Measurement of Effectiveness for Training Simulations

Dr. J.E. (Hans) Korteling, Dr. E.A.P.B. (Esther) Oprins, Dr. V.L. Kallen
TNO Human Factors, P.O. Box 23, 3769 ZG, Soesterberg, The Netherlands
hans.korteling@tno.nl

Abstract

This paper presents and discusses experimental designs, measures, and measurement methods for determining the effectiveness of training simulators. First, we describe experimental designs in which training effects of training simulators are compared to those of conventional training. Next, the most commonly used metrics for quantifying the potential beneficial effects of training applications are explicated. We also present and discuss three main categories of measurement methods that may be used to assess the beneficial effects of new ways of training on transfer or training effectiveness; that is, methods based on measurement of learning performance of trainees, methods focusing on the synthetic training device or overall training program itself, and ratings or questionnaires focusing on the subjective evaluations of trainees. All designs, metrics, and measurement methods have their specific advantages and limitations, which may make them highly complementary. In general, one should always be aware of the advantages and drawbacks of each method and consider the most appropriate combination of methods for each study, given the main research questions. Therefore, various types of measurement techniques should be used in combination with each other for effective results in order to meet reliability and validity requirements of training effectiveness studies. Finally, we give some examples of practical approaches and draw conclusions on best practices.

1 Introduction

People have long used synthetic environments to simulate reality for training purposes. The first instances of these “training simulators” or “simulations” were primitive and mainly used for flight training before and during World War 1. Since then, training simulators have expanded to include transport and vehicle systems, military sensor and weapon systems, and complex processes, such as military combat, genocide, and the training of first responders (Schollmeyer, 2006).

Training simulators are here defined as devices generating a synthetic and interactive environment on the basis of mathematical, physical, or logical models representing (aspects of) the real (operational) world, in order to obtain training goals. The most prominent examples of training simulators are training simulations, instructional computer games, and simulations in e-learning environments. These synthetic training environments may have many potential advantages over real-task equipment and, if adequately designed and used, can substantially reduce the cost of military training and enhance training effectiveness. Since training simulators may also require substantial investment, not only in the device itself but also in instructional personnel and infrastructure, an important question concerns the degree to which these systems are beneficial and cost-effective.

The benefit and cost effectiveness of training systems, however, is often only partially investigated (Cohn, et al, 2009). Many studies have been limited in their design. They focus, for instance, only on fidelity of the simulation (e.g., Allan, Hays & Buffardi, 1986), or on trainee reactions or subjective opinions of experts or instructors (e.g., Stehouwer et al., 2005). A meta-analysis by Alliger et al. (1998) has shown that evaluations of training by trainees do not correlate strongly with subsequent performance on the job. In addition, results of many studies are confounded by characteristics of the educational context in which the simulator is embedded, including the target group, instructional practices, and preconceived opinions of personnel. In addition, not all training simulations, or simulated instructional games, seem to live up to their potential. With regard to instructional games, Hays (2005) has reviewed 48 empirical research articles on training effectiveness. This report also includes summaries of 26 other review articles and 31 theoretical articles on instructional gaming. The relevant major conclusions and recommendations from their report are as follows:

- Empirical research on the instructional effectiveness of games is fragmented, filled with ill-defined terms, and plagued with methodological flaws.
- Some games provide effective instruction for some tasks some of the time, but these results may not be generalizable to other games or instructional programs.
- No evidence exists that games are the preferred instructional method in all situations.

On the basis of a meta-analysis of the literature on instructional effectiveness of computer-based games, Sitzmann (2011) has drawn similar overall conclusions.

The objective of training effectiveness measurement is acquiring knowledge concerning the degree to which a training system accomplishes the purposes for which it was developed. These purposes are usually related to real job activities and organizational goals, and thus lie outside the training game or simulation itself. In many studies, measurements focus on the experiences and opinions of users or trainees with regard to the effectiveness of a training device. Next to that, many measurements and evaluations focus on learning and trainee performance in the training device itself. Both approaches lack the measurement of *real* transfer and retention of training results to the workplace, i.e., the situation for which the training was actually intended. Without such measurement, one remains ignorant of the tool's effectiveness or quality with regard to enhanced task performance in the operational environment and its organizational impact (Cohn et al, 2009; Kirkpatrick 1959, 1998). In addition, the factors responsible for the tool's success or failure will remain obscure.

At present, there is no generic framework predicting either the amount of transfer of training or the cost of obtaining it. Generic knowledge is sparse on the determining factors, as affected by situational factors (such as target group or type of task). This, however, is crucial for decision makers who have to decide about the purchase and application of training simulators. Therefore, in the present paper we will focus on the *designs*, *measures*, and *methods* available to determine the quality or effectiveness of training simulators in comparison with more conventional training environments, such as military training in the field and/or (embedded) training using operational equipment. In the domain of learning and modeling and simulation (M&S) for training purposes, the concepts of training output are often captured in the term *transfer*. Transfer denotes the ability to flexibly apply (parts of) what has been learned to new tasks and/or new situations, i.e., real world tasks on the job [see, e.g., Baldwin and Ford (1988); Detterman & Sternberg, 1993; Gielen (1995); Mayer & Wittrock, (1996)]. The degree to which training leads to enhancement of actual behavior on the job is the gold standard of training (Alvarez et al, 2004). In this contribution, we will discuss experimental designs, time- and cost-based measures, and evaluation instruments (questionnaires, checklists) for determining the benefit of training simulators relative to other forms of training.

2 Experimental designs for studies on transfer of training

2.1 Seven designs

This section describes designs that can be used in measuring the training effectiveness of training simulators. The descriptions are based on the most common designs, first described by Campbell and Stanley (1963). These designs focus on the comparison of the effect of a treatment with that of no treatment in an experimental setting. Below we have translated these designs into experimental designs, in which the training effects of training simulators are compared to those of conventional training.

Experimental-versus-control-group method
The experimental-versus-control-group method uses a design in which the experimental group is trained with the simulator and the control group is trained on real-task equipment only. Afterwards, task performance is measured on real-task equipment on a predetermined criterion task resembling operational task performance. Preferably, performance is also measured before the training (pre-) to get clear data on the actual learning performance of the trainees. In this case, the experimental-versus-control-group method is generally thought to be the most appropriate study design to determine whether the simulator has improved real-life performance (Caro, 1977).

Self-control-transfer method
According to this method, the experimental group is also the control group. A group of subjects already receiving real-task training would train for a given time on a simulator. Data from subject performance on the real task before synthetic training started are obtained. These data are compared to data of performance obtained on the real task after synthetic training. The difference between these datasets could be attributed to the simulator. The mayor flaw in this design lies in the absence of a genuine control group. One cannot draw hard 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.

Pre-existing-control-transfer method
There are studies in which a concurrently trained control group might not be necessary. For instance, synthetic training can be introduced in an existing training program. Learner performance data from the older or on a predetermined criterion task can be compared to data of performance by the new experimental group who trained with the simulator. This method is called the pre-existing-control-transfer method. Conclusions based on this method are tentative because of time-related changes; for example, in the trainee group, in training methods or circumstances, or in the training staff.

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 training effectiveness of the simulator can be established by determining whether subjects can perform the learned task on a real-life system the first time they perform this task. This is called first-shot performance and the method that is based on this kind of measurement is called the uncontrolled-transfer method. Data collected from such studies are tentative, since it cannot be conclusively shown that simulator training has had an effect on the real-task operations performed by the subjects (Caro, 1977).

Quasi-transfer-of-training method
Because of efficiency (or financial) reasons, one method often applied in validation of training simulators is the quasi-transfer-of-training method (QToT). The difference between the experimental-versus-control-group method and the QToT method is that real-task training occurs in the former (until criterion performance is reached), while it does not in the latter. Experimental groups receive training in the simulator or with the 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 contribution in learning results of the simulator to be evaluated relative to the high-fidelity simulator. Of course, the major limitation of this design is the absence training and performance measurement under real-task conditions.

Backward-transfer method
In a backward transfer study, an operator who has already shown sufficient performance on the relevant task has to perform in the simulator or serious game. If he 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 learners who have been training on the simulator will also occur.

Simulator-performance-improvement method

In the simulator-performance-improvement method, the performance of a learner is measured in a number of subsequent sessions. An essential premise of an effective synthetic training program is improvement in performance by the learners over several sessions of training. If this does not occur, there would be little expectation of improvement in executing the real task. However, the existence of learning in the training simulator or game does not necessarily mean that what is learned is relevant and, thus, transferred to the real, operational-task environment. In general, the assumption of transfer is only plausible if the training environment is a high-fidelity environment with a high degree of physical, functional, and psychological fidelity (similarity) with regard to the real-task environment (Korteling, Helsdingen & Theunissen, 2012).

2.2 Discussion

Except for the experimental-versus-control-group method, all other (quasi-experimental) methods may be susceptible to questions about their internal validity. This means that these methods have major limitations for drawing certain conclusions about effects on performance of training manipulations. Generally, strictly controlled experiments permit strong inferences about the effects. However, it is generally difficult to execute these experiments in practical settings and the degree to which the results can be generalized will be lower (low external validity). While quasi-experiments may be susceptible to threats of internal validity because of less rigorous control, they allow the researcher to apply less controlled and more realistic contexts (see also discussion in Section 4.4) and, thus, have a higher external validity, i.e., better generalization, or translatability of results, to operational settings.

3 Time- and cost-based metrics

3.1 Four metrics

Here we present four metrics that have been introduced previously to quantify the potentially beneficial effects of training applications, mainly by Roscoe (e.g., Roscoe & Williges, 1980). These metrics can be adopted for use in the determination of transfer of training, training effectiveness, and cost-effectiveness in training simulators. The type of experimental design needed to do that is the experimental-versus-control-group method, which was presented above as the most appropriate study design to determine whether the training has improved real-life performance (Caro, 1977). In this kind of experiment, an experimental group is trained with a simulator. After a certain period of time, the group receives additional training in the real-task environment, until the real-task performance of this group reaches a predetermined criterion level. The time needed for the experimental group to reach this criterion performance is then compared with the time needed by a control group that has been trained only on the real task. The basic computation for T (percentage of transfer) is

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

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

$\mathbf{T_e}$ Time needed for on-the-job training by the experimental group after training with a simulator or serious game.

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

For (expensive) training simulators, this percentage of transfer formula has a large flaw, because it fails to consider the previously provided amount of practice with the training environment by the experimental group. Because the percentage of transfer 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).

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

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

where,
T_c Time needed for on-the-job training by a control group to reach the criterion level.
T_e Time needed for on-the-job training by the experimental group after completing synthetic training.
T_s Synthetic training time by the experimental group.

A TER of 1.0 indicates that time savings on training for the real task are equal to the amount of time spent training in the 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 on the real task. When TER is lower than 1.0, real-task training is more effective.

One should keep in mind that there is a maximum on the transfer of training in most training simulators, because all the skills needed on the real task cannot be learned on a simulator. Therefore, TER is a negatively decelerated function of the training time with simulator or serious game.

A measure for expressing the effectiveness of financial training cost has also been developed, because synthetic training in general is less costly than real-task training. It is expressed via the cost effectiveness ratio (CER), which is a ratio of TER and the training cost ratio (TCR). The computation for TCR is

$$\mathbf{TCR} = \frac{\mathbf{C_s}}{\mathbf{C_c}} \qquad \text{(Equation 3)}$$

where,
C_s financial cost of simulator or game training (per time unit).
C_c financial cost of control group training (per time unit).

The formula for the CER is as follows:

$$\mathbf{CER} = \frac{\mathbf{TER}}{\mathbf{TCR}} = \frac{\mathbf{C_c\,(T_c - T_e\,)}}{\mathbf{T_s \times C_s}} \qquad \text{(Equation 4)}$$

For different durations of synthetic training, CER, TER, as well as T will change. A fictional example illustrates this:

A control group needs 20 hours of on-the-job training to reach the predetermined

criterion level on a given task. After completing 8 hours of simulator training, an experimental group only needs 16 hours of additional on-the-job training to reach the criterion level. In that case,
T = 20%
TER = 0.50

Suppose that operating cost of simulator training has been determined to be 15% of costs associated with the real-task equipment.
TCR = 0.15
CER = 0.50/ 0.15 = 3.33

If only 15 hours of additional on-the-job training are needed in another situation, where the experimental group gets 11 hours of simulator training
T = 25%
TER decreases to 0.45
CER = 0.45 / 0.15 = 3
Cost-effectiveness is still achieved.

Cost-effective training can be achieved with CER values above 1.

3.2 Discussion

The pure transfer-of-training measure (T) provides information about the amount of transfer in terms of training time savings, but without considering the amount of training effort to obtain this result. So, it fails to consider the previously provided amount of practice with the simulator or serious game by the experimental group and, thus, precludes conclusions about the *effectiveness* of the simulator as a training tool. This flaw is especially relevant for expensive, high-end simulations. For serious gaming or simple PC simulations, this is less of a problem, since these kinds of training simulators are considered cheap. Moreover, playing instructional games may be entertaining and, therefore, done in private or during spare time. The transfer effectiveness ratio (TER) is an adequate measure, which reckons with the time spent in the simulator. It takes into consideration the amount of synthetic training time required to obtain the time savings in conventional (on-the-job) training to reach the training objectives. The cost effectiveness ratio (CER) goes a step further by also taking into account the cost of simulator versus conventional training.

One important nuance should be mentioned here with regard to the time- and cost-based measures. Outcomes showing limited values for T or a TER value below 1 may indicate that real-task training is more effective or efficient. However, this does not necessarily mean that simulator training has little added value. Although it may not be as effective or efficient as training in real, on-the-job training settings, simulated training environments can still be cost-efficient or valuable for various other reasons:

- It may be *very inexpensive* relative to training with real equipment and/or under real training conditions.
- It may provide an alternative training solution when real equipment is unavailable and/or training under real task conditions is *dangerous* or restricted due to regulations.
- It may be preferred because of *environmental and sustainability* issues.
- It offers the possibility of training under certain relevant conditions that *rarely occur* at the working place, such as emergency situations.
- It can be done in *leisure* time, which may make it very cost-effective.
- It still may *save on the cost* of instructional personnel.
- It may encourage people to engage in new initiatives or *stimulate interest* for new tasks or knowledge areas.

4 Methods for training effectiveness measurement

There are many types of measurement methods that may be used to assess the possible beneficial effects of new ways of training on transfer or training effectiveness. Below, we will briefly present and discuss three main categories, i.e., methods based on measurement of trainee (learning) performance, methods focusing on the synthetic training device or overall training program itself, and ratings or questionnaires focusing on subjective evaluations by trainees.

4.1 Measurement of trainee performance

When using the (preferred) experimental-versus-control-group method in order to calculate measures such as T or TER, we need to measure additional training time needed in the real-task environment, until the real-task performance of the training groups reaches a predetermined criterion level. This means that during this additional on-the-job training time, performance measurements will have to be made in a standardized way, in order to determine whether or not the criterion level is reached. So, this method entails several groups training under carefully controlled conditions without artifacts due to, for example, repeated testing (so-called test-effect), selection, or instrumentation effects. In addition, the performance measures should be ecologically valid, that is, they must be relevant for real-task performance. Therefore, the measurements usually have to be carried out in practical situations at schools and training sections or areas. This may lead to several common difficulties:

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 fielded systems (Cohn et al., 2009). Lack of control of all these factors may severely threaten the validity of inferences based on objective measurements of performance (Boldovici, Bessemer & Bolton, 2002). Usually, fulfillment of all these (basic) requirements in combination can only be accomplished by creating a specific training program just for the experiment.

Measurement problems
Other difficulties are caused by the fact that it is often hard to measure *what* and *how much* exactly is learned with respect to the (real) task or job for which the training is intended. Real operational situations and even many normal job situations do not always easily allow the objective measurement of performance of former learners. And even when these real-world measures can be collected, it remains questionable as to what respect the (confounding) training has contributed to that performance level, and to what respect performance effects can be attributed to other factors (such as the measurements itself).

Limited availability of control conditions
Finally, measurements of training effectiveness are traditionally performed after a training simulator is fully developed and instantiated in the training curriculum. As such, it has already replaced a legacy training system, such that comparisons between an experimental and control group cannot be made (Cohn, et al., 2009). In many other cases there is no operational system or prototype available to facilitate empirical evaluation of training results.

4.2 Opinion-based evaluation of the simulator or training program

Due to these factors, studies that directly try to quantify performance-based training effects in training simulators or instructional games are hard to conduct and/or may have a limited scope. Therefore, evaluation studies usually may (or should) include structured evaluation by experts, personnel, or students. This should be done in the context of the training program and skills and competencies to be trained. Hence, next to measurement of trainee performance, other types of methods also have to be applied for evaluating the quality of synthetic training. Below, we present three methods that evaluate various aspects of the training environment and/or the whole program in which the device is embedded.

The opinion survey method

In the opinion survey method, operators, instructors, training specialists, even students are interviewed in order to gather their opinions concerning the training effectiveness of a training device. For instance, questions are posed to determine which aspects of the simulator do or do not contribute to a high transfer of training. Such opinion data often do not guarantee success because the subjects interviewed may have little or no expertise on learning or cues facilitating learning. Therefore, the data gathered may easily lead to erroneous conclusions about the required properties of the trainer under development (Caro, 1977).

The simulator fidelity method

In this method, operational personnel (experts) compare the simulator on its physical, functional, and psychological aspects with the real system (e.g., comparison of the physical, perceptual, cognitive, or affective characteristics of both). Systematic analytic procedures have been developed for the employment of this model, which take into account fidelity of both the stimuli the simulator presents to the trainee and the responses he/she makes 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). Some investigators have argued that a simulator can be a faithful physical copy of the real-life system, but that this, by itself, does not allow any conclusive statement about its effectiveness as a training tool (e.g., Adams, 1972).

The simulator training program analysis method

In this regard, the simulator training program analysis method (STPA-method) may be a substantial improvement, since this method specifically determines whether the *training program* is well-designed. This is or can be done by using standardized checklists (Caro, 1977). The STPA-method involves analysis of the way the simulator is used to determine whether the training program is well-designed. It is directed toward the appropriate attainment of training objectives. This method can pinpoint possible factors limiting the effectiveness of a simulator under particular circumstances. However, it cannot determine the extent of training effectiveness.

The main deficiencies of these evaluation methods lie in their one-sidedness. For instance, the simulator fidelity method does not take into account contextual factors, such as didactical, motivational, and organizational aspects of training aids; e.g., 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 real training value because the quality of the simulator itself is not sufficiently taken into account (i.e., the various aspects of functional and physical fidelity). In general, the combination of the training environment and the didactical and organizational quality of the context in which the system is embedded determine training effectiveness.

4.3 Ratings and questionnaires focusing on the trainee

The limited scope of the (often opinion-based) simulator evaluation methods may be extended by using structured ratings or questionnaires on learning processes and competencies in the trainee. These may provide valuable additional information about the effectiveness and limitations of training devices. These methods focus more on generic learning processes in the *trainee* than on measurable learning performance of trainees or characteristics of the simulator or training program. This additional information may include ratings or questionnaires on knowledge and skills, self-efficacy, situational awareness, flow, stress, motivation, experienced problems that remain after finishing the training, etc. These aspects may provide insight into the underlying generic factors related to human performance and learning determining outcomes of simulator evaluation studies.

4.4 Discussion

With regard to all evaluation methods, a distinction has to be made between process measures and outcome measures (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 a 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 simulator that should be either maintained, improved, or further developed to ensure that training goals are met (Cohns et al, 2009; Fowlkes et al., 1999). All aforementioned methods may focus on process- or outcome measures.

In general, there is an inverse relationship between the practicability of a particular method and the reliability of that method's results. Opinion-based (subjective) evaluation measures—surveys, ratings, questionnaires, and checklists—provide elaborate information concerning training processes. They provide insight into the underlying learning processes and intervening factors that may determine training outcomes. Compared to (more objective) measurements of training and simulator performance, these evaluation measures are also more easily applied when complete experimental control is difficult to achieve. However, the more subjective methods may provide limited or false information about the quality or effectiveness of a particular simulator. They may reflect personal opinions, expectations, biases, or preferences, instead of measuring the effectiveness of training of a particular simulator. Indeed, the ability to report about experiences varies substantially between subjects (Sander et al., 2005). In addition, experts (and students) often have preconceived opinions about simulators and games 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 therefore important to use as many "blind" procedures as possible and tests that do not enable individuals to identify desired response behavior. One should also limit the use of distorted retrospective reports and the need to disrupt an operator performing his or her task in order to ask questions.

In general, opinion-based surveys, questionnaires, ratings, and checklists are best used in combination with objective measurements (e.g., time, speed, error) of trainee (learning) performance. This may provide the best combination of reliable and relevant information on training effectiveness in a relatively pragmatic way. Of course, the trainee-performance measures should aim to reflect the most crucial and relevant skills and competences with regard to the operational task. Therefore, these measures should preferably be adopted from on-the-job training performance that mimics operational situations, or they must be determined as much as possible under real operational task conditions.

5 Two practical examples

Following are examples of parsimonious, cheap, and simple ways to assess the outcomes of simulator training, still using on-the-job performance by subjects.

A low-cost driving simulator (Kappé, 2001) has been used an alternative method for studying training effectiveness. The method was comparable to the uncontrolled transfer method in that it can be relatively valid without needing a real control group. In this study, the experimental group of trainees started training a number of driving skills for which this low-cost simulator was intended to be an adequate training tool (e.g., start-up procedures, driving on straight and curved rural roads, driving on highways, traffic insight). After a couple of simulator lessons, trainees had to show their skills to an experienced instructor in the real lesson vehicle. On the basis of this assessment, the instructors could estimate the number of real lessons it would take for the average person who is trained in the conventional way to reach a level of performance equal to that of this simulator-trainee. This is a very parsimonious measure because it is simple, straightforward, and capitalizes on the existing knowledge and experience of the experts. The independent variable is the amount of simulator practice, the dependent variable is the estimation made by the expert about the number of real lessons it would take to get to this level of performance (which implicitly is based on knowledge concerning the performance of conventional trainees, i.e., a kind of virtual control group). In addition, the assessors were able provide performance scores on different parts or aspects of vehicle driving, which gives information concerning the strong and weak aspects of the simulator training. It may be expected that these kinds of judgments can be made by experts like experienced instructors. It is recommended that the assessors remain ignorant about the types and number of simulator lessons provided to the trainees. This way, training effectiveness and efficiency are rapidly and cheaply assessed without a control group. In addition, inter-rater reliability could be easily measured and verified by using two or more experienced assessors. Of course, the value of such judgments is less reliable than the outcome of a transfer of training study in which a "harder" measurable criterion performance level on the real system is determined, or when a control group is used.

In general, it is always beneficial to use several methods to evaluate the effectiveness of a training simulator. This minimizes the impact of the artifacts of each individual method. It is not so difficult to apply several methods in one transfer-of-training study. Interviewing subjects on their opinion after receiving synthetic training is easy to do. This may issue in a more valid outcome of the study. An example is a study conducted by TNO, using this method for validating the Link-Miles Leopard 2 driving simulator (Moraal & Poll, 1979). One group was trained on the simulator, one on the real tank. After acquiring a certain level of performance on the tasks, the experimental group of soldiers had to reach a predetermined criterion level on the real tank. The T formula was used to define the effectiveness of training on the simulator. Because the researchers where still unsure about the validity of the outcome of the experiment, they applied 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 opinion on the simulator (simulator fidelity method). All this was done to ensure a valid (internal and external) outcome of the experiment, with useful suggestions to improve the effectiveness of simulator training. The results of the different methods applied in this experiment did not contradict each other.

6 Recommendations and final notes

This review was meant to give researchers in the field of simulator research an overview of the different designs, metrics, and methods available for determining training effectiveness of training simulators or games. It will be clear that this measurement of transfer-of-training simulators is a rather complicated issue and prone to methodological flaws and confounding factors. Therefore, we provide some recommendations and notes on evaluating simulator effectiveness.

1. Apply more than one method to assess the effectiveness of a simulator or serious game. The use of several methods in the same study will eliminate the disadvantages of each single method and will reduce the risk of erroneous conclusions.

2. Since a simulator in isolation is only a tool designed to attain specific training objectives, the quality of the whole training context in which this tool is embedded must be taken into consideration. In other words, the effectiveness of a training system is always a function of the training system itself and its training context. Therefore, evaluating training effectiveness (and its economic dimensions, utility of training) must always occur in combination with the specific characteristics of mission, tasks, goals, and operational environment. If these aspects are not considered, training effectiveness evaluations are useless (Muth & Switzer, 2009).

3. Next too that motivational, organizational, cultural, and infrastructural features, may have huge impact on the transfer and (cost)-effectiveness of training simulators. If, for example, instructors do not *believe* in the simulator, fear for their job, or are not motivated to work with new and complicated

products of technology, transfer of training can be severely and negatively affected (Emmerik & Korteling, 2012).

4. Measuring physical fidelity of a simulator is the most precise and reliable validation method because it is based on objective physical measurements. Without physical data, however, one cannot predict the behavioral characteristics of humans in the simulator. It is therefore useful to use task-specific formulas that relate physical simulator variables to psycho-physical variables and human performance variables (e.g., the relation between a display and object detectability). This may reduce the need to measure human task performance in the validation studies of a simulator. Future investigations into research training simulators can be used to pinpoint the relevant simulator performance relationships.

5. Transfer-of-training studies are time-consuming and costly. Therefore, when designing a transfer study, devote substantial attention to finding relatively simple methods and procedures providing transfer-of-training data. For instance, it is sometimes possible to use as control data the existing performance registrations of groups that have previously been educated or trained on the basis of conventional methods and/or equipment.

References

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

Allan, J.A.,Hays, J.T., Buffardi, L.C. (1986). Maintenance Training Simulator Fidelity and Individual Differences in Transfer of Training. *Human Factors*, 28, 297-509.

Alliger, G.M., & Janak, E.A. (1989). Kirkpatrick’s levels of training criteria: Thirty years later. *Personell Psychology*, 42, 331-342.

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

Baldwin, T.T, & Ford, J.K. (1988). transfer of training: a review and directions for future research. *Personnel Psychology, 41*, 63-105.

Boldovici, J.A., Bessemer, D.W., and Bolton, A.E. (2002). The Elements of Training Evaluation. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Breda, L. van & Burry, S. (1991). *Meting van de rijeigenschappen van de Leopard 2 tank en simulator* [Measurements of the vehicle characteristics of Leopard 2 and simulator] (Report IZF 1991 A-56). Soesterberg, the Netherlands: TNO Institute for Perception.

Campbell, D.T. & Stanley J.C. (1963). *Experimental and Quasi-Experimental Designs for Researc.* Chicago, Illinois: Rand McNally & Company.

Caro, P.W. (1977). *Some factors influencing air force simulator training effectiveness* HUMRRO Technical Report tr-77-2. Alexandria, Virginia: Human Resources Research Organisation.

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.

Detterman, D.K., & Sternberg, R.J. (Eds.) (1993). *Transfer on trial: intelligence, cognition, and instruction.* Norwood, NJ: Ablex Publishing.

Emmerik, M.L. van, Korteling, J.E. (2012). *Optimale inzet van synthetische trainingsomgevingen voor defensie.* Optimal use of synthetic training environments for the Defence [in Dutch]. Concept Report, Soeterberg NL: TNO Human Factors.

Gielen, E.W.M. (1995). *Transfer of training in a corporate setting* (doctoral thesis). Enschede; UniversiteitTwente.

Hays, R.T. (2005). *The effectiveness of instructional games: a literature review and discussion.* Technical Report 2005-004. Naval Air Warfare Training Systems Division. Orland, USA.

Kappé, B. (2001). *Validation of the prototype of the INTRASIM rijsimulator* (Memo in Dutch TNO-TM 2001-M020). Soesterberg, TNO Human Factors.

Kirkpatrick, D.L. (1959). *Evaluating training programs* (2nd ed.) San Francisco: berrett Koehler.

Kirkpatrick, D.L. (1998). *Another look at evaluating training programs*. Alexandria, VA: American Society for Training and development.

Korteling, J.E. & van den Bosch, K. & van Emmerik, M.L. (1997). *Low-cost simulators 1a: literature review, analysis of military training, and selection of task domains* (Report TNO-TM-97-A035). Soesterberg, The Netherlands: TNO Human Factors Research Institute.

Korteling, J.E., Helsdingen, A.S., Theunissen, N.C.M. (2012). Serious Games @ Work: Learning job-related competences using serious gaming. In A. Bakker & D. Derks (Eds) *The Psychology of Digital Media at Work*. Psychology Press LTD / Taylor & Francis Group.

Mayer R.E., & Wittrock, M. C. (1996). Problem-solving transfer. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp.47-62). New York: Simon & Schuster Macmillan.

Moraal, J. & Poll, K.J. (1979). *The Link-Miles driving simulator for armoured vehicles; report of a validation experiment* [De Link-Miles rijsimulator voor pantservoertuigen; verslag van een validatie-onderzoek] (Report IZF 1979-23) Soesterberg, The Netherlands: TNO Institute for Perception.

Muth, E., & Switzer, F. (2009). Training effectiveness and evaluation. In: D. Schmorrow, J. Cohn, D. Nicholson (Eds). *The PSI Handbook of Virtual Environments for Training and Education*. pp 147-156

Oprins, E.A.P.B. & Korteling, J.E., (in press). Transfer of gaming: effectiveness of a cashier trainer.

Roscoe, S.N. & Williges, B.H. (1980). Measurement of transfer of training. in S.N. Roscoe (Ed.), *Aviation Psychology*. Iowa: The 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*, 3-16.

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

Schollmeyer, J. (2006). Games get serious. *Bulletin of the Atomic Scientists*, *62*, 34-39.

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology*, 64, 489-528.

Stehouwer, M., Serné, M. & Niekel, C. (2005). A tactical trainer for air defence platoon commanders. In: *Proceedings of the Interservice/Industry, Training, Simulation, and Education Conference*. Orlando I/ITSEC 2005, Paper no. 2066