variables

Effectiveness evaluation: (syn. outcome evaluation; summative evaluation) evaluation which determines whether a safety initiative had the effect (e.g., decrease injuries) it was intended to have

Evaluation design: the general plan for taking measurements during an evaluation; i.e., from how many group(s) of workers / workers are measurements taken and when

Experimental design: a research design with both intervention and control groups created through a randomization process

History threat (to internal validity): when some other influential event happens during the intervention and evaluation period

Human sub-system (in the workplace): human knowledge, competencies, attitudes, perceptions, motivations, behaviors

Implementation: putting the intervention in place

[1] This glossary is not intended to be used as a general reference for evaluation terms. In many cases, terms have been expressed in the same context in which they are used in the guide. The definitions appearing here might therefore be more restrictive than they might be found in a more general reference.

Instrumentation threat (to internal validity): when the measurement method changes during the intervention and evaluation period

Inter-rater reliability: degree of agreement in scores between two different people rating the same phenomenon

Intervening outcomes: outcomes which result from the intervention but precede the final outcome; includes implementation, short-term outcomes, and intermediate outcomes

Intervention: see Safety intervention

Intervention group: group which undergoes the intervention; not a control group

Moderating variable: see effect modifying variable

P-value: in statistical analysis, the probability that a difference at least as large as the one seen could have occurred simply by chance, if there really is no effect of the intervention

Power: see Statistical power

Program logic model: diagram depicting the linkage of intervention components to implementation objectives to short-, intermediate- and long-term outcome objectives

Qualitative methods: research methodology which yields non-numerical data; includes interviews, document analysis, observations

Quantitative methods: research methodology which yields numerical data

Quasi-experimental design: research design which permits more confident causal inference than a before-and-after design; often includes a non-randomized control group

Random number tables: Tables consisting of randomly generated digits, 0 to 9, with each digit having a probability 1:10 of being selected. Used to select random samples or to randomize participants to intervention and control groups.

Random sampling: technique of selecting a study sample so that the choice is made randomly (using random number tables, etc.) and each participant has a known probability of being selected

Randomization: method of selecting participants for intervention and control groups such that the probability of being selected into one or the other groups is the same for all participants; method of forming intervention and control groups in experimental designs

Regression-to-the-mean threat (to internal validity): when a pre-intervention measurement of safety for a group is atypical and later measurements over the course of the intervention and evaluation are more similar to mean values

Reporting threat (to internal validity): when something changes the validity of (injury) reporting over the course of the intervention and evaluation

Reliability: degree to which the values measured for a certain concept are consistent

Safety intervention: any attempt to change how things are done in order to improve safety (e.g., engineering intervention, training program, administrative procedure)

Sample: see study sample

Sample size: the number of experimental units (people, workplaces, etc.) in the study sample

Sampling frame: the group within the target population from which you draw the study sample

Selection threat (to internal validity): when the apparent effect of the intervention could be due to differences in the participants' characteristics in the groups being compared

Selection interaction threats (to internal validity): when the apparent effect of the intervention could be due to something happening to only one of the groups being compared in a experimental or quasi-experimental design

Statistical power: Likelihood of detecting a meaningful effect if an intervention is truly effective

Study sample: participants selected to undergo either intervention or control conditions in a research design

Target population: larger group from which the study sample is selected; larger group to which evaluation results should be generalizable

Technical sub-system (in the workplace): the organization, design and environment of work, including hardware, software, job procedures, etc.

Testing threat (to internal validity): when the taking of the (pre-intervention) safety measurement for a group has an effect on the subsequent measurements of safety for the group

Threats to internal validity: possible alternative explanations for observed evaluation results; typically, experimental designs have less threats than quasi-experimental designs, which have less than a before-and-after design

Unintended outcomes: outcomes of the intervention besides the intended ones; can be desirable or undesirable

Validity: degree to which we measure the concept we intend to measure

Variable: any attribute, phenomenon or event that can have different quantitative values

Appendix A

Some models to assist in planning

Chapter 2 emphasized the importance of having an explicit conceptual model and/or program logic model related to the intervention. Since there are three levels of interventions (organization of safety management, technical sub-system and human sub-system), we present different models corresponding to each of these three levels, and which can either be applied or adapted for your own use.

A.1 A model for interventions in the technical sub-system

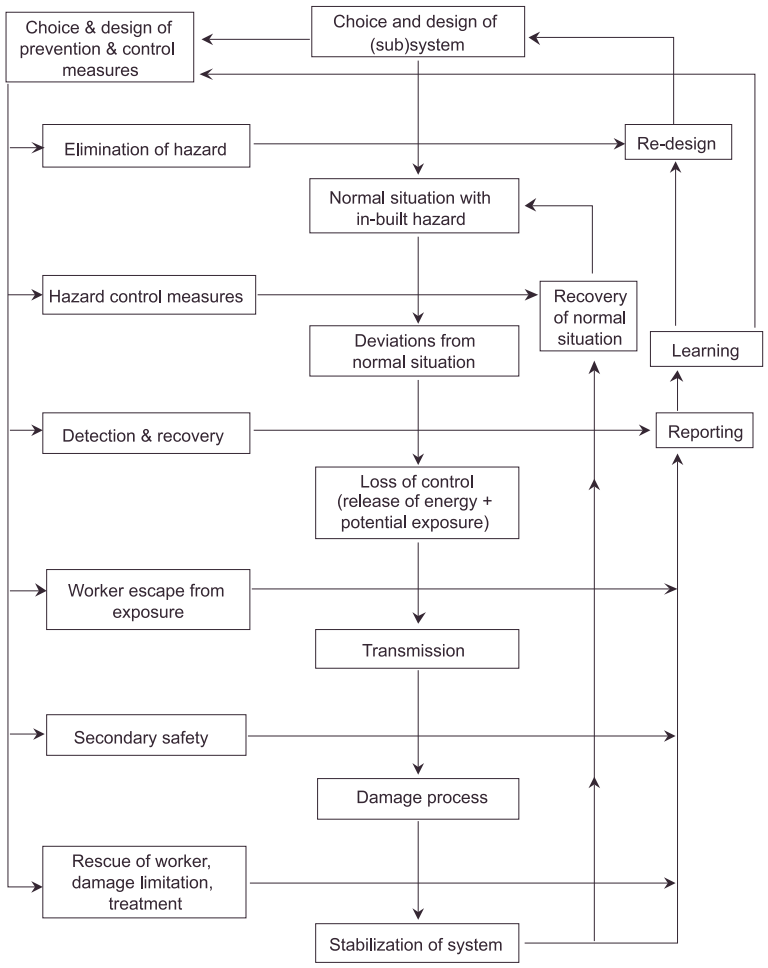
The first type of model is one most applicable to interventions in the technical sub- system; i.e., interventions concerned with the organization, design or environment of work, or with secondary safety or emergency measures. Here, the harm process is seen as a deviation, which, if unchecked, develops into an exposure to dangerous energies because of a loss of control of the work process (see Figure A.1). Another model of this nature is by Kjellén [1984].

As an illustration of how models can assist in designing an evaluation, consider the application of the first one to an engineering intervention where one style of machine guarding is replaced by a new one. Machine guarding is a specific example of a "hazard control measure" depicted in the model. (It would also be considered an independent variable in the context of the proposed intervention.) A typical evaluation measures the frequency of damage processes or injuries. (Injuries are a dependent variable and a final outcome measure). By referring to the model, one can identify intermediate variables to measure (i.e., deviations from normal situation) which here could be the percentage of times during a given observation schedule the guard

was not in place. The distinction in the model between "damage process" and "reporting" reminds us that injuries are not necessarily reported, and so, any changes in other variables affecting reporting should be accounted for during the evaluation.

In general, when dealing with interventions in the technical-subsystem, one also needs to think about the possibility of compensatory behavior on the part of managers or workers and, if necessary, account for this in the evaluation. For example, interventions to reduce the levels of noise emission from machines could be evaluated by measuring the actual noise emissions. However, there may be no effect of this intervention on audiometric measures of hearing loss if one result of quieter machines is that workers use less hearing protection. Ideally, one wants to include measures of noise emission, protection equipment use and audiometric measures in the evaluation.

Figure A.1: Deviation model[2]



[2] Adapted from Hale and Glendon [1987]. Horizontal arrows from prevention or recovery activities indicate at which point they can curtail or stop the harm development processes, which are represented by the vertical arrows running from the top to the bottom in the figure. If the prevention/recovery activities are successful, movement down a vertical line shifts with a right angle to a horizontal prevention/recovery pathway; if not, movement continues in a vertical direction.

A.2 Models for interventions in the human sub-system

When interventions are planned that do not directly intervene in the work process, but are designed to modify human knowledge, competence, attitudes or behavior, it is valuable to use a specific model to help guide the research. Even when an intervention in the technical sub-system is planned, behavioral models could be relevant for the steps in the deviation which involve human activity. We present two behavioral models. One is concerned with errors made without the intention to take risk; the other is concerned with intentional risk taking.

Model relevant to unintentional risk-taking

Figure A.2[3] shows a model of three types of error mechanisms that can occur without the intention of taking risk. "Slips" (of action) and "lapses" (of memory) can happen in the absence of a problem in the environment and result in the failure to execute a plan as intended. When operators realize that a problem exists, theory suggests that people will most likely search for familiar patterns and try to apply a known problem-handling rule. Only if the "rule-based level" approach fails, do people then resort to a "knowledge-based level" approach. This involves a wider range of cognitive strategies. In this model, errors arising at the rule-based and knowledge-based levels are called "mistakes". They result from carrying out an inadequate plan in the face of a problem. A model from Hale and Glendon [1987] develops the problem-solving aspects of Reason's model into a number of specific steps which can also be useful for evaluation planning. Both models indicate which intervening variables are possibly relevant to interventions aimed at modifying normal working behavior.

Model relevant to intentional risk-taking

The other type of model relevant to the human sub-system is one concerned with intentional risk-taking. The Health Belief Model[4] (Figure A.3) has been frequently used in the health promotion field in relation to health-related behaviors, such as smoking. However, the underlying theory is likely relevant to decisions and behaviors in safety and occupational health contexts. It can be applied to observing safety rules, using personal protective equipment and dismantling safety guards, etc. by workers; and, with modification, to observing safety rules and regulations by managers designing, planning and monitoring work processes.

The model shows that the likelihood of undertaking the recommended health (or safety) action depends on the individual's perceptions of their susceptibility to disease/injury, its seriousness, the benefits of taking the preventive action and the barriers to taking such action. Benefits are things like saving time and effort, and approval from a production-oriented supervisor, etc. Barriers are things like inconvenience, lack of knowledge or skill in undertaking the new behavior, fear of countering local norms, etc. All of these categories provide ideas about the intervening attitude measures that can be taken in evaluations of behavioral interventions.

[3] Figure from Reason [1990] is reprinted with the permission of Cambridge University Press.

[4] Becker MH, Haefner KP, Kasl SV, Kirscht JP, Maiman LA, Rosenstock IM [1977]. Selected psychosocial models and correlates of individual health-related behaviors. Med Care 15:27-46. With permission of Lippincott Williams & Wilkins.

Figure A.2: Generic error-modeling system [Reason 1990]

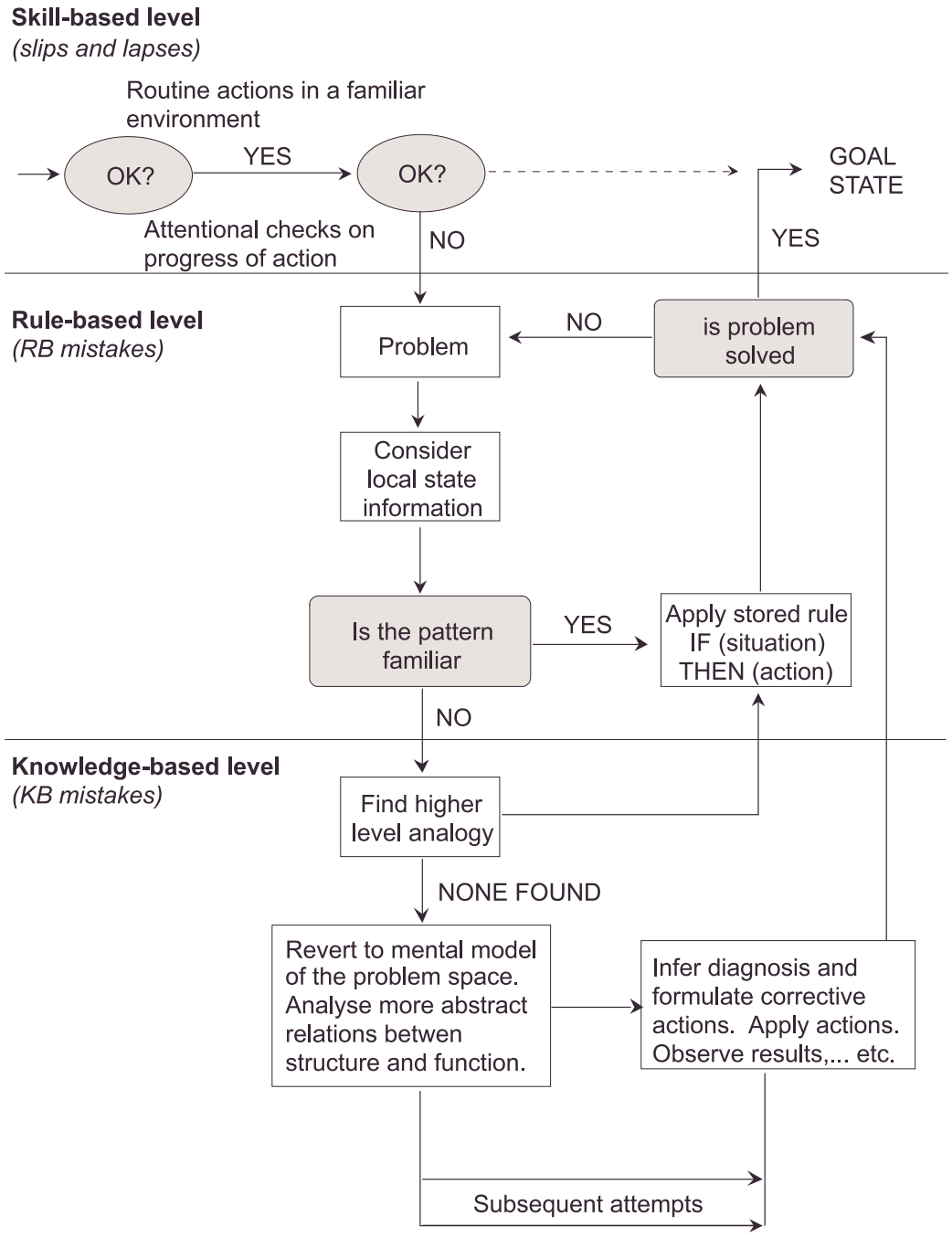
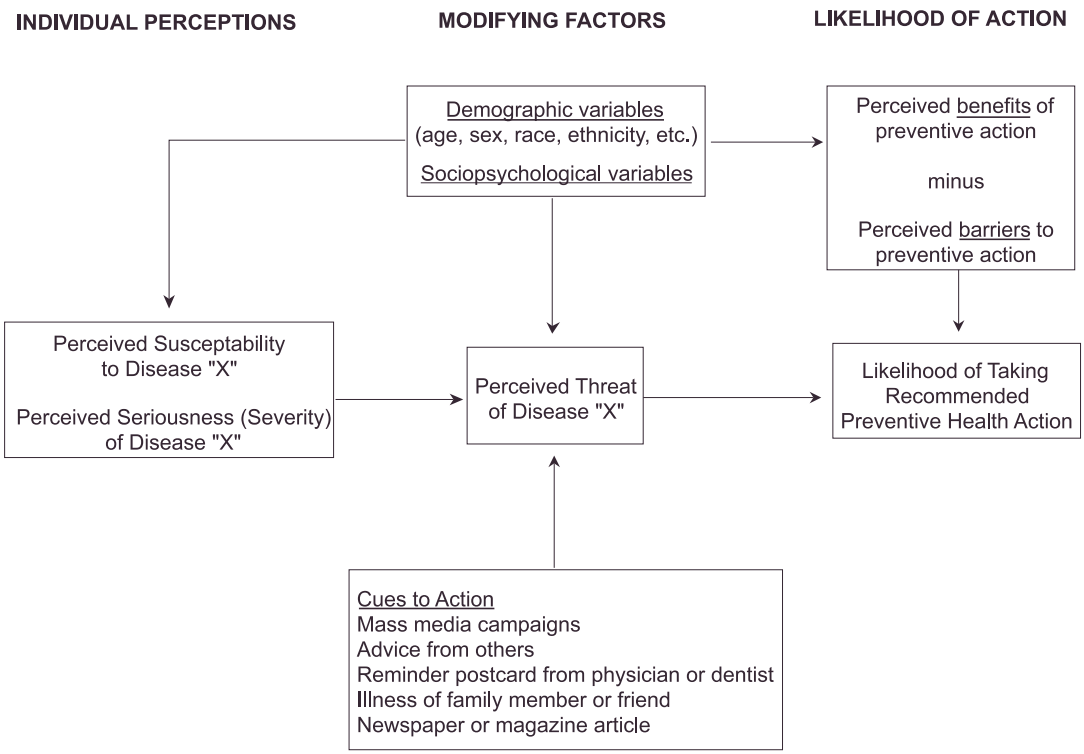


Figure A.3: Health belief model[5]



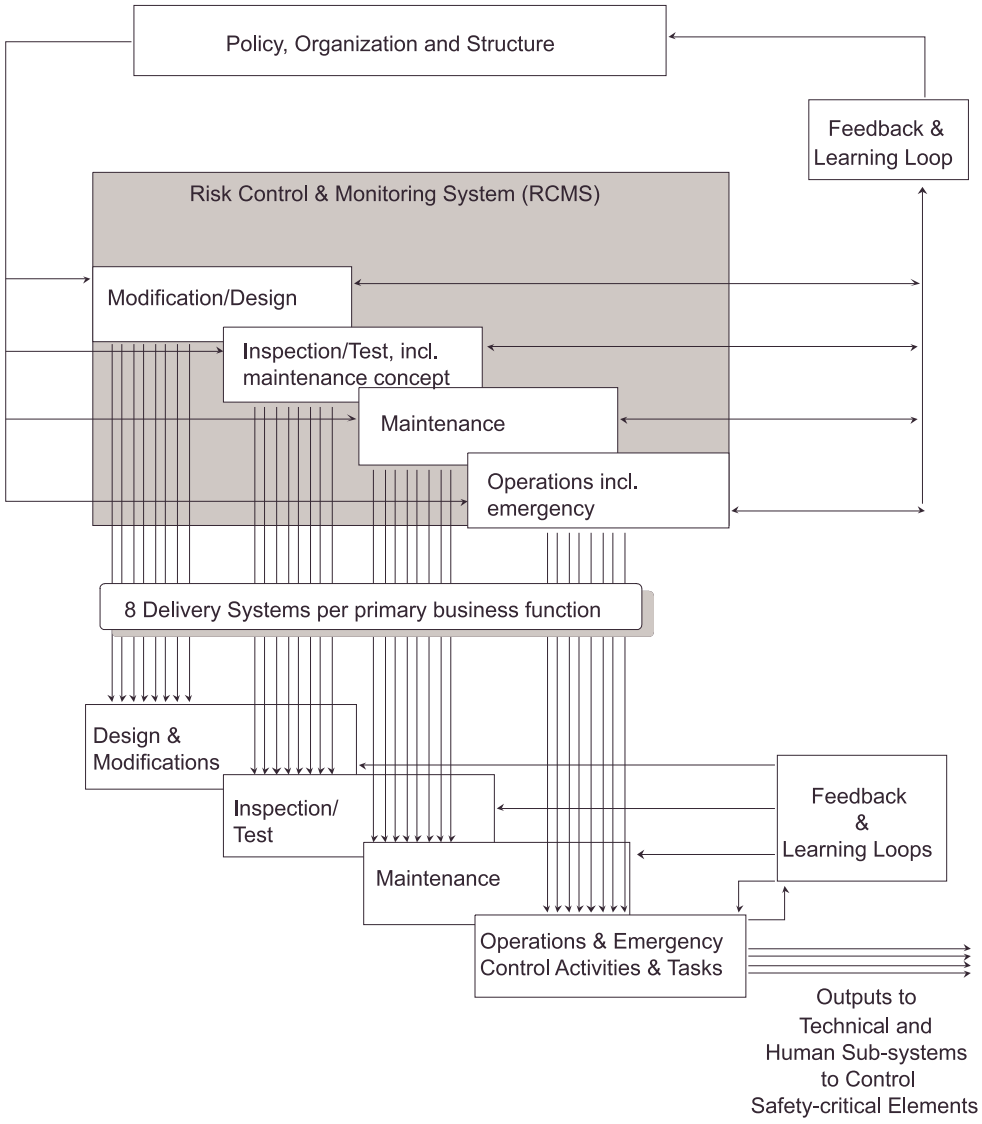
A.3 Models for interventions in the safety management system

When we move to interventions at the level of the organization (i.e., interventions to change workplace safety policies, procedures, structures, organization), the causal chain between the intervention and the final outcome measures of, for example, injury, becomes even longer. It is therefore much more difficult to find evidence for this link in a convincing way. Thus, there is an even greater need to measure intermediate outcomes as well. Few comprehensive organizational models exist, linking aspects of management structure all the way through to the injury process. The model in Figure A.4 is one which attempts to do this.

The model shows a management system as an interacting set of tasks, controls and resources, linked by communications and feedback loops, that develops, operates, monitors and improves a risk control and monitoring system (RCMS). The RCMS carries out risk analysis and plans all of the control functions and activities, including the system for monitoring the performance of the risk control system. The outputs to the technical and human sub-systems in the management model can be seen as the links to the earlier models (Sections A.1 and A.2). The management policy-making system sets up the RCMS, alongside the other aspect systems of the company (such as quality, environment or productivity), reviews it and gives it signals to improve. These loops make it a dynamic, learning system.

[5] Becker MH, Haefner KP, Kasl SV, Kirscht JP, Maiman LA, Rosenstock IM [1977]. Selected psychosocial models and correlates of individual health-related behaviors. Med Care 15:27-46. With permission of Lippincott Williams & Wilkins.

Figure A.4: Model of a safety management system[6]



[6] Model from Hale et al. [1999]

The eight delivery systems, referred to in the figure, deliver for each primary business function the following generic controls and resources to the requisite safety-critical tasks:

- people *available* when required to perform safety-critical tasks
- *competent* in performing them and
- *committed* to safety;
- clear *output goals, procedures, rules* and *plans* for safety;
- hardware resources of well designed *work interfaces, tools* and *equipment* that can be worked safely;
- *spares, replacements* and *modifications to plant and equipment* that maintain safety;
- *communication and coordination* channels for ensuring that information about safety is disseminated to the right people, and that tasks performed by different people are safely coordinated;
- *mechanisms for resolving conflicts* among safety and other criteria.

Each delivery system actually involves a number of steps or sub-tasks. For example, the first delivery system mentioned - delivering people - involves analysing their tasks; specifying the personnel requirements; selecting and training the right people; and allocating them to the work at the appropriate times to perform their safety-critical functions.

Interventions at the level of the management system can be done to introduce or improve the functioning of any one of the elements of the model, or a combination of them. Examples include involving operators in writing safety procedures; adopting a new safety audit system to review the RCMS; appointing a review committee to check proposed plant modifications for safety and health implications; and introducing ergonomic standards for purchasing tools and equipment. This model therefore provides possible sets of intervening variables to link the management interventions to the ultimate change in injury experience in the deviation model.

Appendix B

Examples of statistical analyses

This appendix shows some simple statistical analyses. We assume that you have a computer with a statistical software package, or if not, can get a book that explains statistical calculations. Some well-liked basic texts are Altman [1991]; Armitage and Berry [1994]; Clarke [1980]; Colton [1974]; Freedman et al. [1998]; Healey [1984]; Norman and Streiner [1994]; Siegel and Castellan [1988]; Swinscow [1978]; Weinberg and Goldberg [1990]. Some of the techniques you might need are available in spreadsheet software packages - but they have their limitations. We demonstrate how to do some analyses not typically found in most packages.

Reasonably good statistical applications are fairly cheap (many universities have site licences) and some can be downloaded free from the Internet. Two products, available at the time of writing, include: the widely used Epi Info, a DOS-based word processing (questionnaire design), database and statistical program (http://www.cdc.gov/epo/epi/epiinfo.html); and PEPI (http://www.usd-inc.com/pepi.html), a statistical package with many useful statistical procedures.

Keep in mind that most statistical tests make certain assumptions about the real world, which may or may not be true in your situation. We mention a few of these when they are particularly relevant.

The examples included in this chapter are organized according to the evaluation designs presented in Chapters 3 and 4. For some, we show how to do the calculations. For others we show examples from published papers. In some of these cases (e.g. paired t-test, two-sample t-test), any statistical package will do the computations.

B.1 Analyses for before-and-after designs

B.1.1 Before-and-after design with injury rate data[7]

Calculating a p-value

Much injury data is in the form of rates. Typically they are expressed in the form:

$$\textit{Injury rate} = \frac{\textit{Number of Injuries}}{\textit{Number of hours worked}}$$

The denominator could also be the number of people working in a plant in a given year. You may want to compare two rates in one workplace, before and after an intervention. The hypothesis here is that the intervention has no effect, i.e., no difference between the two measurements, apart from chance variability. The data are in Table B.4. (Note that it is important to do the calculations as accurately as possible, so do not round off numbers on your calculator until the very end of the analysis. For ease of reading we round off below, but we have already done the calculations more precisely.)

Are the rates 70 (per 100,000 hours worked) and 37 significantly different? The approach is to see how many injuries are expected to occur if the total rate, in both time periods combined, is applied to the Before and After groups - as if there is no true difference between groups. You then compare the observed (actual) numbers of injuries with those expected. In this case, the overall rate is 50 (per 100,000 hours). So the expected number of injuries in the Before group is (50/100,000) x 40,000 = 20, and in the After group is (50/100,000) x 60,000 = 30.

[7] The method described here assumes that the risks of an injury before-and-after are "independent." Strictly speaking, with many of the same people working that may not be true, but we are reasonably safe in using this approach. If most injuries were occurring to the same few people before and after the intervention, we would need to use a more sophisticated statistical technique.

Table B.4: Injury rate data from a before-and-after evaluation design

	Before Intervention	*After Intervention*	*Total*
Number of injuries	28	22	50
Employee hours	40000	60000	100000
Number of injuries / 10^5 hours	70	37	50

Now you calculate the test statistic (X^2):

$$X^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

where $\sum$ (the Greek letter sigma) means you add up the quantities (Observed - Expected)^2 / Expected for all the groups. For the data here,

$X^2 = [(28 - 20)^2 / 20] + [(22 - 30)^2 / 30] = 5.33.$

You compare the calculated value of 5.33 with critical values of something called the *chi- squared* (χ^2) distribution with *one degree of freedom*[8]. For 5% significance (i.e., $\alpha = 0.05$), $\chi^2 = 3.84$, and the calculated value 5.33 is larger than this. This means that our result is statistically significant, i.e., the probability of getting a difference in rates as large or larger than was found is less than 5%, if there really is no effect of the intervention. (A computer program gives the precise p-value: p = 0.021.)

Note of caution: This method works well when the numbers of injuries are not too small. A reasonable rule here is that the number of injuries in each group should be at least five. If it is less, you can use a different statistical approach known as an *exact* method, which would likely require a computer.

Calculating rate ratios and rate differences

One limitation of this method is that it does not indicate the strength of the effect of the intervention. How can you measure that? We describe two obvious ways. The first is to look at the relative injury rate. Take the rate after the intervention and divide by the rate before. Call the result RR (for Rate Ratio), which in this case is 36.7 / 70 = 0.52, or 52%. You could also say that the rate has dropped by 100 - 52 = 48%. The second measure is simply the difference in the rates, RD. Here RD is 70 - 36.7 = 33.3 per 100,000 hours worked.

Calculating a confidence interval for a rate ratio

You can calculate a confidence interval (CI) for these estimates. As described in Section 8.5, a confidence interval is a range of values within which the true value of the parameter lies, with a probability of (100 - α)%, usually 95%. Let us start with the RR. For reasons we will not go into, the analysis uses natural logarithms, abbreviated as "ln" on your calculator. The CI for ln(RR) is given by the limits:

$$\ln(RR) \pm Z \times SE,$$

where $\pm$ means plus or minus. Z is a number you can look up in a statistical table of critical values from the normal distribution. Its value depends on whether you want a 95% CI or 90% or some other value. [We use the conventional 95%, for which the appropriate value of Z is 1.96.]

[8] Degrees of freedom are used a lot in statistics. We will not go into any detail, but simply point out that in this situation the number of degrees of freedom is one less than the number of groups.

SE is something called the Standard Error and in this case it refers to the Standard Error of ln(RR).

The Rate Ratio = 36.7 / 70 = 0.52, and a calculator shows ln(0.52) = -0.65. You now need to calculate the Standard Error (SE). It is not too complicated. Take the reciprocal (1 divided by the number) of the number of injuries in each time period, add the reciprocals up, and take the square root of the sum. The reciprocals are 1/28 and 1/22. Adding them gives 0.08, and the square root of that is 0.28. With our data, the 95% CI for ln(RR) is then:

$$-0.65 \pm (1.96 \times 0.28) = -1.2 \text{ to } -0.09.$$

Since this is the CI for ln(RR), you now need to take antilogs to get the CI for the Rate Ratio (RR) itself. (On your calculator, you get antilogs with the button that reads "e^x"). This gives the CI for RR as 0.30 to 0.92. In other words, you are 95% sure that the true value for the Rate Ratio is between 0.30 and 0.92.

Calculating a confidence interval for a rate difference

Now let us work out the CI for the RD, the difference in rates. The CI for RD is given by the limits:

$$RD \pm (Z \times SE).$$

You earlier calculated rates per 100,000 hours (not per hour) to avoid lots of zeros after the decimal point. This means you can use time units of 100,000 hours, which makes things a little easier on your calculator. You found earlier that the RD is 33.3 per 100,000 hours. You again need to get the SE - the Standard Error of RD in this case. This time you calculate # injuries / (# time units)^2 for each of the time periods, add them up and take the square root of the sum.

Thus,

$$SUM = [28 / 0.4^2] + [22 / 0.6^2] = 236.11$$
$$SE = \sqrt{SUM} = \sqrt{236.11} = 15.37$$

For the 95% CI,

$$RD \pm (1.96 \times SE) = 33.33 \pm (1.96 \times 15.37)$$
$$= 3.22 \text{ to } 63.45.$$

The CIs for RR and RD show that you cannot rule out that the effect could be quite small or very large. Your best estimate of each, though, is still the value you actually found from the data, e.g., 33.3 is the best estimate of RD.

What is the appropriate "before" measurement when you have historical data?

In our example above, we have a situation where the periods of time before and after the introduction of the intervention are similar in length. What do you do in a situation where, for example, you have injury rate data for several years before the intervention, as well as for one year after the intervention's introduction? Some people would propose calculating the "before" measure from the data of several years. This is generally <u>not</u> recommended, since many things can change over such a long period of time.

On the other hand, the historical data are useful for judging whether or not the most recent rate measurement for the period just before the intervention - your candidate "before" measurement - is typical of the preceding years. If it is atypical, then consider regression-to-the-mean as a possible threat to internal validity. If there is any suggestion from the historical data that a trend over time is occurring, you would be well-advised to use the time series methods described in Section B.4 for your analysis instead of the above test.

B.1.2 Before-and-after design with continuous data

The next illustration is also concerned with before-and-after designs, but in this case the data are continuous. We will refer to a paper by Robins et al. [1990] which used several statistical methods. The study was of a large U.S.

manufacturing firm with 50 plants. The intervention was designed to bring the company into compliance with the U.S. Federal Hazards Communications Standard (HCS), informing workers about work-related hazardous substances, the potential consequences of exposure, detection and protection. The ultimate goal of the HCS is to reduce chemically related occupational illnesses and injuries.

Trainers were trained, and worked with union and management, who jointly developed and implemented the programs at each plant. The evaluation was designed to assess the planning and implementation, attitudes and knowledge, work practices, working conditions and organizational impacts. (The authors did examine changes in rates of illness and injury, but because of a change in the classification system, decided they could draw no conclusions.) Five plants, representing the variety of manufacturing processes, were chosen as sites for data collection. At each plant, data were collected in three phases - phase one, as the training was being completed; phase two, one year later; and phase three, a further year later, i.e., two years after the training.

Although the design had three measurement times, the data were analysed by comparisons of the measures at two times - e.g., phase one with phase two. Information was collected from various sources, including semi-structured interviews, feedback sessions, observations and questionnaires. The questionnaires were given to about 50 hourly paid employees at each of the five plants. One hundred and twenty-five employees answered questions on work practices at both phases one and two. A composite score was calculated as the average response to the questions, which were rated from 1 = never to 4 = always. The hypothesis being tested involved whether the program was effective in changing work practices from phase one to phase two. Statistically, you start with a null hypothesis that the difference in practices between the phases should be zero. The observed mean score for this group was 2.80 at phase one and 2.92 at phase two.

Thus, the observed difference between phases was not zero but rather 0.12. The next step was to see if this difference was statistically significant or not. Since each person's composite score was the average of responses to several questions, it was reasonable to consider the data as continuous. As well, since the data were *paired* - each individual provided a pair of scores, with one at each phase - the appropriate statistical test was the *paired t-test*. This method produces a *t-statistic*, which can be looked up in a table to determine the p-value - computer printouts give this automatically. The authors reported the p-value to be 0.07. Using a cut point (significance level) of 0.05, the result was thus not statistically significant. Therefore, the null hypothesis could not be rejected.

B.2 Analyses with pre-post measures and a control group

B.2.1 Pre-post with control group and rate data

The wrong way to compare pre-post rate changes in two groups

As noted in earlier chapters, you should always use a control group - randomized or non-randomized - if at all possible, since rates before and after an intervention can differ for reasons that have nothing to do with the intervention. Suppose you collect the data in Table B.5. You might think you could do a before-after comparison for each group, using the method shown in Section B.1.1, and see if there was a significant drop in the rate in the intervention group, but a non-significant one in the controls.

The problem is, it is WRONG! To see why, look at Table B.5. In the intervention group, the "before" rate is 7.3 injuries per 100 FTE workers. The "after" rate is 4.1. Calculations, as shown in section B.1.1, might find p = 0.048, i.e., just statistically significant. In the controls, the "before" rate is 7.7, and the "after" rate is 4.6, giving p = 0.052, not quite significant. This could lead you to think that the intervention is effective.

Table B.5: Injury rate data from a pre-post with control group evaluation design

(injuries per 100 FTE workers)

Period of Injury Rate Measurement	Intervention Group	Control Group
Pre-intervention	7.3	7.7
Post-intervention	4.1	4.6

In fact, the appropriate statistical test compares the before-after difference in the intervention group (3.2) with the difference in the controls (3.1). (You are examining a difference of differences!) If you do the calculations (we show you how in another example below), you would find that it is NOT significant, showing you cannot claim the intervention works - something else likely led to the drop in rates in both groups.

Now there is another way to think of these data. You could look at the *relative* (or *proportional*) values of the rates as we did in section B.1.1. The rate ratios are 56% in the intervention group and 60% in the controls. (You could also say that there was a 44% drop in rate in the intervention group and 40% in the controls.) Some statisticians view this as a better way to consider the data. One reason for this is that if the groups have quite different initial rates, then one of the groups can improve a lot more than the other based on the difference in rates. For example, if Group One has a Before rate of 4.3 and an After rate of 1.2, a difference of 3.1, then if Group Two's Before rate is 2.9, its rate cannot possibly be reduced as much as Group One's, even if it has no injuries at all! This problem does not apply to relative changes, which cannot exceed 100%.

Comparing the pre-post rate changes in two groups using rate ratios

We now show you the two methods for comparing the change in an intervention group versus the change in a control group. The statistical methods for analyzing rate differences and rate ratios are not the same. First, let us look at rate ratios. We will use the data in Table B.6 and show you how to calculate the test statistic, z, using rate ratios

We start by going through some calculations similar to those in Section B.1.1. Get the ln(RR) for each group. (We use subscripts 1 and 2 to represent the two groups.)

$$\ln(RR_1) = \ln(5.74/6.00) = -0.043,$$
$$\ln(RR_2) = \ln(2.03/6.40) = -1.149.$$

Then calculate the difference between these, D:

$$D = \ln(RR_1) - \ln(RR_2) = -0.043 - (-1.149) = 1.106.$$

You need the Standard Error (SE) of this difference as well. Simply take the reciprocal of the number of injuries in each of the four pre/post, control/intervention categories, add them up and take the square root of the sum. The reciprocals are 1/49, 1/46, 1/26, and 1/8. Adding them gives 0.206, and the square root of that is 0.453.

Now calculate the test statistic, z:

$$z = D / SE = 1.106 / 0.453 = 2.44.$$

When z is bigger than 1.96 - the critical value from the normal distribution, when $\alpha = 0.05$ - as it is here, the difference in (the ln of) the rate ratios is statistically significant. Thus, the data suggest the intervention works.

Table B.6: Injury rate data from a pre-post with control group evaluation design

		Pre-intervention	Post-intervention
Control (Group 1)	Injuries	49	46
	Hours	817000	801000
	Rate per 100,000 hrs.	6	5.74
Intervention (Group 2)	Injuries	26	8
	Hours	406000	394000
	Rate per 100,000 hrs.	6.4	2.03

Comparing the pre-post rate changes in two groups using rate differences

Let's do the calculations based on the difference in rates, rather than the ratio. You used rates per 100,000 hours so you can use time units of 100,000 hours for the rate differences.

$$RD_1 = 6.00 - 5.74 = 0.26,$$
$$RD_2 = 6.40 - 2.03 = 4.37.$$

And the difference between them is calculated by subtracting the RD for the control group from the RD for the intervention group.

$$D = RD_2 - RD_1 = 4.37 - 0.26 = 4.11.$$

Again you need the SE of the difference. In this case, calculate # injuries / (#time units)^2 for each of the four categories, add them up and take the square root of the sum:

$$SUM = 49/8.17^2 + 46/8.01^2 + 26/4.06^2 + 8/3.94^2 = 3.54$$
$$SE = \sqrt{SUM} = \sqrt{3.54} = 1.88.$$

As before, calculate the test statistic, z:

$$z = D / SE = 4.11 / 1.88 = 2.19.$$

This z is bigger than 1.96, so this analysis also provides evidence that the intervention works.

Notice that the z values calculated by the rate ratio and rate difference methods are not quite the same. This is because the analyses have actually tested different hypotheses. If the pre-intervention values are different, then it is possible for the hypothesis about rate ratios to be true, but that for rate differences to be false; or for the rate difference hypothesis to be true, but the rate ratio one to be false. For example, the pre and post rates for controls could be 12 and 9 respectively; those for the intervention group might be 6 and 3. The rate difference is the same, but the ratios are quite different - 75% and 50%. Likewise, suppose the control group's rates pre and post were 12 and 6, respectively, compared with 6 and 3 for the intervention group. The rate ratios are the same, but the rate differences are not - they are 6 and 3.

In practice, if the pre-intervention rates are very different, you should be concerned about a potential selection threat to internal validity. When you have chosen the controls well, the pre-intervention rates in the two groups should be similar, in which case the two analyses should give reasonably similar conclusions. If you are concerned about moderate differences between the groups, consult a statistician - there are ways to make adjustments to your analysis through multiple regression analyses.

B.2.2 Pre-post with control group and continuous data

Section B.1.2 demonstrated how to compare before and after scores on a continuous scale for a single group. Essentially, you take the difference score (pre - post) for each person and work with those values in the analysis. With two groups (intervention and control), you can do the same thing. Then for each group you have the difference scores for all the people in the group and you have reduced the data to a single difference score for each person. If the intervention is ineffective, then there should be no difference in the means of these difference scores between the groups. The way to see if any difference is statistically significant or not is to do another type of t-test. You have two groups; so this is called the *two-sample t-test*. It is in any computer package or textbook.

What do you do when there are differences in the characteristics of participants in the intervention and control groups that might influence how they respond to an intervention? In such a situation, apply a more sophisticated technique that allows a correction of these differences, such as some form of multiple regression.

B.3 Analyses for designs with after-only measures and a control group

When you only obtain measurements after the intervention and not beforehand, you can have a problem if the subjects are not randomized to intervention or control group. This is because the groups may have been quite different at the start. For this reason, in chapter 4, we recommended "after"-only designs only when using groups formed through randomization. We therefore expect the statistical methods of this section to be used primarily in such situations. However they can sometimes be used - with caution - in the absence of randomization, especially if there is information on potential *confounders* post-intervention. If so, statistical techniques are available that can use this extra data. As you can imagine, they are too complicated for this appendix - you will have to talk to a statistician - but we mention some of them here.

B.3.1 After-only measurements with two groups and rate data

The same statistical test, χ^2, is used as in the before-and-after design with rate data (Section B.1.1), but because you have two different groups of workers, the cautionary footnote no longer applies.

B.3.2 After-only measurements with several groups and rate data

Sometimes you might have several groups, for example one control and others in which different interventions are carried out (Table B.7). This time our hypothesis is that none of the interventions work; so we expect the rates in all the groups to be the same. Notice that if even one of the interventions works, our hypothesis is contradicted. A simple approach might be to compare each intervention group with the control group, using the test indicated in the preceding section (B.3.1). However, this would be invalid because it involves multiple testing.[9] Instead, we use an approach similar to the one for two groups.

Again, use the overall rate to estimate the expected number of injuries in each group, assuming the overall rate applies to all groups. For example, for the control group, the expected number of injuries = (150 / 250,000) x 60,000 = 36.0. The equivalent values for the other three groups are 42.0, 30.0, and 42.0.

[9] Multiple testing means you are doing more than one test at the same time. If you do each one using an alpha of 0.05, the probability that at least one of the tests is significant is more than 0.05, sometimes considerably more. You have to make an adjustment or use a different method, as is done here, to keep the overall alpha at 0.05.

Table B.7: Example injury data from an experimental design with post-only measurements and multiple intervention groups

	Control	Intervention 1	Intervention 2	Intervention 3	Calculated Total
Injuries	43	47	36	24	150
Hrs.	60000	70000	50000	70000	250000
Rate/10^5 hrs.	71.7	67.1	72	34.3	60
Expected number of injuries	36	42	30	42	

Again, calculate X^2:

$$X^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

This time there are 4 quantities to add up, each corresponding to one of the four groups:

$$X^2 = [(43-36)^2 / 36] + [(47-42)^2 / 42] +$$
$$[(36-30)^2 / 30] + [(24-42)^2 / 42] = 10.87.$$

Again, compare this with the chi-squared distribution. The number of degrees of freedom is one less than the number of group measures being compared, i.e., , 4 - 1 = 3. A chi-squared table shows the 5% significance point is 7.81. Our X^2 is bigger than this, so again, it is statistically significant, i.e., , $p < 0.05$. (The actual p-value is 0.012). Note that this method does not work properly if the number of injuries is small (less than 5) in one or more of the groups. In such cases, you need to use an exact method.

B.3.3 After-only measurements with two groups and continuous data

If there are two groups with after-only measures, you have the same situation as described in Section B.2.2, with a series of scores for each of the two groups. You may want to see if any difference is statistically significant. Again do a *two-sample t-test*.

B.3.4 After-only measurements with several groups and continuous data

Survey questions in the Robins et al. study described earlier also asked about the extent to which workers followed certain work practices. The responses for each item ranged from 1 = "never" / "almost never" to 4 = "always", and an average was calculated to give a scale score. For each plant, the mean scores of individuals were calculated. These means ranged from 2.42 to 3.06.

The hypothesis tested was that there was no difference in work practices between plants. Once again, the data were treated as continuous, but this time with the means of several groups compared. Whereas the two-sample t-test applied to a comparison of the means of two groups, the generalization of this to several groups is called a one-way Analysis of Variance (ANOVA). ANOVA is actually a generic name for a range of procedures, of which this is a particular example. The method produces an *F-statistic*, and a subsequent p-value. In this case, $p = 0.004$. Thus, the result is statistically significant, and we conclude that the differences in observed means are not due simply to chance.

As with the 2 x 5 table discussed earlier, the method does not tell us which means are different from others. In fact, with five groups,

there are ten possible comparisons between pairs of groups. This raises the issue of multiple testing, described earlier. You should only consider testing individual pairs if the "overall" F-test is significant. Various text books demonstrate how these pairwise comparisons, allowing for multiple testing, are made[10].

In a quasi-experimental (and sometimes in a truly experimental) setting, it is important to remove the effects of any characteristics of individuals that may affect the outcome measure. For example, older workers may behave more safely on the job. If this is true, and the intervention group contains more older workers, it will likely have better outcomes following the intervention than the comparison group, regardless of the value of the intervention. Likewise, if the comparison group is older, we may fail to see a real effect of the intervention. Statistically, you can allow for the age effect, to obtain a more proper estimate of the impact of the intervention. (The technical term for these adjustment variables is *confounders*.)

Robins et al, in the example above, noted known differences in the people surveyed in each plant. Several variables were listed where the respondents differed in occupation, degree of job hazard, age, education level, and number of years in the plant. There was obvious concern that if work practices varied by age and more older people worked at some plants, then this disparity (and others like it) - rather than the training program - could have accounted for the significant differences. This would threaten internal validity through a selection bias. A statistical approach that combines ANOVA with allowance for these confounders (covariates) is needed. The answer is a technique called analysis of covariance (often abbreviated as ANCOVA). The authors reported that in their study the new analysis did not change the basic conclusions about differences in work practices.

B.4 Multiple measurements over time

Komaki et al. [1980] reported the effect of an intervention involving training and feedback. Concerned that training alone was not sufficient to produce behavioral change, a "multiple baseline design across groups" was applied with four sections of a city's vehicle maintenance division. (This design was described in section 4.2.3.) Five "conditions", each lasting for several weeks, were used: 1) Baseline; 2) Training only I; 3) Training and Feedback I; 4) Training only II; 5) Training and Feedback II. (i.e., after a Baseline period, there was a period when only training was given (Training only I), followed by one that also included feedback (Training and Feedback I). Feedback was then dropped (Training only II), before being reinstated (Training and Feedback II). Since the study was done in four sections of the workplace, the times for the changeover from one condition to another differed. (This allowed the authors to check on whether or not other changes at the work-site unrelated to the intervention might have influenced behavior.)

The main outcome measure was of the safe performance of tasks, measured by behavioral observations. Observations were made several times a week and plotted to give a weekly average of the percentage of incidents performed safely in each maintenance section. Since the conditions each lasted from five to eleven weeks, there were multiple measures before and during each condition, with approximately 45 observations for each section over the course of the evaluation. Such data are a form of *repeated measures* known as *time series data*.

The authors wanted to allow for general trends in safe behavior, as well as see if the change from one condition to the next led to a switch in behavior. The appropriate method for this form of time series is known as ARIMA - auto regressive integrated moving averages. Now the

[10] For example Kleinbaum et al. [1988]

name might seem enough to put you off statistics for life. What is important is that it takes account of correlations from measurement to measurement (auto correlation) - if the percentage was high in one week, it was likely to have been relatively high the next week. The data are too complex to describe in detail here. Nevertheless, the general message from the evaluation was clear. Without feedback as well, training showed relatively little effect.

Time series analyses are appropriate in a situation where there are repeated measurements on the same subjects, e.g., when taking behavioral measurements. They are also appropriate when there are trends in workplace data due to changes in the business cycle, weather etc.

There are occasions, perhaps after an intervention is in place and running well, when the injury rate is expected to remain stable over time. Yet because a single plant may experience only a few injuries per month, the monthly rate may vary considerably simply because of random variability. To check if the results for a single month are significantly out-of-line with previous experience, you can use control charts. They are used in quality control settings, perhaps to make sure that the size of ball bearings is within a small range. They can be readily adapted to workplace safety.

As an example, suppose that on average there are three injuries per month. (We assume the number of hours worked is constant; if not, you can use the average rate with its standard deviation, a measure of month-to-month variability in the rate.) Sometimes, there will be only one or two injuries in a month, or maybe 4 or 5. In fact, the probability of only one in any month is about 10%, while the probability of five is about 10%. Even 6 in a month will happen about 5% of the time, so might well occur at some point in a two-year period. This means you shouldn't be too quick to push the panic button when one month's figures are somewhat higher than normal. But by the same token, you shouldn't be smug if in one month there are no injuries at all.

Control chart methodology will alert you when the number of injuries in a month is so high that there seems to be a real problem, or when the pattern over two or three months is a cause for concern.

Appendix C

Reporting your evaluation results

C.1 Introduction

Most of this guide has focused on the methodology required to do a good intervention effectiveness evaluation. This appendix focuses on what to do with the results of the evaluation. Written reports are the usual way to summarize them, even when an evaluation is done in-house. Not only does this provide a record for the organization, the process of writing the report also encourages a critical examination and synthesis of the evaluation activities and results. We will describe the sections that people typically include in a report. We will also discuss how your communication strategy should extend beyond the report itself.

C.2 Evaluation report

C.2.1 Structure of the report

Table C.1 lists what you would typically include in a report. First is the abstract/executive summary, which incorporates the main points of the introduction, methods, results, discussion and conclusion sections. This is typically one or two pages in length. This summary is an important since, for many readers, this might be the only section they read in its entirety.

- The introduction presents the goals of the intervention, the intervention itself and the general approach taken in the evaluation.
- Methods/procedures then describe the evaluation methods in detail.
- Results present the data gathered through the evaluation which address the evaluation

questions. This section should present the results not only in text, but also through figures, graphs and tables. These visual summaries facilitate uptake of the information for many readers.

- Many reports include a discussion section, which should lead the reader from the results to the conclusion. Whereas the results section gives a logical presentation of the results, the discussion synthesizes and interprets them with reference to current theory and understanding. The discussion section is also the place to consider threats to the internal validity of the evaluation, including any reasoning based on theory or data from outside of the evaluation.
- The conclusions summarize what is concluded from the data and, possibly, any resulting recommendations. Conclusions should address the main evaluation questions.
- In fact, as much as possible, the entire report should be constructed so that the relationship of the various methods and results sub-sections to the evaluation questions and conclusions is clear.

Table C.1 What to include in the evaluation report [11]

Sections of report	Content of sections
Abstract/executive summary	• Overview of the program and evaluation • General results, conclusions and recommendations
Introduction	• Purpose of the evaluation • Program and participant description (including staff, materials, activities, procedures, etc.) • Goals and objectives • Evaluation questions
Methods/procedures	• Design of the evaluation • Target population • Instruments (e.g., questionnaire) • Sampling procedures • Data collection procedures • Validity and reliability • Limitations • Data analyses procedures
Results	• Description of findings from data analyses • Answers to evaluation questions • Charts and graphs of findings
Discussion	• Explanation of findings • Interpretation of results • Consideration of threats to internal validity
Conclusions/ recommendations	• Conclusions about program effectiveness • Program recommendations

[11] Table from McKenzie and Smeltzer, Planning, Implementing, and Evaluating Health Promotion Programs: A Primer, 2nd ed. Copyright (c) 1997 by Allyn & Bacon. Adapted by permission.

C.2.2 Audience specificity

One of the key principles in communicating a report is to tailor it to the audience. It should have the audience's level of education and interests in mind. Key messages should be formulated in the conclusion and abstract so that they answer the questions most pertinent to the audience. Conceivably you might have more than one report - preparing both technical and lay person versions is common.

C.2.3 Clear language

The report should be written in clear language if a non-technical audience is planned. This means it will be quite different from the style found in many academic publications. Guidelines for clear language have been developed by many organizations[12]. The following is a compilation from some of these .

Guidelines for writing in clear language

Overall

- Write with your audience's needs, knowledge and abilities in mind

Document organization

- Include a table of contents for longer documents
- Divide document into sections of related information, using headings and sub-headings
- Include detailed or technical material in an Appendix

Paragraphs

- Limit each paragraph to one idea
- Avoid paragraphs of more than five sentences
- Consider using point form for a list of related items
- Use left justification, but not right justification; i.e., leave a ragged right margin

Sentences

- Limit each sentence to one point
- Sentences should be no more than 20 words on average and, typically, not exceed 25 to 30 words
- Use a subject-verb-object order for most sentences

Words

- Avoid jargon and technical words; explain them when used
- Eliminate unnecessary words (e.g., replace "in view of the fact" with "because")
- Use the active voice instead of the passive voice. (e.g., replace "The requirement of the workplace was that employees....." with "The workplace required employees....)
- Avoid chains of nouns (e.g., resource allocation procedures)

Font

- Use a serif style of font (with hooks on the end of characters) instead of a sans serif style
- Do not use all upper case (i.e., all capital) letters for anything longer than a brief statement
- 12 point type is recommended for the main text

[12] For example: Baldwin R [1990]. Clear writing and literacy. Toronto: ON Literacy Coalition; Canadian Labour Congress [1999]. Making it clear: Clear language for union communications. Ottawa: Canadian Labour Congress; Gowers E (revised by Greenbaum S, Whitcut J) [1986]. The complete plain words. London: HMSO; Ministry of Multiculturalism and Citizenship [1991]. Plain language clear and simple. Ottawa: Ministry of Supply and Services.

C.3 Communication beyond the report

Communicating the evaluation results involves more than producing the evaluation report. Relationships between would-be users of the evaluation results and the evaluators should be established early on, because the successful uptake of results often depends on a few key individuals who understand and support the evaluation. For this reason we recommend forming an evaluation committee at the outset that includes key stakeholders (Chapter 2). This committee should be involved at all stages of the evaluation: development of questions; selection of design and methodology; and interpretation of results. An ongoing engagement of stakeholders fosters trust, understanding and ownership of the results. It also helps ensure that the results are appropriate to their needs.

At the point of release of the final results, you will ideally include several interactive means of presenting the report's results, involving either larger verbal presentations or small group meetings. Interaction of the audience with the presenters should be encouraged in both cases. Make sure the key messages of the report are emphasized and give people the opportunity to voice any doubts or lack of understanding. A variety of written, verbal and visual presentations might be needed for various audiences.

C.4 Summary

Communication of the evaluation results involves, at the very least, a clear, well-organized, audience-specific evaluation report. Other strategies, including the ongoing engagement of an appropriately structured evaluation committee can further the use of an evaluation's results.

Bibliography

Aday LA [1996]. Designing and conducting health surveys: a comprehensive guide. 2nd ed. San Francisco: Jossey-Bass.

Altman DG [1991]. Practical statistics for medical research. London, New York: Chapman & Hall.

Armitage P, Berry G [1994]. Statistical methods in medical research. 3rd ed. Oxford, Boston: Blackwell Scientific Publications.

Becker MH, Haefner KP, Kasl SV, Kirscht JP, Maiman LA, Rosenstock IM [1977]. Selected psychosocial models and correlates of individual health-related behaviors. Med Care 15:27-46.

Cherry N [1995]. Evaluation of preventive measures. In: McDonald C, ed. Epidemiology of work related diseases. London: BMJ.

Clarke GM [1980]. Statistics & experimental design. London: Arnold.

Cohen J [1988]. Statistical power analysis for the behavioral sciences. New York: Academic Press.

Colton T [1974]. Statistics in medicine. Boston: Little, Brown & Company.

Conrad KM, Conrad KJ, Walcott-McQuigg J [1991]. Threats to internal validity in worksite health promotion program research: common problems and possible solutions. Am J Health Promot 6:112-122.

Cook TD, Campbell DT [1979]. Quasi-experimentation: design & analysis issues for field settings. Chicago: Rand McNally College Publishing Co.

Cook TD, Campbell DT, Peracchio L [1990]. Quasi experimentation. In: Dunnette JD, Hough LM, eds. Handbook of industrial and organizational psychology. 2nd ed., vol. 1. Palo Alto, CA: Consulting Psychologists Press, Inc., pp. 491-576.

Cresswell JW [1994]. Research design. Qualitative & quantitative approaches. Thousand Oaks, CA, London, New Delhi: Sage Publications.

Dignan MB [1995]. Measurement and evaluation of health education. 3rd ed. Springfield, Illinois: Charles C. Thomas.

Drummond MF, Stoddart GL, Torrance GW [1994]. Methods for the economic evaluation of health care programmes. Oxford, Toronto, New York: Oxford University Press.

Earp JA, Ennett, ST [1991]. Conceptual models for health education research and practice. Health Educ Res 6:163-171.

Fleiss JL [1981]. Statistical methods for rates and proportions, New York: John Wiley & Sons.

Freedman D, Pisani R, Purves R [1998]. Statistics. 3rd ed. WW Norton & Co.

Glassock DJ, Hansen ON, Rasmussen K, Carstensen O, Lauritsen J [1997]. The West Jutland study of farm accidents: a model for prevention. Safety Sci 25:105-112.

Glendon I, Booth R [1995]. Risk management for the 1990s: Measuring management performance in occupational health and safety. J Occup Health Safety - Aust NZ *11:559-565*.

Gold RG, Siegel JE, Russell LB, Weinstein MC, eds. [1996]. Cost-effectiveness in health and medicine. New York: Oxford University Press.

Goldenhar LM, Schulte PA [1994]. Intervention research in occupational health and safety. J Occ Med 36:763-775.

Goldenhar LM, Connally LB, Schulte PA, eds. [1996]. Intervention research in occupational health and safety: science, skills and strategy. Special Issue of Am J Ind Med 29(4).

Green LW, Lewis FM [1986]. Measurement and evaluation in health education and health promotion. California: Mayfield Publishing Co.

Greene JC, Caracelli VJ, Graham WF [1989]. Toward a conceptual framework for mixed-method evaluation designs. Educ Eval & Policy Analysis *11:255-274*.

Guastello SJ [1993]. Do we really know how well our occupational accident prevention programs work? Safety Sci *16: 445-463*.

Haddix AC, Teutsch SM, Shaffer PA, Dunet DO [1996]. Prevention effectiveness: a guide to decision analysis and economic evaluation. Oxford, New York: Oxford University Press.

Hale AR [1984]. Is safety training worthwhile? J Occup Accid *6:17-33*.

Hale AR, Glendon AI [1987]. Individual behavior in the control of danger. Amsterdam, New York: Elsevier.

Hale AR, Guldenmund F, Bellamy L [1999]. Annex 2: management model. In: Bellamy LJ, Papazoglou IA, Hale AR, Aneziris ON, Ale BJM, Morris MI & Oh JIH I-Risk: Development of an integrated technical and management risk control and monitoring methodology for managing and quantifying on-site and off-site risks. Den Haag, Netherlands, Report to Ministry of Social Affairs and Employment, European Union, Contract ENVA-CT96- 0243.

Hauer E [1980]. Bias-by-selection: Overestimation of the effectiveness of safety countermeasures caused by the process of selection for treatment. Accid Anal Prev *12:113-117*.

Hauer E [1986]. On the estimation of the expected number of accidents. Accid Anal Prev *18:1- 12*.

Hauer E [1992]. Empirical Bayes approach to the estimation of “unsafety”: the multivariate regression method. Accid Anal Prev *24:457-477*.

Hawe P, Degeling D, Hall J [1990]. Evaluating health promotion: a health worker's guide. Sydney, Philadelphia, London, Australia: MacLennan & Petty.

Healey JF [1984]. Statistics: a tool for social research. Belmont: Wadsworth.

Hugentobler MK, Israel BA, Schurman SJ [1992]. An action research approach to workplace health: integrating methods. Health Educ Q *19*:*55-76*.

Johnston JJ, Cattledge GTH, Collins JW [1994]. The efficacy of training for occupational injury control. Occup Med *9*:147-158.

Kjellén U [1984]. The role of deviations in accident causation and control. J Occup Accidents *6*:117-126.

Kjellén U [2000]. Prevention of accidents through experience feedback. London, New York: Taylor & Francis.

Kleinbaum DG, Kupper LL, Muller KE [1988]. Applied regression analysis and other multivariable methods, 2nd ed. Boston: PWS-Kent.

Komaki JL, Jensen M [1986]. Within-group designs: an alternative to traditional control-group designs. In: Cataldo MF, Coates TJ, eds. Health and industry: a behavioral medicine perspective. New York, Toronto, Chichester, Brisbane, Singapore: John Wiley & Sons.

Komaki J, Barwick KD, Scott LR [1978]. A behavioral approach to occupational safety: pinpointing and reinforcing safe performance in a food manufacturing plant. J Appl Psychol *63*: 434-445.

Komaki J, Heinzmann AT, Lawson L [1980]. Effect of training and feedback: component analysis of a behavioral safety program. J Appl Psychol *65*: 261-270.

Krause TR [1995]. Employee-driven systems for safe behavior: integrating behavioral and statistical methologies. New York: Van Nostrand Reinhold.

Laitinen H, Marjamaki M, Paivarinta K [1999a]. The validity of the TR safety observation method on building construction. Accid Anal Prev *31*:463-472.

Laitinen H, Rasa P-L, Resanen T, Lankinen T, Nykyri E [1999b]. The ELMERI observation method for predicting the accident rate and the absence due to sick leaves. Am J Ind Med *1*(Suppl):86- 88.

Lipsey MW [1990]. Design sensitivity. Newbury Park, CA: Sage Publications.

Mason ID [1982]. An evaluation of kinetic handling methods and training. [PhD thesis]. Birmingham: University of Aston.

McAfee RB, Winn AR [1989]. The use of incentives / feedback to enhance work place safety: a critique of the literature. J Safety Res *20*:7-19.

McKenzie JF, Smeltzer JL [1997]. Planning, implementing, and evaluating health promotion programs: a primer. 2nd ed. Boston: Allyn and Bacon.

Menckel E, Carter N [1985]. The development and evaluation of accident prevention routines: a case study. J Safety Res *16*:73-82.

Miles MB, Huberman AM [1994]. Qualitative data analysis. 2nd ed. Thousand Oaks, CA, London, New Delhi: Sage Publications.

Mohr DL, Clemmer DI [1989]. Evaluation of an occupational injury intervention in the petroleum drilling industry. Accid Anal Prev *21*:263-271.

Morgan DL [1998]. Practical strategies for combining qualitative and quantitative methods: applications to health research. Qual Health Res *8*:362-376.

Morgan DL, Krueger RA [1998]. The focus group kit. Thousand Oaks, CA., London, New Delhi: Sage Publications.

Needleman C, Needleman ML [1996]. Qualitative methods for intervention research. Am J Ind Med *29*:329-337.

Norman GR, Streiner DL [1994]. Biostatistics: the bare essentials. St. Louis: Mosby.

Orgel DL, Milliron MJ, Frederick LJ [1992]. Musculoskeletal discomfort in grocery express checkstand workers. J Occup Med *34*:815-818.

Patton MQ [1986]. Utilization-focused evaluation. 2nd ed. Beverly Hills: Sage.

Patton MQ [1987]. How to use qualitative methods in evaluation. Newbury Park, CA, London, New Delhi: Sage Publications.

Patton MQ [1990]. Qualitative evaluation and research methods. 2nd ed. Newbury Park, CA: Sage Publications.

Pekkarinen A, Anttonen H, Pramila S [1994]. Accident prevention in reindeer herding work. Arctic 47:124-127.

Reason J [1990]. Human error. Cambridge: Cambridge University Press.

Rivara FP, Thompson DC, eds. [2000]. Systematic reviews of strategies to prevent occupational injuries. Special issue of Am J Prev Med 18(4)(Suppl).

Robins TG, Hugentobler MK, Kaminski M, Klitzman S [1990]. Implementation of the federal hazard communication standard: does training work? J Occup Med *32*:1133-1140.

Rossi PH, Berk RA [1991]. A guide to evaluation research theory and practice. In: Fisher A, Pavlova M, Covello V, eds. Proceedings of the Evaluation and Effective Risk Communication Workshop. Interagency Task Force on Environmental Cancer and Heart and Lung Disease Committe on Public Education, EPA / 600 / 9-90-054, pp. 205-254.

Rush B, Ogborne A [1991]. Program logic models: expanding their role and structure for program planning and evaluation. Can J Program Eval 6:95-106.

Siegel S, Castellan NJ Jr. [1988]. Nonparametric statistics for the behavioral sciences. 2nd ed. New York: McGraw-Hill.

Sizing up safety: how to measure where your organization has been, where it's at and where it's going. [1995]. Occup Health Safety 11(Mar / Apr):54-60

Steckler A, McLeroy KR, Goodman RM, Bird ST, McCormick L [1992]. Toward integrating qualitative and quantitative methods: an introduction. Health Educ Q *19*:1-8.

Stewart AL [1990]. Psychometric considerations in functional status instruments. In: Lipkin M Jr., ed. Functional status measurement in primary care. New York: Springer-Verlag.

Streiner DL, Norman GR [1989]. Health measurement scales: a practical guide to their development and use. Oxford: Oxford University Press.

Swinscow TDV [1978]. Statistics at square one. London: British Medical Association.

Tarrants WE [1987]. How to evaluate your occupational safety and health program. In: Slote L, ed. Handbook of occupational safety and health. New York: John Wiley & Sons, ch. 8.

Vojtecky MA, Schmitz MF [1986]. Program evaluation and health and safety training. J Safety Res *17*:57-63.

Walsh NE, Schwartz RK [1990]. The influence of prophylactic orthoses on abdominal strength and low back injury in the workplace. Am J Phys Med Rehabil 69:245-250.

Webb GR, Redman S, Wilkinson C, Sanson-Fisher RW [1989]. Filtering effects in reporting work injuries. Accid Anal Prev *21*:115-123.

Weinberg SL, Goldberg RP [1990]. Statistics for the behavioral sciences. Cambridge: Cambridge University Press.

Weiss CH [1988]. Evaluation for decisions: is anybody there? Does anybody care? Eval Practice 9:5-19.

Yin RK [1984]. Case study research: design and methods. Applied social research methods series. Vol. 5. Beverly Hills, London, New Delhi: Sage Publications.

Zwerling C, Daltroy LH, Fine LJ, Johnston JJ, Melius J, Silverstein BA [1997]. Design and conduct of occupational injury intervention studies: a review of evaluation strategies. Am J Ind Med *32*: 164-179.