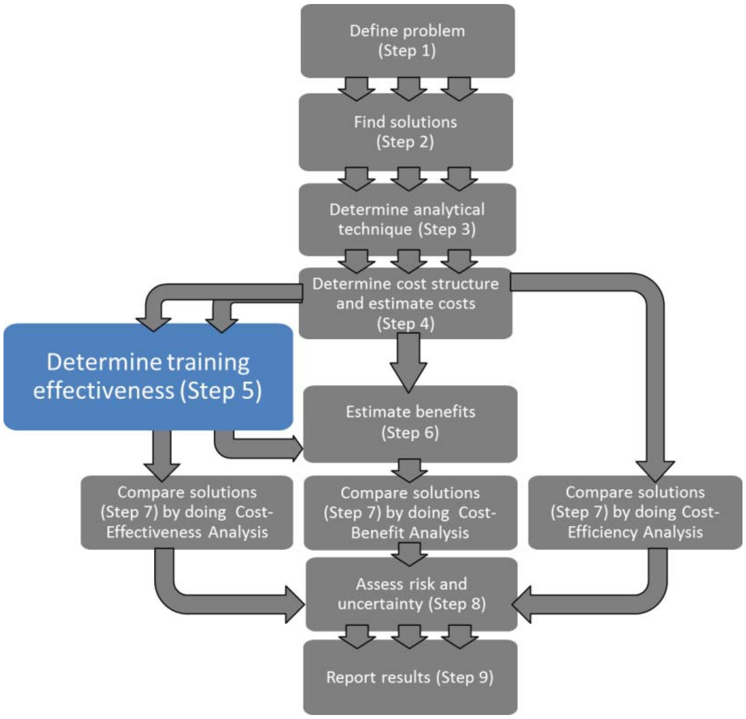
Chapter 6 – DETERMINING TRAINING EFFECTIVENESS[1]

Korteling, J.E. (Hans), Ph.D.
TNO – Human Factors
Kampweg 5, P.O. Box 23, 3769 ZG, Soesterberg, The Netherlands

hans.korteling@tno.nl


Define problem (Step 1)
Find solutions (Step 2)
Determine analytical technique (Step 3)
Determine cost structure and estimate costs (Step 4)
Determine training effectiveness (Step 5)
Estimate benefits (Step 6)
Compare solutions (Step 7) by doing Cost-Effectiveness Analysis
Compare solutions (Step 7) by doing Cost-Benefit Analysis
Compare solutions (Step 7) by doing Cost-Efficiency Analysis
Assess risk and uncertainty (Step 8)
Report results (Step 9)

6.1 KEY QUESTIONS

- What basic information is already available and can be easily collected?
- What kind of performance measures should be chosen to indicate training effectiveness?
- Which relevant (objective) human performance aspects are difficult to obtain and should be inventoried using structured checklists or questionnaires?
- Is it possible to gain more insight into the process aspects of the training that may explain measured training outcomes (e.g., by using questionnaires or observation checklists)?
- Given the objectives and constraints, what is the best experimental design for measuring and comparing training effects in different training solutions?
- Is it possible to combine different kinds of data and measures?
- How can the effectiveness of a military training be estimated before funds are committed?

[1] This chapter is developed upon Korteling, Oprins, & Kallen' paper (2012) published in the Proceedings of NATO Workshop on Cost-Benefit Analysis of Military Training (Wang, 2012).

6.2 INTRODUCTION

Benefits and costs have to be estimated, calculated, weighed, and compared to justify or to evaluate an investment in a new military training intervention or system. CBA, cost-effectiveness analysis, or cost-efficiency analysis have to be conducted for at least two alternative training solutions: the existing training system and a new version or some other alternatives (proposed by the analyst or a subject-matter expert) are needed to determine the most effective solutions in compliance with the stated aim and boundary conditions. Since training aims at improving students' skills, knowledge, or attitudes with regard to mission readiness, the outcome of a training investment has to be expressed in these terms. If such improvement results, we may simply say that training has had an effect. An example of a measurable effect is the number of detections performed by an operator in a surveillance task. When a CBA can be conducted, these training effects or outcomes form the basis for the estimation of the financial training benefits. If the determination of (financial) benefits is too difficult or requires too many assumptions for a CBA, the measurement or estimation of training effects may form the basis for a cost-effectiveness analysis or cost-efficiency analysis (see Chapter 4).

This chapter describes Step 5 of the nine-step economic evaluation framework, focusing on the main aspects of determining (measuring or estimating) the effects of alternative training solutions as a basis for a CBA or cost-effectiveness analysis.

6.3 BASIC QUESTIONS

Effectiveness is not an all-or-nothing phenomenon, but something varying on a continuum from zero effect (or even negative) to 100% effect (Orlanski, 1989). In order to calculate the benefits of a capital investment on the basis of training outcomes, therefore, measurement of training effectiveness is the first and the preferred option. Measurement provides data that can be used to make absolute judgments about the outcome; i.e., about the degree to which the training (or mission-readiness) objectives are met. In other cases, measurement will provide the basis for making relative judgments; i.e., a comparison of two or more training solutions or systems developed to reach the same objectives. Absolute and relative judgments require reliable and valid measures and evaluations of training effects, which can be generated in many different combinations of ways (Orlanski, 1989). Before choosing the optimal approach, several basic questions always have to be answered first (see also Chapter 2):

- What is the objective of the effectiveness measurement or estimate? This may involve setting the budgets for acquisition of advanced training aids, comparing the benefits or utility of alternative solutions, comparing different types or different combinations of training, or just estimating training effectiveness or the benefits of one or more alternative training solutions.
- What are the external and internal constraints related to the estimate? These may include available time, personnel, and trainees, and limitations in the external resources required to support the study and measurements, etc. The possible internal constraints include infrastructure constraints, tools and technology, data availability, training system availability, personnel availability, limited suitable internal resources to carry out the study and its measurements, and financial assets.
- Which data are already available or easily made available (e.g., from records on incidents or accidents or time or numbers of trials needed to obtain the training objectives)?
- Which data are not yet available, but can be (easily) measured objectively or easily captured (e.g., data that can be gathered from trainees using questionnaires or paper-and-pencil tests)?

Answers on these questions are fundamental to decisions on how to proceed further and to choose among the different methods (e.g., quantitative, qualitative, checklists, questionnaires, etc.) to be used for measuring or estimating training effects. For example, a high number of constraints or a dynamic training environment may entail that only semi-objective and subjective evaluation methods are feasible. These alternatives and their advantages and disadvantages are discussed below.

6.4 TYPES OF EFFECTIVENESS MEASURES

There are many types of measurement methods for assessing the potential benefits or training effectiveness of the new training interventions and training environments (e.g., simulators, games or other kinds of training devices). Three of the main global categories are methods based on the objective, instrumented measurement of the trainees' (learning) performance, methods focusing on the structured observations or evaluations of the training environments or training programs (semi-objective or subjective), and ratings or questionnaires focusing on subjective evaluations by trainees, instructors, or training experts.

6.4.1 Objectively Measuring Trainee Performance

Direct quantitative measurements of training effects always have to be performed under carefully controlled conditions that are designed to avoid artefacts caused by, for example, repeated testing (so-called test-effect), selection, or instrumentation effects. In addition, the performance data should be representative of the task at hand in order to be relevant for real-task performance. Sometimes, it may be useful for these task variables to have clear performance criteria in order to obtain the most representative training results. The measurements have to be carried out properly and be standardized, which often has to be done in practical situations at schools and training sections. As Korteling, Oprins, and Kallen (2012) report, the following three common difficulties arise when one attempts to create standardized measurements.

6.4.1.1 Lack of Control

Measures may be hampered by rigid training schedules, lack of control over events, logistical constraints and circumstances, limited numbers of trainees available, or lack of access to the fielded systems (Cohn et al., 2009). Lack of control of all these factors may severely threaten the validity of inferences based on the objective measurements of performance (Boldovici, Bessemer, & Bolton, 2002). Usually, fulfilling all these basic requirements in combination can only be accomplished by creating a training program tailored to the experiment.

6.4.1.2 Measurement Problems

It is often hard to measure what and how much exactly is learned about the real task or job for which the training is intended. Real operational situations, and even many normal job situations, do not easily allow the objective measurement of former trainees' performance. Even when these real-world data can be collected, it remains questionable as to how much the training has contributed to their performance and how much performance effects can be attributed to other factors.

6.4.1.3 Limited Availability of Control Conditions

Finally, measurements of training effectiveness are traditionally performed after a new training device or method is fully instantiated in the training curriculum. As such, it has already replaced a legacy training system, making it hard to compare an experimental and control group (Cohn et al., 2009). In many other cases, there is no operational system or prototype available to allow empirical evaluation of its value in training.

6.4.2 Semi-Objective Structured Evaluation of the Training Intervention Program

As a result of the difficulties mentioned above, it is difficult to quantify performance-based training effects in training simulators or instructional games. This limits the feasibility of research designs that objectively compare an experimental group trained with new methods or devices with a control group trained on existing training equipment. Such experimental or quasi-experimental methods are only feasible if it is possible to control for confounding variables (e.g., Hawthorne effects). For these reasons, measurements should include other methods, such as structured evaluation by various stakeholders (e.g., experts, personnel, or students) and should be carried out in the context of the training program. Three other methods that evaluate training environments are presented below. These methods have been discussed by Korteling, Oprins, and Kallen (2012).

6.4.2.1 The Opinion Survey Method

Operators, instructors, training specialists, and even students can be interviewed about the effectiveness of a training method or device. For instance, questions are posed to determine which aspects of a simulator do or do not contribute to a high transfer of training. Opinion data often do not guarantee success because the subjects interviewed may have little or no expertise on learning or on the cues that facilitate learning. Therefore, the data may easily lead to erroneous conclusions about the properties required of the trainer under development (Caro, 1977).

6.4.2.2 The Simulator Fidelity Method

In this method, operational personnel (i.e., experts) compare the simulator in its physical, functional, and psychological aspects with the real system (e.g., comparison of the physical, perceptual, cognitive, or affective characteristics of both). Systematic analytical procedures have been developed for this model, and these assure the fidelity of both the stimuli the simulator presents to the trainee and his or her responses to these stimuli. This method is based on the assumption that when physical, functional, and psychological fidelity is high, transfer will also be high, and when fidelity is low, transfer will be low (Caro, 1977) or *negative transfer* may occur (i.e., counter-productive learning). Some investigators have argued that a simulator can be a faithful physical copy of the real-life system, but that this is not a conclusive overall statement about its ultimate effectiveness as a training tool (e.g., Adams, 1972; van Emmerik & Korteling, 2013).

6.4.2.3 The Simulator Training Program Analysis Method

The simulator training program analysis (STPA) method may be a substantial improvement over simulator fidelity testing, because this method determines whether a training program is well designed by analyzing the way a simulator is used. This analysis looks at whether the simulator is being used for its intended purpose. Questions are answered about whether or not training sessions are carried out in such a way that trainees optimally benefit from the advantages of simulation and how well the simulator sessions are connected to the training program as a whole. This can be done with standardized checklists (Caro, 1977), which can pinpoint factors limiting the effectiveness of a simulator under particular circumstances (or explain success). However, this method cannot be used to directly measure the extent of training effectiveness.

6.4.2.4 Combination of Multiple Perspectives in Semi-Objective Determinations

The main deficiencies of semi-objective evaluation methods lie in their one-sidedness (see Box 6.1). For instance, the simulator fidelity method does not take into account contextual factors, such as the didactical, motivational, and organizational aspects of training aids or, more specifically, the quality and completeness of the training program, instruction, and feedback, or aspects of game play. In contrast, the STPA method yields an outcome that may be unrelated to the real training outcome because the technical quality of the simulator is not sufficiently taken into account (i.e., on account of the various aspects of functional and physical fidelity). In general, training effectiveness is determined by the combination of the training environment and the didactical and organizational quality of the context in which the training system is embedded. It is therefore better to combine these different semi-objective, opinion-based evaluation methods.

Box 6.1: Examples of Semi-Objective Structured Evaluations or Comparisons of Alternatives

The quality of training simulations used by the Netherlands Armed Forces was recently evaluated using a comprehensive tool called TNO-CEES 3.0 (*TNO Checklist for the Evaluation of Educational Simulations*; van Emmerik & Korteling, 2013). This checklist included all relevant aspects of training simulation and consists of more than 350 items that have been defined on the basis of an extensive study of the relevant literature (e.g., Korteling, Padmos, Helsdingen, & Sluimer, 2001; van Emmerik & Korteling, 2002). The checklist included:

- the comprehensiveness of the specification, design, and validation process
- fidelity (e.g., models, databases, image presentation, motion, sound)
- didactics (e.g., user interface, scenarios and scenario-management, intelligent tutoring and other kinds of instructor support, testing and feedback facilities, and motivational features)
- organizational aspects (e.g., education of simulation personnel, technical support, knowledge management, acceptation, interoperability)

Relatively simple standalone simulations were first evaluated using this checklist (e.g., the CV90 Driving Simulator, the Small Arms Weapon Simulator, and a CV90 Turret Trainer). In addition, two large-scale networked battlefield simulators were evaluated (i.e., the virtual Tactical Indoor Simulator [TACTIS] and the life simulation Mobile Combat Training Centre [MCTC]). Next, two serious games were evaluated: *Virtual Battle Space 2* (VBS2) for dismounted infantry and *Steel Beasts Pro* for mounted infantry.

The results of this kind of evaluation may show minor and large deficiencies, but also may indicate training gaps and overlap between the different simulation systems used for training a certain military task. Apart from improving the quality of simulator use by the armed forces, this kind of method can be used for making quantitative cost-effectiveness comparisons between different training alternatives—for example, when comparing two alternative kinds of training simulations (full-scale battlefield simulation vs. PC-based simulation or gaming) to be used for similar purposes (training for intermediate or higher infantry training levels). Such comparisons demand careful weighing of scores on the many different and often partly overlapping or interdependent items.

A simplified spreadsheet tool (CONCERT) has been developed to weigh all scores on the many different measures, such that overlap between measures can be taken into account. Weighing also captures effects of the kind of simulator (or training task) and characteristics of the target group (van Emmerik & Korteling, 2003). These latter aspects are very important since they depends on the kind of task for which the simulator is intended (e.g., perceptual-motor or cognitive) to which degree a certain deficiency or possibility will impact the training outcome. For instance, physical fidelity is usually much more critical for perceptual-motor tasks, such as shooting with a hand weapon, than for tasks on a tactical or strategical level. Likewise, this impact may be highly affected by the level of proficiency of the trainees (e.g., initial versus recurrent training).

6.4.3 Subjective Ratings and Questionnaires

According to Korteling, Oprins, and Kallen (2012), the limited scope of opinion-based simulator evaluation methods may be extended by using structured ratings or questionnaires on the learning processes and competencies in the trainee. These methods focus more on the generic learning processes in the trainee than on the measurable learning performance of the trainees or the characteristics of the simulator or training program. This additional information may include ratings or questionnaires for instructors or trainees on knowledge and skills, self-efficacy, situational awareness, flow, stress, motivation, experienced problems that remain after finishing the training, etc. These aspects may provide useful insight into the underlying factors that determine human performance. These kinds of subjective data are less useful measures of the value or effectiveness of a training solution.

With any evaluation method, a distinction has to be made between process measures and outcome measures (e.g., Alvarez, Salas, & Garofano, 2004; Salas, Milham, & Bowers, 2003). Process measures examine the manner in which a task is performed by the trainee, whereas outcome measures focus on how well the trainee accomplishes the overall task. Process measures can be useful diagnostic tools explaining certain outcomes; i.e., why it happened, illustrating strengths and weaknesses of the training program or of the simulator that should be maintained, improved, or further developed to ensure that training goals are met (Cohn et al., 2009; Fowlkes et al., 1999). All the aforementioned methods (the subjective measures as well as the objective data, e.g., based on eye movements or neuro-physiological parameters) may focus on process or outcome measures (Korteling, Oprins, & Kallen, 2012).

6.5 DESIGNS FOR MEASURING EFFECTIVENESS

Various attempts have been made to design a generic evaluation model for learning interventions. The most popular evaluation model is Kirkpatrick's (1976, 1998; Kirkpatrick & Kirkpatrick, 2006), which comprises four levels (Figure 6.1) for evaluating a learning intervention: (1) direct *reactions* of the trainees, including their opinion on the attractiveness of the (new) learning intervention; (2) *learning effects* of the intervention, measured directly after education or training; (3) *transfer of learning* to another context (e.g., job performance); and (4) *results* for the organization (e.g., reduced cost, increased productivity, decreased injuries, etc.). Later on, a fifth level was added to measure *return on investment*, the costs and benefits of the learning intervention.

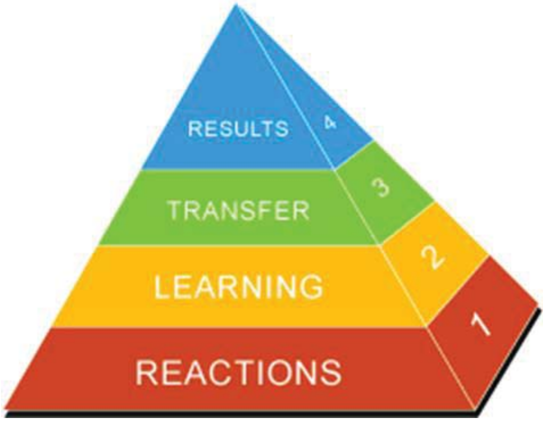


Figure 6.1: The Kirkpatrick Model

According to Cohn and colleagues (2009), many studies of training effectiveness fall short in measuring learning effects beyond Kirkpatrick's Levels 1 (reactions) and 2 (learning). Yet the virtue of this model is that it measures the effectiveness of learning interventions at different levels, and measurement goes beyond the outcomes observed immediately after the intervention. The model's popularity can also be attributed to its systematic approach to evaluation and to simplifying the complex process of evaluation

(Alliger & Janak, 1989; Bates, 2004).

We next describe research designs that can be used for all levels, but that are specifically suited for Level 3 (transfer of skills to the workplace) and Level 4 (results for the organization as a whole), which are the most relevant for a defence organization. Level 4 results of a training intervention are difficult to distinguish from other factors that shape organizational effectiveness, like human resource factors or quality of leadership. With that in mind, the following research designs can be used to measure the training value (e.g., benefits, utility, effectiveness, transfer) of training devices, such as simulators. The most common designs are described by Campbell and Stanley (1963). Their designs focus on comparing the effect of a treatment with that of no treatment in an experimental setting. Korteling, Oprins, and colleagues (2012) have translated these descriptions into seven experimental designs in which the training effects of new training solutions are compared to those of conventional training.

6.5.1 Experimental-Versus-Control-Group Method

The experimental-versus-control-group method uses an experimental and control group with randomly allocated subjects. The experimental group is trained with a new training solution, such as a simulator, and the control group is trained on real-task (or conventional) equipment only. Afterwards, task performance is measured on real-task equipment, using a predetermined criterion task resembling operational task performance. Preferably, performance is also measured before the training in order to get clear data on the actual learning performance of the trainees. The experimental-versus-control-group method is generally thought to be the most appropriate way to determine whether a new training solution has improved real-life performance (Caro, 1977). All other (quasi-experimental) methods may be susceptible to questions about their internal validity.

6.5.2 Self-Control-Transfer Method

According to this method, the experimental group is also the control group. Data are collected on subject performance on the real task and after real-task training. The same group then undergoes advanced training on a synthetic training device. Post-synthetic-training data are then compared to real task performance data. The difference between these datasets is attributed to the simulator. The major flaw in this design lies in the absence of a genuine control group. One cannot draw firm conclusions about the effectiveness of the training device because the effect of synthetic training is not compared to a control group that is completely trained on real-life equipment.

6.5.3 Pre-Existing-Control-Transfer Method

On this method, concurrent training of comparable groups is not necessary. Synthetic training is introduced into an existing training program. The job performance of employees trained prior to the introduction of the synthetic training can be compared to the performance of those trained in the new environment. Conclusions based on this method are tentative because of time-related changes (e.g., changes in the trainee group, training methods or circumstances, or the training staff).

6.5.4 Uncontrolled-Transfer Method

There are also circumstances where no control group exists. Such a condition can occur when safety plays a role (e.g., forced landing by an airplane). When no control group can be formed, the effectiveness of a training solution is established by determining whether subjects can perform the learned task on a real-life system the first time they perform it. This is called first-shot performance, and the method based on it is called the uncontrolled-transfer method. Data collected from such studies are tentative, since one cannot conclusively show that the new training has had an effect on the real-task operations performed by the subjects (Caro, 1977).

6.5.5 Quasi-Transfer-of-Training Method

The quasi-transfer-of-training method (QToT) is often applied in validating training systems because it is efficient. The difference between the experimental-versus-control-group method and the QToT method is that real-task training only occurs in the former (until criterion performance is reached). Experimental groups receive training in a new training environment, like a simulator or with an instructional game that has to be evaluated. The control group is trained on a fully operational high-fidelity simulator. Eventually, both groups are evaluated on a criterion task in this fully operational simulator. The difference in performance reveals the simulator's the contribution to learning results. Of course, the major limitation of this design is the absence of training and performance measurement under real-task (i.e., operational) conditions.

6.5.6 Backward-Transfer Method

In a backward transfer study, a proficient operator performs the task in the new training environment (e.g., simulator or instructional game). If he or she can perform the task on the synthetic device, backward transfer has occurred. The assumption here is that transfer of training in the other direction (forward transfer) for trainees who have been training on the simulator will also occur.

6.5.7 Simulator-Performance-Improvement Method

On the simulator-performance-improvement method, the performance of a trainee is measured across a number of successive sessions. The working assumption is that the trainee should improve over several sessions of training if the new training program is effective. If this does not occur, there would be little expectation of improvement in executing the real task. Improvement in learning in the training simulator or game, however, does not necessarily mean that what is learned is relevant and, thus, can be transferred to the real, operational-task environment. In general, the assumption of transfer is only plausible if the training environment has a high degree of physical, functional, and psychological fidelity to the real-task environment (Korteling, Helsdingen, & Theunissen, 2012).

6.6 SOME OVERALL CONSIDERATIONS

Except for the experimental-versus-control-group method, these quasi-experimental methods are susceptible to questions about their internal validity, which means they have major limitations for drawing certain conclusions about outcome effects on performance of training manipulations. Generally, a strictly controlled experiment permits strong inferences about the behavioural effects of training interventions (i.e., high internal validity). However, it is often difficult to execute these experiments in practical settings, and the degree to which the results can be generalized will be lower (i.e., low external validity). While quasi-experiments may lack rigorous controls, they allow researchers to apply a more realistic context (Korteling, Oprins, & Kallen, 2012) and, thus, the results have a higher external validity—i.e., their results are easier to generalize to operational and on-the-job settings.

Opinion-based evaluation measures—surveys, ratings, questionnaires, and checklists—may provide a lot of information about the underlying learning processes and the intervening factors that may explain training outcomes (Alvarez, Salas, & Garofano, 2004; Salas, Milham, & Bowers, 2003). Compared to direct and objective measurements of training and simulator performance, these measures are more easily applied when complete experimental control is difficult to achieve. However, these more subjective methods may provide limited or false information about the effectiveness of a particular training method or environment. They may reflect personal opinions, expectations, biases, or preferences, instead of measuring the effectiveness of training. Indeed, the ability to report experiences varies substantially between subjects (Sander, Grandjean, &

Scherer, 2005). In addition, experts often have preconceived opinions about new methods or equipment that may compromise their objectivity, and professional crews working with training simulators mostly have some interest in the outcomes of evaluation studies. All these factors may degrade the reliability and the internal validity of opinion-based results. When relying solely on these subjective evaluation methods, it is important to use as many objective and "blind" procedures as possible, meaning tests that do not allow individuals to identify desired response behaviours. One should also limit the use of retrospective reports and avoid disrupting an operator performing his or her task in order to ask questions.

In short, surveys, questionnaires, ratings, and checklists are best used in combination with quantitative measurements (e.g., time, speed, error) of trainee performance. A combined approach may provide the best mix of reliable and relevant information on training effectiveness in a relatively pragmatic way. In general, sound conclusions of effectiveness studies require multiple sources of data converging on the same outcomes. Examples of such converging measurement methodologies are described by Bell and Waag (1998), Schreiber and Bennett (2006), Schreiber, Bennett, and Stock (2006), and Schreiber, Schroeder, and Bennett (2011). For example, Schreiber and colleagues (2011) employed three major types of measurements that represented combat skills in flight training: *objective data* to quantify training effectiveness by measuring improvements on outcomes and skill proficiency (e.g., enemy strikers reaching their target, number of mortalities); *expert observation data* providing assessment of competency; and *user opinion data* that captured opinions on the usefulness of the training system, its pros and cons, and which tasks are best suited for the system. Of course, all data should reflect the crucial and relevant skills and competences (i.e., mission essential competences, or MECs) of the operational task (Alliger et al., 2012). For example, objective data should preferably be adopted from on-the-job training performance that mimics operational situations, or they must be determined as much as possible on the basis of real operational task conditions.

In many cases, all relevant data cannot be measured. Estimates of effectiveness will have to be made. These estimates have to be based on assumptions that are tested or substantiated by data. The effectiveness estimate should be broken down into relevant aspects, such as required training time, type of competences or performance, level of competence, savings (e.g., equipment, personnel), or operational output. The complete results (i.e., the measured and estimated training effectiveness) become inputs in the estimation of training benefits.

6.6.1 Example

Training evaluation requires careful selection and weighing of the measures. Using several methods to evaluate the effectiveness of a training system is usually beneficial because it minimizes the impact of the artefacts and limitations of each individual method. For instance, it is not very difficult to apply several methods in one transfer-of-training study. Interviewing trainees or instructors about the synthetic training is relatively easy to do. Although interviews alone are insufficient, they may result reinforce the conclusions of the study. On the other hand, employing too many measures in an effort to capture training effects as accurately as possible may also be detrimental. Multiple measures can lead to problems because of interdependences among them. Separating the effect of one measure from those related to it can be difficult, and the importance of individual measures can be diluted. Where one skill or competence impacts two different measures, for example, it is possible to measure the same training effect twice, doubling the theoretical significance of its impact.

An example of a multi-method and multi-level approach is described in Box 6.2, which describes a study conducted by TNO to validate a tank driving simulator (Moraal & Poll, 1979). The preferred experimental design for this validation study was the experimental-versus-control-group method, where an experimental group is trained with a new method or tool; for example, a training simulator (i.e., advanced or synthetic training). After a period of time, the group receives additional training in the real-task environment (e.g., on-the-job training), until the real-task performance of this group reaches a predetermined level. The time needed for the experimental group to reach criterion performance is then compared with the time needed by a control group that has been conventionally trained only on the real task or with on-the-job training. Box 6.2 shows how the percentage of transfer and the training effectiveness ratio can be calculated on the basis of this design.

In this study (Moraal & Poll, 1979), the researchers were still not sure about the validity of the outcome of the experiment; therefore, they introduced a second method. A group of experienced drivers was asked to execute the same tasks as the trainees on the simulator (the backward transfer method). The experienced drivers were then interviewed about their opinions of the simulator (simulator fidelity method). In order to get more insight into the underlying factors that contributed to the quantitative results of the study, instructors and students were also interviewed. These interviews provided more insight into the strengths and weaknesses of conventional versus simulator training. All this was done to ensure a solid outcome of the experiment, with useful suggestions to improve the effectiveness of simulator training and reduce cost.

Box 6.2: Validation of the Link-Miles Leopard 2 Driving Simulator

Using the experimental-versus-control-group method, one group was trained on the driving simulator and one on the real tank. After acquiring a certain level of performance in the simulator, the experimental group had to train to reach a predetermined level on the real tank. On the basis of the training time on the real tank needed for both groups to reach this performance criterion, transfer of training was calculated. The basic computation for percentage of transfer, %*T*, is:

$$\%T = \frac{T_c - T_e}{T_c} \times 100\% \qquad \text{(Equation 1)}$$

$\boldsymbol{T_c}$ Time needed for (conventional) on-the-job training by a control group to reach the criterion level.

$\boldsymbol{T_e}$ Time needed for this conventional training by the experimental group after training with a new, advanced training system (e.g., synthetic training).

From Equation 1, it can be derived that when ***T*** of a training program using an advanced training system is 100%, no additional field training is needed by the experimental group to reach the same criterion performance as the control group. When $\boldsymbol{T_e}$ increases, ***T*** decreases; hence, when ***T*** is 0%, training with the new system does not produce any effect. ***T*** can even become negative. Negative transfer means that training with the new system (e.g., a simulator) interferes with the development of proper performance.

However, this percentage of transfer formula does not account for the amount of practice that the experimental group received within the synthetic training environment. Because the *T*-formula ignores the amount of synthetic training prior to on-the-job training, it permits no conclusions about the *effectiveness* of the simulator as a training tool (Roscoe & Williges, 1980).

Continued on the next page…

An adequate measure, one that incorporates the time spent in the simulator, is the transfer effectiveness ratio (TER). The computation for TER is:

$$TER = \frac{T_c - T_e}{T_s} \qquad \text{(Equation 2)}$$

where,

$\boldsymbol{T_c}$ Time needed for conventional training by a control group to reach the criterion level.

$\boldsymbol{T_e}$ Time needed for conventional training by the experimental group after completing synthetic training with the advanced training system.

$\boldsymbol{T_s}$ Synthetic training time by the experimental group.

A TER of 1.0 indicates that time savings on conventional on-the-job training are equal to the amount of time spent training in the advanced, synthetic training environment. When TER is larger than 1.0 (T_s + T_e is smaller than T_c), synthetic training is more effective than training in the conventional, on-the-job training environment. When TER is lower than 1.0 (T_s + T_e is higher than $T_{c)}$, conventional on the job training is more effective.

6.7 ESTIMATING EFFECTIVENESS

Relevant and reliable data on comparable cases are sometimes lacking or difficult to collect and measure (e.g., data on a training prototype). In these cases, effectiveness estimations have to be made: the analyst must estimate the effectiveness of a new training system against some alternatives in order to determine the most effective solution, given the boundary conditions of the military training system. Accordingly, this section focuses on the estimation of performance outcomes: the training outcome or the training effectiveness of alternative training solutions. The previous chapter also describes methods to estimate cost when real cost data are not available. As can be seen, there is substantial overlap between possible methods for benefit and cost estimation.

6.7.1 Expert judgment

On this method, the experience and knowledge of experts is used to estimate the effectiveness of military training or to rank the effectiveness of various alternatives. The judgment is based on the specific expertise of an individual or a group. Sometimes these are people from the project team; however, expert judgment typically requires an expertise that is not present in the project team. In this case, external experts are brought in to consult. This technique is used when other methods are not feasible (data are limited or not available) or an additional validation is required. Generalization and induction are two techniques that can be used by the experts (Basnett, Medhurst, & Irwin, 2012).

6.7.2 Analogous Estimating

Sometimes called top-down estimating, this method is used when information about a new training system is limited. The method relies on expert judgment based on historical data from similar training. Historical data is used to create a generic set of characteristics (e.g., performance of a specific task) that are then used to evaluate various training options. The results are not as accurate as other estimation techniques because the characteristics of the analogous training will not completely match. However, results provide a ballpark figure for estimating the effectiveness of the training.

6.7.3 Parametric Estimating

The estimated effectiveness is statistically determined using historical data about the relationships between certain characteristics of a training environment (e.g., a moving-base for a simulator or a training program based on educational gaming), and it measures training outcomes associated with these components or characteristics. The accuracy of the estimate depends on the quality of the data.

6.7.4 Engineering or Bottom-up Method

This method is based on detailed estimates or measures of the quality of individual elements of a training solution. These separate elements are summed to obtain an estimate of the training outcome. This is done by summing the effects of elements of the training solution (e.g., effects of a motion characteristic of a moving-base) on different part-tasks (e.g., driving sharp curves) of the overall task that has to be learned. The engineering or bottom-up method is the most detailed of all techniques.

6.7.5 Three-Point Estimate

The Program (or Project) Evaluation and Review Technique (PERT), is a statistical mathematics tool, used in general project management, which was designed to analyze and represent the tasks involved in completing a given project. It is a method for analyzing the tasks involved in completing a given project, especially the time needed to complete each task, and to identify the minimum time needed to complete the total project. PERT was developed primarily to simplify the planning and scheduling of large and complex projects. It is able to incorporate uncertainty by making it possible to schedule a project while not knowing precisely the details and durations of all the activities. It is more of an event-oriented technique than a start- and completion-oriented one, and it is used more in projects where time rather than cost is the major factor. It is applied to very large-scale, one-time, complex, non-routine infrastructure and research and development projects.

The PERT method can be used for estimating effectiveness by constructing an approximate probability distribution. The method makes use of three estimates to determine the bandwidth. The three estimates include a “most likely,” an “optimistic,” and a “pessimistic” scenario. An average is then calculated. It applies weighting so that the most-likely estimate is weighted four times more than the other two estimates (i.e., optimistic and pessimistic). For training that is similar to previous training, and where there is good historical data and expert experience, the formula is less useful because one could use other techniques, like parametric estimating.

It is important to understand the applicability and boundaries of each method in order to use them appropriately. In the early phases, user requirements and training development knowledge of a new system and its application for training may be limited; thus data on training effectiveness may be lacking or difficult to acquire. In these cases, expert judgment and analogy will be most suitable. At later stages of the lifecycle (e.g. when the system is in use and decisions on maintenance and upgrades have to be made) a more detailed method, like parametric estimating or the engineering method, will then be very suitable.

6.7.6 Benefit- and Cost-Reduction Drivers

Some elements or characteristics of a training solution may have a high impact on training effectiveness and efficiency, depending on the training goals, target group, and other intervening circumstances. Examining these benefit drivers may shed light on how training effectiveness is affected by the characteristics of training alternatives (e.g., in “what-if” analyses).

Effectiveness drivers that have a relatively high impact on training effectiveness include:

- Flexible training scenario management facilities
- Adaptive and personalized training fitted to the needs of the individual
- Better performance monitoring and feedback and better facilities for after-action review
- Proficiency of training staff with effective usage of an advanced training system ("train the trainer")
- Awakening awareness or encouraging interest in new challenges or initiatives (e.g., by educational gaming)

These effectiveness or benefit drivers may be distinguished from the so-called cost-reduction drivers (i.e., factors that may produce a relatively large decrease in the cost of training). For instance, the cost-reduction drivers of educational simulations are the most prominent practical advantages of simulation: simpler logistics, maintenance, safety, sustainability, and automation (Korteling & van den Bosch, 2014). Examples of such practical advantages may be:

- Reduced costs of transportation and/or accommodation
- Better system availability, less attrition
- Less costs related to risk-prevention and/or pollution
- Intelligent tutoring, automatic performance measurement, and feedback facilities
- Use of artificial intelligence and computer-generated forces instead of real persons
- Education and training in leisure time (educational gaming)

Box 6.3: Summary

When measuring or estimating training effectiveness, some basic information has to be collected first. This includes the objective of the effectiveness study, any external and internal constraints, and the availability of relevant data. When effectiveness measurement based on trainee (learning) performance is chosen, the most representative human performance variables of training effectiveness or transfer have to be chosen—taking into account Kirkpatrick's levels of the data. Then, if possible, performance criteria for obtaining training goals have to be clearly defined. When relevant human performance aspects (e.g., process, outcome) are not easy to obtain or measure objectively and should be inventoried otherwise (e.g. by checklists or questionnaires). These checklists and questionnaires can focus on (1) the training device, (2) the training context, or (3) subjective reactions of the trainees or training staff.

It is also recommended, however, that more insight into the *process* aspects of the training be sought out. This can be done using questionnaires or observation checklists, but also by more objective behavioural measures. In the latter case, use an experimental-to-control-group design should be used—if this is feasible—to objectively compare transfer of training in different training solutions; e.g., simulation versus field training. If such an experimental-to-control-group design is not feasible, try to include a quasi-experimental research design in the study. When conducting a (quasi-)experimental design, confounding variables that may disturb drawing sound conclusions (artefacts) need to be taken into account. Always try to combine multiple methods (i.e., objective, subjective) and measures of the outcome of processes. In the case of subjective methods, consider that the objectivity of people involved in the study may be compromised by preconceived notions about, or an interest in, the outcomes of new methods or equipment. If measurement of training effects is not feasible, try to estimate the effectiveness of a military training solution before funds are committed.

6.8 REFERENCES

Adams, J. A. (1972). Research and the future of engineering psychology. *American Psychologist*, *27*(7), 615–622.

Alliger, G. M., Beard, R., Bennett, W., Jr., & Colegrove, C. M. (2012). Mission essential competencies: An integrative approach to job and work analysis. In M. J. Wilson, W. Bennett, Jr., S. G. Gibson, & G. M. Alliger (Eds.), *The handbook of work analysis in organizations: The methods, systems, applications, & science of work measurement in organizations*. Mahwah, NJ: Taylor Francis.

Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personnel Psychology*, *42*(2), 331–342.

Alvarez, K., Salas, E., & Garofano, C. M. (2004). An integrated model of training evaluation and effectiveness. *Human Resource Development Review*, *3*(4), 385–416.

Basnett R., Medhurst J., & Irwin, C. (2012). *Application of heuristics to high level operational analysis* (Report no. CR58929). Oxford, UK: DSTL.

Bates, R. (2004). A critical analysis of evaluation practice: The Kirkpatrick model and the principle of beneficence. *Evaluation and Program Planning*, *27*, 341–347.

Bell, H. H., & Waag, W. L. (1998). Evaluating the effectiveness of flight simulators for training combat skills: A review. *The International Journal of Aviation Psychology*, *8*(3), 223–242.

Boldovici, J. A., Bessemer, D. W., & Bolton, A. E. (2002). *The elements of training evaluation.* Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Campbell, D.T., & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Caro, P.W. (1977). *Some factors influencing air force simulator training effectiveness* (Report No. TR-77-2). Alexandria, VA: Human Resources Research Organization.

Cohn, J., Kay, S., Milham, L., Bell Carroll, M., Jones, D., Sullivan, J., & Darken, R. (2009). Training effectiveness evaluation: From theory to practice. In D. Schmorrow, J. Cohn, & D. Nicholson (Eds.), *The PSI handbook of virtual environments for training and education* (pp. 157–172).

Emmerik, M. L. van, & Korteling, J. E. (2002). *Certificering van trainings simulatoren 2: De TNO-TM checklist [Certification of training simulators 2: The TNO-TM checklist]* (Report No. TM-02-D010). Soesterberg, The Netherlands: TNO Human Factors Research Institute.

Emmerik, M. L. van, & Korteling, J. E. (2003). *Certificering van trainingssimulatoren 3: Computer-gebaseerd ONdersteuningmiddel voor CERtificering van Trainingssimulatoren [Certification of training simulators 3: Computer-based support tool for certification of training simulators]* (Report No. TM-03-D005). Soesterberg, The Netherlands: TNO Human Factors Research Institute.

Emmerik, M. L. van, & Korteling, J. E. (2013). *Optimale inzet van simulatiemiddelen voor O&T bij Defensie [Optimal use of simulation devices for education and training for the Defence]* (Report No. TNO-R10590). Soesterberg, The Netherlands: TNO Behavioral & Societal Sciences.

Fowlkes, J. E., Dwyer, D. J., Milham, L. M., Burns, J. J., & Pierce, L. G. (1999). Team skills assessment: A test and evaluation component for emerging weapon systems. *Paper presented at the Interservice/Industry Training, Simulation, and Education Conference*, Orlando, FL

Kirkpatrick, D. L. (1976). Evaluation of training. In R. L. Craig (Ed.), *Training and development handbook: A guide to human resource development* (pp. 18–27). New York: McGraw Hill.

Kirkpatrick, D. L. (1998). *Evaluating training programs: The Four Levels* (2nd ed.). San Francisco, CA: Berrett-Koehler.

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs*: *The Four Levels* (3rd ed.). San Francisco, CA: Berrett-Koehler.

Korteling, J. E., & Bosch, K. van den (2014). Effectiviteitsfactoren van simulatiemiddelen. *[Effectiveness factors of simulation systems]*. Soesterberg, The Netherlands: TNO Behavioral & Societal Sciences.

Korteling, J. E., Helsdingen, A. S., & Theunissen, N. C. M. (2012). Serious games at work: Learning job-related competencies using serious gaming. In A. Bakker, & D. Derks (Eds.), *The psychology of digital media at work* (pp. 123–144). London, UK: Taylor & Francis.

Korteling, J. E., Oprins, E. A. P. B., & Kallen, V. L. (2012). Measurement of effectiveness for training simulations. In Z. Wang (Ed.), *Cost-benefit analysis of military training* (Report No. MP-SAS-095). Amsterdam, The Netherlands: NATO RTO.

Korteling, J. E., Padmos, P., Helsdingen, A. S., & Sluimer, R. R. (2001). *Certificering van trainingssimulatoren 1: kennisinventarisatie [Certification of training simulators 1: Knowledge inventarization]* (Report No. TM-01-D003). Soesterberg, The Netherlands: TNO Human Factors Research Institute.

Moraal, J., & Poll, K. J. (1979). *De Link-Miles rijsimulator voor pantservoertuigen; verslag van een validatie-onderzoek [The link-miles driving simulator for armoured vehicles: Report of a validation experiment]* (Report No. IZF 1979-23). Soesterberg, The Netherlands: TNO Institute for Perception.

Orlansky, J. (Ed.). (1989). *The military value and cost effectiveness of training.* Panel 7, on the Defence Applications of Operational Research, RSG-15 on the Military Value and Cost Effectiveness of Training, AC/243 (Panel 7/RSG-15) D/4. Brussels, Belgium: NATO Defence Research Group.

Roscoe, S. N., & Williges, B. H. (1980). Measurement of transfer of training. In S. N. Roscoe (Ed.), *Aviation psychology* (pp. 182–193). Ames: Iowa State University Press.

Salas, E., Milham, L. M., & Bowers, C. A. (2003). Training evaluation in the military: Misconceptions, opportunities, and challenges. *Military Psychology*, *15*(1), 3–16.

Sander, D., Grandjean, D., & Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural Networks*, *18*(4), 317–352.

Schreiber, B. T., & Bennett, W., Jr. (2006). *Distributed mission operations within-simulator training effectiveness baseline study: Summary report* (Report No. AFRL-HE-AZ-TR-2006-0015, Vol. 1). Mesa, AZ: Warfighter Readiness Research Division.

Schreiber, B. T., Bennett, W., Jr., & Stock, W.A. (2006). *Distributed mission operations within-simulator training effectiveness baseline study: Metric development and objectively quantifying the degree of learning* (Report No. AFRL-HE-AZ-TR-2006-0015, Vol. 2). Mesa, AZ: Warfighter Readiness Research Division.

Schreiber, B. T., Schroeder, M., & Bennett, W., Jr. (2011). Distributed mission operations within-simulator training effectiveness. *The International Journal of Aviation Psychology, 21*(3), 254–268. doi: 10.1080/10508414.2011.582448

Wang, Z. (Ed.) (2012). *Proceedings of NATO Workshop on Cost-Benefit Analysis of Military Training*. North Atlantic Treaty Organization [NATO] Research & Technology Organisation [RTO].