TNO-report
TM-99-A023

TNO Human Factors
Research Institute

Kampweg 5
P.O. Box 23
3769 ZG Soesterberg
The Netherlands

Phone +31 346 35 62 11
Fax +31 346 35 39 77

title

A critical review of validation methods for man-in-the-loop simulators

authors

J.E. Korteling
R.R. Sluimer

date

16 March 1999

number of pages : 32 (incl. appendices, excl. distribution list)

TNO Human Factors Research Institute
is part of TNO Defence research
to which also belong:
TNO Physics and Electronics Laboratory
TNO Prins Maurits Laboratory

Netherlands Organization for
Applied Scientific Research

Management uittreksel TNO Technische Menskunde, Soesterberg

titel:	Een kritisch overzicht van validatiemethoden voor man-in-the-loop simulatoren
auteurs:	Dr. J.E. Korteling en R.R. Sluimer
datum:	16 maart 1999
opdrachtnr.:	A98/KL/301
IWP-nr.:	730.3
rapportnr.:	TM-99-A023

Simulatoren worden steeds meer en steeds effectiever toegepast in allerlei situaties waarbij mensen voor (moeilijke) taken worden getraind of waarbij de menselijke taakprestatie wordt onderzocht. Een belangrijke vraag die steeds weer opduikt betreft de kwaliteit van deze 'man-in-the-loop' simulatoren. In dit verband richt een deel van het onderzoek waarbij simulatoren worden gebruikt zich op de simulator zelf. Dit rapport behandelt de methodologische concepten, paradigma's en valkuilen die gerelateerd zijn aan dergelijk onderzoek *naar* de validiteit en natuurgetrouwheid van simulatoren. Er wordt onderscheid gemaakt tussen onderzoeksmethoden voor trainingssimulatoren en voor researchsimulatoren. Validatiemethoden voor trainingssimulatoren hebben betrekking op de effecten van simulatorvariabelen (bv. resolutie van de display, moving base karakteristieken) die de effectiviteit van de simulator als een *trainingsinstrument* bepalen. Validatiemethoden voor researchsimulatoren betreffen de effecten van simulatorvariabelen op de effectiviteit van de simulator als *researchinstrument*. Artefacten die de uitkomst van een experiment beïnvloeden worden apart beschreven voor alle validatiemethoden die in dit rapport worden behandeld. Geconcludeerd wordt dat validatie van simulatoren een zeer complexe zaak is en onderhevig aan vele methodologische valkuilen en storende factoren. Bovendien bestaan er veel misverstanden op dit gebied. Na de uitleg van alle gebruikelijke methoden en hun voor- en nadelen worden de volgende aanbevelingen voor toekomstig onderzoek gedaan:

1 De terminologie is ambigu. Standaardiseer de terminologie zodat spraakverwarring wordt voorkomen.
2 Validiteit is geen eenduidig, onafhankelijk kenmerk. Het begrip validiteit in het simulatoronderzoek is alleen zinvol als deze in verband wordt gebracht met functionele aspecten zoals: de functie van de simulator (training of research) en de betreffende deeltaken of trainingsmethodes. Dit zal de neiging tot over-generalisatie op basis van afgebakende onderzoeksresultaten beperken.
3 Let altijd op de face validiteit. Als mensen niet in het systeem geloven zullen ze het evenmin op de juiste wijze gebruiken.
4 Pas zoveel mogelijk verschillende methoden toe bij het verrichten van validatie studies. Combinatie van bijvoorbeeld objectieve met subjectieve methodes zal de kans op foutieve conclusies verminderen terwijl de voordelen van beide methodes worden opgeteld.
5 Meer onderzoek zou er gedaan moeten worden naar het creëren van taakspecifieke formules die fysische data van de simulator relateren aan psychofysische data en de prestatie van proefpersonen (bv. contrast ratio van het beeldscherm relateren aan detecteerbaarheid van objecten door proefpersonen). Dit zal de noodzaak van het meten van de prestatie van proefpersonen in validatie studies van simulatoren verminderen.
6 Besteedt veel aandacht aan het vinden van validatiemethodes waarbij simulatorprestaties op een eenvoudige manier kunnen worden vergeleken met prestaties op het werkelijke systeem. Een relatief eenvoudige en praktische methode wordt voorgesteld om de effectiviteit van een rijsimulatortraining objectief te bepalen.

CONTENTS Page

Rep.No. TM-99-A023

TNO Human Factors Research Institute
Soesterberg, The Netherlands

A critical review of validation methods for man-in-the-loop simulators

J.E. Korteling and R.R. Sluimer

SUMMARY

This review examines the methodological concepts, paradigms and pitfalls related to validation- and fidelity studies of man-in-the-loop simulators. A distinction is made between validation methods for training simulators and for research simulators. Validation methods for training simulator are applied in experiments which assess effects of simulator variables (e.g. resolution of the display, cue augmentation, moving base characteristics) on the effectiveness of a simulator *as a training device*. Validation methods for research simulators are applied in experiments which assess the effects of simulator variables on the effectiveness of a simulator *as a research tool*. The review is particularly focussed on the various artefacts that may affect the outcome of such validation experiments. The artefacts are separately described for each single validation method. It will be demonstrated that validation of simulators is a very complicated matter and prone to various methodological flaws and confounding factors.

After the discussion of the common methods including their advantages and disadvantages the following recommendations for future research are given:

1 Terminology in the field of simulator research is ambiguous. It is advised to standardise terms, which will lead to more comprehensible communication among researchers.

2 Validity is not a single, independent attribute. The term validity in simulator research only makes sense if related to functional aspects of simulators, such as the purpose of the simulator (training, research) and the tasks and training methods involved. This will reduce the amount of overgeneralizations that are now encountered too frequently.

3 Always take *face validity* into consideration. If people do not believe in the simulator they are not very likely to use it properly.

4 Apply more than one method in a simulator validity study. Combination of e.g. objective with subjective methods reduces the risk of erroneous conclusions and combines the benefits of both kinds of methods.

5 Aim more research at creating task-specific formulas, which relate physical simulator variables to psycho-physical and human performance variables. This will reduce the need to measure human task performance in simulator validation studies.

6 Always allocate substantial effort to find a practical method that still compares simulator performance with on-the-job performance by subjects. A relatively simple and practical method is proposed to assess the effectiveness of a driving simulator training.

Rap.nr. TM-99-A023

TNO Technische Menskunde
Soesterberg

Een kritisch overzicht van validatiemethoden voor man-in-the-loop simulatoren

J.E. Korteling en R.R. Sluimer

SAMENVATTING

Dit rapport behandelt de methodologische concepten, paradigma's en valkuilen die gerelateerd zijn aan onderzoek naar de validiteit en natuurgetrouwheid van man-in-the-loop simulatoren. Er wordt onderscheid gemaakt tussen onderzoeksmethoden voor trainingssimulatoren en voor researchsimulatoren. Validatiemethoden voor trainingssimulatoren hebben betrekking op de effecten van simulatorvariabelen (bv. resolutie van de display, moving base karakteristieken) die de effectiviteit van de simulator als *trainingsinstrument* bepalen. Validatiemethoden voor researchsimulatoren betreffen de effecten van simulatorvariabelen op de effectiviteit van de simulator als *researchinstrument*. Artefacten die de uitkomst van een experiment beïnvloeden worden apart beschreven voor alle validatiemethoden die in dit rapport worden behandeld. Geconcludeerd wordt dat validatie van simulatoren een zeer complexe zaak is en onderhevig aan vele methodologische valkuilen en storende factoren. Bovendien bestaan er veel misverstanden op dit gebied. Na de uitleg van alle gebruikelijke methoden en hun voor- en nadelen worden de volgende aanbevelingen voor toekomstig onderzoek gedaan:

1 De terminologie is ambigu. Standaardiseer de terminologie zodat spraakverwarring wordt voorkomen.

2 Validiteit is geen eenduidig, onafhankelijk kenmerk. Het begrip validiteit in het simulatoronderzoek is alleen zinvol als deze in verband wordt gebracht met functionele aspecten zoals: de functie van de simulator (training of research) en de betreffende deeltaken of trainingsmethodes. Dit zal de neiging tot over-generalisatie op basis van afgebakende onderzoeksresultaten beperken.

3 Let altijd op de *face validiteit*. Als mensen niet in het systeem geloven zullen ze het evenmin op de juiste wijze gebruiken.

4 Pas zoveel mogelijk verschillende methoden toe bij het verrichten van validatie studies. Combinatie van bijvoorbeeld objectieve met subjectieve methodes zal de kans op foutieve conclusies verminderen terwijl de voordelen van beide methodes worden opgeteld.

5 Meer onderzoek zou er gedaan moeten worden naar het creëren van taakspecifieke formules die fysische data van de simulator relateren aan psychofysische data en de prestatie van proefpersonen (bv. contrast ratio van het beeldscherm relateren aan detecteerbaarheid van objecten door proefpersonen). Dit zal de noodzaak van het meten van de prestatie van proefpersonen in validatie studies van simulatoren verminderen.

6 Besteedt veel aandacht aan het vinden van validatiemethodes waarbij simulatorprestaties op een eenvoudige manier kunnen worden vergeleken met prestaties op het werkelijke systeem. Een relatief eenvoudige en praktische methode wordt voorgesteld om de effectiviteit van rijsimulatortraining objectief te bepalen.

1 INTRODUCTION

1.1 Background

For a long time, and for many different reasons, people have used models to simulate reality. With regard to the simulation of man-in-the-loop tasks, the first simulators were primitive devices used for flight training before and during World War 1. Since then, simulators applied in systems other than aircraft have proliferated, including:

- transport systems such as automobiles, ships and railway systems;
- military systems such as command and control systems, weapon and sensor systems;
- systems for air traffic- and (nuclear) power plant control.

Simulators have many potential advantages over real-task equipment:

- cost reduction: simulator training or research may be less expensive than in real task conditions;
- availability: real-task equipment may not always be available;
- safety: it is less dangerous to train or evaluate emergency procedures in a simulator than under real task conditions;
- training and instruction enhancement: e.g. many extra instructional facilities in simulators, such as record-replay, scenario control functions, or augmented cueing will enhance learning in a simulator;
- better possibilities of control and monitoring: training and experimental conditions and events can be exactly defined, manipulated, and recorded.

In addition to their application as training devices, simulators have been used extensively for research into the development and design of complex man-machine systems, investigating how physical simulator properties affect human behaviour compared to real-task equipment (Meister, 1995). The present review will focus on the use of the concept of *validity* in man-in-the-loop simulator research. Since a training- or research simulator is only a tool, its validation is merely verification that it accomplishes the purpose for which it was developed. Without such validation, it remains obscure to what degree this tool is effective and which factors may be responsible in case a simulator does not do what is expected (Moraal & Kraiss, 1981).

In most fields of Human Factors, research paradigms and methods are comprehensively described in textbooks or reviews (e.g. Sanders & McCormick, 1992; Stammers & Shepard, 1995). Not in the field of simulator research, however. Moreover, proper studies measuring validity of simulators are hard to conduct and usually very expensive. This counts especially for the validation of training simulators. In this connection, the present report is a preparation of future simulator validation studies to be carried out by TNO under contract of the Royal Netherlands Army. It will provide a review of methods available on simulator validation. In addition, the report proposes a relatively simple and practical method to assess the effectiveness of a simulator training.

Before explaining the way validity is measured, a list of common terms and concepts regarding the field of simulator research will be discussed. Subsequently, a description of all common methods on effectiveness of training simulators will be given and what should be kept in mind when one conducts such an experiment. The third section describes currently available methods of validating research simulators. Finally, some recommendations are given for future research on simulator validation experiments.

1.2 The 'validity' concept in simulator research

Scientific methodology books rarely provide a general definition of the term *validity* (see for instance: Swanborn, 1993; Sanders & McCormick, 1992). Usually the meaning of the word validity is described in several sentences. Afterwards some definitions of different sorts of validities are described such as: *content validity* or *internal validity*. The reason for this is that the definition of the word *validity* is different for different contexts or applications. The dictionary of Human Factors defines validity as follows: *Validity is the degree to which a test or other measurement device really measures what it was designed to measure* (James & Stramler, 1993). This definition, originating from the field of human performance assessment, does not exactly cover the meaning of validity in simulator research. Simulators are not primarily used as measurement or test devices. Therefore, the validity of simulators should be conceived as the degree to which they fulfill their purpose. For training simulators this is the attainment of certain training objectives and for research simulators this is the adequate exploration of research questions.

Another problem of terminology is how validity is defined and used among researchers within the field of simulator research. When speaking of training simulators, validity is supposed to reflect the quality of the simulator as a training device, but this 'attribute' can only be measured by taking into account the training program and the instruction processes. Probably for practical reasons, many researchers (e.g. Moraal & Van Meeteren, 1990) use the concept of validity as a single property that can be defined independently of the training program and the influence of the instruction processes. One should always be aware of this simplification because these factors may substantially affect the degree to which experimental results can be generalized to different training conditions. As an example: a low-cost driving simulator may be valid to train traffic insight and traffic interactions, whereas it may be completely unfit to train advanced vehicle control skills. Likewise, a simple flight simulator may be valid for the training of basic flight skills such as instrument flying for beginning pilots, whereas it may be completely unsuited for conversion training of skilled professional pilots to another kind of aircraft. Hence, the concept of validity has always to be related to the functional characteristics of the simulator, such as the purpose for which the simulator is used (training, research), the tasks involved, trainees (or experimental subjects) and training methods and aids involved.

Not only in scientific methodology books but also within the field of simulator research, several kinds of validity are discerned. Below, the terminology used in simulator research will be described.

The term *face validity* refers to the extent of subjectively experienced similarity between the simulator and the real-life situation. When the appearance of a simulator resembles that of the real-life situation it makes subjects in general more motivated to execute the task (Korteling, Van den Bosch & Van Emmerik, 1997).

To what extent skills, learned on a simulator, are transferred to the real task is called *functional validity* (Korteling, Van den Bosch & Van Emmerik, 1997). In many reports functional validity is called *transfer of training*. For readability reasons both terms will be used throughout this review.

Another important concept in simulator research is *fidelity*. Fidelity is the amount of similarity between the simulator and real-task equipment. There is a distinction between *physical fidelity* and *functional fidelity*. Physical fidelity denotes to what extent the simulator mimics the real equipment and environment in terms of physical measurable characteristics (e.g. the resistance of the brake pedal). Functional fidelity is defined as to what extent the behaviour (perceptual, cognitive and motor processes) of a person in the simulator resembles his or her behaviour on the real task under the same conditions (Korteling, Van den Bosch & Van Emmerik, 1997). Functional fidelity is an important concept in the field of research simulators. The term can be divided in *absolute functional fidelity* and *relative functional fidelity*. These two concepts are used rarely in simulator research, however. Instead the concepts of *absolute validity* and *relative validity* are usually employed. *Absolute validity* refers to the degree of correspondence between provoked subject behaviour in the simulator and the behaviour in practice in a quantitative way. For example: real drivers and drivers in a simulator both show a mean driving speed of 110 km/h on a four-lane highway. Relative *validity* denotes the correspondence of the effects of experimental variables on subject behaviour. If operator behaviour in the simulator is affected the same way as on real-life equipment by a certain task variable, the system is relatively valid (Korteling & Van Randwijk, 1991). For example: a driving simulator would be relatively valid on speed choice when both in the simulator and in real vehicles, drivers choose a much higher driving speed on highways than on urban roads, although the absolute driving speeds in the simulator and in real vehicles in these situations differ substantially, say more than 15%.

1.3 Scope of this review

In the book *Experimental and Quasi-Experimental Designs for Research* written by Campbell and Stanley (1963), artefacts affecting experimental designs are thoroughly described. This book has been used to critically examine the different simulator validation methods. When methods are correctly applied, there still may be artefacts that are inherent to the method used (inherent artefacts) affecting the outcome of an experiment. There are, however, other artefacts (non-inherent artefacts) introduced by incorrectly executed research methods, that can also jeopardise the validity of the experiment. This review will discuss both inherent and non-inherent artefacts belonging to the different validation methods discussed. These different artefacts affecting the outcome of an experiment are summarised in Appendix A. Appendix B presents a table showing which artefacts may affect the outcome of a particular method discussed in this review.

The present report will discuss only validation methods that are applied in *simulator validation* research. This means that some of the most widely used methods in the field of Human Factors are not reviewed, like the walkthrough, user-trial or task analysis method. This last method sets out to represent information that is to be used in either the design of a new human/machine system or in the evaluation of an existing system design (Stammers & Shepard, 1995). Knowing the requirements of the task, i.e. which tasks have to be performed by subjects in terms of inputs, processing operations and outputs (including performance criteria), and the characteristics of the system in terms of a mathematical model, one can formulate simulator specifications. In simulator design, task analysis is an extensively used tool. Task analysis can also serve in making suggestions for improvement of the simulator (e.g. Wertheim, 1984).

2 TRAINING SIMULATOR VALIDATION METHODS

2.1 Introduction

This chapter will describe the different methods used in studies on the validation of training simulators. Studies investigating the effectiveness and/or efficiency of training on a simulator measure the *transfer of training (ToT)*. The next paragraph describes objective measures used to quantify ToT. These measures can be applied in studies in which subjects practice on real-life equipment and/or simulators. First methods in which subjects practice on real-life equipment and on simulators will be discussed. After that, methods that measure training effectiveness on the simulator only, and finally subjective measurement methods are discussed. These latter methods express the validity of a training simulator in a qualitative way.
The names for the different methods described below are derived from Caro (1977). He described these methods in a study done in commission of the US air force. The only exception is the term: *quasi-training-of-transfer method*, which is widely used in the literature, whereas Caro used the term simulator-to-simulator method.

2.2 Transfer of training measures

In the seventies objective measures have been introduced to quantify ToT, mainly by Stanley Roscoe (see e.g. Korteling & Van Randwijk, 1991). In experiments using these measures an experimental group is trained on a simulator. After a certain period the group gets additional training on the real task (on-the-job training) until the real task performance of this group reaches a predetermined criterion level. The time needed for the experimental group to reach the real task performance on this criterion is then compared to the time needed by a control group, who has been trained on the real task only (Roscoe & Williges, 1980). The basic computation for %**T** (*Percentage of transfer*) is:

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

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 the simulator training program

From equation 1 it can be derived that when %T of a given simulator training program is 100% no additional field training is needed by the experimental group to reach the same criterion performance as the control group. When T_e increases, %T decreases, hence when %T is 0% training on the simulator does not produce any effect. %T can even become negative. This means that training on the simulator interferes with acquiring the necessary skills for executing the real task (Korteling, Van den Bosch & Van Emmerik, 1997).
The percentage of transfer formula has a big flaw, as it fails to consider the *amount of practice* on the simulator by the experimental group. The transfer is a negatively decelerated function of the simulator training (Boer, 1991). Because the percentage of transfer formula does not consider the *amount* of simulator training prior to on-the-job training it permits no conclusions about the effectiveness of the simulator as a training tool (Roscoe & Williges, 1980). Therefore a more adequate measure is the *Transfer Effectiveness Ratio* (TER) or graphically the *Cumulative Transfer Effectiveness Function* (CTEF) which reckons with the time spent in the simulator. The computation for TER is:

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

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 the simulator training program
T$_s$ Simulator training time by the experimental group

A TER of 1.0 indicates that time savings for on-the-job training are equal to the amount of time spent in the simulator. When TER is larger than 1.0 (T_s + T_e is smaller than T_c) simulator training is more effective than training on the real task. When TER is lower than 1.0 the real task training is more effective. This does not necessarily mean that training on the simulator has to be stopped. Simulator training can still continue for a number of reasons:

Training on the simulator is less costly than training with real equipment

Training on the simulator is less dangerous than training with real equipment

Training on the simulator is preferred because of environmental issues
Training on the simulator gives the possibility of training under certain relevant conditions that rarely occur in real life such as emergency situations (Korteling, Van den Bosch & Van Emmerik, 1997).

One should keep in mind that there is a maximum on the transfer of training in a simulator. Not all skills needed on the real task can be trained on a simulator. Therefore TER is a negatively decelerated function of the simulator training time. In the following graph TER and %T are projected as a function of training time.

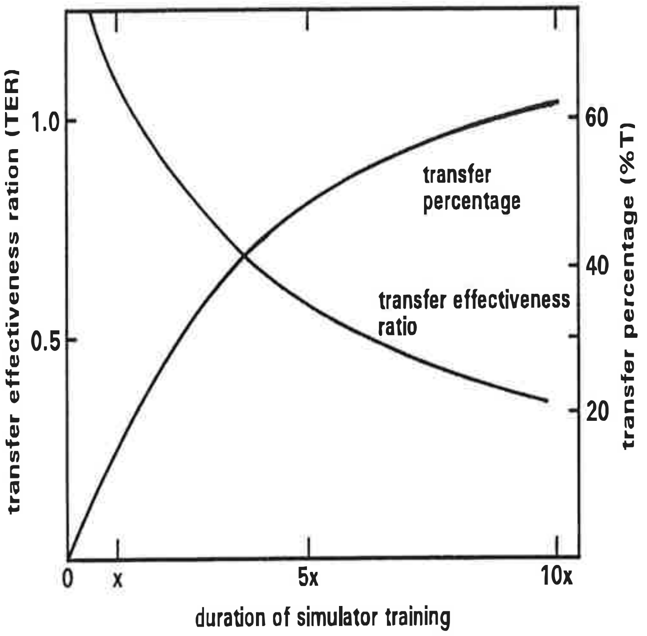

Fig. 1 Transfer Effectiveness Ration (TER) and Percentage of Transfer (%T) as a function of duration of simulator training.

A measure for expressing the effectiveness of financial training cost has also been developed, because simulator training in general is less costly then 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:

$$\mathrm{TCR} = \frac{C_s}{C_c} \qquad (3)$$

where:
C_s financial cost of simulator group training (per time unit)
C_c financial cost of control group training (per time unit)

The formula for the CER is as follows:

$$\text{CER} = \frac{\text{TER}}{\text{TCR}} = \frac{C_c\ (T_c - T_e)}{T_s \cdot C_s} \tag{4}$$

Cost effective training can be achieved with CER values above 1. For a CER smaller than 1, simulator training might still be effective for safety or environmental reasons. Ratio's to calculate safety or environmental effectiveness are more difficult to construct. It requires estimations of accident probabilities or damage to the environment. This issue is, however, beyond the scope of this present review. For different duration's of simulator training, CER, TER, as well as %T will change. A small fictional example will illustrate 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.
>
> **%T = 20%; TER = 0.50**
>
> Operating cost of the simulator has been figured out to be 15% of costs associated with the real-task equipment.
>
> **TCR = 0.15; CER = 0.50 / 0.15 = 3.33**
>
> Only 15 hours of additional on-the-job training are needed if the experimental group gets 11 hours of simulator training.
>
> **%T increases to 25%; TER increases to 0.45; CER = 0.45 / 0.15 = 3**
>
> Cost effectiveness is still achieved.

Roscoe and Williges also introduced the Incremental Transfer Efficiency Ratio (ITER) (Roscoe & Williges, 1980). With this technique the performance of an experimental group in real life is compared to another experimental group with less simulator experience. The Ratio defines the amount of extra time needed on the real task by the experimental group with less training on the simulator compared to another experimental group to reach the same criterion level. The computation of ITER is:

$$\text{ITER} = \frac{T_{e-\ X} - T_e}{X} \tag{5}$$

where:

- T_e Time trained on the simulator by an experimental group
- $T_{e-\Delta x}$ Time trained on the simulator by another experimental group with ΔX less training time
- ΔX Difference in time trained on the simulator between the two groups

ITER is, just like TER, a negatively decelerated function of simulator training time. The difference in time needed on the real task to reach the criterion level between two experimental groups, when a lot of training on the simulator has taken place, will approach 0.

2.3 Simulator to real task methods

2.3.1 The Experimental-versus-Control-Group Method

This method uses the experimental setup described in § 2.2, in which the experimental group is trained on the simulator and (afterwards) on the real system to reach the criterion level; the control group is trained on real-task equipment until criterion level. Both groups are tested on the real system. The experimental-versus-control-group method is generally thought to be the most appropriate experimental design to determine whether simulator training has improved subsequent real-life performance (Caro, 1977). There are no *inherent artefacts* connected to this method. When the method is properly applied, the results cannot be alternatively explained by methodological deficiencies that form a necessary part of the method.

Lintern et al. (1990) conducted a study in which beginning flight students were given two sessions of simulator training, before they commenced landing practice in the real aeroplane. A relatively inexpensive computer-animated landing display was used to train the experimental group. For each experimental student there was a control student, paired with the same instructor, who received no simulator training. Experimental students required significantly fewer presolo landings in the aeroplane compared to their control students.

Usually, it is very hard to conduct an experiment in which this method is applied, because it entails a great deal of expenses relative to the amount of (objective) knowledge that is acquired. For instance: the cost of one hour of simulator training in a B-52 bomber is $500 and one hour of on-the-job training costs $9000 (these figures include instructors salaries, fuel, maintenance etc.). For this practical reason, the experiment is often organised such that it fits in a running training program. Unfortunately, this may be hampered by logistical- and randomisation problems (Orlansky et al., 1994). In addition, the ToT ratio's of § 2.2 only provide objective and quantitative information on transfer given the *specific training programme* included in the experimental setup, i.e. the chosen amount of training on the simulator, the training content, the training methods, etcetera. Because of these (insuperable) problems and limitations, experiments in which this method is applied are often not correctly executed. Boldovici (1987) describes a number of practical factors causing incorrect validation experiments:

- Small number of subjects or crews may be used in the comparison. It is often not possible to find enough subjects to fill the two groups sufficiently for statistical purposes.
- Because of the relatively high costs of this method, insufficient amounts of practice may be provided, which will affect proficiency. Therefore, the groups should not be judged too early in the course. This may result in insufficient learning such that one may not find performance difference between the two groups.
- In order to reduce costs, experimental tasks are often simplified, which may result in ceiling effects. These effects may mask differences between groups. That is: in a simple task, performance of both groups will easily reach the maximum performance level on the real task. This low amount of variance will make it difficult to reveal statistically significant differences.

When the experimental-versus-control-group method is correctly applied, no artefacts should affect the outcome of the experiment. Nevertheless artefacts may still be provoked by the fact that the method is often applied with existing simulator training programmes in order to reduce cost. These kinds of artefacts, that are provoked by the method but not necessarily connected to it, are called *non-inherent* artefacts. Such artefacts, described by Campbell and Stanley (1963), can still affect on the outcome of a validation study.

Non-inherent artefacts:

- Selection. When subjects are not matched or randomly assigned to groups, the observed difference may be a confounding effect of a difference already existing between the groups (e.g. age). In some research the experimental group and control group already exist. For example in one experiment two groups of soldiers companies were assigned to an evaluation study for tank gunnery (Boldovici, 1987). But pre-experimental gunnery scores already showed one company performing better than the other company. Maintaining unit integrity may be desirable in training or in combat but seldom is desirable in device evaluations.
- Instrumentation. Objective comparison requires that the experimental group is treated the same way as the control group during the real task execution. If one group has favourable conditions during performance on the real task compared to the other group, judgements of achievement are not reliable.
- Mortality. Loss of subjects in one of the groups can negatively affect the outcome of the experiment. For instance: Subjects who perform poor during training may drop out of the course. This will negatively affect the validity of the experiment, because only the best trainees are tested on the predetermined criterion level.

As shown, it is difficult to create an experimental group and a control group in training courses for device evaluation purposes. Several other methods to measure the functional validity of a training simulator will now be discussed. These methods still involve some kind of measurement on the real task.

2.3.2 The 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* simulator training started is obtained. This data is compared to data of performance obtained on the real task *after* simulator training. The difference between these data sets could be attributed to the simulator training, but concluding this is a hazardous act. The mayor flaw in this design lies in the absence of a genuine control group. One cannot draw any conclusion about the simulators efficiency, because the effect of simulator training is not compared to an experimental group trained on real-life equipment. There are a number of artefacts that are necessarily connected to this method *(inherent artefacts)*. These can provide alternative explanations for performance differences between the groups.

Inherent artefacts:

- Maturation. The difference in performance could be generated from confounding effects of forgetting or skill decay, particularly if the time interval between real task training (before and after simulator training) is great (Caro, 1977).
- History. Between the two measurements of subject performance, equipment might have changed which influences the test results (e.g. ageing of real-task equipment).
- Testing. If the test to obtain data on the real task is standardised, differences in performance on the two tests could be the result of learning from the first test.
- Instrumentation. There may be a difference in the method of measuring performance data. This will also result in differences in performance data.

The latter two artefacts may not both be avoidable: Assume that one artefact is controlled. This implies that one runs the risk of the other artefact having a jeopardising effect on the internal validity of the experiment. Hence, it will be difficult to conduct an experiment applying this method without the jeopardising effect of one of the two last-mentioned artefacts.

2.3.3 The Pre-existing-Control-Transfer Method

There are instances in which a concurrently trained control group might not be necessary. For instance: a simulator is introduced in a training program or a new simulator is replacing the old one. Student performance data from the older or existing program on the predetermined criterion task can be compared to data of performance by the new experimental group who has been trained on the simulator. Veltman and Korteling (1993) conducted an experiment in which the training effectiveness of a new simulator for the Leopard 2 driver training course was evaluated. Two groups of trainees first trained on the simulator. Subsequently, training on the real-task was implemented until the criterion level was reached. The time on the real-task, needed to complete the course was compared with the time of former trainees who did not receive simulator training. It was concluded that the implementation of the simulator reduced real-task practice with 50% in terms of covered distance. The implementation of the simulator did not lead to a longer overall training time compared to former trainees. There are three artefacts threatening the validity of this approach.

Inherent artefact:

- Selection. In this design there is no randomisation or matching of subjects over the control and experimental group. The population under training may have a different background compared to the group previously trained. Thus, differences in performances may be the results of a cohort effect.

Non-inherent artefacts:

- Instrumentation. The existing data must have been gathered under the same conditions and the same calibration of measuring instruments should be used as used in data obtained from the experimental group (Caro, 1977). This might not be the case when there is a large time interval between data gathering.
- Mortality. It is possible that in one of the courses trainees dropped out, due to bad performances. This may affect the outcome of the experiment.

2.3.4 The Uncontrolled Transfer Method

There are circumstances where no control group exists. Such a condition can occur when safety plays a role (forced landing by an aeroplane) or it may be impossible to create a control group (lunar landings). When no control group can be formed, simulator training effectiveness can be established by determining whether subjects can perform the learned task on a real-life system the first time they (have to) perform this task. This is called *first shot performance*. Data collected from such studies will be suspect, since it cannot be conclusively shown that the simulator training has had any effect on the real task operations performed by the subjects (Caro, 1977). There are two artefacts affecting the validity of an experiment using this method.

Inherent artefact:

- Maturation. Subjects may cumulate relevant skills by operating with real-task equipment on the job. These skills might enhance performance on the criterion task. It may, for instance, be supposed that pilots (who receive emergency training in the simulator) who are confronted with a real forced landing, after flying the real plane for a couple of years, will benefit from their experience with the handling characteristics of the real aeroplane.

Non-inherent artefacts:

- Mortality. Trainees who performed best during simulator training are the only ones assigned to use real-task equipment. This might positively affect the outcome of the measurement of first shot performance.

One could compare subject performance on the real-life equipment with subjects who did not receive training on the simulator or are trained on a different simulator. This would slightly resemble the first method described in this chapter, the experimental-versus-control-group method. However, this comparison has also some big flaws.

Inherent artefacts:

- Selection. Subjects in the two groups are not randomly assigned to the groups. This might affect the measurement. For instance: older operational personnel who did not receive any simulator training perform worse than younger personnel who did receive simulator training.
- Instrumentation. The criteria might have changed over time and conditions might have changed. For instance: the great time interval between two groups of astronauts landing on the moon could contribute to different measurement standards or changed equipment. Performance measurement criteria of the two groups should be the same.

All methods described above involve measurement of performance on real-life equipment. However, real task performance is not necessarily needed in an evaluation study. One can assess the effectiveness of a training simulator without on-the-job performances by human subjects. Next, three methods of Transfer-of-Training assessment are described where simulator performance only is used to determine the value of the simulator in a training program. The fact that simulator training cannot always involve transfer to a real-life system is not necessarily wrong. Many research issues can by investigated in a laboratory with the advantage of a higher control of particular independent variable(s) (Caro, 1976).

2.4 Within Simulator validation methods

2.4.1 The Quasi-transfer-of-training Method

Because of efficiency (or financial) reasons a method often applied in validation of a training simulator is the Quasi-Transfer-of-training method (QToT). The difference between the QToT method and ToT method is that in the former no training on the *real* equipment occurs while in the latter it does. The experimental group(s) get(s) training on a simulator with one or more simulator variables (colour, sound, mechanical motion) being omitted. The control group is trained on the fully operational simulator. Eventually both groups are evaluated on the fully operational simulator. The difference in performance reveals the relative contribution of the manipulated variable on the effectiveness of a simulator.

Strictly speaking, the QToT method does not reveal *absolute* information on functional validity, i.e. the effectiveness of a training simulator (Korteling & van Randwijk, 1991). There are four artefacts threatening the validity of the outcome of the experiment.

Inherent artefact:

- Reactive effect of experimental arrangements. The assumption on which this design is based states that there is equivalence between the fully operational simulator and the real life system as far as the criterion performance is concerned. This is a tenuous assumption.

Non-inherent artefacts that might influence the experiment are the same artefacts that jeopardise the validity in the experimental-versus-control-group method. Of course, in the laboratory, these artefacts can be more easily controlled compared to the experimental-versus-control-group method.

Non-inherent artefacts:

- Selection. When subjects are not matched or randomly assigned to groups, the observed difference may be a confounding effect of difference already existing between the groups.
- Mortality. Loss of subjects in one of the groups can degrade the outcome of the experiment. In a training situation there is usually drop out of trainees who are not capable of executing the task.
- Instrumentation. Objective comparison requires that the experimental group is treated the same way as the control group during performance measurement on the fully operational simulator. If one group has favourable conditions during performance on the fully operational simulator compared to the other group, judgements of achievement are not reliable.

There is one situation where this method can be very adequate. When a part-task simulator is used to reduce the time and costs spend in the full-mission simulator. In such a situation performance on the part-task simulator might involve inter-mediate training objectives on the way to the final objectives to be reached with subsequent training on real life equipment. (Caro, 1977).

2.4.2 The Backward Transfer Method

In a backward transfer study, an operator, who has already shown sufficient performance on the relevant task, is placed in a simulator. If he can perform the task on the simulator, backward transfer has occurred. Truijens conducted an experiment in which experienced skippers had to navigate a ship, passing a bridge (Truijens & Schuffel, 1978). These skippers were familiar with the conditions in the real environment, which was reconstructed in the simulator. The skippers had to indicate, previous to simulator performance, how they would operate the ship under different circumstances, e.g. current flow deviations and wind conditions. The predictions of the skippers of handling characteristics and their actual simulator performance showed no significant differences. Therefore, it was concluded that the ship simulator was valid.
This method assumes that there is transfer of training in the opposite direction (forward transfer) for trainees on such a simulator. However, drawing conclusions on the effectiveness of training with this design is risky on account of one artefact.
Inherent artefact:

- Interaction of selection and experimental variables. The simulator will give cues to experienced personnel, which will produce certain behaviour necessary to perform well. This does not mean these cues are appropriate for *learning* such behaviour (Caro, 1977).

2.4.3 The Simulator Performance Improvement Method

The simulator performance improvement method resembles an *equivalent time samples design*. Each training session, the performance of a trainee on the simulator is measured. An essential feature of an effective simulator training program is improvement in performance by the trainees over several sessions of training. If this does not occur, there would be little expectation of improvement in executing the real task. The one artefact that may contribute to a faulty conclusion is.
Inherent artefact:

- Reactive effects of experimental arrangements: if performance on the simulator improves, one can not conclusively state that significant transfer of training to real-task equipment should occur. For instance: improvement of performance on the simulator result could also in negative transfer to real-task equipment.

This method is most useful in a negative sense: if no improvement occurs in the simulator, none should be expected on the real task (Caro, 1977).

2.5 Subjective validation methods

Measurement of functional validity of training simulators can also be accomplished by asking experts, personnel or students on their opinion on the simulator or assessing the effectiveness of the training program concerning the simulator. In the following paragraphs three methods will be discussed in which these measurements are used.

2.5.1 The Opinion Survey Method

When no data is available on trainee performance or the simulator is under development, other kinds of methods have to be applied for evaluating simulator training performances. In the opinion survey method, operators, instructors, training specialists, even students may be interviewed regarding their opinions concerning simulator training effectiveness i.e. which aspects of the simulator contribute to a high transfer of training. Such opinion data often does not guarantee success, because the people 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 training simulator under development (Caro, 1977).

2.5.2 The Simulator-training-program-analysis method

The simulator-training-program-analysis method (STPA-method) determines whether the training program is well-designed by using a standardised checklist (Caro, 1977). The training program in which the simulator is embedded can have a significant impact on the indices reflecting functional validity of a training simulator. 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. Like the simulator fidelity method discussed below, its main deficiency is that it yields a measure that may be unrelated to real-task trainee performance.

2.5.3 The Simulator Fidelity Method

In this method operational personnel (experts) compare the simulator on its physical aspects with the real system (e.g. comparison of handling characteristics of the real-task equipment with the simulator equipment like resistance of the steering wheel, gear box etc.). Systematic, analytic procedures have been developed for the employment of this model that take into account fidelity of both the stimuli the simulator presents to the trainee and the responses he makes to these stimuli. This method measuring face validity is based on the assumption that when physical fidelity is high, transfer will also be high, and when fidelity is low, transfer will be low. The method has a wide appeal among operational personnel who are not familiar with transfer of training studies. It can be employed by anyone who is familiar with the real-task equipment and does not require any subjects or objective measurement (Caro, 1977). Some investigators have argued that a simulator can be a faithful copy of the real-life system but that this, in itself, does not allow any conclusive statement about its effectiveness as a training tool. Adams (1972) stated that equating fidelity with transfer of training leads to the development of unnecessarily costly devices. The basic deficiency of the method is the impossibility of generalising the result into the training effectiveness of the simulator.

2.6 Discussion

On the basis of the aforementioned methods it will be clear that the proper objective measurement of ToT (i.e. the purpose of the training simulator) is difficult and that other, more practical methods, are subject to quite a lot of methodological flaws. The three subjective methods described all have one major deficiency, they are subjective and thus reflect personal opinions expectations or preferences instead of measuring the effectiveness of training of a particular simulator is a training program. However, subjective evaluation methods may be useful in combination with methods described in §§ 2.3 and 2.4 as will be shown below.

The, from a methodological viewpoint optimal method, i.e., the *experimental-versus-control-group method*, entails a great deal of expenses (cost of simulator and real equipment, cost of instructors, trainees etc.), relative to the generality of knowledge that is acquired. The results are limited to a specific simulator configuration in combination with the used specific training methods (trajectory, scenario's, instruction protocols, exercises, feedback methods), trainee- and instructor characteristics, instruction facilities amount of training provided, and the chosen experimental (subtasks and task conditions). Each of these variables may have substantial impact on the results and thus degrades the *generality* of the conclusions. For practical reasons and in order to reduce expenses, experiments often need to fit in a relevant existing training program, which may produce logistical problems and methodological problems (Orlansky et al., 1994). In order to ensure that the experimental group and the control group train the same tasks, specific training programs have to be developed for the experiment. Also the QToT method needs a special training program only for experimental purposes. This training program may have to be implemented during ongoing training sessions. This will be expensive or will be logistically difficult. This issue also holds for the backward transfer method. Experts performing certain tasks on a particular simulator may interfere with the continuing training program. The performance improvement method is just like most simulator to real tasks methods an observational method. It is easy to carry out, however results of such an experiment may be feeble. This will lead to dubious conclusions about training effectiveness of the simulator.

In general, the experimental-versus-control-group method is more difficult to apply than the other simulator tot real task methods (i.e. Self-control-transfer method, Pre-existing-control-transfer method, uncontrolled-transfer method). Here, no unique training programmes and randomized experimental groups are involved. Observation of existing groups is possible, which of course is more cost-effective. The results, however, are less hard and reliable that the experimental-versus-control-group method.

The subjective measurements methods are the most easy to be applied. They do not need training on the simulator or the real task. However, subjective methods may provide limited and probably even false information about the functional validity of a particular simulator.

A list of the practicability of the different methods is presented in Appendix B. It may be concluded that the more objective validation methods are the most complicated and may in some cases produce limited information concerning only a few isolated elements of the whole set of skills that usually can be trained on simulators. For this reason it is not possible to provide general statements concerning the overall quality ("best" or "worst") of a method. An example of a more parsimonious way to assess the effectiveness of a simulator training, still using on-the-job performance by subjects, will be given below, it refers to car driving.

Instead of taking driving lessons from the beginning, one could train elementary tasks on a driving simulator (e.g. changing gear, force needed to apply to brake pedal and throttle to attain a certain acceleration, etc.). After a couple of simulator lessons, trainees will have to show their skills in the real car. Experts (driving instructors for instance) may then estimate the number of real lessons it would take for the average person to reach this level of performance. The independent variable is the amount of simulator practice or simulator characteristics, the dependent variable is the estimation made by the driving instructors about the number of lessons it would take to obtain this level of performance. It may be expected that these kinds of judgements can be made in a reliable way by experienced instructors. The ratio of simulator lessons and the estimated equivalent on the real system indicates the training effectivity ratio (TER), which is rapidly and cheaply assessed this way. It may seem that the value of such expert judgement is less reliable than the outcome of a transfer of training study in which an objectively measurable criterion level is determined. It should be noted, however, that objective criterion measurements often reflect training effects on *isolated* part-task skills. Therefore, subjective skill assessments by instructors will mostly be necessary to complete the objective information concerning ToT.

By using several methods to evaluate the effectiveness of a training simulator, the impact of artefacts of each individual method can be minimised. It is preferred to apply several methods in one experiment. Interviewing subjects on their opinion after receiving simulator training is easy to do. In this way one may obtain a more valid outcome of the study.
We will end this discussion with an example. TNO conducted a study 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, both groups 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 (*Experimental-versus-control-group Method*). Because the researchers were 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*). Subsequently, the experienced drivers were interviewed about their opinion on the simulator (*Simulator Fidelity Method*). All this was done to ensure a valid outcome of the experiment with useful suggestions to improve the effectiveness of simulator training. The results of the different method applied in this experiment did not contradict each other.

3 RESEARCH SIMULATOR VALIDATION METHODS

3.1 Introduction

Research simulators are applied in experimental situations to assess the effects of task-, system- or environmental variables on human performance. Data from such studies are often used to interpret task performance of subjects using real-life equipment.
In the first section methods will be described in which a simulator is compared to real-task equipment. These methods either assess physical- or functional fidelity. Second, a method is described in which task performance is compared between different simulators or between different configurations of one simulator. Finally subjective methods will be discussed. The literature does not provide names for the different methods described below. Therefore the names of the methods stem from concepts already discussed in chapter 2.

3.2 Simulator to real task methods

3.2.1 Functional fidelity assessment method

This method resembles the experimental-versus-control-group method described in § 2.2. A group of subjects performs a task on a simulator. The same (or another) group performs the same task on real-life equipment. The correspondence in behaviour of the two groups of subjects is assessed. Hence one measures the functional fidelity of a simulator. One can assess also absolute- and relative validity of the simulator by comparing the obtained results from simulator performance with results from real task performance. In 1984 a certain road in the Netherlands, was reconstructed. The main purpose was to reduce speed of through-traffic. A couple of measures were taken concerning such issues as road surface colour at the entrance of the village, road width within the village, etc. A field study on speed profiles of the through-traffic was carried out. It concentrated on the difference in driving speeds between the before and the after situation. However, it was impossible to conduct a study where all combinations of the various proposed changes would have been possible. This can only be done in a simulator study. In a study conducted by TNO driving behaviour in the Daimler Benz simulator in Berlin was compared to driving behaviour in the real world (Riemersma, Hoekstra & Van der Horst, 1988). The village through which the reconstructed road layout was modelled in the simulator. This made it possible to experimentally assess the effects of all possible road changes. More importantly, it was also possible to compare the behaviour of subjects on the simulator with that of drivers in the real world. If executed correctly, no inherent artefact would have jeopardised this research method. However, some non-inherent artefacts were still unavoidable.
Non-inherent artefacts:

- Selection. Drivers in the real task experiment were Dutch, but drivers in the simulator experiment were German. Because of the difference in speed limits in the two countries this may have affected the outcome of the experiment.

- Testing. Simulator drivers drove several times the same route through the village and may thus have benefited from learning effects. Dutch real-task drivers drove only once through the scene, which means that in their case learning effects were absent.

3.2.2 Physical fidelity assessment method

The method described above involves measurement of behaviour of subjects executing a task. Another option is to measure the physical aspects of the simulator and compare these to the physical aspects of the real-life system. Differences would imply that manipulation characteristics of the simulator are not the same as those of the real-life system. Hence physical fidelity is not high. One may also investigate those handling characteristics themselves in an experiment in which subjects have to execute a predetermined tasks on a simulator and also on real-life equipment. The differences in handling characteristics in the simulator and in real-task equipment are then examined. In a study executed by TNO in commission of the Dutch army, the Leopard 2 tank simulator was validated using this method (Van Breda & Burry, 1991). Military drivers and instructors drove a Leopard 2 tank on a test course in Germany. A couple manoeuvres were carried out. The same manoeuvres were carried out in the tank simulator in Amersfoort, the Netherlands. With a data-acquisition-system, measurements were taken from dynamic aspects of both the tank and the simulator (e.g. braking deceleration as a function of initial speed etc.). The measurements were then compared, which yielded inside into the physical fidelity of the tank simulator.
However, it should be noted that this simulator was build for training purposes, and that this method used was not ideal for validating the simulator as it does not consider the effectiveness of training. However, with conclusions being drawn from the study, TNO could still give recommendations for improving the simulator. The resulting increase in functional fidelity is likely to yield also a higher transfer of training.
There is however one artefact threatening the validity of an experiment described above.
non-inherent artefact:

- Instrumentation. Simulator and real-life equipment should have the same physical transfer function in terms of input (e.g. forces on the gas pedal and output (speed of the car)). It is nearly impossible for subjects to administer the same changes to the controls of either the simulator or real-life equipment. This may easily result in changes in the measurements. Therefore it is better to use calibrated instruments to control the input to the simulator and real-life equipment. An example of such an experiment is described below.

Measuring physical fidelity with calibrated instruments is objective, precise and reliable. Padmos investigated the image system of the ship simulator of the Royal Institute for the Navy (Padmos & Varkevisser, 1995). They measured the luminance of the visual display. The results were compared to the luminance of the sight on a operational ship under the same circumstances. It was concluded that in foggy conditions the display system of the simulator did not generate an image that was similar to real life conditions, hence the physical fidelity of the visual display was not high.

Measuring physical fidelity according to this method is the most precise and objective validation method in the field of simulator research. However, a limitation of this method is the impossibility of generalisation to human task performance. One can not precisely relate physical data to human performance data. In fact it is the only method in simulator validation research which generates objective results not contaminated by artefacts.

3.3 Within simulator validation method

Do subjects perform significantly better on a simulator with a moving base than on a simulator without one? The within simulator validation method or the simulator to simulator method can answer these kinds of questions. According to this method one experimental group executes a task on a particular (usually degraded) simulator, whereas an other group executes the same task on the same simulator with an optimal configuration or on another simulator. With this method effects of differences in behaviour of subjects between different simulator settings such as motion, field of view etc. is measured.

Hogema conducted an experiment to investigate the effectiveness of *compensation for delay* in the visual display of a driving simulator (Hogema, 1993). In the study, subjects had to drive in the middle of a road as accurately as possible, while side wind was disturbing the yaw angle. From the results of this study it was concluded that driving performance in the simulator was better when the compensation technique was implemented. Hence, the functional fidelity of the simulator with the compensation technique was higher than the functional fidelity of the simulator without the compensation technique.

The method is ideal for comparing two (or more) simulators and decide which configuration has the highest functional fidelity. There are three artefacts which may affect the interpretation.

Inherent artefact:

- Reactive effects of experimental arrangements. No comparison is made with real-task equipment. Therefore, conclusions on the behaviour of subjects on real-life equipment are speculative.

Non-inherent artefacts may be the result of comparing two simulators int different settings, these are:

- Selection. Subjects may not be randomised over the experimental groups, as may happen when the two simulators are located at remote sites.
- Instrumentation. This may occur when two simulators are compared and the experimental variable is not the only variable influencing the measurements. For instance: when investigating the differences in functional fidelity between a simulator without a moving base and an other simulator with a moving base but also with a bigger field of view, this artifact hampers a straightforward interpretation of results.

3.4 Subjective validation methods

The subjective methods described in § 2.5 are also applicable to research simulators. In § 3.2 a study was described (Riemersma et al., 1988) in which subjects had to drive through a simulated town. After driving in the simulator subjects had to fill in a questionnaire. The questions concerned subjective sense of reality conveyed by the simulator e.g. how did the vehicle react to steering movement or how realistic did the road environment come across. In general the subjects answered favourable, which indicated a good face validity. However, as mentioned before, on the basis of such opinions one can not conclude that the simulator also has a high physical or functional fidelity.

All subjective methods have one mayor deficiency: they fail to measure fidelity. Instead they reflect the opinion of subjects. This can be interpreted as a face validity measurement of a particular simulator. Like stated in chapter 2, drawing conclusions from only a questionnaire is a hazardous act. Nevertheless, they can be useful in combination with other methods.

3.5 Discussion

Validation studies on research simulators concerns knowledge of particular simulator variables affecting human task performance. Like validation methods of training simulators some validation methods of research simulators are easier applicable than others.

The simulator-to-real-task methods are prone to similar problems as the experimental-versus-control-group method discussed in chapter 2. Thus, the methods are costly and time consuming. The only advantage these methods for research simulators have on the experimental-versus-control-group method is that there is no need to include a training program.

The physical assessment method is the most easily applicable, but this method fails in generalising the results to the consequences for human performance.

The simulator to simulator method is the most feasible method compared to all methods discussed in this chapter, while it does not need real-life equipment. When a research institute has the necessary equipment (simulators), it is relatively easy to conduct an experiment using this method. Still, the method is costly and time consuming.

The subjective method is the most easy applicable method, but it probably fails to measure what it has to measure: fidelity.

Just like research on training simulators it is recommended to use several methods in the validation of research simulators. This will enhance the validity of the conclusions of an experiment.

4 CONCLUSION AND RECOMMENDATIONS

4.1 Conclusions

The present review was carried out to provide an overview of the different methods available for the validation of simulators. In order to properly classify all different methods, a distinction was made between training simulators and research simulators. As was shown in chapter 3 (research simulator validation methods) training simulators can be validated to a certain degree with methods that do not involve measurement of transfer of training. In that case the validity is limited to the physical similarity and the similarity in human behaviour, which in itself does not necessarily provide information about training effectivity.

The main conclusion is: validation of simulators is often hampered by practical complications and (therefore) prone to various methodological flaws and confounding factors. In general, it has been shown that, there is an inverse relationship between the practicability of a particular method and the validity and reliability of the results of that method. Methods that are easily applied usually provide ambiguous or sloppy data and/or data that are difficult to interpret. The best validation methods are the most complicated and may produce limited information concerning only a few isolated elements of the whole set of skills or tasks for which a simulator can be used. Hence, the generality of the experimental-versus-control-group and the functional fidelity assessment method is limited by the manifold constraints of the experimental design. These constraints include the training objectives and/or used performance criteria, and the characteristics of the trainees/experimental subjects under consideration. In addition, the generality of experimental results is limited by the tasks or subtasks selected (and dependent variables) for the study, the task conditions and skills under consideration. Finally, in case of training simulators, generality is limited by additional factors such as the type of training, the training methods, and additional instructional aids. This means that the concept of validity cannot be used as a single, independent attribute that characterizes the general quality of a simulator or a class of simulators. Nevertheless, it is common to use the concept this way.

4.2 Recommendations

1. Terminology in the field of simulator research is ambiguous. The best example of this is the confusion of the term functional fidelity with functional validity. Interchanging these similar terms easily leads to utter nonsense and false conclusions. It is advised to standardise terms, which will lead to more comprehensible communication among researchers.

2. Speaking about the validity of a simulator only makes sense if functional aspects are taken into account, such as the purpose of the simulator, the tasks, trainees (or subjects), training methods, and additional training aids. Therefore, in simulator validation studies, the distinction between training- and research simulators will always have to be made very clear and explicit. For training simulator studies, the potential benefits of additional

instructional facilities, such as automated performance measurement and feedback and scenario management, must be taken into consideration.

3. Always take *face validity* seriously. It is known that the motivation of instructors affects the motivation of trainees. If instructors do not *believe* in the simulator or are not motivated to work with new and complicated products of technology, transfer of training can be influenced in a negative way. The same counts for research personnel and experimental subjects working with research simulators.

4. It is advised to apply more than one method to assess validity. Using several methods in the same study combines the advantages of each individual method and reduces the potential impact of artifacts and thereby the risk of erroneous conclusions. In general, objective measurements should be combined with subjective expert assessments.

5. Measuring physical fidelity of a simulator is the most precise and reliable validation method because it is based on objective physical measurement. However, with these kinds of physical data one can not predict the behavioural characteristics of humans in the simulator. It is therefore useful to use task-specific formulas which relate physical simulator variables to psycho-physical variables and human performance variables (e.g. the relation between a display characteristics and object detectibility). This may reduce the need to measure human task performance in validation studies of a simulator. Future investigations into research simulators can be used to pinpoint which are the relevant simulator performance relationships for different task-categories.

6. The most valid simulator-to-real-task methods are practically the most complicated whereas the produced information may be rather limited. It is therefore recommended to always allocate effort to find a practical method to assess the effectiveness of a simulator training, still comparing simulator performance with on-the-job performance. For example, in a training simulator validation study, after a couple of simulator lessons the resulting skills may be assessed at the real system. This assessment could be done by experienced personnel who can estimate the number of *real* lessons it would take for the average person to reach the observed level of performance. The ratio of simulator lessons and the estimated equivalent on the real system indicates the training effectivity ratio (TER). This method capitalizes on the existing experience of the instructors with training results on the real system such that a control group is not needed. This will substantially decrease time and costs.

REFERENCES

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

Boer, J.P.A. (1991). *Het gebruik van simulatoren voor opleiding en training 1: Bepalende factoren voor de waarde van een simulator als leermiddel* [The use of simulators for education and training 1: Factors that determine the value of a simulator as a learning tool] (Report IZF 1991 A-48). Soesterberg, The Netherlands: TNO Institute for Perception.

Boldovici, J.A. (1987). Measuring Transfer in Military Settings. In S.M. Corbier & J.D. Hagman (Eds.), *Transfer of Learning: Contemporary Research and Applications*. London, Academic Press.

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

Caro, P.W. (1976). *Some factors influencing transfer of simulator training* (HUMRRO Technical Report TR-1-76). Alexandria, VA: Human Resources Research Organisation.

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

Hogema, J.H. (1993). *Compensation for delay in the visual display of a driving simulator effects on lane keeping* (Report IZF 1993 C-21) Soesterberg, The Netherlands: TNO Institute for Perception.

James, H. & Stramler, J.R. (1993). The Dictionary for Human Factors/Ergonomics. Boca Raton, FL: CRC Press, Inc.

Korteling, J.E. & Van Randwijk, M.J. (1991). *Simulatoren en verkeersoefenterreinen in de militaire rijopleiding; literatuurstudie en advies* [Simulators and training grounds for military driver training; literature survey and advice] (Report IZF 1991 A-11). Soesterberg, The Netherlands: TNO Institute for Perception.

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 TM-97-A035). Soesterberg, The Netherlands: TNO Human Factors Research Institute.

Lintern, G., Roscoe, S.N. & Koonce, J.M. & Segal, L.D. (1990). Transfer of landing skills in beginning flight training. *Human Factors, 32 (3),* 319-327.

Meister, D. (1995). Simulation and Modelling. In J.R. Wilson & E.N. Corlett (Eds.), *Evaluation of Human Work: A practical ergonomics methodology* (Chapter 8). London: Taylor & Francis.

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 IZF 1979-23) Soesterberg, The Netherlands: TNO Institute for Perception.

Moraal, J. & Kraiss, K.F. (1981). Manned Systems Design: methods, equipment and applications (NATO Conference series III: Human Factors, Volume 17). London: Plenum Press.

Moraal, J. & Van Meeteren, A. (1990). *Validatie van trainingssimulatoren* [Validation of Training Simulators]. Soesterberg, The Netherlands: TNO Institute for Perception.

Orlansky, J., Dahlman, C.J., Hammon, C.P., Metzko, J., Taylor, H.L. & Youngblut, C. (1994). *The value for simulation for training* (IDA-Paper p-2982). Alexandria, VA: Institute for Defense Analysis.

Padmos, P. & Varkevisser, J. (1995). *Misteffecten in het beeldsysteem van de scheepssimulator op het KIM* [Fog effects in the image system of the ship simulator of the Royal Institute for the Navy] (Report TNO-TM-1995 A-59) Soesterberg, The Netherlands: TNO Human Factors Research Institute.

Poll, K.J. (1980). *Invloeden op de trainingsoverdracht* [Influences on transfer of training] (Memo IZF 1980-M17). Soesterberg. The Netherlands: TNO Institute for Perception.

Riemersma, J.B.J., Hoekstra, W. & Van der Horst, A.R.A. (1988). *The validity of the effects of speed reducing measures obtained in a simulator* (Report IZF 1988 C-18). Soesterberg, The Netherlands: TNO Institute for Perception.

Roscoe, S.N. & Williges, B.H. (1980). *Measurement of transfer of training*. In S.N. Roscoe (Ed.), *Aviation Psychology* (Chapter 16). The Iowa State University Press.

Sanders, M.S. & McCormick, E.J. (1992). *Human Factors in Engineering and Design: 7th edition*. Singapore: McGraw-Hill, Inc.

Stammers, R.B. & Shepard, A. (1995). *Task analysis*. In J.R. Wilson & E.N. Corlett (Eds.), *Evaluation of Human Work: A practical ergonomics methodology* (Chapter 6). London: Taylor & Francis.

Swanborn, P.G. (1993). *Methoden van sociaal- wetenschappelijk onderzoek: nieuwe editie* [Methodology of social-scientific research: new edition]. Meppel: Boom.

Truijens, C.L. & Schuffel, H. (1978). *Ergonomisch onderzoek "Open Hartelkanaal" Deel 3; Validering van de simulatie van duwvaart in het Hartelgebied* [Ergonomical research "Open Hartelkanaal" Part 3; Validating of simulation of push-towing in the Hartel area] (Report IZF 1978-C6). Soesterberg, The Netherlands: TNO Institute for Perception.

Van Breda, L. & 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.

Veltman, J.A. & Korteling, J.E. (1993). *Validatiestudie Rijsimulatoren Leopard 2 en YPR-765* [Validation study driving simulators Leopard 2 and YPR-765] (Report IZF 1993 A-43). Soesterberg, The Netherlands: TNO Institute for Perception.

Wertheim, A.H. (1984). *Over de mogelijkheid om de MECH LUA-Trainer te valideren* [About the possibility to validate the MECH LUA-Trainer] (Report IZF 1984-1). Soesterberg, The Netherlands: TNO Institute for Perception.

Soesterberg, 16 March 1999

Dr. J.E. Korteling
(First author)

Dr. K. van den Bosch
(Project leader)

APPENDIX A List of artefacts which have confounding effects on an experiment

Artefacts contaminating internal validity:

1. Selection. Occurs when the assignment of subjects to the different groups have an effect on the outcome of an experiment.
2. *Instrumentation*. Occurs when changes in the calibration of a measuring instrument or changes in the observers or scores used, produces changes in the obtained measurements.
3. *History*. Occurs when specific events occurring between the first and the second measurement in the addition of the experimental variable influence the dependent variable(s).
4. *Maturation*. Occurs when a process (not specific to a particular event, including: growing older, growing hungrier, growing more tired, etc.) within the subjects, operating as a function of the passing of time per se, has an effect on the outcome of an experiment.
5. *Testing*. Occurs when the effects of taking a test upon the scores of a second testing selectively influences the obtained measurements.
6. *Mortality*. Occurs when loss of subjects from the comparison groups affects the outcome of the experiment.

Artefacts contaminating external validity:

1. Reactive effects of experimental arrangements. Occurs when a faulty generalisation, about the effect of the experimental variable(s) upon persons being exposed to it in a non-experimental setting, is made.
2. *Interaction effect of selection and experimental variables*. Occurs when selection effects interact with the independent variable in such a way that generalisation to other circumstances is impossible.

Note: Campbell and Stanley describe more artefacts in their book *Experimental and Quasi-Experimental Designs for Research*. These other artefacts are not of interest to simulator research and therefore not included in this list.

APPENDIX B Table of the different methods and artefacts which affect them

Research methods		**Artefacts influencing internal validity***						**Artefacts influencing external validity***		**Practica-bility****
Training simulators	**Page**	**1 Sel**	**2 Instr**	**3 Hist**	**4 Matu**	**5 Test**	**6 Mort**	**1 Reactive**	**2 Interaction**	
Exp. versus Cont. group method	10	N	N	C	C	C	N			1
Self-control-transfer method	12	C	I	I	I	I	C			3
Pre-existing-control-transfer method	13	I	N	C	C	C	N			3
Uncontrolled transfer method	13	N	N	C	I	C	N			3
QToT method	14	N	N	C	C	C	N	I	C	2
Backward transfer method	15	C	C	C	C	C	C		I	3
Performance improvement method	16	C	C	C	C	C	C	I		4
Simulator fidelity method	16							I		5
Training program analysis method	17							I		5
Opinion survey method	17							I		5
Research simulators	**Page**	**1**	**2**	**3**	**4**	**5**	**6**	**1**	**2**	
Functional fidelity assessment method	20	N	C	C	C	N	C			2
Physical fidelity assessment method	21				N					3
Simulator to simulator method	22	N	N	C	C	C	C	I		4
Subjective method	23							I		5

* The numbers of the different artefacts correspond with the numbers of the artefacts in Appendix A.

** The numbers indicate the practicability of a particular method; 1 indicates what seems to be a difficult method to apply, 5 an easy method.

Note: In the tables, an I indicates an inherent artefact, a C indicates that the factor is controlled, an N indicates a non-inherent artefact and a blank indicates that the factor is not relevant. It is with extreme reluctance that this table is presented, because they are apt to be *too helpful* and to be depended upon, instead of the more complex and qualified presentation in the text. No I or C should be respected unless the reader comprehends why it is placed there. This table is made for review purposes, not to create fears, or confidence in, a specific method.

REPORT DOCUMENTATION PAGE

1. DEFENCE REPORT NO. TD 99-0026	2. RECIPIENT ACCESSION NO.	3. PERFORMING ORGANIZATION REPORT NO. TM-99-A023
4. PROJECT/TASK/WORK UNIT NO. 730.3	5. CONTRACT NO. A98/KL/301	6. REPORT DATE 16 March 1999
7. NUMBER OF PAGES 32	8. NUMBER OF REFERENCES 27	9. TYPE OF REPORT AND DATES COVERED Interim

10. TITLE AND SUBTITLE

A critical review of validation methods for man-in-the-loop simulators

11. AUTHOR(S)

J.E. Korteling and R.R. Sluimer

12. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

TNO Human Factors Research Institute
Kampweg 5
3769 DE SOESTERBERG

13. SPONSORING AGENCY NAME(S) AND ADDRESS(ES)

Director of Army Research and Development
Van der Burchlaan 31
2597 PC DEN HAAG

14. SUPPLEMENTARY NOTES

15. ABSTRACT (MAXIMUM 200 WORDS (1044 BYTES))

This review examines the methodological concepts, paradigms and pitfalls related to validation- and fidelity studies of man-in-the-loop simulators. A distinction is made between validation methods for training simulators and for research simulators. Validation methods for training simulator are applied in experiments which assess effects of simulator variables (e.g. resolution of the display, cue augmentation, moving base characteristics) on the effectiveness of a simulator as a training device. Validation methods for research simulators are applied in experiments which assess the effects of simulator variables on the effectiveness of a simulator as a research tool. The review is particularly focussed on the various artefacts that may affect the outcome of such validation experiments. The artefacts are separately described for each single validation method. It will be demonstrated that validation of simulators is a very complicated matter and prone to various methodological flaws and confounding factors. After the discussion of the common methods including their advantages and disadvantages the following recommendations for future research are given:

1. Terminology in the field of simulator research is ambiguous. It is advised to standardise terms, which will lead to more comprehensible communication among researchers.
2. Validity is not a single, independent attribute. The term validity in simulator research only makes sense if related to functional aspects of simulators, such as the purpose of the simulator (training, research) and the tasks and training methods involved. This will reduce the amount of overgeneralizations that are now encountered too frequently.
3. Always take face validity into consideration. If people do not believe in the simulator they are not very likely to use it properly.
4. Apply more than one method in a simulator validity study. Combination of e.g. objective with subjective methods reduces the risk of erroneous conclusions and combines the benefits of both kinds of methods.
5. Aim more research at creating task-specific formulas, which relate physical simulator variables to psycho-physical and human performance variables. This will reduce the need to measure human task performance in simulator validation studies.
6. Always allocate substantial effort to find a practical method that still compares simulator performance with on-the-job performance by subjects. A relatively simple and practical method is proposed to assess the effectiveness of a driving simulator training.

16. DESCRIPTORS Simulators Training Transfer of Training	IDENTIFIERS Fidelity Validity

17a. SECURITY CLASSIFICATION (OF REPORT)	17b. SECURITY CLASSIFICATION (OF PAGE)	17c. SECURITY CLASSIFICATION (OF ABSTRACT)

18. DISTRIBUTION AVAILABILITY STATEMENT Mailing list only	17d. SECURITY CLASSIFICATION (OF TITLES)